PROCESS DATA IN EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 2nd Edition

EDITED BY: Hong Jiao, Qiwei He and Bernard Veldkamp
PUBLISHED IN: Frontiers in Psychology, Frontiers in Education and Frontiers in Applied Mathematics and Statistics

frontiers Research Topics

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88974-138-0
DOI 10.3389/978-2-88974-138-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

PROCESS DATA IN EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 2nd Edition

Topic Editors:
Hong Jiao, University of Maryland, United States
Qiwei He, Educational Testing Service, United States
Bernard Veldkamp, University of Twente, Netherlands

Publisher's note: In this 2nd edition:

The following article has been added: Jiao H, He Q and Veldkamp BP (2021) Editorial: Process Data in Educational and Psychological Measurement. *Front. Psychol*. 12:793399. doi: 10.3389/fpsyg.2021.793399

The following article has been added: Reis Costa D, Bolsinova M, Tijmstra J and Andersson B (2021) Improving the Precision of Ability Estimates Using Time-On-Task Variables: Insights From the PISA 2012 Computer-Based Assessment of Mathematics. *Front. Psychol*. 12:579128. doi: 10.3389/fpsyg.2021.579128

The following article has been removed: Minghui L, Lei H, Xiaomeng C and Potměšilc M (2018) Teacher Efficacy, Work Engagement, and Social Support Among Chinese Special Education School Teachers. *Front. Psychol*. 9:648. doi: 10.3389/fpsyg.2018.00648

Citation: Jiao, H., He, Q., Veldkamp, B., eds. (2021). Process Data in Educational and Psychological Measurement, 2nd Edition. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-138-0

Table of Contents

06 ***Editorial: Process Data in Educational and Psychological Measurement***
Hong Jiao, Qiwei He and Bernard P. Veldkamp

LEVERAGING PROCESS DATA TO EXPLORE TEST-TAKERS' BEHAVIORS AND STRATEGIES

11 ***Exploring Multiple Goals Balancing in Complex Problem Solving Based on Log Data***
Yan Ren, Fang Luo, Ping Ren, Dingyuan Bai, Xin Li and Hongyun Liu

21 ***Validating Test Score Interpretations Using Time Information***
Lena Engelhardt and Frank Goldhammer

39 ***Taking a Closer Look: An Exploratory Analysis of Successful and Unsuccessful Strategy Use in Complex Problems***
Matthias Stadler, Frank Fischer and Samuel Greiff

49 ***How Do Test Takers Interact With Simulation-Based Tasks? A Response-Time Perspective***
Yi-Hsuan Lee, Jiangang Hao, Kaiwen Man and Lu Ou

66 ***Detecting Examinees With Pre-knowledge in Experimental Data Using Conditional Scaling of Response Times***
Sarah L. Toton and Dennis D. Maynes

84 ***Understanding Test Takers' Choices in a Self-Adapted Test: A Hidden Markov Modeling of Process Data***
Meirav Arieli-Attali, Lu Ou and Vanessa R. Simmering

99 ***Data Mining Techniques in Analyzing Process Data: A Didactic***
Xin Qiao and Hong Jiao

110 ***Analysis of Process Data of PISA 2012 Computer-Based Problem Solving: Application of the Modified Multilevel Mixture IRT Model***
Hongyun Liu, Yue Liu and Meijuan Li

122 ***Mapping Background Variables With Sequential Patterns in Problem-Solving Environments: An Investigation of United States Adults' Employment Status in PIAAC***
Dandan Liao, Qiwei He and Hong Jiao

JOINT MODEL FOR RESPONSE ACCURACY AND RESPONSE TIMES

154 ***Variable Speed Across Dimensions of Ability in the Joint Model for Responses and Response Times***
Peida Zhan, Hong Jiao, Kaiwen Man, Wen-Chung Wang and Keren He

165 ***Improving the Precision of Ability Estimates Using Time-On-Task Variables: Insights From the PISA 2012 Computer-Based Assessment of Mathematics***
Denise Reis Costa, Maria Bolsinova, Jesper Tijmstra and Björn Andersson

178 ***A Speed-Accuracy Tradeoff Hierarchical Model Based on Cognitive Experiment***
Xiaojun Guo, Zhaosheng Luo and Xiaofeng Yu

184 ***Bayesian Covariance Structure Modeling of Responses and Process Data***
Konrad Klotzke and Jean-Paul Fox

199 ***Invariance of the Response Processes Between Gender and Modes in an Assessment of Reading***
Ulf Kroehne, Carolin Hahnel and Frank Goldhammer

215 ***An Overview of Models for Response Times and Processes in Cognitive Tests***
Paul De Boeck and Minjeong Jeon

226 ***Modeling Response Time and Responses in Multidimensional Health Measurement***
Chun Wang, David J. Weiss and Shiyang Su

237 ***Modeling Learner Heterogeneity: A Mixture Learning Model With Responses and Response Times***
Susu Zhang and Shiyu Wang

251 ***Modeling Nonlinear Conditional Dependence Between Response Time and Accuracy***
Maria Bolsinova and Dylan Molenaar

STATISTICAL MODEL ON RESPONSE PROCESS

263 ***Understanding Therapeutic Change Process Research Through Multilevel Modeling and Text Mining***
Wouter A. C. Smink, Jean-Paul Fox, Erik Tjong Kim Sang, Anneke M. Sools, Gerben J. Westerhof and Bernard P. Veldkamp

277 ***Scaling the Variance of a Latent Variable While Assuring Constancy of the Model***
Karl Schweizer, Stefan J. Troche and Christine DiStefano

287 ***Compromised Item Detection for Computerized Adaptive Testing***
Cheng Liu, Kyung T. Han and Jun Li

303 ***A Multidimensional IRT Approach for Dynamically Monitoring Ability Growth in Computerized Practice Environments***
Jung Yeon Park, Frederik Cornillie, Han L. J. van der Maas and Wim Van Den Noortgate

313 ***Statistical Analysis of Complex Problem-Solving Process Data: An Event History Analysis Approach***
Yunxiao Chen, Xiaoou Li, Jingchen Liu and Zhiliang Ying

ADVANCEMENT IN COGNITIVE DIAGNOSTIC MODEL WITH PROCESS INFORMATION

323 ***Termination Rules for Variable-Length CD-CAT From the Information Theory Perspective***
Lei Guo and Chanjin Zheng

336 ***Integrating Differential Evolution Optimization to Cognitive Diagnostic Model Estimation***
Zhehan Jiang and Wenchao Ma

345 ***Joint Testlet Cognitive Diagnosis Modeling for Paired Local Item Dependence in Response Times and Response Accuracy***
Peida Zhan, Manqian Liao and Yufang Bian

USING DATA STREAMS FOR ESTIMATING COLLABORATIVE PROBLEM-SOLVING SKILLS

359 ***Characterizing Interactive Communications in Computer-Supported Collaborative Problem-Solving Tasks: A Conditional Transition Profile Approach***

Jiangang Hao and Robert J. Mislevy

368 ***Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data***

Jianlin Yuan, Yue Xiao and Hongyun Liu

frontiers in Psychology

EDITORIAL
published: 03 December 2021
doi: 10.3389/fpsyg.2021.793399

Editorial: Process Data in Educational and Psychological Measurement

Hong Jiao[1]*, *Qiwei He*[2]* and *Bernard P. Veldkamp*[3]*

[1] Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, United States, [2] Educational Testing Service, Princeton, NJ, United States, [3] Department of Learning, Data Analytics and Technology, Faculty of Behavioral Management and Social Sciences, University of Twente, Enschede, Netherlands

Keywords: process data, psychological assessment, data mining, educational assessment, computer based assessment

Editorial on the Research Topic

Process Data in Educational and Psychological Measurement

The increasing use of computer-based testing and learning environments is leading to a significant reform on the traditional form of measurement, with tremendous extra available data collected during the process of learning and assessment (Bennett et al., 2007, 2010). It means that we can learn and describe the respondents' performances not only by their responses, but also their responding processes, in addition to the response accuracy in the traditional tests (Ercikan and Pellegrino, 2017).

The recent advances in computer technology enhance the convenient collection of process data in computer-based assessment. One such example is time-stamped action data in an innovative item which allow for the interaction between a respondent and the item. When a respondent attempts an interactive item, his/her actions are recorded, in the form of an ordered sequence of multi-type, time-stamped events. These sorts of data stored in log files, referred to as *process data* in this book, provide information beyond response data that typically show response accuracy only. This additional information holds promise to help us understand the strategies that underlie test performance and identify key actions that lead to success or failure of answering an item (e.g., Han et al., 2019; Liao et al.; Stadler et al., 2019; He et al., 2021; Ulitzsch et al., 2021a; Xiao et al., 2021).

With the availability of process data in addition to response data, the measurement field is becoming increasingly interested in borrowing additional auxiliary information from the responding process to serve different assessment purposes. For instance, recently researchers proposed different models for response time and the joint modeling of responses and response time (e.g., Bolsinova and Molenaar; Costa et al.; Wang et al.). In addition, other process data such as the path collected based on eye-tracking devices (e.g., Zhu and Feng, 2015; Maddox et al., 2018; Man and Harring, 2021), action sequences in problem-solving tasks (e.g., Chen et al.; Tang et al., 2020; He et al., 2021; Ulitzsch et al., 2021b), and processes in collaborative problem solving (e.g., Graesser et al., 2018; Andrews-Todd and Kerr, 2019; De Boeck and Scalise, 2019), are also worthy of exploration and integration with product data for assessment purposes.

This Research Topic (formed in this edited e-book) intends to explore the forefront of responding to the needs in modeling new data sources and incorporating process data in the statistical modeling of multiple possible assessment data. This edited book presents the cutting-edge research related to utilizing process data in addition to product data such as item responses in educational and psychological measurement for enhancing accuracy in ability parameter

OPEN ACCESS

Edited and reviewed by:
Huali Wang,
Peking University Sixth Hospital, China

Correspondence:
Hong Jiao
hjiao@umd.edu
Qiwei He
qhe@ets.org
Bernard P. Veldkamp
b.p.veldkamp@utwente.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *12 October 2021*
Accepted: *25 October 2021*
Published: *03 December 2021*

Citation:
Jiao H, He Q and Veldkamp BP (2021) Editorial: Process Data in Educational and Psychological Measurement. Front. Psychol. 12:793399. doi: 10.3389/fpsyg.2021.793399

estimation (e.g., Bolsinova and Molenaar; De Boeck and Jeon; Engelhardt and Goldhammer; Klotzke and Fox; Liu C. et al.; Park et al.; Schweizer et al.; Wang et al.; Zhang and Wang), cognitive diagnosis facilitation (e.g., Guo and Zheng; Guo et al.; Jiang and Ma; Zhan, Liao et al.; Zhan, Jiao et al.), and aberrant responding behavior detection (e.g., Liu H. et al.; Toton and Maynes).

Throughout the book, the methods for analyzing process data in technology-enhanced innovative items in large-scale assessment for high-stakes decisions are addressed (e.g., Lee et al.; Stadler et al.). Further, the methods for the extraction of useful information in process data in assessments such as serious games and simulations were also discussed (e.g., Liao et al.; Kroehne et al.; Ren et al.; Yuan et al.). The interdisciplinary studies that borrow data-driven methods from computer science, machine learning, artificial intelligence, and natural language processing are also highlighted in this Research Topic (e.g., Ariel-Attali et al.; Chen et al.; Hao and Mislevy; Qiao and Jiao; Smink et al.), which provide new perspectives in data exploration in educational and psychological measurement. Most importantly, the models presenting the integration of the process data and the product data in this book are of critical significance to link the traditional test data with the new features extracted from the new data sources. Meanwhile, the papers included in the book provide an excellent source for data and coding sharing, which entails significant contributions to the applications of the innovative statistical modeling of assessment data in the measurement field.

The book chapters demonstrate the use of process data and the integration of process and product data (item responses) in educational and psychological measurement. The chapters address issues in adaptive testing, problem-solving strategy, validity of test score interpretation, item pre-knowledge detection, cognitive diagnosis, complex dependence in joint modeling of responses and response time, and multidimensional modeling of these data types. The originality of this book lies in the statistical modeling of innovative assessment data such as log data, response time data, collaborative problem-solving tasks, dyad data, change process data, testlet data, and multidimensional data. Further, new statistical models are presented for analyzing process data in addition to response data such as transition profile analysis, the event history analysis approach, hidden Markov modeling, conditional scaling, multilevel modeling, text mining, Bayesian covariance structure modeling, mixture modeling, and multidimensional modeling. The integration of multiple data sources and the use of process data provides the measurement field with new perspectives to solve assessment issues and challenges such as problem-solving strategy, cheating detection, and cognitive diagnosis.

An overview of all the papers included in this Research Topic is summarized in **Table 1** with respect to their key features. The scope of the Research Topic can be classified into five major categories:

(1) leveraging process data to explore test-takers' behaviors and problem-solving strategies,
(2) proposing joint modeling for response accuracy and response times,
(3) proposing new statistical models on analyzing response processes (e.g., time-stamped sequential events),
(4) advancing cognitive diagnostic models with new data sources, and
(5) using data streams in estimating collaborative problem-solving skills.

The above categorization focused on each paper's core contribution though some papers can be cross-classified. The papers' key findings and advancements impressively represent the current state-of-the-art methods in the field of process data analysis in educational and psychological assessments. As topic editors, we were happy to receive such a great collection of papers with various foci and submit these publications right as digital assessments are booming. The papers collected in this Research Topic are also diverse in data types, statistical approaches, and assessment with an extensive scope in both high-stake and low-stake assessments, covering research fields in education, psychology, health, and other applied disciplines.

As one of the first comprehensive books addressing the modeling and application of process data, this e-book has drawn great attention since its debut was cross-loaded on three journals in *Frontiers in Psychology, Frontiers in Education,* and *Frontiers in Applied Mathematics and Statistics.* With 29 papers from 77 authors, this book enhances interdisciplinary research in fields such as psychometrics, psychology, statistics, computer science, educational technology, and educational data mining, to name a few. As highlighted on the e-book webpage, (https://www.frontiersin.org/research-topics/7035/process-data-in-educational-and-psychological-measurement#impact) on November 13, 2021, this e-book has accumulated 115,069 total reviews and 17,940 article downloads since the Research Topic project launched in 2017. This number keeps growing on a daily basis. The diversified demographics provide convincing evidence that the papers in this book reached the global research community, addressing the critical issues of statistical modeling of multiple types of assessment data in the digital era. This book is just on time to provide tools and methods to shape this new measurement horizon.

As more and more data are being collected in computer-based testing, process data will become a very important source of information to validate and facilitate measuring response accuracy and provide supplementary information in understanding test-takers' behaviors, the reasons of missing data, and links with motivation studies. There is no doubt that there is high demand of such research in the large-scale assessment, both high-stake and low-stake, as well as in the personalized learning and assessment to tailor the best source and methods to help people learn and grow. This book is a timely addition to the current literature on psychological and educational measurement. It is expected to be applied

TABLE 1 | An overview of papers collected in this Research Topic.

References	Areas of advancement	Data types	Statistical approaches	Assessment domains
Leveraging process data to explore test-takers' behaviors and strategies				
Ren et al.	Exploring multiple goals in interactive problem-solving items	Extracted response process variables, correctness of responses	Cluster analysis, logistics, and least-squares regression	Interactive problem-solving in PISA 2012
Engelhardt and Goldhammer	Proposing a validity research that uses processing times to provide both convergent and discriminant validity evidence for the construct interpretation of reasoning and reading ability scores	Response data, response times	MLR estimator (maximum likelihood estimation with robust standard error)	PIAAC 2012 literacy assessments
Stadler et al.	Exploring successful and unsuccessful strategies with process data in complex problem-solving items	Response process data, correctness of responses	N-grams model	Interactive problem-solving items
Lee et al.	Exploring response times in complex simulation-based tasks to understand test-takers' interactions	Response data, response times	Cluster analysis and hierarchical framework for joint modeling item responses and response times	Interactive problem-solving items
Toton and Maynes	Detecting examinees with pre-knowledge in experimental data with conditional scaling of response times	Item scores, response times	Cluster analysis, factor analysis	Simulation study and empirical study in GRE quantitative testing
Arieli-Attali et al.	Understanding test-takers' choices using hidden Markov modeling of process data	Response data, answer change, item difficulty	Hidden Markov model	Self-adapted tests
Qiao and Jiao	Using data mining techniques in analyzing process data and making comparisons among machine-learning algorithms in exploring problem-solving items	Extracted response process variables, correctness of responses	Multiple machine learning algorithms: supervised techniques (CART, gradient boosting, random forest, and SVM), unsupervised techniques (SOM, k-means)	Interactive problem-solving in PISA 2012
Liu H. et al.	Exploring test-takers' problem-solving strategies with a modified multilevel mixture IRT model	Extracted response process variables, correctness of responses	Modified multilevel mixture IRT model, latent class analysis	Interactive problem-solving in PISA 2012
Liao et al.	Exploring sequential patterns in problem-solving items and relationship with individual differences in background variables	Extracted response process variables, response data, background variables	N-grams model, feature selection model, regression analysis	PIAAC 2012 problem-solving in technology-rich environment
Joint model for response accuracy and response times				
Zhan, Jiao et al.	Proposing a joint model for multidimensional abilities and multifactor speed	Response data, response times	Joint modeling of response and response time, exploratory factor analysis	Simulation study and empirical study in computer-based math assessment (PISA 2012)
Costa et al.	Proposing a joint model for item response and time-on-task to increase the precision of ability estimates	Response data, response times	Multidimensional latent model for response and response time	Interactive problem-solving in PISA 2012
Guo et al.	Proposing a joint model for a speed-accuracy tradeoff hierarchical model based on cognitive experiment	Response data, response times	Bayesian MCMC algorithm, speed-accuracy hierarchical model	Simulation study and empirical study in Raven's Standard Progressive Matrices
Klotzke and Fox	Proposing a Bayesian modeling framework for response accuracy, response times, and other process data variables	Response data, response times, extracted response process variables	Bayesian covariance structure models	Simulation study and empirical study in PIAAC 2012 cognitive assessments
Kroehne et al.	Proposing a parameterized joint model of response data and response time to detect invariance by gender and mode between computer-based and paper-based tests	Response data, response times	Bivariate generalized linear IRT model framework (B-GLIRT)	PISA 2012 and PISA 2009 reading assessments

(Continued)

TABLE 1 | Continued

References	Areas of advancement	Data types	Statistical approaches	Assessment domains
De Boeck and Jeon	An overview of models for joint modeling of response times and response accuracy in cognitive tests	Response data, response times	Multiple response models and joint models of response data and response times	Literature review
Wang et al.	Modeling response time and responses in multidimensional health measurement	Response data, response times	Multidimensional-graded response model, hierarchical joint model of responses and response times	Health measurement
Zhang and Wang	Proposing a mixture learning model that utilizes the response times and response accuracy in learning progression	Response data, response times	Diagnostic classification model framework, Bayesian estimation	Simulation study and empirical study in a computer-based learning environment
Bolsinova and Molenaar	Proposing a joint model for response accuracy and response times with consideration on non-linear conditional dependence	Response data, response times	Joint model for quadratic conditional dependence, joint model for multiple-category conditional dependence, indicator-level non-parametric moderation method	Simulation study and empirical study in high-stakes arithmetic assessment
Statistical model on response process				
Smink et al.	Therapeutic change process research through multilevel and text mining	Life narratives textual data and response data	Multilevel models, text mining	Epidemiologic Studies Depression Scale and life narratives (CES-D)
Schweizer et al.	Investigating how the major outcome of a confirmatory factor investigation is preserved when scaling the variance of a latent variable by the various scaling methods	Scaling data	Multiple confirmatory factor analysis	Simulation study and empirical study in Multitrait-Multimethod (MTMM) design
Liu C. et al.	Proposing a model with a leakage parameter to better characterize the item leaking process and develop a more generalized detection method by monitoring responses of test-takers	Response data	Generalized linear model for detection, leakage simulation model	Simulation study and empirical study in operational computerized adaptative testing
Park et al.	Proposing a multidimensional IRT approach for dynamically monitoring ability growth in adaptive learning systems	Response data, response times	Multidimensional IRT	Simulation study and web-based learning platform
Chen et al.	Proposing an event history analysis approach to predict duration and outcome of solving a complex problem by making use of process data	Time-stamped sequential events data, correctness of responses	Regression model	Interactive problem-solving in PISA 2012
Advancement in cognitive diagnostic model with process information				
Guo and Zheng	Comparing termination rules for variable-length CD-CAT from the information theory perspective	Response data, test construction variables	Multiple cognitive diagnostic models	Simulation study
Jiang and Ma	Proposing a model to integrate differential evolution optimization into the EM framework in the log-linear cognitive diagnostic model estimation	Response data	Log-linear cognitive diagnostic model with EM algorithm, differential evolution	Simulation study and empirical study in assessment of a health profession
Zhan, Liao et al.	Proposing a joint testlet cognitive diagnostic model for paired local item dependence using response time and response accuracy	Response data, response times	Joint testlet cognitive diagnosis modeling	PISA 2015 computer-based math assessment

(Continued)

TABLE 1 | Continued

References	Areas of advancement	Data types	Statistical approaches	Assessment domains
Using data streams for estimating collaborative problem-solving skills				
Hao and Mislevy	Characterizing interactive communications in collaborative problem-solving using a conditional transition profile approach	Conversations collected in a computer-based collaborative problem-solving platform	Conditional transition profile, cluster analysis	Collaborative problem-solving platform
Yuan et al.	Assessing collaborative problem-solving competence by extracting indictors from process stream data and modeling dyad data	Process stream data in collaborative problem solving, response data	Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM)	Collaborative problem-solving platform adapted from a problem-solving task in PISA 2012

more extensively in educational and psychological measurement, such as in computerized adaptive testing and dynamic learning.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

QH was partially supported by the National Science Foundation grants IIS-1633353.

ACKNOWLEDGMENTS

We thank all authors who have contributed to this Research Topic and the reviewers for their valuable feedback on the manuscript.

REFERENCES

Andrews-Todd, J., and Kerr, D. (2019). Application of ontologies for assessing collaborative problem-solving skills. *Int. J. Testing* 19, 172–187. doi: 10.1080/15305058.2019.1573823

Bennett, R. E., Persky, H., Weiss, A., and Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *J. Technol. Learn. Assessment* 8, 1–44. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1627

Bennett, R. E., Persky, H., Weiss, A. R., and Jenkins, F. (2007). *Problem Solving in Technology-Rich Environments. A Report From the NAEP Technology Based Assessment Project, Research and Development Series* (NCES 2007-466). National Center for Education Statistics.

De Boeck, P., and Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Front. Psychol.* 10:1280. doi: 10.3389/fpsyg.2019.01280

Ercikan, K., and Pellegrino, J. W. (Eds.). (2017). *Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes*. New York, NY: Routledge.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., and Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychol. Sci. Public Interest* 19, 59–92. doi: 10.1177/1529100618808244

Han, Z., He, Q., and von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: an application of random forests. *Front. Psychol.* 10:2461. doi: 10.3389/fpsyg.2019.02461

He, Q., Borgonovi, F., and Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Identifying generalized behavioral patterns with sequence mining. *Comp. Educ.* 166:104170. doi: 10.1016/j.compedu.2021.104170

Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G., and Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: evidence from the OECD PIAAC assessment. *Eur. J. Psychol. Educ.* 33, 543–558. doi: 10.1007/s10212-018-0380-2

Man, K., and Harring, J. R. (2021). Assessing preknowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educ. Psychol. Meas.* 81, 441–465. doi: 10.1177/0013164420968630

Stadler, M., Fischer, F., and Greiff, S. (2019). Taking a closer look: an exploratory analysis of successful and unsuccessful strategy use in complex problems. *Front. Psychol.* 10:777. doi: 10.3389/fpsyg.2019.00777

Tang, X., Wang, Z., He, Q., Liu, J., and Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika* 85, 378–397. doi: 10.1007/s11336-020-09708-3

Ulitzsch, E., He, Q., and Pohl, S. (2021a). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *J. Educ. Behav. Statistics* 45, 59–96. doi: 10.3102/10769986211010467

Ulitzsch, E., He, Q., Ulitzsch, V., Nichterlein, A., Molter, H., Niedermeier, R., et al. (2021b). Combining clickstream analyses and graph-modeled data clustering for identifying common response process using time-stamped action sequence. *Psychometrika* 86, 190–214. doi: 10.1007/s11336-020-09743-0

Xiao, Y., He, Q., Veldkamp, B., and Liu, H. (2021). Exploring latent states of problem-solving competence using hidden markov modeling on process data. *J. Comp. Assisted Learn.* 37, 1232–1247. doi: 10.1111/jcal.12559

Zhu, M., and Feng, G. (2015). "An exploratory study using social network analysis to model eye movements in mathematics problem solving," in *LAK '15: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 383–387. doi: 10.1145/2723576.2723591

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Jiao, He and Veldkamp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 September 2019
doi: 10.3389/fpsyg.2019.01975

Exploring Multiple Goals Balancing in Complex Problem Solving Based on Log Data

*Yan Ren[1], Fang Luo[1], Ping Ren[2], Dingyuan Bai[2], Xin Li[2] and Hongyun Liu[1]**

[1] School of Psychology, Beijing Normal University, Beijing, China, [2] Collaborative Innovation Center of Assessment Towards Basic Education Quality, Beijing Normal University, Beijing, China

Multiple goals balancing is an important but not yet fully validated dimension of complex problem solving (CPS). The present study used process data to explore how solvers clarify goals, set priorities, and balance conflicting goals. We extracted behavioral indicators of goal pursuit from the log data of 3,201 students on the third subtask of the "Ticket" task in the PISA 2012 CPS test. Cluster analysis was used to identify 10 groups that varied in goal pursuit behavior. Logistics and least-squares regression analysis were used to explore how goal pursuit affected task scores and CPS proficiency. The results showed that competent solvers clarified goals and weighed priorities more effectively. They also made trade-offs between conflicting goals. The importance of theoretically-driven log data analysis and coping strategies in the face of multiple goals conflict scenarios was discussed.

Keywords: complex problem solving, multiple goals balancing, log data analysis, educational data mining, K-means cluster analysis

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Vanessa R. Simmering,
ACT, Inc., United States
Susu Zhang,
University of Illinois
at Urbana-Champaign, United States

***Correspondence:**
Hongyun Liu
hyliu@bnu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 December 2018*
Accepted: *12 August 2019*
Published: *27 September 2019*

Citation:
Ren Y, Luo F, Ren P, Bai D, Li X and Liu H (2019) Exploring Multiple Goals Balancing in Complex Problem Solving Based on Log Data. Front. Psychol. 10:1975. doi: 10.3389/fpsyg.2019.01975

INTRODUCTION

Science and technology are developing in the current information explosion era. People are facing an increasing number of complex problems in daily life, many of which involving the simultaneous pursuit of multiple goals. Therefore, complex problem solving (CPS) becomes common in real life, such as the use of complex technology (e.g., mobile phones, personal computers, and vending machines), the management of complex organizations (e.g., companies and departments), and the prediction of complex environments (e.g., weather and stock prices; Funke, 2003, 2010).

Complex problem solving refers to successful interaction with a dynamic task environment, wherein all or some rules in the environment can only be learned by exploring and integrating information (Buchner, 1995). Many researchers have suggested that CPS should be assessed in a simulated problem scenario (a complex system) where has a plurality of variables. In the scenario, solvers are asked to manipulate some of the variables to explore effective rules of describing relationships among all variables (knowledge acquisition), and then solvers need to use the learned knowledge of rules to achieve specific goals (knowledge application; Funke, 2001).

Multiple Goals in Complex Problem Solving

Blech and Funke (2010) verified that the presence of conflicting goals affects the difficulty of a complex system. They found that increasing the number of goals in CPS – especially with respect to conflicting goals – increases the cognitive and emotional challenges faced by solvers (Funke, 1992). Therefore, in the case of complex problems involving multiple goals, solvers may not be

able to fully account for each goal. Thus, they might weigh the priority of each. They may first achieve one goal and then find the next one; they may sacrifice one goal in exchange for another; or they may choose to achieve a complementary goal. These strategies emphasize the importance of goal priority (Funke, 2010). Dörner and Kreuzig (1983) proposed that operative intelligence involves the skills of goal elaborating and goal balancing. Later, Dörner proposed the CPS action theory, which divides the CPS solution process into six characteristic phases, of which one is exploring and determining important parts of the system (e.g., such as defining and balancing conflicting goals; Dörner and Wearing, 1995).

The five-dimensional model of CPS consists of system exploration, information reduction, model formation, control considering dynamic change, and prioritization of goals, in which the ability to clarify, prioritize, and balance goals is an important dimension (Funke, 2001; Fischer et al., 2012; Greiff and Fischer, 2013; Schoppek and Fischer, 2015; Herde et al., 2016). However, when assessing CPS, researchers usually develop complex systems based on simplified models. For example, complex systems like Genetic Lab (Sonnleitner et al., 2012), MicroDYN and MicroFIN (Greiff et al., 2012) were developed based on the three-dimensional model comprising information retrieval, model building, and forecasting or the two-dimensional model including knowledge acquisition and knowledge application. The assessment of problem solving in PISA 2012[1] is based on a four-dimensional CPS model consisting of exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting (OECD, 2013). The use of this simplified model neglects other CPS dimensions that have been raised in the literature. For example, no researchers have developed complex systems that directly measure the skill of multiple goals balancing, nor have any researchers attempted to extract behavioral indicators from the log files of a complex system to evaluate solvers' competency in this skill.

Data Mining of the Log Data

Analysis of log data has become an important method for revealing how CPS proficiency might be improved (Herde et al., 2016; Dörner and Funke, 2017). Comparing to results data, process data contain more information about the problem solving process, which can better represent solvers' actual CPS proficiency. Some researchers have attempted to conduct preliminary analyses of log data recorded by complex systems. It is found that the "vary one thing at a time" (VOTAT) strategy is an effective problem solving strategy in complex systems (Vollmeyer et al., 1996; Wüstenberg et al., 2014). Greiff et al. (2016) collected the log data of MicroDYN and counted solvers' non-interfering observations, intervention frequency, VOTAT strategy, and time on task. In this study, solvers were instructed to manipulate the input variables and then to click on "apply" to activate it, and each round represented an intervention. The results showed that good problem solvers were not only good at using VOTAT, but also had moderately frequent intervention, highly frequent non-intervention observation, and a moderate response time; poor problem solvers tended to demonstrate little intervention or to constantly intervene.

However, most of the process indicators extracted from log data proposed by the previous research are simple indicators (e.g., time on task, intervention frequency, etc.) and fail to delve into meaningful CPS behavior sequences. To date, VOTAT is the only CPS strategy based on the behavior sequences that has been verified by log data analysis (Kröner et al., 2005; Wüstenberg et al., 2012; Müller et al., 2013; Greiff et al., 2015).

The purpose of the present study is to test and supplement the theoretical discussion of multiple goals balancing in the literature based on the meaningful CPS behavior sequences contained in the log data.

This Study

Although researchers generally believe that multiple conflicting goals is an important feature of CPS systems, not all CPS systems have this feature, such as MicroDYN (Schoppek and Fischer, 2015). If the problem solving goals of a complex system are clear and equally important or independent and non-conflicting, solvers' ability to balance multiple goals cannot be assessed. After considering the above-mentioned issues, the present study selected a proper CPS task that can meet the research purpose. Log data were analyzed to explore solvers' processes of clarifying goals, setting priorities, and making trade-offs between conflicting goals. Thereby, the study aimed at confirming the important role of multiple goals balancing in CPS and increasing emphasis placed on this skill.

With the overarching goal of exploring the role of multiple goal balancing ability in CPS, the research questions were as follows:

(1) When solving complex problems, how do solvers clarify and weigh goal priorities to achieve better scores?
(2) Which goal pursuing strategies are more productive to solving a complex problem?

MATERIALS AND METHODS

Task Descriptions

In the study, data from the "Ticket" task of the PISA 2012 CPS log data were selected as the analytical subject. This task requires solvers to purchase a ticket using a virtual ticket vending machine. Ticket type is determined by three attributes [as shown in **Figure 1A**: train network (city subway or country train), fare type (full fare or concession), and trip (daily or individual)]. Daily tickets can be used an unlimited number of times on the day of purchase. Individual tickets can be used on different dates. If the latter is selected, the number of trips must be determined (from 1 to 5). Therefore, solvers have a total of $2 \times 2 \times (1 + 5) = 24$ ticket types to choose from. When the ticket type is determined, the ticket price in zed (virtual currency unit) will be shown on the vending machine (see **Figure 1B**). Solvers then have two options:

[1]The Program for International Student Assessment (PISA) is a comprehensive large-scale international comparative test project conducted by the Organization for Economic Cooperation and Development (OECD) (OECD Indicators, 2009).

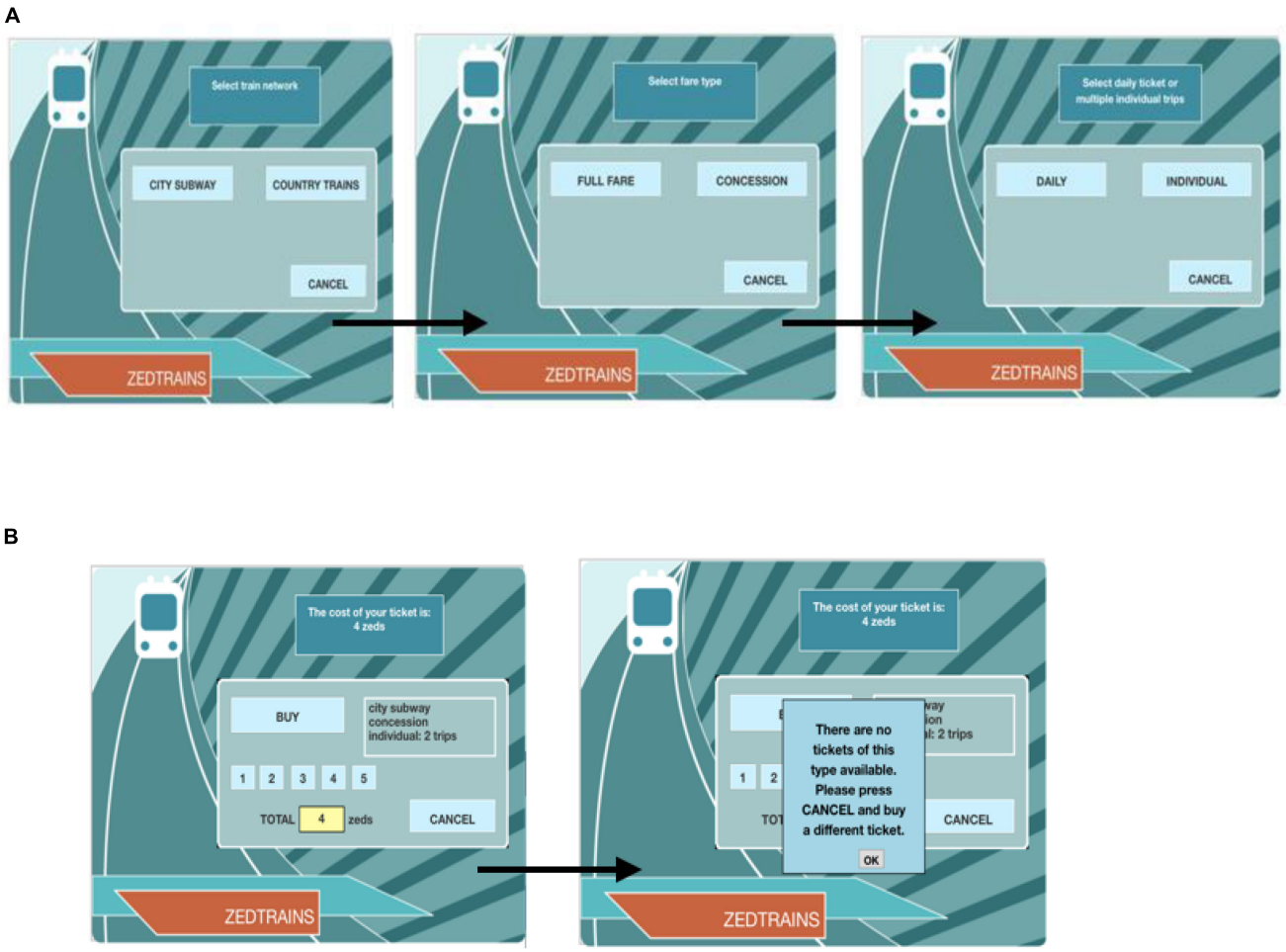


FIGURE 1 | Interface of the "Ticket" task. **(A)** Selecting stage. **(B)** Purchasing stage.

purchase the ticket or cancel the purchase and return to the initial selection screen.

The subject of the present study was the third subtask of the "Ticket" task, wherein solvers must purchase a city subway ticket that includes two trips. More than one choice is available to meet the ride need. Thus, solvers must consider the two goals of ride demand and price discount simultaneously, in order to find an optimal ticket. However, when solvers choose to purchase concession tickets, they receive notice that "There are no tickets of this type available. Please press CANCEL and buy a different ticket." Because solvers only get feedback that the ticket is not available and receive no further information about the reasons why, they must try more ticket types until they are successful in securing a ticket. Thus, the third goal of the subtask is to find an available ticket.

The optimal purchase plan meets the three goals simultaneously: it satisfies the ride demand (demand goal), has the lowest price (price goal), and is available (availability goal). Since no concession tickets meet the needs of city travel, if solvers always work toward the price goal; they will repeatedly encounter a situation in which they cannot buy a concession ticket. Therefore, solvers must buy a slightly more expensive ticket in order to perform the task. Therefore, there is a direct conflict between the price goal and the availability goal. The demand and availability goals are superior to the price goal, and solvers should give priority to achieving the first two goals before striving to fulfill the price goal.

In summary, this third subtask of the "Ticket" task contains multiple conflicting goals with varying priority. Therefore, the log data of this subtask were deemed suitable for exploring solvers' ability to balance multiple goals in CPS.

Log Data Sample

The study was based on a secondary analysis of previously collected and publicly available data. The data selected from the PISA 2012 CPS log data were de-identified. The fare system of subway ticket vending machines is common in developed countries. We selected students from six developed countries: Austria, Japan, Australia, Ireland, Germany, and France. Because it is reasonable to believe that they share similar behavioral patterns given that their countries have similar economic backgrounds, and the students could be coded with the same

coding scheme. A total of 40,217 students in these countries participated in the test, and the log data of 3,896 (9.69%) students were made available for analysis.

Partial log data are displayed in **Figure 2A**, below. Each row records an operation of an individual solver. Solvers' cancelation or purchase operations after generating a complete ticket purchase were treated as segmentation marks between plans, and operations were divided into different ticket purchase plans. For example, **Figure 2** shows the first 14 operations of a particular student. In the first six trials, the student attempted to purchase the ticket plan of (city subway, concession, individual ticket, two trips), and then canceled the purchase. This represented a complete purchase plan. The last six actions constitute the solver's second purchase plan. In addition, a purchase plan may be incomplete. For example, the first three operations in **Figure 2B** showed that he/she chose the country train, and then clicked the cancel button. Because the solver clicked CANCEL without selecting fare type and trip, this purchase plan was considered as incomplete. A total of 5,933 incomplete operations were excluded from the log data, resulting in 113,707 actions in the final data set (95.04%).

Six hundred and ninety-five students with less than two plans after the first purchase plan were excluded because there were not enough plans to analyze their competence of balancing goals. Data from the remaining 3,201 students (82.16%) were used in the final analysis. Among these students, there were 1,594 males and 1,607 females, with an average age of 15.82 years ($SD = 0.39$).

Indicators of Goal Pursuit

In order to evaluate whether problem solvers clearly identified and pursued the goals, two psychometricians and three undergraduate psychology students were invited to identify solvers' behaviors that can represent goal pursuit. If no sufficient reasons can support the behavior is goal pursuit, then the conclusion cannot be made. This means that goal pursuit was defined as a deliberate strategy that solvers pursue rather than the solvers' willingness in pursing the goal.

The following coding schemes were ultimately developed:

(1) Demand goal pursuit: If the plan met the ride demand as required by the task (city subway with two trips or an unlimited ticket), it would be coded as "1" (pursuing the demand goal); otherwise, it would be coded as "0" (abandoning the demand goal).
(2) Price goal pursuit: Solvers could only see the ticket price (the only output variable of the "Ticket" task) after completing their ticket selection. Thus, if a solver canceled the ticket after seeing the price, we would conclude that the solver was not satisfied with the price. That is, if the plan ended in "cancel," the plan would be coded as "1" (pursuing the price goal).

If the plan ended with "purchase," further judgment would be needed to determine whether the price of the plan was the cheapest among the available plans known to the solver (all previous plans except failed ones). If all of the plans that were known to the solver were failed, because there were no enough

A

FullID	cnt	schoolid	StIDStd	event	time	event_number	event_value	network	fare_type	ticket_type	number_trip
MNE0000...	MNE	0000011	00851	START_ITEM	1035.5000	1	NULL	NULL	NULL	NULL	NULL
MNE0000...	MNE	0000011	00851	ACER_EVENT	1060.8000	2	city_subway	city_subway	NULL	NULL	0
MNE0000...	MNE	0000011	00851	ACER_EVENT	1064.5000	3	concession	city_subway	concession	NULL	0
MNE0000...	MNE	0000011	00851	ACER_EVENT	1067.7000	4	individual	city_subway	concession	individual	0
MNE0000...	MNE	0000011	00851	ACER_EVENT	1069.0000	5	trip_2	city_subway	concession	individual	2
MNE0000...	MNE	0000011	00851	ACER_EVENT	1073.5000	6	Buy	city_subway	concession	individual	2
MNE0000...	MNE	0000011	00851	click	1079.0000	7	ok	NULL	NULL	NULL	NULL
MNE0000...	MNE	0000011	00851	ACER_EVENT	1089.2000	8	Cancel	NULL	NULL	NULL	0
MNE0000...	MNE	0000011	00851	ACER_EVENT	1092.7000	9	city_subway	city_subway	NULL	NULL	0
MNE0000...	MNE	0000011	00851	ACER_EVENT	1094.5000	10	concession	city_subway	concession	NULL	0
MNE0000...	MNE	0000011	00851	ACER_EVENT	1100.5000	11	individual	city_subway	concession	individual	0
MNE0000...	MNE	0000011	00851	ACER_EVENT	1101.5000	12	trip_2	city_subway	concession	individual	2
MNE0000...	MNE	0000011	00851	ACER_EVENT	1102.5000	13	Buy	city_subway	concession	individual	2
MNE0000...	MNE	0000011	00851	click	1105.8000	14	ok	NULL	NULL	NULL	NULL

B

FullID	cnt	schoolid	StIDStd	event	time	event_number	event_value	network	fare_type	ticket_type	number_trip
BRA13	BRA		13	START_ITEM	2241.6000	1	NULL	NULL	NULL	NULL	NULL
BRA13	BRA		13	ACER_EVENT	2323.3000	2	country_trains	country_trains	NULL	NULL	0
BRA13	BRA		13	ACER_EVENT	2333.3000	3	Cancel	NULL	NULL	NULL	0
BRA13	BRA		13	ACER_EVENT	2336.2000	4	city_subway	city_subway	NULL	NULL	0
BRA13	BRA		13	ACER_EVENT	2337.4000	5	full_fare	city_subway	full_fare	NULL	0
BRA13	BRA		13	ACER_EVENT	2338.2000	6	daily	city_subway	full_fare	daily	0
BRA13	BRA		13	END_ITEM	2367.3000	7	NULL	NULL	NULL	NULL	NULL

FIGURE 2 | Example of log data for the third subtask of the "Ticket" task. **(A)** Example of complete plan log data. **(B)** Example of incomplete plan log data. The variables (from left to right) are full ID code, country code, school code, student code, event type (task start, task end, or intermediate event), time point of the event, serial number of the event, value of the event input variable, type of train network, type of price, type of ticket, and number of trips.

evidences to illustrate if the solver considered the price, we applied strict rules, which was the plan had to be the cheapest among all previous failed plans. In either scenario, if the price of the plan ended with "purchase" was the cheapest, then it would be coded as "1" (pursuing the price goal), otherwise it would be coded as "0" (abandoning the price goal).

(3) Availability goal pursuit: Participants would know that some tickets could not be bought but had to be got through strategic plans only after they clicked "purchase" at the first time but did not receive tickets, so all the plans before the first purchase plan would be coded as missing values for the unawareness of the availability goal.

When the solver found that the ticket is not exist, he/she usually changed one attribute or just changed the number of trips. However, this change could not reveal the solver's deliberate efforts to avoid failure. The availability goal pursuit indicator should reflect the search strategy of available ticket type between attributes rather than within attributes. Therefore, we used three consecutive plans to code the indicator. Plan adjustment made on the basis of a solver's initial purchase failure would be recorded as "Adjustment 1–2," and the next plan adjustment would be recorded as "Adjustment 2–3." When "Adjustment 1–2" and "Adjustment 2–3" pertained to different ticket attributes, the third plan would be coded as "1," (pursuing the availability goal); otherwise, it would be coded as "0" (abandoning the availability goal). Similarly, "Adjustment 3–4" and so on would be coded according to the previous "Adjustment."

The ratios of the number of the plans pursued for each of the three goals to the total number of plans would be considered goal pursuit indicators. We developed an autoscoring program in the R language that divided the operations into mark plans, judged whether the plans were complete, and coded the goal pursuit of plans. The autoscoring program in R has been double-checked by two undergraduate psychology students.

Statistical Analysis

Cluster analysis was carried out on the goal pursuit indicators in order to identify groups with different set goal priorities. These groups were then compared to the task score used in the present study (with "1" meaning success and "0" meaning failure in problem solving) and CPS proficiency (solvers' ability estimated from their performance on the PISA 2012 CPS test) in order to explore whether groups with better priority setting showed better performance. We chose the partitioning around medoids (PAMs) algorithm for cluster analysis, because it was more robust than K-Means against noise and outliers (Kaufman and Rousseeuw, 1987, 2009). The package 'FPC' ('Flexible Procedures for Clustering') of the statistical software R 3.4.4. was used to carry out PAM algorithm (Hennig, 2007).

Regressions analyses were used to analyze the effects of goal pursuit indicators and their interactions on problem solving performance. A simple slope test of the interaction was applied to explore the impact of different combinations of the three goal pursuit behaviors on problem solving and to probe problem solvers' strategies for choosing among conflicting goals. The standardized z-scores of goal pursuit indicators and their interactions were predictors, and the task score and CPS proficiency on the PISA 2012 test were outcome variables. Regression analyses were performed using SPSS 23.0 (IBM, 2015).

RESULTS

Descriptive Statistics

The descriptive statistics are presented in **Table 1**. Results showed that the third "Ticket" subtask score was moderately correlated with CPS proficiency (r = 0.39). Three goal pursuit indicators were correlated with task score and CPS proficiency.

Cluster Analysis of Goal Pursuit Indicators

The PAM algorithm was applied to cluster the dataset. The numbers of clusters were determined at 10 when the average silhouette width reached the maximum value. **Table 2** lists the number of people in each group, within-cluster sum of squares and their average values for task score, CPS proficiency, and the three goal pursuit behaviors (columns 6–8, respectively) and their z-scores (columns 9–11, respectively).

The z-scores of the three goal pursuit behaviors were compared across groups. The results showed that the three groups with the highest task scores showed positive demand and availability goals pursuit (0.68, 0.59, and 0.69; and 1.33, 0.62, and 1.10, respectively). In other groups, at least one of the demand and availability goals was negative. This showed that successful problem solvers worked hard on both of these goals. The price goal pursuit of Groups 1, 3, and 4 were negative (−0.74, −2.73, and −0.27, respectively), and that of Group 2 was positive (1.07), which indicated that price goal pursuit was not important for solving the problem.

The task score of Group 5 (0.61) was 0.2 lower than that of Group 4 (0.81). In this group, demand goal pursuit was positive, but availability goal pursuit was negative (−0.94), that is the availability goal was ignored, and the price goal was much more highly prioritized than that in Group 4. Group 10 pursued a higher demand goal but neglected the availability goal (−2.02); it also excessively pursued the price goal (1.27), which greatly reduced the task score (0.18). This showed that pursuit of the demand goal on its own could not effectively solve the problem; rather, solvers had to also pursue the availability goal, without strongly pursuing the price goal.

The demand goal pursuit of all other groups was negative, and this affected their task scores (0.24–0.50) (e.g., Groups 6–9).

Logistic Regression of Goals Pursuit on Task Score

Logistic regression was used to test whether task score could be predicted by three types of goal pursuits and their interactions. The results of the Omnibus Test compared to the previous model were significant (χ^2 = 770.433, df = 3, $p < 0.001$; χ^2 = 112.543, $df = 2, p < 0.001$) when adding all goal pursuits and their two-way interaction terms into the model in sequence. The results showed

TABLE 1 | Descriptive statistics of variables.

Variable	*M*	*SD*	Minimum	Maximum	1	2	3	4
(1) Task score	0.63	0.48	0	1	–			
(2) CPS proficiency	544.06	86.75	214.9	802.5	0.39**	–		
(3) Demand goal pursuit (%)	90.12	13.65	0	100	0.38**	0.23**	–	
(4) Price goal pursuit (%)	68.30	23.18	0	100	−0.04*	−0.02	0.10**	–
(5) Availability goal pursuit (%)	61.91	27.79	0	100	0.25**	0.09**	−0.05**	−0.31**

*$p < 0.05$, **$p < 0.01$. The columns names 1,2,3,4 refer to the four variables in the rows.*

TABLE 2 | Cluster results of goal pursuit.

Group	Number of people	Within-cluster sum of squares	Task score	CPS proficiency	Demand goal (%)	Availability goal (%)	Price goal (%)	Demand goal (z)	Availability goal (z)	Price goal (z)
1	447	197.61	0.88	557.04	99.37	98.99	51.03	0.68	1.33	−0.74
2	411	214.50	0.88	575.15	98.19	79.16	93.13	0.59	0.62	1.07
3	73	181.26	0.82	561.16	99.57	92.45	4.88	0.69	1.10	−2.73
4	501	153.07	0.81	563.47	97.46	56.07	61.99	0.54	−0.21	−0.27
5	558	375.86	0.61	542.74	95.98	35.9	83.63	0.43	−0.94	0.66
6	369	313.75	0.50	530.97	82.29	45.59	45.27	−0.57	−0.59	−0.99
7	456	204.50	0.43	528.00	76.9	64.05	75.73	−0.97	0.08	0.32
8	117	84.88	0.24	494.11	67.32	91.23	34.78	−1.67	1.05	−1.44
9	117	307.71	0.24	498.81	48.48	76.34	78.26	−3.05	0.52	0.43
10	152	104.38	0.18	507.83	99.15	5.87	97.72	0.66	−2.02	1.27
overall	3201	2137.52	0.63	544.06	90.12	61.91	68.30	0	0	0

that the availability goal pursuit and the demand goal pursuit as well as the interactions between the availability goal pursuit and the other two goal pursuits, respectively, were significant and had effects that differ from zero.

The Nagelkerke R^2 of the final model was 0.329, demonstrating a medium effect on task score by goal pursuit. The prediction accuracies of the correct and incorrect responses were 87.2 and 54.7%, respectively, and total accuracy was 75.3% (see **Table 3**).

Demand goal pursuit ($B = 1.041$, Wald $\chi^2_{(df=1)} = 394.670$, $p < 0.001$) and availability goal pursuit ($B = 0.702$, Wald $\chi^2_{(df=1)} = 198.667$, $p < 0.001$) significantly positively predicted task score. The demand goal resulted in a larger regression coefficient than the availability goal (see **Table 4**).

The interaction between demand goal pursuit and availability goal pursuit was significant ($B = 0.277$, Wald $\chi^2_{(df=1)} = 25.187$, $p < 0.001$). The results of the simple slope test of availability goal pursuit at high/low demand goal pursuit (the mean plus/minus a standard deviation) are shown in **Figure 3**. On the whole, solvers with high demand goal pursuit scored higher than those with low demand goal pursuit. This suggested that the demand goal was more important than the availability goal. For those with high demand goal pursuit, availability goal pursuit significantly positively predicted task score to a greater extent ($B = 0.559$, $p < 0.001$). For those with low demand goal pursuit, availability goal pursuit significantly positively predicted task score to a smaller extent ($B = 0.246$, $p < 0.001$) (see **Figure 3**). If the solver pursuing the demand goal pursued the availability goal at the same time, he/she would efficiently improve task scores and solve the problem best. However, if the solver neglected the demand goal, even though pursuing availability goal was beneficial to improve task score, the improvement would be limited.

The interaction between price goal pursuit and availability goal pursuit was significant ($B = 0.319$, Wald $\chi^2_{(df=1)} = 56.857$,

TABLE 3 | Logistic regression of goal pursuit on task score.

		Correct or incorrect prediction results		Predicted accuracy (%)
		0	1	
Observation data of correct and incorrect answers	0	644	533	54.7
	1	259	1765	87.2
	Overall			75.3

TABLE 4 | Logistic regression coefficient of goal pursuit on task score.

Predictor	Model 1			Model 2		
	B	*SE*	Wald χ^2	*B*	*SE*	Wald χ^2
Demand	1.013	0.049	425.54***	1.041	0.052	394.670***
Availability	0.726	0.047	235.425***	0.702	0.050	198.667***
Availability ∗ demand				0.277	0.055	25.187***
Availability ∗ price				0.319	0.042	56.857***
Nagelkerke R^2	0.292			0.329		
The Omnibus Test (χ^2/df)	770.433(3)			112.543(2)		

***$p < 0.001$. The variables not significant was excluded in the model.*

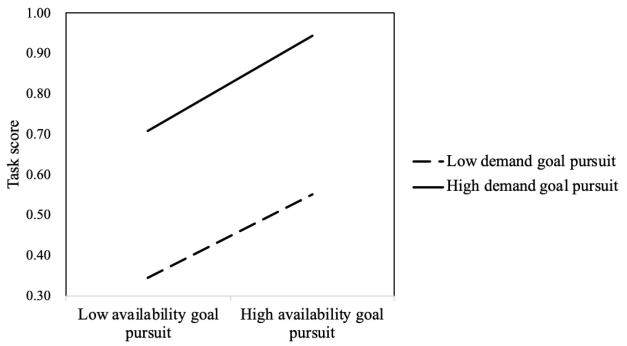


FIGURE 3 | Interaction between availability goal pursuit and demand goal pursuit.

$p < 0.001$). The results of the simple slope test are shown in **Figure 4**. On the whole, solvers who pursued the availability goal scored higher than problem solvers who did not pursue the availability goal. This suggested that the availability goal was prioritized over the price goal. For those with high availability goal pursuit, price goal pursuit significantly positively predicted task score ($B = 0.201$, $p < 0.001$). However, for those with low availability goal pursuit, price goal pursuit significantly negatively predicted task score ($B = -0.172$, $p < 0.001$) (see **Figure 4**). These results indicated that, when solvers did not work hard to pursue the availability goal and pursued the price advantage blindly, their task score dropped significantly. If solvers worked hard toward achieving the availability goal pursued price advantage, they would achieve higher scores.

Linear Regression of Goals Pursuit on CPS Proficiency

A linear regression model was built to examine goal pursuit as the predictor of CPS proficiency using the least-squares regression method. All of the model tests were significant, with $R^2 = 0.078$. Demand goal pursuit ($B = 20.700$, $t = 12.832$, $p < 0.001$) and availability goal pursuit ($B = 8.204$, $t = 5.234$, $p < 0.001$) significantly positively predicted CPS proficiency (see **Table 5**).

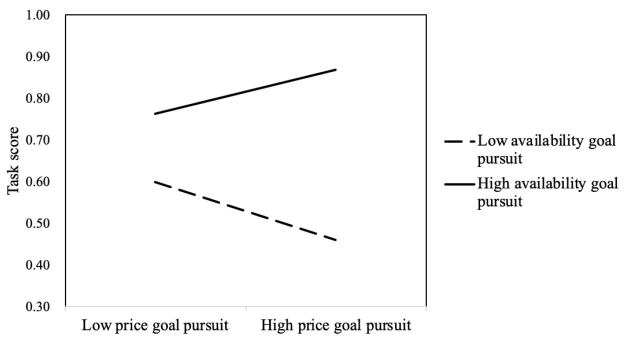


FIGURE 4 | Interaction between availability goal pursuit and price goal pursuit.

The interaction between demand goal pursuit and availability goal pursuit also significantly negatively predicted CPS proficiency ($B = 4.345$, $t = 2.637$, $p < 0.01$). The simple slope test showed that, for those with high demand goal pursuit, the availability goal significantly positively predicted CPS score ($B = 12.557$, $p < 0.05$), and for those with low demand goal pursuit, the availability goal did not significantly predict CPS score ($B = 3.810$, $p = 0.058$). Solvers who pursued the demand goal and the availability goal achieved the highest CPS proficiency scores (see **Figure 5**).

The interaction between price goal pursuit and availability goal pursuit significantly positively predicted CPS proficiency ($B = 7.329$, $t = 5.309$, $p < 0.001$). The results of the simple slope test showed that, for those with high availability goal pursuit, price goal pursuit significantly positively predicted CPS proficiency ($B = 8.411$, $p < 0.001$); for those with low availability goal pursuit, price goal pursuit significantly negatively predicted CPS proficiency ($B = -6.275$, $p < 0.01$) (see **Figure 6**).

DISCUSSION

Although CPS assessment has been researched for more than 40 years, its development as an assessment tool is far from mature. Many researchers have proposed rich theoretical models of CPS, but none of these suggested models has been fully validated (Fischer et al., 2012; Greiff and Fischer, 2013; Greiff et al., 2013).

The present study explored multiple goals balancing behavior shown in the log data of the third subtask of the "Ticket" task of PISA 2012. The purpose of the study was to test and supplement the theoretical discussion of previous studies through an analysis of measured data. The main findings are described in two sections below.

Coping Strategies in Multiple Goals Conflict Scenarios

The Importance of Clarifying Goals and Working Hard for Them

Dörner placed great emphasis on the importance of clarifying goals in a vague situation, noting that this is often the first step in solving a problem (Dörner and Wearing, 1995). The present study analyzed the log data of a CPS task to explore solvers' ability to clarify and pursue each goal. In the task, solvers needed to first clarify the demand and price goals, then identify the availability goal following failure to secure a ticket. The demand and availability goals had to be met to solve the problem, but their difficulty levels obviously differed. The result showed that, on average, 90.12% of solvers' plans met the demand goal, but only 61.91% met the availability goal. As solvers had to sacrifice the price goal to get the ticket, only 68.30% met this goal. In fact, unless goals were in direct conflict, solvers' goal-oriented actions toward as many goals as possible resulted in more effective problem solving. For example, amongst solvers who gave up the demand goal and turned to pursue the availability goal, task

TABLE 5 | Linear regression coefficients for goal pursuit on CPS proficiency.

	Model 1				Model 2			
Predictor	*B*	*SE*	*t*	β	*B*	*SE*	*t*	β
Demand	20.585	1.491	13.808***	0.237	20.700	1.613	12.832***	0.239
Availability	9.450	1.558	6.064***	0.109	8.204	1.568	5.234***	0.095
Availability ∗ Demand					4.345	1.647	2.637**	0.050
Availability ∗ price					7.329	1.380	5.309***	0.096
R^2			0.066***				0.078***	
ΔR^2							0.012***	

$^{***}p < 0.001$. The variables not significant was excluded in the model. B stands regression coefficients, and β stands standardized regression coefficients.

scores were higher for those who also pursued the price goal than for those who ignored it.

The Importance of Prioritizing Goals and Balancing Strategies

In a CPS task with conflicting goals, solvers must do more than clarify goals to solve the problem; they must also correctly prioritize the goals and execute strategies to deal with conflicting goals. For this reason, Funke (2001) strongly emphasized the cognitive process of assessing and prioritizing goals in CPS.

In the task used in the present study, because tickets that met the ride demand and were also available were not concession tickets, solvers had to buy a slightly more expensive ticket in order to solve the problem. Therefore, the demand and availability goals were superior to the price goal, and solvers should have given priority to achieving the first two goals before striving to fulfill the price goal. The regression and cluster analyses showed that if solvers tried hard to pursue both goals, they solved the problem well; if they did not work hard toward one of the goals, they were almost unable to solve the problem. Fischer et al. (2012) suggested that, when there is a conflict between goals, solvers must find a satisfactory trade-off by only partially achieving some goals. This is indeed an effective strategy; it is very important to prioritize goals and pursue the most important ones, rather than all goals.

The results also showed that the demand goal resulted in a larger regression coefficient than that of the availability goal. Further, the cluster analysis demonstrated that solvers who pursued the demand goal scored highest and solvers who gave up the demand goal scored lowest. This suggested that, although the demand goal and the availability goal were both important, the demand goal had priority over the availability goal. Indeed, in the task, since "the availability of such a ticket" was unpredictable, solvers should have first ensured that each solution met the demand goal before working toward the availability goal. Effective solvers prioritized goals in the following sequence: demand goal, availability goal, and price goal. Thus, when solvers face multiple conflicting goals in a problem, they should first consider the most important goals and gradually explore ways to balance them and then consider the secondary goals. If the process is reversed, with the important goals sacrificed for the secondary goals, problem solving will fail.

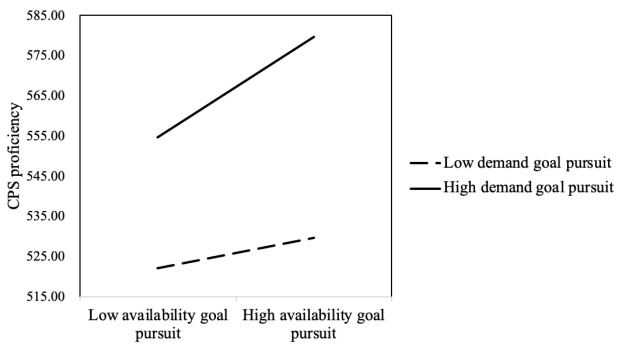


FIGURE 5 | Interaction between availability goal pursuit and demand goal pursuit.

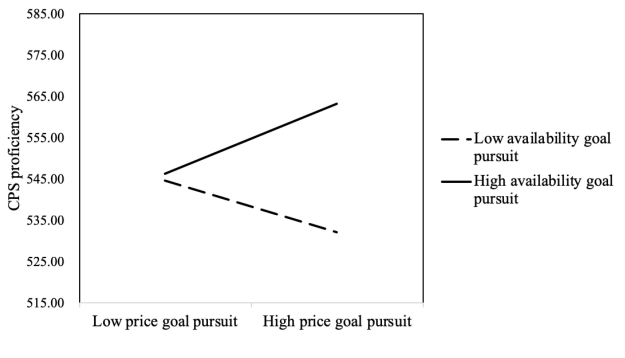


FIGURE 6 | Interaction between availability goal pursuit and price goal pursuit.

The Importance of Theoretically-Driven Log Data Analysis

In the present study, the extraction of goal pursuit indicators from the log data were based on the task feature of multiple goals conflict. The resulting indicators were meaningful and easy to facilitate the next step of analysis. This process of identifying indicators fundamentally differs from that of data-driven analysis, which commonly obtains indicators that are huge and uninterpretable. Researchers have to use complex data mining techniques such as machine learning or deep learning to analyze a large number of features or indicators in the data-driven analysis. Therefore, using an appropriate theory to analyze log data files is often a multiplier.

We used cluster and regression analysis to explore solvers' goal prioritization and balancing strategies in the context of multiple goals conflict. The results strengthen our understanding of multiple goals balancing behavior and support and complement the theoretical elaboration of multiple goals balancing in CPS. They also demonstrate that theoretically-driven log data analysis cannot only make log analysis more concise, efficient, and interpretable, but also contribute to the confirmation, improvement, and promotion of CPS theory.

Limitations

The present study analyzed the log data of a task with multiple conflicting goals of varying priorities. It demonstrated the importance and necessity of multiple goals balancing in CPS. However, there are still some shortcomings in this study, which should be supplemented and improved in further research.

Firstly, the findings of this study is only applicable to participants with three or more purchase attempts (3,201 students, 82.16% of all participants) because those with a lower number of purchase attempts were not included in the analyses. Secondly, this study used a subtask of the PISA test that was not designed originally for multiple goals balancing. Further research should develop more complex systems with multiple goals conflict in order to fully explore strategies of multiple goals balancing. These complex systems should involve an increasing number of complex goals, in order to better reveal how people solve problems effectively in complex problem scenarios. Otherwise, it should ensure that the complex system includes a more flexible "confirm submit" button. In such a scenario, every time a solver attempts a purchase plan, he/she will click a button to get information about the availability and price of the ticket. This will enable researchers to easily segment hidden plans, as the "confirm submit" button will automatically indicate the end of a complete plan in the log data.

CONCLUSION

Overall, theoretically-driven log data analysis of CPS process data can extract valuable information from messy process data, and this information can contribute to the improvement of CPS cognitive theory. Competent solvers identify and clarify goals more effectively and ensure that each step of their action plan has a clear goal orientation. More importantly, successful coping with multiple goals in tasks requires proper goal prioritizing and balancing.

ETHICS STATEMENT

The study did a secondary analysis of data, therefore: ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants was not required to participate in this study in accordance with the national legislation and institutional requirements.

AUTHOR CONTRIBUTIONS

FL designed the study, supervised the quality in the process of researching, and provided financial support. HL supervised the quality in the process of researching, revising it critically for important intellectual content, and provided financial support. XL collected and analyzed the data. YR analyzed the data. PR and DB wrote the manuscript.

FUNDING

This project was supported by the National Natural Science Foundation for Young Scientists of China (Grant No. 61807005).

REFERENCES

Blech, C., and Funke, J. (2010). You cannot have your cake and eat it, too: how induced goal conflicts affect complex problem solving. *Open Psychol. J.* 3, 42–53. doi: 10.2174/1874350101003010042

Buchner, A. (1995). "Basic topics and approaches to the study of complex problem solving," in *Complex Problem Solving: The European Perspective*, eds P. A. Frensch, and J. Funke, (Hillsdale, NJ: Lawrence Erlbaum), 27–63.

Dörner, D., and Funke, J. (2017). Complex problem solving: what it is and what it is not. *Front. Psychol.* 8:1153. doi: 10.3389/fpsyg.2017.01153

Dörner, D., and Kreuzig, H. W. (1983). Problemlösefähigkeit und intelligenz. [Problem solving ability, and intelligence]. *Psychol. Rundschau.* 34, 185–192.

Dörner, D., and Wearing, A. (1995). "Complex problem solving: toward a (computer-simulated) theory," in *Complex Problem Solving: The European Perspective*, eds P. A. Frensch, and J. Funke, (London: Psychology Press).

Fischer, A., Greiff, S., and Funke, J. (2012). The process of solving complex problems. *J. Probl. Solving* 4, 19–42.

Funke, J. (1992). Dealing with dynamic systems: research strategy, diagnostic approach and experimental results. *German J. Psychol.* 16, 24–43.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046

Funke, J. (2003). *Problemlösendes Denken [Problem-Solving Thinking]*. Stuttgart: Kohlhammer.

Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cogn. Proc.* 11, 133–142. doi: 10.1007/s10339-009-0345-0

Greiff, S., and Fischer, A. (2013). Measuring complex problem solving: an educational application of psychological theories. *J. Educ. Res. Online* 5, 38–58.

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., and Martin, R. (2013). A multitrait-multimethod study of assessment instruments for complex problem solving. *Intelligence* 41, 579–596. doi: 10.1016/j.intell.2013.07.012

Greiff, S., Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Comput. Human Behav.* 61, 36–46. doi: 10.1016/j.chb.2016.02.095

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Meas.* 36, 189–213. doi: 10.1177/0146621612439620

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Comput. Stat. Data Anal.* 52, 258–271. doi: 10.1016/j.csda.2006.11.025

Herde, C. N., Wüstenberg, S., and Greiff, S. (2016). Assessment of complex problem solving: what we know and what we don't know. *Appl. Meas. Educ.* 29, 265–277. doi: 10.1080/08957347.2016.1209208

IBM, (2015). *IBM SPSS Statistics for MAC, Version 23.0.0*. Armonk, NY: IBM.

Kaufman, L., and Rousseeuw, P. J. (1987). "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, ed. Y. Dodge, (Basel: Birkhäuser Basel), 405–416.

Kaufman, L., and Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Kröner, S., Plass, J. L., and Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence* 33, 347–368. doi: 10.1016/j.intell.2005.03.002

Müller, J. C., Kretzschmar, A., and Greiff, S. (2013). "Exploring exploration: inquiries into exploration behavior in complex problem solving assessment," in *Proceedings of the 6th International Conference on Educational Data Mining*, eds S. K. D'Mello, R. A. Calvo, and A. Olney, 336–337.

OECD Indicators (2009). *Education at A Glance*. Available at: https://www.oecd.org/education/skills-beyond-school/43636332.pdf

OECD, (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD.

Schoppek, W., and Fischer, A. (2015). Complex problem solving-single ability or complex phenomenon? *Front. Psychol.* 6:1669. doi: 10.3389/fpsyg.2015.01669

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., et al. (2012). The genetics lab: acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychol. Test Assess. Model.* 54, 54–72.

Vollmeyer, R., Burns, B. D., and Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cogn. Sci.* 20, 75–100. doi: 10.1016/s0364-0213(99)80003-2

Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving – more than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003

Wüstenberg, S., Stadler, M., Hautamäki, J., and Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technol. Knowl. Learn.* 19, 127–146. doi: 10.3389/fpsyg.2017.02145

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Ren, Luo, Ren, Bai, Li and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 28 May 2019
doi: 10.3389/fpsyg.2019.01131

Validating Test Score Interpretations Using Time Information

Lena Engelhardt[1]* *and Frank Goldhammer*[1,2]

[1] DIPF – Leibniz Institute for Research and Information in Education, Frankfurt, Germany, [2] Centre for International Student Assessment (ZIB), Frankfurt, Germany

A validity approach is proposed that uses processing times to collect validity evidence for the construct interpretation of test scores. The rationale of the approach is based on current research of processing times and on classical validity approaches, providing validity evidence based on relationships with other variables. Within the new approach, convergent validity evidence is obtained if a component skill, that is expected to underlie the task solution process in the target construct, positively moderates the relationship between effective speed and effective ability in the corresponding target construct. Discriminant validity evidence is provided if a component skill, that is not expected to underlie the task solution process in the target construct, does indeed not moderate the speed-ability relation in this target construct. Using data from a study that follows up the German PIAAC sample, this approach was applied to reading competence, assessed with PIAAC literacy items, and to quantitative reasoning, assessed with Number Series. As expected from theory, the effect of speed on ability in the target construct was only moderated by the respective underlying component skill, that is, word meaning activation skill as an underlying component skill of reading competence, and perceptual speed as an underlying component skill of reasoning. Accordingly, no positive interactions were found for the component skill that should not underlie the task solution process, that is, word meaning activation for reasoning and perceptual speed for reading. Furthermore, the study shows the suitability of the proposed validation approach. The use of time information in association with task results brings construct validation closer to the actual response process than widely used correlations of test scores.

Keywords: validity evidence based on response processes, speed, component skills, time on task effect, processing times

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Yang Jiang,
Educational Testing Service,
United States
Daniel Bolt,
University of Wisconsin-Madison,
United States

***Correspondence:**
Lena Engelhardt
engelhardtl@dipf.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *14 September 2018*
Accepted: *29 April 2019*
Published: *28 May 2019*

Citation:
Engelhardt L and Goldhammer F (2019) Validating Test Score Interpretations Using Time Information. Front. Psychol. 10:1131. doi: 10.3389/fpsyg.2019.01131

INTRODUCTION

Assessing the validity of the intended test score interpretation is critical when drawing conclusions based on test scores. Various sources of validity evidence were described in the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). One such source is evidence based on response processes. Information on the response process is now more easily available than ever before due to computer-based assessments and is also closer to the actual response process than test scores (Kane and Mislevy, 2017). For instance, information from the response process, namely processing times, has been used to support

construct interpretation for mental rotation tasks: Spatial rotation theory postulates that mental rotation should proceed similarly to physical rotations. Indeed, the physical angle of the rotation object predicted not only item difficulty but also processing time (Bejar, 1990; Embretson, 1994). However, referring to the response process can be challenging if no single process model exists and various cognitive processes are involved in the task solution, as is the case for reading and reasoning tasks (Kane and Mislevy, 2017, p. 11).

The aim of this paper is to propose a construct validation approach that uses information from the response process, namely, processing times. This approach does not require complete process models, but simply assumptions about underlying component skills of the response process that are related to automation of information processing elements. These component skills have previously been used in classical approaches like the nomothetic span approach (Embretson, 1983), which investigates relations between test scores and other constructs as validity evidence (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). Our proposed construct validation approach combines the nomothetic span approach with the relation of speed to ability.

Relations of speed to ability can be considered at different levels. The within-person level refers to the relation of effective speed to effective ability within a person, which can typically be investigated by observing a person completing a task under multiple experimental speed conditions (e.g., Goldhammer et al., 2017b). The obtained speed-ability relation is always negative as predicted by the speed-ability tradeoff (van der Linden, 2009). In contrast, our proposed validation approach is based on the speed-ability relationship at the between-person (or population) level. That is, persons complete a test without any speed manipulation. The observed speed-ability relation can be positive, zero, or negative depending on characteristics of the person and item level.

SPEED AND ABILITY

Recent technologies offer the opportunity to record not just the product of task performance, that is the task solution, but also aspects of the behavioral process, for example, by recording time information or eye movements. Previous research using process data indicates that experts' task solution process tends to differ from that of novices. Higher reading skills are associated with less and shorter fixations, longer saccades and fewer regressions (Rayner, 1998). Chess experts detect relevant information on a chess board faster, on average, than chess novices (Sheridan and Reingold, 2014). Chess experts were also four times faster than novices in a visual chess task and times for this task even correlated with the degree of expertise, measured as Elo ratings for experts or hours practicing chess per week for novice players (Sheridan and Reingold, 2017). In matrices tasks, total test scores were correlated with different task solution behaviors; they were positively related to the proportion of total time spent on inspecting matrices and negatively related to the proportion of total time spent on the response options (Vigneau et al., 2006). According to these studies, the task solution behavior of more proficient persons tends to differ across domains from that of less proficient persons, indicating differences in the cognitive processes underlying task solution.

Information about the time test-takers spend on each task is available by default nowadays in computer-based assessments. Time information carries information about the duration of the performed cognitive processes, with the limitation that the time a person spends on a given task might not only reflect task-related cognitive processes, but also non-task-related processes; for instance, it might also be affected by engagement (cf. response time effort; Wise and Kong, 2005). However, correct solutions do indicate a "successful mental process" (Hornke, 2000, p. 182), making it reasonable to interpret time as the duration of task-related cognitive processes, especially in the case of fast and correct responses. Also, rapid guessing may be associated with correct responses although not consistently and by chance, respectively. In this study, we interpret time information as the duration of the cognitive processes but consider processing times only in relation to the outcomes of this cognitive processing, namely response accuracy as an indicator of ability. Note that when referring to speed and speed-ability relationships, higher speed always means shorter time. If results are reported from studies in which response times were used, we have reversed the effects to also interpret them consistently in terms of speed.

Relations Between Speed and Task Success at the Between-Person Level

The relation between speed in a task and the probability of task success is described as the 'time on task effect' (Goldhammer et al., 2014) and has been investigated in various studies with regard to item difficulty, person ability, and different domains (Goldhammer et al., 2014, 2015; Becker et al., 2016; Naumann and Goldhammer, 2017; see also Weeks et al., 2016). The time on task effects is modeled as the (average) effect of speed in an item on task success and the effect may vary across persons and items (fixed and random effect).

The average effect of speed in a task on the probability of a correct task solution has been found to be positive, zero, or negative in different studies. This means that in some assessments, more speed was associated with a higher probability of task solution, while in other assessments, less speed was associated with a higher probability of task solution. The direction of the relation depends *first* on the kind of cognitive processes required by a task. Goldhammer et al. (2014) reported that speed in a problem-solving task was associated with a lower probability of task solution, and for a reading task with a higher probability of task solution. The different directions were explained by differences in the task demands. Problem solving was assumed to require more controlled processing; thus, higher speed in a task was associated with a lower probability of a correct task solution. Reading was assumed to be based more on automatic processes; thus, higher speed in a task was associated with a higher probability of a correct task solution. Hence, the relation between speed in a task and the probability

of a correct task solution was considered to depend on the cognitive processes performed in a task: whether they were more automatic or controlled.

The direction of the relation depends *secondly* on the interaction between person ability and item difficulty. Higher speed in a task is associated with a higher probability of task solution for more able persons working on rather easy items, and with a lower probability of task solution for less able persons working on harder items. Irrespective of whether the average effect of speed in a task on the probability of a correct task solution is positive, zero, or negative, across domains the effect varies consistently in that it is more positive, or less negative, for persons with higher abilities, compared to persons with lower abilities, and for easier items compared to harder items. Such variations have been found across domains, for instance in reading, problem solving, and reasoning (Goldhammer et al., 2014, 2015; Becker et al., 2016; Naumann and Goldhammer, 2017; see also Weeks et al., 2016; Bolsinova et al., 2017a; De Boeck et al., 2017; Chen et al., 2018). Thus, the relation between speed in a task and the probability of a correct task solution for a specific test in a certain domain can be positive for one group and negative for another. Different relations for speed in a task and the probability of a correct task solution indicate that the performed cognitive processes differ.

Theoretical Models for Speed in a Task and Task Success

A number of different – possibly domain-specific – models explain why persons differ in their cognitive processes when they solve a task depending on their proficiency.

The distinction between the two kinds of cognitive processes explaining variation in the time-on-task effect stems from dual processing theory (Schneider and Shiffrin, 1977; Schneider and Chein, 2003): Automatic processes are well learned, run in parallel, and are unaffected by cognitive load. Controlled processes require attention, run serially, and depend on cognitive load. Controlled processes can also run automatically when they are well learned (cf. Ackerman, 1988). Persons who cannot solve tasks in automatic mode need to perform controlled processes, which leads to higher cognitive load and exceeds cognitive resources at some point (Sweller et al., 1998). Persons with a high proportion of automatized processes will solve items with high speed and high accuracy and solve even hard items correctly, because working memory can handle items with a higher cognitive load in the presence of more automatized processes. Persons with fewer automatized processes will need more time for correct solutions and will not be able to solve hard items correctly, because controlled processes are impaired by cognitive load.

Becker et al. (2016) also referred to cognitive processes performed in an automatic or controlled mode for matrices tasks. They stated that for very easy tasks, task complexity is low, which leads to a low cognitive load and automatic processing. In very hard tasks, task complexity is high, which leads to a high cognitive load and controlled processing. For items that are in between, more able respondents will be able to solve them in an automatic mode, while less able respondents will need to solve them in a controlled mode. If mental load is too high, working memory operates at its capacity limit increasing the probability that the task cannot be solved correctly.

Naumann and Goldhammer (2017) explained the difference between more and less proficient readers with reference to the compensatory-encoding model (Walczyk, 1995). This model posits that automatic reading processes on the word level are important for text comprehension. Readers with less automatized processes need to compensate for this deficit by performing these processes in a controlled mode. For less proficient readers, this might still lead to a correct, but slower, solution for relatively easy items and will burden working memory. As a consequence, as cognitive load increases, working memory will at some point reach its limit, meaning that the respondent will not be able to solve a task with a high cognitive load correctly (Sweller et al., 1998).

We assume that dual processing theory, extended by cognitive load theory (Sweller et al., 1998) as described by Becker et al. (2016), explains the relation between speed in a task and response accuracy in tasks where more than a single cognitive process is involved in task solution. Complex tasks like reading comprehension or reasoning (Kane and Mislevy, 2017) are based on numerous processes, and component skills enable the automation of information processing elements. Domain-specific theories explain which component skills are important for the automatization of tasks in a given domain. For instance, the automatic lexical access to word meaning is essential for reading comprehension (cf. Perfetti, 2007).

The proposed validation approach is suitable for tasks following the dual processing theory classification. There might be tasks for which this assumption does not hold, for example tasks which mainly require knowledge, in which domain-specific component skills might not be involved in the task solution process and automatization does not provide any advantage. In some tasks, other factors like decision-making speed might also matter, and models other than the dual processing framework are more suitable for describing the relation between task speed and accuracy. For example, diffusion models are typically used for very easy two-choice response tasks requiring short response times. Such models represent the response process as an information accumulation process the proceeds until enough evidence for one of the two choices is collected (van der Maas et al., 2011).

VALIDITY APPROACH USING TIME INFORMATION

Although validating test score interpretations based on response processes (cf. American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) is closer to the actual cognitive processes than merely using assessment results, providing this kind of validity evidence can be challenging if no single process model is available (Kane and Mislevy, 2017), as is the case for reading tasks

(Kintsch, 1998), for instance. The following validation approach allows for investigating the validity of test score interpretations from assessments of complex constructs such as reading comprehension by using processing times as generic information about the response process.

The validation approach is based on the latent effect of the person variable speed on ability. Following van der Linden (2007), speed in a task (i.e., processing time observed for a task) depends on a person-specific and an item-specific component. The person-specific component (effective) speed represents inter-individual differences in time use, and it is assumed to be the same across all items (although the weighting can vary across items, see Klein Entink et al., 2009). The item-specific component describes an item's time intensity and the difference between the observed processing time and the expected processing time given the person-/item-specific components represents the residual. Just as ability is estimated based on all item responses, the speed is estimated according to time use across all items.

The effect of speed on ability is assumed to be more positive (or less negative) in a group of strong test-takers and vice versa. This kind of moderation has been shown by previous empirical studies which revealed differences in the relationship between speed (e.g., average item response time or time for a specific processing behavior) and ability (e.g., test score) depending on item difficulty or person ability (Neubauer, 1990; Knorr and Neubauer, 1996; Rayner, 1998; Vigneau et al., 2006; Sheridan and Reingold, 2014, 2017).

We assume that in complex tasks, differences in the speed-ability relation depend on automatized sub-processes, and thus on well-developed component skills. Persons with better component skills will be able to perform sub-processes in an automatized mode, while persons with weaker component skills will perform these sub-processes in a controlled mode (cf. *dual processing theory*; Schneider and Shiffrin, 1977; Schneider and Chein, 2003). The automatized mode enables fast correct solutions. In contrast, the controlled mode allows correct but slower task solutions and is affected by a high cognitive load. Consequently, solving tasks with a high cognitive load in a controlled mode will exceed working memory capacity at some point (Sweller et al., 1998), and prevent the respondent from solving these tasks successfully. Based on the dual-processing theory together with cognitive load theory it is expected that the strength of the component skill has an impact on the observed speed-ability relation. If these differences in the speed-ability relation depend on certain component skills predicted by domain-specific theories, the relation between speed and ability in a sample would be positively moderated by these component skills. In turn, *a component skill is involved in a task's response process and supports a fast and correct task solution process if it positively affects the relation between speed and ability in the target domain* (cf. **Figure 1**). *Such a positive interaction effect supports the validity of the construct interpretation assuming that a task's solution process requires these component skills.*

The presented validation approach focuses on person-level variables. It is based on the assumption that differences in the relation between effective speed and effective ability describe differences in cognitive processes. Person-level variables, for example component skills that theoretically underlie the task solution process, are assumed to refer to those differences in cognitive processes and should hence moderate the speed-ability relation.

Interpretations of Moderation Effects

Before the moderation effects are discussed, the main effects of effective speed are focused. *A positive main effect of effective speed on the effective ability* (cf. **Figure 1A**) means that persons with a higher effective speed show also a higher effective ability. This positive effect suggests that processes amenable to automation were involved in performing these tasks and that, in persons working both fast and successfully, these processes are automatized to a large extent. Lower speed and lower ability would result for those persons who compensate non-automated processes by controlled processes. While for simple tasks controlled processing would lead to correct and slower task solutions, difficult tasks would be wrong due to limited working memory capacity. *A negative main effect of effective speed on the effective ability* (cf. **Figure 1B**) means that persons with a higher effective speed have a lower effective ability and vice versa. This suggests that controlled processes were more likely to be used to perform these tasks. A lower speed along with higher ability would result for those who performed these tasks in a controlled mode. For those who did not thoroughly engage into the solution process and/or gave up at an early stage, this would result in higher speed and lower performance at the same time. Such a disengaging behavior could be driven by individual expectations of task success (cf. intentional omissions, Mislevy and Wu, 1996; see also Goldhammer et al., 2017a). Thus, potential sources of variations in effective speed may originate not only from the performed task-specific cognitive processes but also from the test-taker's meta-cognition about performing the task.

A positive moderation effect of a component skill on the speed-ability relationship means that the speed-ability relation becomes more positive or less negative. The positive moderation of the positive speed effect (see **Figure 1A**) indicates that for persons being in command of well-automatized procedures (e.g., lexical access), effective speed (e.g., in reading) more strongly reflects individual differences in ability (e.g., reading comprehension). For persons with weak component skills, however, effective speed is less (or even negatively) related to ability, since time-consuming controlled or strategic processes have to be performed to obtain a correct response. The positive moderation of the negative speed effect (see **Figure 1B**) indicates that for persons with strong component skills effective ability would be less impaired in the situation of high effective speed; put differently, such persons may afford to work fast to some extent given highly automatized elements of cognitive processing. The opposite is true for persons with weak component skills. Here, the detrimental effect of fast controlled processing is strengthened. A positive moderation effect supports that the respective component skill is associated with automatic processing in the target domain and would thus provide convergent validity evidence for the construct interpretation of the ability test score, if this component skill is theoretically assumed to underlie the task

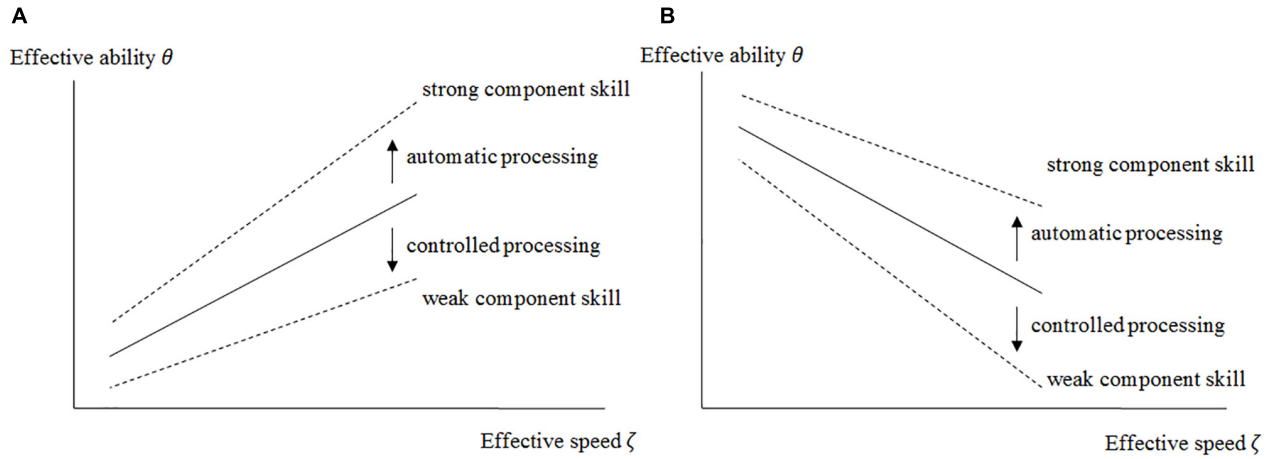


FIGURE 1 | Illustration of the relation of effective speed to effective ability at the between person level (**A:** positive relation, **B:** negative relation). The upper/lower dashed line indicates how a strong/weak construct-related component skill is expected to moderate the speed-ability relation. For persons with strong component skill it is expected to be more positive and less negative, respectively (and vice versa).

solution process. Conversely, a component skill that underlies the task solution process according to an alternative theory (cf. Kane, 2013), but does *not* moderate the speed-ability relation in a positive direction, would support the intended theory-based interpretation and provide discriminant validity evidence for the construct interpretation of the test score.

A negative moderation effect of a component skill on the speed-ability relationship would indicate that the difference in effective ability between persons with strong vs. weak component skills becomes even smaller for persons with higher effective speed (the lines in **Figure 1** would converge at high effective speeds). As described, if a component skill was responsible for performing processes in automated mode, the difference should be higher at higher speed. A negative moderation effect would therefore support that this component skill was *not* associated with automatic processing in the target domain. Instead, the advantage for persons with strong component skills in the event of a negative moderation effect is greater for those who worked slowly. In the case of a negative main effect of speed, for example, the component skill could be a resource that contributes to the correct solution of the task, particularly when working at low levels of speed.

Assumptions

Please note that the conclusions that are drawn from these moderation effects are different from those that can be drawn based on the pure main effects of component skills on target ability: The main effects of component skills on target ability describe the relationships between two ability variables being defined by item response variables (i.e., correctness of task results). The moderation effects, however, consider not only the task outcomes but also the speed at which these results were achieved, that is, the process of task completion. Consequently, a failure to find a main effect would mean that persons with high component skills do not reach higher test scores in the target ability. A failure to find a moderation effect would mean that persons with higher component skills do not show a more positive relation of ability and speed in the target construct, hence, that this component skill is not related to underlying automated processes. The time a person takes on an item also depends – especially when the item is solved incorrectly – on motivational factors like the willingness to perform the tasks as instructed (cf. test-taking effort; Wise and DeMars, 2005). This means that a respondent who takes a relatively short time across items can be indicative of a high degree of automated processes, but also low engagement throughout the test (cf. Goldhammer et al., 2017a). However, only persons with automated processes will be able to solve tasks in a domain correctly and with high speed. Hence, we assume that considering speed across tasks, together with ability, allows for interpreting differences in the relation between speed and ability in terms of differences in cognitive processing. Still, this approach depends on the assumption that most persons perform task-related processes. If many persons do not behave as intended, it will be hard to detect moderating variables for the speed-ability relationship. This can be especially problematic when many respondents perform rapid guessing.

Just as in other correlational approaches, person variables might moderate the speed-ability relation not because they are part of the assumed task solution process but because they correlate with other third variables that describe why persons work faster and rather correctly. It clearly follows from this that the construct validity of the interpretation of the component skill score is a crucial precondition for the suggested approach.

Furthermore, the extent to which the target ability and the component skill overlap is crucial for our approach: The component skill should describe a relevant sub-process of the cognitive processes in the target construct that can be automated. If the component skill represents only a rather irrelevant aspect of the cognitive processes in the target ability, performing this process automatically will not significantly moderate the speed-ability relationship in the target construct. If the component skill

represents the cognitive processes in the target ability to a very large extent, the component skill would strongly predict the target ability, making an interaction effect between target speed and component skill unlikely. Furthermore, our approach is based on differences in cognitive processes as represented by the speed-ability relation in the target construct and in the respective component skills. Therefore, the sample needs to include such differences. Heterogeneous samples, such as those in the PIAAC study, are likely to meet these preconditions.

Classification of the Validity Approach

In terms of classical validation approaches, the proposed validation approach of examining relations with component skills can be seen as similar to the nomothetic span approach (Embretson, 1983) or collecting validity evidence based on relations with other variables (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). However, our approach does not focus on relating test scores of the target construct (e.g., reading) and component skills (e.g., lexical access) to each other. Instead, we analyze whether component skills moderate the relation between speed and ability. If a component skill that is theoretically assumed to be elicited by the task actually moderates the relation between speed and ability (cf. **Figure 1**), it supports the notion that this component skill is indeed involved in the response process of this task. Consequently, such a result provides validity evidence for the construct interpretation of the test scores based on response processes (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). The nomothetic span approach focuses on individual differences rather than on differences between items (cf. *construct representation approach*; Embretson, 1983). This is also why our suggested approach focuses on the person-specific time component that underlies response times in all items (van der Linden, 2007) rather than on the time a person takes on single items.

RESEARCH QUESTIONS

The overall empirical goal of this study is to test the proposed validity approach based on processing times. Two cognitive constructs, reading comprehension and reasoning, were selected to investigate the validity of the construct interpretation of related test scores. The literacy competence test from the Program for the International Assessment of Adult Competencies (PIAAC; OECD, 2016) was used for assessing reading comprehension and the Number Series Test (McArdle and Woodcock, 2009) for assessing reasoning. Competencies such as reading comprehension are assumed to matter for the handling of very specific situations, whereas general cognitive skills can be applied to a wide range of situations (Klieme et al., 2008).

The following sections describe component skills underlying reading and reasoning that are thus critical for automated processing. They are assumed to moderate the speed-ability relationship in a positive direction. A positive moderation by the component skill that theoretically underlies the task solution process would provide convergent validity evidence. Kane (2013) argues that test score interpretations should be put to the test. In traditional approaches, such as analyzing relations to other variables (e.g., American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014), this type of challenging analysis would be done to provide discriminant validity evidence for the construct interpretation. In a similar vein, in our approach, discriminant validity evidence would be provided if a component skill that is believed to underlie the task solution process according to alternative theories does not moderate the relation between speed and ability in a positive direction.

Reading

Literacy items in PIAAC are assumed to involve "a range of skills from the decoding of written words and sentences to the comprehension, interpretation, and evaluation of complex texts" (OECD, 2016, p. 18). Kintsch (1998) describes reading as the interplay of bottom-up and top-down processes in his construction-integration model. Bottom-up processes are performed to process words in order to build a propositional representation of the text. Then, knowledge is integrated in top-down processes to construct a situation model. One bottom-up-process that is theoretically involved in reading (Kintsch, 1998) and also an empirical predictor of reading comprehension (Perfetti, 2007) is the activation of word meanings. Word reading is a process that can be automatically performed (Augustinova and Ferrand, 2014). Thus, the relationship between speed and ability in reading is assumed to be influenced by the extent to which readers activate word meanings from the text in an automatic or controlled mode and should be more positive for automatic activation. Knowing more words might prevent a person from encoding letters separately or from guessing the meaning from the context. In the context of cognitive load (Sweller et al., 1998), not knowing words might burden working memory capacity and might not only prevent faster task solution in easy items but might even more so prevent correct task solution on harder items.

If the relation between speed and ability in reading is more positive among persons with greater word meaning activation, this provides convergent validity evidence for the construct interpretation, because it indicates that the solution process in reading tasks requires reading-specific component skills (Hypothesis 1a).

The wide range of situations to which general cognitive skills can be applied bring them into play as an alternative interpretation of competence scores. Whether competence tests used in large-scale assessments are also based on general cognitive skills and to what extent they represent the outcomes of learning processes have been investigated in numerous studies based on item scores (Brunner, 2005; Nagy, 2006; Rindermann, 2006; Prenzel et al., 2007; Baumert et al., 2009; Rindermann and Baumeister, 2015; Saß et al., 2017). The construct interpretation

should be challenged through alternative interpretations that see general cognitive skills as also involved in literacy items. An important component skill of general cognitive skills is perceptual speed (e.g., Vernon et al., 1985).

If the relation between speed and ability in reading is not more positive among persons with higher perceptual speed, this provides discriminant validity evidence for the construct interpretation, as it indicates that the solution process in reading tasks does not involve reasoning-specific component skills (Hypothesis 1b).

Reasoning

Fluid reasoning is assumed to be a good indicator for general cognitive skills (Vernon, 1965). Reasoning requires controlled mental operations to solve novel problems. Deductive/inductive reasoning and quantitative reasoning are considered to belong to the broad category of fluid reasoning alongside other constructs. Fluid reasoning is required to accomplish cognitively complex tasks and is hence based on various elementary cognitive processes (McGrew, 2009). The elementary cognitive processes underlying reasoning processes include working memory capacity and perceptual speed (Vernon et al., 1985; Neubauer, 1990; Schweizer and Koch, 2002; Altmeyer et al., 2009). Perceptual speed describes the ability to perform easy and elementary cognitive tasks automatically, and is one of the specific, narrower abilities involved in processing speed (McGrew, 2009). Higher perceptual speed can lead to faster solutions on easy fluid reasoning tasks and correct solutions on demanding tasks, because it allows a greater amount of information to be processed despite limited working memory capacity. Slow processes, in contrast, may lead to a loss of information and a slow or even incorrect task solution (Jensen, 1982; Vernon et al., 1985; Sweller et al., 1998).

If the relation between speed and ability in reasoning is more positive among persons with greater perceptual speed, this provides convergent validity evidence for the construct interpretation, as it indicates that the solution process in reasoning tasks requires reasoning-specific component skills (Hypothesis 2a).

Although it is has been shown that schooling can affect reasoning (Ceci and Williams, 1997; Guill et al., 2017), we assume that highly specific component skills of education-related competencies, such as word meaning activation, do not moderate the relation between reasoning speed and reasoning ability in a positive direction.

If the relation between speed and ability in reasoning is not more positive among persons with higher word meaning activation, this provides discriminant validity evidence for the construct interpretation, as it indicates that the solution process in reasoning tasks does not involve component skills that are related to reading (Hypothesis 2b).

MATERIALS AND METHODS

Sample

This study is based on data from the PIAAC-L study (GESIS – Leibniz Institute for the Social Sciences, German Socio-Economic Panel (SOEP) at DIW Berlin and LIFBI – Leibniz Institute for Educational Trajectories, 2017; Rammstedt et al., 2017). In PIAAC-L, all German respondents from the PIAAC study were re-contacted in 2015 and received either (randomly selected) PIAAC literacy items (N = 1423) or other instruments. One year later, respondents from the 2015 assessment were re-contacted again and all received measures from the Socio-Economic Panel (SOEP; Schupp et al., 2008), the Symbol-Digit Test (Schupp et al., 2008) and a multiple-choice vocabulary intelligence test (Lehrl, 2005). Some of those respondents were also selected to complete the Number Series Test in 2016 (McArdle and Woodcock, 2009; Engelhardt and Goldhammer, 2018) based on the instruments they had received in 2015. The data set used for the analyses in this study consists of N = 1588 respondents. Of those, N = 744 respondents completed the PIAAC literacy items and the Number Series, N = 679 only the PIAAC literacy items, and N = 165 only the Number Series. In the whole data set, respondents were M = 42.41 years old (SD = 13.72; Min = 19, Max = 69) on average in 2015, and 48.55% were male (51.45% female).

Measures

The PIAAC literacy test[1] included a total of 49 dichotomously scored items and is assumed to assess reading competence (cf. OECD, 2016). PIAAC is an OECD study that aims to assess adults' competencies in literacy, numeracy, and problem-solving in an international comparison. These "key information-processing competencies" (OECD, 2016, p. 16) are necessary, for example, to participate in social life or the labor market. In addition, they are also assumed to be transferable to different situations and learnable.

In PIAAC, literacy is defined as "understanding, evaluating, using and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential" (OECD, 2016, p. 19). The reading tasks can contain a continuous text, a non-continuous text (e.g., form), or both, and can even contain more than one text. Each reading task also requires one of three cognitive strategies (access and identify, integrate and interpret, or evaluate and reflect), and can address topics related to work, personal matters, society and community, or education and training. The example item (see footnote 1) "preschool rules" requires the cognitive strategy "access and identify" and takes up a personal topic. In this item, test-takers have to take, from a text, by what time, at the latest, children have to arrive at preschool. The text contains nine bullet points from one or two short sentences and each bullet point describes a rule. Two of these rules contain time information, which makes it necessary to also read the text in which the time information is embedded in order to solve the item. Regardless of the type of text (continuous texts, non-continuous texts, mixed texts, or multiple texts), cognitive strategy (access and identify, integrate and interpret, or evaluate and reflect), or topic (work-related, personal, community and society, or education and training), test takers must comprehend text in each PIAAC literacy item. It is assumed that word activation skills support the task solution

[1]For example items, see www.oecd.org/skills/piaac/Literacy%20Sample%20Items.pdf (accessed October 26, 2018).

in terms of speed and accuracy, as the fast retrieval of word meaning from memory supports the correct semantic integration of words and, in turn, the comprehension of text. In addition, the automatic retrieval of words from memory reduces cognitive load and does not compromise the cognitive processes that are required for task solution. For this reason, across all items, word meaning activation skill is expected to moderate the relationship between speed and ability in reading. The literacy test was administered in a two-stage adaptive test design (Kirsch and Yamamoto, 2013, p. 10), with 9 out of 18 items administered in the first items and 11 out of 31 items administered in the second stage (OECD, 2016). The assessment had no time restriction. Respondents had an increased probability of receiving a testlet appropriate for their skill level depending on three variables (education level, native speaker, passing score on computer-based assessment core tasks; for more details, see OECD, 2016). Two separate latent factors were modeled on the basis of the literacy items: reading ability and reading speed. Fitting 2-parameter IRT models in Mplus (Muthén and Muthén, 2015) based on N = 1423 respondents revealed that all literacy item response variables loaded significantly on a joint latent ability factor [standardized loadings (variance of the latent variable fixed to one): M = 0.60; SD = 0.11; Min = 0.37; Max = 0.82; see the **Appendix** for further information]. Note that the MLR estimator (maximum likelihood estimation with robust standard errors), which was used to deal with the missing data structure, does not provide absolute model fit information. We refrain from presenting additional information on item fit given that both the PIAAC literacy test and the Number Series Test are well-established and trialed instruments. The reading speed factor was obtained as follows: item-level processing times, that is the total time a person spent on an item including editing or reviewing their answer, were at first log-transformed and then subjected to a confirmatory factor analysis. On average, respondents spent M = 72.87 s on an item (SD = 28.66; Min = 25.59; Max = 129.04). The log-transformed processing times for all items loaded on a joint latent factor (cf. **Figure 2**, which describes the model for data analyses) representing person-specific time use (standardized loadings: M = 0.59; SD = 0.07; Min = 0.46; Max = 0.75; see the **Appendix** for further information). For easier interpretation in terms of reading speed, we switched the polarity of the processing time results (from positive to negative and vice versa), such that higher values indicate greater speed and thus less time spent on an item. The model fit of the measurement model for the processing times was acceptable (CFI = 0.879; TLI = 0.872; RMSEA = 0.029; SRMR = 0.072).

Fifteen dichotomously scored Number Series items (McArdle and Woodcock, 2009) were used to measure fluid reasoning ability, because quantitative reasoning is a specific ability within the broader domain of fluid reasoning (McGrew, 2009). Each of the fifteen number series consisted of between four and seven numbers. One number was missing in each of the first 14 number series. Only in the fifteenth number series were two numbers missing. The missing numbers were either located at the beginning, in a middle position, or at the end of the number series. After 16 min, test-takers were navigated not to the next number series but to the end of the assessment (cf. Engelhardt and Goldhammer, 2018). This only happened to seven respondents. Two separate latent factors were modeled on the basis of the number series items: reasoning ability and reasoning speed. A 2-parameter IRT model was fitted, with all items loading significantly on a joint latent ability factor [standardized loadings (variance of the latent variable fixed to one): M = 0.69; SD = 0.11; Min = 0.52; Max = 0.87; N = 909 respondents; see the **Appendix** for further information]. In order to test the measurement model for a joint latent reasoning speed factor, the total processing times, that is, the time in seconds a person spent on a single item including editing or reviewing their answer (M = 28.40; SD = 26.33; Min = 5.36; Max = 110.61), were first log-transformed. Then, a confirmatory factor analysis of the log-transformed processing times was conducted to model a latent factor (cf. **Figure 2**) representing person-specific time use. The polarity of the processing time results was changed for the results presented in the tables. That is, higher values indicate less time spent on an item (higher speed). Due to an unacceptable model fit (CFI = 0.707; TLI = 0.658; RMSEA = 0.129; SRMR = 0.103), correlations between items were allowed back into the model one-by-one according to the modification indices until the model fit reached an acceptable level (CFI = 0.966; TLI = 0.948; RMSEA = 0.050; SRMR = 0.043). Correlations were only added between very easy items and between hard items. This indicates that not only did persons differ in their general reasoning speed, but that there were also differential differences in reasoning speed for easy and harder items. In the final model, all items loaded on the reasoning speed factor (standardized loadings: M = 0.51; SD = 0.15; Min = 0.16; Max = 0.68; see the **Appendix** for further information).

For the two component skills, word meaning activation and perceptual speed, one latent factor was modeled per test. Perceptual speed is frequently assessed using the Symbol-Digit Test[2] (Ackerman, 1988). This test was also part of the Socio-Economic Panel (Schupp et al., 2008) and was thus used in

[2]For an example, see p. 10 https://www.diw.de/documents/publikationen/73/diw_01.c.570984.de/diw_ssp0339.pdf (accessed October 26, 2018).

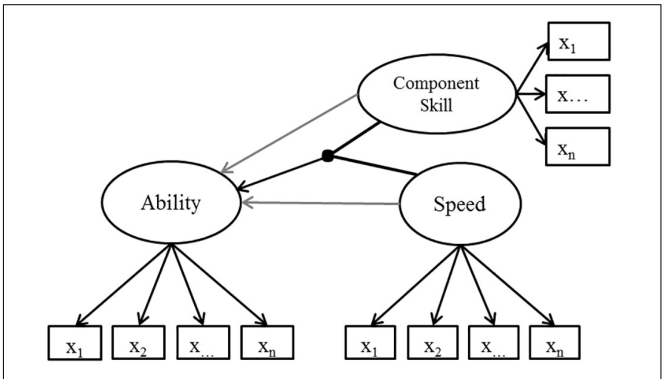


FIGURE 2 | Structural model to investigate how the component skills [perceptual speed (PS) and word meaning activation (WMA)] moderate the relation between speed and ability in reading and reasoning tests (latent correlations are omitted).

this study. To complete the test, respondents recoded symbols into digits according to a legend. The legend consisted of nine symbols corresponding to the digits one to nine. Participants had 90 s to recode as many symbols as possible one after the other on the computer. Of these 90 s, the total number of correctly coded digits was recorded for three time intervals of 30 s each, known as parcels. A latent perceptual speed factor was modeled on the basis of these three parcels, leading to a fully saturated and therefore perfectly fitting model (see the **Appendix** for further information).

The Multiple-Choice Vocabulary Intelligence Test[3] (Lehrl, 2005) from the Socio-Economic Panel (Schupp et al., 2008) was used to measure a component skill specific to reading. This test requires respondents to identify the existing word in 37 word groups of increasing difficulty. Each group consists of five potential German words, four of which are fictitious. Task completion time was not restricted (Zabal et al., 2016). Kintsch (1998) describes this skill as word meaning activation. Because we wanted subsequent analyses to be based on relations among latent variables, a single indicator model was used to estimate a latent variable for word meaning activation. In order to ensure model identification, the single indicator variable (i.e., the number of correct answers across all 37 items) was standardized, the variance of the latent factor was fixed to 1, and the factor loading was fixed to the root of 0.76 (see the **Appendix** for further information). This loading served as a proxy for the estimated reliability of this test and is based on its correlation with a similar test (Satzger et al., 2002).

Data Analyses

Data was analyzed using Mplus (Muthén and Muthén, 2015). As the literacy items were administered in an adaptive design, the three context variables involved in testlet selection (education level, native speaker, passing score on computer-based assessment core tasks) were included as correlated variables for Hypotheses 1a and 1b in order to make it justifiable to assume that the not-administered items were missing at random (MAR; cf. Enders, 2010). The MLR estimator (maximum likelihood estimation with robust standard error) can be used to test structural equation models using categorical items. It also has the advantage of being able to consider all information under the missing at random (MAR) assumption despite the presence of missing data, making it suitable for the present study. Structural equation models for each domain were tested to analyze the hypotheses. First, only main effects for reasoning/reading speed were modeled as predictors of reading/reasoning ability (baseline model). To test the hypotheses, latent interaction terms (cf. latent moderated structural equations; Klein and Moosbrugger, 2000) of reasoning/reading speed and perceptual speed and of reasoning/reading speed and word meaning activation were included in the model for literacy (Hypotheses 1a and 1b) and number series (Hypotheses 2a and 2b). These models also contained the main effects of perceptual speed and word meaning activation. The model for the hypotheses is visualized in **Figure 2**.

[3]For an example, see p. 11 https://www.diw.de/documents/publikationen/73/diw_01.c.570984.de/diw_ssp0339.pdf (accessed October 26, 2018).

RESULTS

To test the hypotheses, we analyzed whether the relation between speed and ability was positively moderated only by the domain-specific component skill, meaning that the relation between speed and ability is assumed to be more positive for persons with higher domain-specific component skills. The results are presented separately for reading (**Table 1**) and reasoning (**Table 2**).

Reading

Reading speed and reading ability were not significantly related in the sample (**Table 1**: $\beta = -0.08$, $p = 0.053$). When including perceptual speed in the analyses (cf. Results for Hypothesis 1b), the relation between reading speed and reading ability became more negative (**Table 1**: $\beta = -0.28$, $p < 0.001$). This might be because reading speed and the component skill perceptual speed were positively correlated ($r = 0.37$; $p < 0.001$).

Word meaning activation, as a domain-specific component skill, was positively associated with reading ability (main effect: $\beta = 0.49$, $p < 0.001$). As expected, the relation between reading speed and reading ability was positively moderated by word meaning activation (Hypothesis 1a; interaction effect: $\beta = 0.12$, $p = 0.001$). This indicates that persons with higher word meaning activation skills had higher reading ability scores (main effect) and those who worked faster had, in addition, higher reading ability scores, compared to those with lower word meaning activation skills (interaction effect; see **Figure 3**, Hypothesis 1a), because the relation between reading speed and reading ability was more positive for these persons, providing convergent validity evidence that word meaning activation is important for automated reading processes.

Perceptual speed, as a component skill of fluid reasoning, was also positively associated with reading ability (main effect: $\beta = 0.49$, $p < 0.001$). As expected, perceptual speed did not moderate the relation between reading speed and reading ability in a positive direction (Hypothesis 1b; interaction effect: $\beta = -0.07$, $p = 0.023$), but did so in negative direction. This indicates that persons with higher perceptual speed had higher reading ability scores (main effect) compared to persons with lower perceptual speed. But this advantage of having higher perceptual speed was smaller for those who worked faster (interaction effect, see **Figure 3**, Hypothesis 1b). The fact that no positive moderation effect was found provides discriminant validity evidence. Further interpretations of the negative interaction effect are presented in the discussion section.

Reasoning

Reasoning speed and reasoning ability were positively related (**Table 2**: $\beta = 0.27$, $p < 0.001$) in this sample. The component skill perceptual speed was positively correlated with reasoning speed ($r = 0.48$; $p < 0.001$), which led to a less positive but still significant relation between reasoning speed and reasoning ability (**Table 2**: $\beta = 0.11$, $p = 0.030$) when the component skill perceptual speed was included in the model.

As expected, perceptual speed, a domain-specific component skill, was positively related to reasoning ability (main effect: $\beta = 0.35$, $p < 0.001$) and moderated the relationship between

TABLE 1 | Interaction effects of component skills – word meaning activation (WMA) and perceptual speed (PS) – and reading speed on reading ability.

	Baseline model				Hypothesis 1a (convergent)				Hypothesis 1b (discriminant)			
	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Reading speed	−0.08[1]	0.04	−1.94[1]	0.053	−0.04[1]	0.04	−1.02[1]	0.310	−0.28[1]	0.04	−7.18[1]	<0.001
WMA					0.49	0.05	10.41	<0.001				
WMA × Reading speed					0.12[1]	0.03	3.89[1]	<0.001				
PS									0.49	0.03	16.32	<0.001
PS × Reading speed									−0.07[1]	0.03	−2.27[1]	0.023

Baseline model: N = 1423; Hypotheses 1a and 1b: N = 1587. [1]Data analysis was based on processing times (higher values indicating more time on an item). In order to interpret the obtained latent variables as "speed" (higher values indicating less time on an item), we inverted the polarity of the β- and z-values for speed in the table.

TABLE 2 | Interaction effects of component skills – word meaning activation (WMA) and perceptual speed (PS) – and reasoning speed on reasoning ability.

	Baseline model				Hypothesis 2a (convergent)				Hypothesis 2b (discriminant)			
	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Reasoning speed	0.27[1]	0.04	6.16[1]	<0.001	0.11[1]	0.05	2.17[1]	0.030	0.24[1]	0.04	6.15[1]	<0.001
WMA									0.46	0.04	10.78	<0.001
WMA × Reasoning speed									0.06[1]	0.04	1.43[1]	0.152
PS					0.35	0.04	8.36	<0.001				
PS × Reasoning speed					0.10[1]	0.04	2.50[1]	0.012				

Baseline model: N = 909; Hypotheses 2a and 2b: N = 1410. [1]Data analysis was based on processing times (higher values indicating more time on an item). In order to interpret the obtained latent variables as "speed" (higher values indicating less time on an item), we inverted the polarity of the β- and z-values for speed in the table.

reasoning speed and reasoning ability in a positive direction (Hypothesis 2a; interaction effect: $\beta = 0.10$, $p = 0.012$). This indicates that persons with higher perceptual speed had higher reasoning ability scores (main effect) and those who worked faster had, in addition, higher reasoning ability scores compared to those with lower perceptual speed for (interaction effect; see **Figure 3**, Hypothesis 2a), because the relation between reasoning speed and reasoning ability was more positive for these persons. This provides convergent validity evidence that the component skill of perceptual speed is important for automated processes in reasoning tasks.

Word meaning activation was also positively associated with reasoning ability (main effect: $\beta = 0.46$, $p < 0.001$), but did not positively moderate the relation between reasoning speed and reasoning ability (Hypothesis 2b; interaction effect: $\beta = 0.06$, $p = 0.152$), providing discriminant validity evidence. This suggests that persons with higher word meaning activation skills had higher reasoning ability scores but this difference did not increase for those who worked faster (interaction effect; see **Figure 3**, Hypothesis 2b). Word meaning activation was as expected not important for automated cognitive processes in reasoning tasks.

DISCUSSION

Main Findings

The results provided both convergent and discriminant validity evidence for the construct interpretation of reasoning and reading ability scores. Convergent evidence was provided because the relations between speed and ability were more positive among persons with stronger domain-specific component skills (word meaning activation for reading and perceptual speed for reasoning). This means that people with stronger component skills that were theoretically assumed to be relevant for the target ability did indeed obtain higher ability scores and this advantage was even more explicit when they worked faster, which supports that the component skills were indeed involved in automated task solution processes. Discriminant validity evidence was provided because the component skills that were assumed to be irrelevant for automated task solution processes in each domain did not moderate the speed-ability relation in a positive direction. Persons with higher scores on the irrelevant component skills and who worked faster did not show relative higher ability compared to persons with lower scores on the irrelevant component skills.

Interpretation of Empirical Findings

Although empirical support was found for the hypothesized moderation effects, the two component skills (word meaning activation and perceptual speed) were positively associated with both abilities examined (reading and reasoning). One could ask whether these correlations between test scores for one ability and component skills for the other ability call the validity of the intended construct interpretation into question. We argue that this is not the case, because competencies and general cognitive skills are assumed to be related, for instance because schooling may affect reasoning, and there is ongoing discussion about the extent to which those skills can be separated (Brunner, 2005; Nagy, 2006; Rindermann, 2006; Prenzel et al., 2007; Baumert et al., 2009; Rindermann and Baumeister, 2015; Saß et al., 2017). Hence, it is not surprising that component skills for reading correlate with tests scores

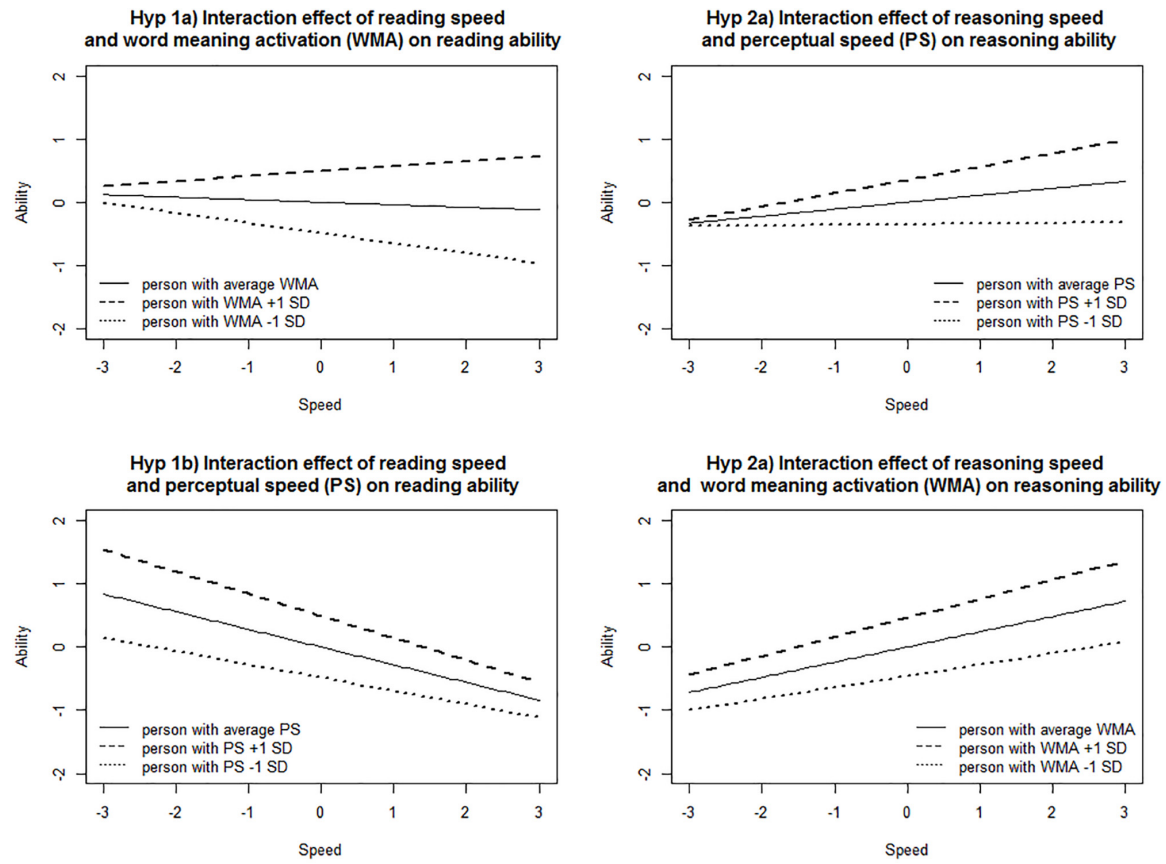


FIGURE 3 | Moderation effects of perceptual speed and word meaning activation for the relation between speed and ability in the reading and reasoning tests.

for reasoning tasks and vice versa, and this does not call the validity of the test score interpretation into question. Perceptual speed is considered to be a general and domain-unspecific skill. According to cognitive load theory (Sweller et al., 1998), fast processing might reduce cognitive load in complex tasks, which could in turn help with task solution even if a task's cognitive load is high. Hence, persons with higher perceptual speed might also have advantages in tasks from other domains. Moreover, according to Cattell's (1963) investment theory, fluid intelligence (e.g., reasoning) is important for the acquisition of crystallized abilities (e.g., reading). The reverse is also posited: Educational processes are assumed to affect fluid intelligence (Ceci and Williams, 1997; Guill et al., 2017). Hence, such correlations can actually be expected on the basis of empirical findings and theoretical assumptions. The advantage of the suggested validity approach is that it helps distinguish the roles of different component skills for different domains by determining whether they are related to fast and correct task solution processes or not. We conclude that not only ability score differences should be focused on when investigating the validity of the construct interpretation of ability scores, but also differences in the speed-ability relation and how they are affected by component skills considered relevant for the construct.

Interpreting the speed-ability relation in terms of construct-related response processes requires ruling out alternative explanations. For instance, test-wiseness could explain the speed-ability relation, as greater test-wiseness presumably makes test-takers both faster and more successful. However, one would assume that test-wiseness has a consistent impact across domains. This is obviously not the case given the differences between reasoning and reading in the relation between speed and ability. Moreover, the pattern of interaction effects for construct-related component skills speaks against this assumption as well.

Unpredicted but interestingly, the relation between reading speed and reading ability was also moderated by perceptual speed, but in a negative direction. This does not contradict our hypothesis, which was that perceptual speed does *not* moderate the speed-ability relation in a *positive* direction. The negative interaction effect does not mean that perceptual speed is involved in performing reading processes in the automated mode (what would have been supported by a positive moderation effect, as was the case for word meaning activation). It rather indicates that higher perceptual speed is associated with higher ability for persons working more slowly. Perceptual speed could have functioned as a resource, for instance to compensate for non-automatized processes. Word identification during reading can be based on different processes (cf. Perfetti, 2007): on extracting

word meanings or decoding single letters. While word meaning activation might matter for the first process, perceptual speed might matter for decoding when the meaning cannot be directly retrieved. Higher perceptual speed might thus increase the probability of correct item solution when respondents invest time into decoding single words in order to understand the text as well as possible.

What do the results mean for the pursued research questions for reading? The results first indicate that solving PIAAC literacy items is rooted in typical reading-specific processes like word meaning activation. Word meaning activation was not only predictive of reading ability (main effect) but was also related to fast and correct solutions and thus involved in the response process (interaction effect). Secondly, higher perceptual speed, a component skill of reasoning, predicted reading ability (main effect). The unexpected negative interaction effect indicated that the speed-ability relation differed depending on a persons' perceptual speed. The results suggested that perceptual speed was not related to automated cognitive processes but may, rather, represent a resource for compensating behavior. This result is highly interesting because it indicates that perceptual speed plays a different role in PIAAC literacy items than reading-specific component skills. Thus, examining the relations between component skills and speed-ability relations in tasks from a given domain can reveal whether and how component skills are involved in the solution process for complex tasks like reading.

The results for reasoning need to be interpreted in light of the general reasoning speed factor we modeled. Correlations between items were allowed in order to achieve an acceptable model fit for the measurement model. Correlations between easy items and between hard items were necessary, but not between items with a medium level of difficulty. This indicates that speed in reasoning tasks can be multidimensional, perhaps because strategies might change from easy to hard items. In this study, the latent reasoning speed factor is dominated by the speed respondents exhibited on items of medium difficulty. As a consequence, the results should primarily be interpreted in this respect, that is, perceptual speed might particularly play a role for items of medium difficulty. Other component skills might potentially be involved in easier or harder items. Hence, it would be interesting to investigate in future studies what other component skills are crucial in reasoning items depending on the level of item difficulty.

Regarding the research question on interaction effects, it can be concluded for reasoning that the component skill of perceptual speed seems to be important for reasoning tasks at least at a medium level of difficulty, while word meaning activation as is not related to automated processes when solving reasoning items.

Limitations of the Study

We applied our validation approach to a subsample of the German PIAAC sample that was re-assessed within a longitudinal setting. This longitudinal setting was highly advantageous because it meant that the same individuals completed reading tasks, reasoning tasks, and both component skills tasks (i.e., word meaning activation and perceptual speed). However, there were also some limitations. First, each construct was only operationalized with one measure, which means that the latent variables we obtained may also reflect properties of the measure (e.g., the ability to deal with numerical material in the case of reasoning). Second, the selection of domain-specific component skills was limited. For instance, in the case of literacy, the component skill semantic integration (Richter and Naumann, 2009) would have been another candidate to affect the speed-ability relation. Third, the PIAAC study is a low-stakes assessment study, and test-takers' motivation might have varied. Low motivation elicits cognitive processes that may be unrelated to the task solution. By contrast, the described approach is based on the assumption that speed in an item can be interpreted as the duration of a person's cognitive processes and differences in the speed-ability relation as an indicator for differences in cognitive processing. Fourth, test-takers in the PIAAC-L assessment completed the PIAAC reading items in both 2012 and 2015. Although there was some time in between, test-takers might have gotten used to these kinds of tasks and have remembered seeing the same items 3 years ago. Thus, carry-over effects could have affected the results.

Strengths and Limitations of the Validation Approach

In our view, the strength of the proposed validation approach is that information from the response process is used to support the validity of construct interpretations. Relations between component skills and the two ability tests (cf. main effects) did not reveal any differences between reading and reasoning, because word meaning activation and perceptual speed were positively predictive for both constructs. Only when considering component skills as a moderator of the relation between speed and ability were differences revealed between the reading and the reasoning tests. In the reasoning test, the relation between reasoning speed and reasoning ability was more positive for persons with higher perceptual speed. In the reading test, the relation between reading speed and reading ability was more negative for persons with higher perceptual speed. Hence, higher perceptual speed supports a higher degree of automatization in reasoning but plays a different role in reading – one possible explanation being to compensate for non-automated processes. However, these different roles only become visible when examining not just the relations between component skills and the product of task completion (relation between component skill and ability), but also the process of task completion (relation between component skill and the speed-ability relation). Although the data analysis is based on regression analyses with speed 'predicting' ability, we assume no causal direction in the relation between speed and ability. However, we do assume that differences in component skills can indeed 'cause' a different relation of speed and ability.

We assume that this approach is especially useful for validating the construct interpretation of constructs for which no single process model exists. A number of different processes are involved in tasks like reading and reasoning, making it challenging to use process information for validation (Kane and Mislevy, 2017, p. 11). Research on the reading process indicates a web of complex, entangled processes that are both top-down and bottom-up, both controlled and automated. The

described approach only explicitly requires assumptions to be made about the involved component skills. However, it also implicitly assumes constant task-related cognitive processes. For assessments in which items are heterogeneous, the role of various component skills may vary across items, and different ones may even be required for different items. In such cases, it seems reasonable to investigate moderation effects at the item level as well.

An additional advantage of our approach is that it helps to collect not only convergent but also discriminant validity evidence. If we had only focused on sources of convergent validity, positive interaction effects could have also been caused by other factors such as restricted variance (cf. Cortina et al., 2018). When a certain level of a component skill that is also related to the target ability is considered, the variance in the target ability is restricted. Positive interactions might stem from the fact that the predictor restricts variance in the criteria. Thus, the analyses of discriminant sources provided additional support by showing that although the predictors and criteria were related, the positive interaction effects did not occur in all cases, but only for the hypothesized effects based on the theoretical assumptions. This supports the notion that the positive interaction effects for sources of convergent validity evidence are not the result of variance restriction. In addition, the collection of convergent and discriminant validity evidence, by referring to the same component skill for different constructs, allows conclusions about the distinctness or relatedness of two constructs in terms of their underlying processes. Two constructs differ in their underlying processes when a component skill positively moderates the speed-ability relationship of construct A but not of construct B. In the present study, perceptual speed as an underlying skill in terms of automation moderated the speed-ability relation for reasoning, but not for reading competence, and word meaning activation for reading competence, but not for reasoning. In future studies focused on constructs that are less divergent than reading and reasoning, where a component skill of one target ability cannot provide discriminant validity evidence for the other construct, one could also include component skills that can serve as sources of discriminant validity evidence for both constructs.

Furthermore, this approach is based on the assumption that sub-processes of the task solution process can be performed in an automatized mode. Hence, this approach is limited to constructs for which the dual processing framework holds. We assume that our approach is especially applicable to assessments from educational studies (e.g., PIAAC; OECD, 2016), which focus on assessing broad competence domains that require the interplay of various component skills. In any event, a sound theoretical basis concerning the involved component skills is required to derive hypotheses. However, it is conceivable that also variables apart from component skills can be used as moderating variables. Different behavior associated with different speed-ability relations could be originate from tasks that allow different solution strategies with one way being superior to the other, or by differently experienced participants (e.g., test-wiseness; Millman et al., 1965). In both cases, persons who worked faster and used superior solution strategies or had a higher test experience compared to other persons should have even higher scores.

Finally, although our validation approach aims to investigate the response process for validation purposes, it does not capture the intra-individual cognitive information processing for a single person completing a single item. Instead, our validation argument relies on statistical parameters describing relations across persons to infer meaningful characteristics of the response process.

Future Directions

In general, the presented empirical findings can be seen as preliminary and must be supported by future studies. In this study, test score validation was challenged (cf. Kane, 2013) with alternative theories by collecting discriminant validity evidence. However, there was no cross validation checking, for instance. In the future, the proposed validation approach must prove itself with respect to other samples, other constructs and other component skills.

As previously argued, the validation approach is assumed to be especially useful when no single process model exists. From this, it follows that more than one component skill is likely to be involved in task solution in such cases. Hence, more than one component skill may moderate the speed-ability relationship and should be considered when collecting convergent validity evidence. The simultaneously inclusion of multiple, interrelated component skills (e.g., semantic integration and word meaning activation in the case of reading) in the model would affect the interpretation of effects. In this case, the effects would be estimated as partial regression coefficients controlling for the other predictors in the model.

As mentioned above, the role of component skills may vary across items depending on item characteristics. Moderation effects could also be tested at the item level by adding effects of (residual) response time on (residual) response within items (cf. Bolsinova et al., 2017b). Investigating how component skills moderate these item-specific effects would shed light on how the effect of being faster than expected on task success depends on certain component skills. The variation in the moderating effect across items could be related to item difficulty as well as item characteristics that determine it. Thus, our approach could be adapted to the item level in order to provide further insights into the response process at this level.

CONCLUSION

Overall, this study proposed a novel validation approach that allows for investigating the role of component skills in response processes by focusing not only on the main effect of component skills on the targeted ability dimension, but also on how they influence the relation between speed and ability. As shown in the empirical example, including speed and interacting variables in the construct validation allows for testing specific hypotheses about the role of component skills in the task completion process, beyond their role in the task outcome.

ETHICS STATEMENT

This study is based on data from the PIAAC-L study (GESIS – Leibniz Institute for the Social Sciences, German Socio-Economic Panel (SOEP) at DIW Berlin and LIFBI – Leibniz Institute for Educational Trajectories, 2017). As explicated by the Technical Report of PIAAC-L "the PIAAC-L Consortium is committed to adhering to survey ethics and undertaking any appropriate and effective measures required by data privacy concerns. These issues were already addressed by the international PIAAC project, and GESIS developed an elaborate data confidentiality strategy for the German PIAAC 2012 Scientific Use File" (Zabal et al., 2016, p. 26).

AUTHOR CONTRIBUTIONS

LE and FG contributed to the conception and design of the study. LE organized the database. LE and FG performed the statistical analysis. LE wrote the first draft of the manuscript. LE and FG contributed to revision, read, and approved the final version of the manuscript for submission.

FUNDING

This work was supported by the Leibniz Association (Grant Number SAW-2015-GESIS-2), and the Centre for International Student Assessment (ZIB; Grant Number ZIB2016).

REFERENCES

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: cognitive abilities and information processing. *J. Exp. Psychol. Gen.* 117, 288–318. doi: 10.1037/0096-3445.117.3.288

Altmeyer, M., Schreiner, M., and Schweizer, K. (2009). A process-based account of the speed-ability relationship for the posner task. *Psychol. Sci.* 51, 470–482.

American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Augustinova, M., and Ferrand, L. (2014). Automaticity of word reading: evidence from the semantic stroop paradigm. *Curr. Dir. Psychol. Sci.* 23, 343–348. doi: 10.1177/0963721414540169

Baumert, J., Lüdtke, O., Trautwein, U., and Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: evidence in support of the distinction between intelligence and student achievement. *Educ. Res. Rev.* 4, 165–176. doi: 10.1016/j.edurev.2009.04.002

Becker, N., Schmitz, F., Göritz, A. S., and Spinath, F. M. (2016). Sometimes more is better, and sometimes less is better: task complexity moderates the response time accuracy correlation. *J. Intell.* 4:11. doi: 10.3390/jintelligence4030011

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Appl. Psychol. Meas.* 14, 237–245. doi: 10.1177/014662169001400302

Bolsinova, M., de Boeck, P., and Tijmstra, J. (2017a). Modelling conditional dependence between response time and accuracy. *Psychometrika* 82, 1126–1148. doi: 10.1007/s11336-016-9537-6

Bolsinova, M., Tijmstra, J., Molenaar, D., and De Boeck, P. (2017b). Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. *Front. Psychol.* 8:202. doi: 10.3389/fpsyg.2017.00202

Brunner, M. (2005). *Mathematische Schülerleistung: Struktur, Schulformunterschiede und Validität [Mathematics achievement: Structure, differences Between Academic Tracks, and Validity]*. Ph.D. thesis, Humboldt University, Berlin.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: a critical experiment. *J. Educ. Psychol.* 54, 1–22. doi: 10.1037/h0046743

Ceci, S. J., and Williams, W. M. (1997). Schooling, intelligence, and income. *Am. Psychol.* 52, 1051–1058. doi: 10.1037/0003-066X.52.10.1051

Chen, H., De Boeck, P., Grady, M., Yang, C. L., and Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence* 69, 16–23. doi: 10.1016/j.intell.2018.04.001

Cortina, J. M., Koehler, T., Keeler, K. R., and Nielsen, B. B. (2018). Restricted variance interaction effects: what they are and why they are your friends. *J. Manag.* doi: 10.1177/0149206318770735

De Boeck, P., Chen, H., and Davidson, M. (2017). Spontaneous and imposed speed of cognitive test responses. *Br. J. Math. Stat. Psychol.* 70, 225–272. doi: 10.1111/bmsp.12094

Embretson, S. E. (1983). Construct validity: construct representation versus nomothetic span. *Psychol. Bull.* 93, 179–197. doi: 10.1037/0033-2909.93.1.179

Embretson, S. E. (1994). "Application of cognitive design systems to test development," in *Cognitive Assessment: A Multidisciplinary Perspective*, ed. C. R. Reynolds (New York, NY: Plenum Press), 107–135. doi: 10.1007/978-1-4757-9730-5_6

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: Guilford press.

Engelhardt, L., and Goldhammer, F. (2018). *Number Series Study (DIPF).* Technical Report. GESIS Papers 2018/01. Cologne: GESIS.

GESIS – Leibniz Institute for the Social Sciences, German Socio-Economic Panel (SOEP) at DIW Berlin & LIFBI – Leibniz Institute for Educational Trajectories (2017). *PIAAC-Longitudinal (PIAAC-L), Germany. GESIS Data Archive, Cologne. ZA5989 Data file Version 3.0.0.* doi: 10.4232/1.12925

Goldhammer, F., Martens, T., and Lüdtke, O. (2017a). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large Scale Assess. Educ.* 5, 1–25. doi: 10.1186/s40536-017-0051-9

Goldhammer, F., Steinwascher, M. A., Kröhne, U., and Naumann, J. (2017b). Modelling individual response time effects between and within experimental speed conditions: a GLMM approach for speeded tests. *Br. J. Math. Stat. Psychol.* 70, 238–256. doi: 10.1111/bmsp.12099

Goldhammer, F., Naumann, J., and Greiff, S. (2015). More is not always better: the relation between item response and item response time in Raven's matrices. *J. Intell.* 3, 21–40. doi: 10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Guill, K., Lüdtke, O., and Köller, O. (2017). Academic tracking is related to gains in students' intelligence over four years: evidence from a propensity score matching study. *Learn. Instr.* 47, 43–52. doi: 10.1016/j.learninstruc.2016.10.001

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica* 21, 175–189.

Jensen, A. R. (1982). "Reaction Time and Psychometric g," in *A Model for Intelligence*, ed. H. J. Eysenck (Heidelberg: Springer), 93–132. doi: 10.1007/978-3-642-68664-1_4

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000

Kane, M. T., and Mislevy, R. (2017). "Validating score interpretations based on response processes," in *Validation of Score Meaning for the Next Generation of Assessments*, eds K. Ercikan and J. W. Pellegrino (New York, NY: Routledge), 11–24. doi: 10.4324/9781315708591-2

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Kirsch, E., and Yamamoto, K. (2013). *Chapter 1: PIAAC Assessment Design. In OECD Technical report of the survey of adult skills (PIAAC).* Available at: https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf (accessed May 15, 2019).

Klein, A., and Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika* 65, 457–474. doi: 10.1007/BF02296338

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Klieme, E., Hartig, J., and Rauch, D. (2008). "The concept of competence in educational contexts," in *Assessment of Competencies in Educational Contexts*, eds J. Hartig, E. Klieme, and D. Leutner (Boston, MA: Hogrefe Publishing), 3–22.

Knorr, E., and Neubauer, A. C. (1996). Speed of information-processing in an inductive reasoning task and its relationship to psychometric intelligence. *Pers. Individ. Dif.* 20, 653–660. doi: 10.1016/0191-8869(96)00020-7

Lehrl, S. (2005). *Mehrfachwahl-Wortschatz-Intelligenztest MWT-B*. Balingen: Spitta Verlag.

McArdle, J. J., and Woodcock, R. W. (2009). *Adaptive Testing with Selected Woodcock Cognitive Tests [Powerpoint slides]*. Charlottesville, VA: Longitudinal Research Institute.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. *Intelligence* 37, 1–10. doi: 10.1016/j.intell.2008.08.004

Millman, J., Bishop, C. H., and Ebel, R. (1965). An analysis of test-wiseness. *Educ. Psychol. Meas.* 25, 707–726. doi: 10.1177/001316446502500304

Mislevy, R. J., and Wu, P.-K. (1996). *Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing*, Vol. RR96-30. Princeton, NJ: Educational Testing Service.

Muthén, L. K., and Muthén, B. O. (2015). *Mplus user's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén.

Nagy, G. (2006). *Berufliche Interessen, Kognitive und Fachgebundene Kompetenzen: Ihre Bedeutung für die Studienfachwahl und die Bewährung im Studium [Vocational Interests, Cognitive and Scholastic Abilities: Their Role in Choice of Major and Success at University]*. Doctoral dissertation, Free University, Berlin.

Naumann, J., and Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learn. Individ. Dif.* 53, 1–16. doi: 10.1016/j.lindif.2016.10.002

Neubauer, A. C. (1990). Speed of information processing in the hick paradigm and response latencies in a psychometric intelligence test. *Pers. Individ. Dif.* 11, 147–152. doi: 10.1016/0191-8869(90)90007-E

OECD (2016). *The Survey of Adult Skills: Reader's Companion, OECD Skills Studies*, Second Edn. Paris: OECD Publishing, doi: 10.1787/9789264258075-en

Perfetti, C. (2007). Reading ability: lexical quality to comprehension. *Sci. Stud. Read.* 11, 357–383. doi: 10.1080/10888430701530730

Prenzel, M., Walter, O., and Frey, A. (2007). PISA misst kompetenzen [PISA measures Competences]. *Psychol. Rundschau* 58, 128–136. doi: 10.1026/0033-3042.58.2.128

Rammstedt, B., Martin, S., Zabal, A., Carstensen, C., and Schupp, J. (2017). The PIAAC longitudinal study in Germany: rationale and design. *Large Scale Assess. Educ.* 5:4. doi: 10.1186/s40536-017-0040-z

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037//0033-2909.124.3.372

Richter, T., and Naumann, J. (2009). "Was misst der ELVES-Subtest Satzverifikation? Analysen von Mess- und Itemeigenschaften mit hierarchisch-linearen Modellen [What does the ELVES subtest record verification measure? Analysis of measurement and item properties with hierarchical-linear models]," in *Diagnostik und Förderung des Leseverständnisses [Diagnostics and Promotion of Reading Comprehension]*, Vol. 7, eds W. Lenhard and W. Schneider (Göttingen: Hogrefe), 131–149.

Rindermann, H. (2006). Was messen internationale schulleistungsstudien? Schulleistungen, schülerfähigkeiten, kognitive fähigkeiten, wissen oder allgemeine intelligenz? [What do international academic achievement studies measure? Academic achievement, student abilities, cognitive abilities, knowledge, or general intelligence?]. *Psychol. Rundschau* 57, 69–86. doi: 10.1026/0033-3042.57.2.69

Rindermann, H., and Baumeister, A. E. E. (2015). Validating the interpretations of PISA and TIMSS tasks: a rating study. *Int. J. Test.* 15, 1–22. doi: 10.1080/15305058.2014.966911

Satzger, W., Fessmann, H., and Engel, R. R. (2002). Liefern HAWIE-R, WST und MWT-B vergleichbare IQ-Werte? [The Equivalence of Three German Vocabulary Tests and the German Version of the Wechsler Adult Intelligence Scale-Revised (HAWIE-R)]. *Zeitschrift für Differentielle und Diagnostische Psychologie* 23, 159–170. doi: 10.1024//0170-1789.23.2.159

Saß, S., Kampa, N., and Köller, O. (2017). The interplay of g and mathematical abilities in large-scale assessments across grades. *Intelligence* 63, 33–44. doi: 10.1016/j.intell.2017.05.001

Schneider, W., and Chein, J. M. (2003). Controlled & automatic processing: behavior, theory, and biological mechanisms. *Cogn. Sci.* 27, 525–559. doi: 10.1016/S0364-0213(03)00011-9

Schneider, W., and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychol. Rev.* 84, 1–66. doi: 10.1037/0033-295X.84.1.1

Schupp, J., Herrmann, S., Jaensch, P., and Lang, F. R. (2008). *Erfassung Kognitiver Leistungspotentiale Erwachsener im Sozio-oekonomischen Panel (SOEP) [Surveying Cognitive Achievement Potentials of Adults in the Socio-Economic Panel (SOEP)]. Data Documentation 32*. Berlin: DIW Berlin.

Schweizer, K., and Koch, W. (2002). A revision of Cattell's investment theory: cognitive properties influencing learning. *Learn. Individ. Dif.* 13, 57–82. doi: 10.1016/S1041-6080(02)00062-6

Sheridan, H., and Reingold, E. M. (2014). Expert vs. novice differences in the detection of relevant information during a chess game: evidence from eye movements. *Front. Psychol.* 5:941. doi: 10.3389/fpsyg.2014.00941

Sheridan, H., and Reingold, E. M. (2017). Chess players' eye movements reveal rapid recognition of complex visual patterns: evidence from a chess-related visual search task. *J. Vis.* 17:4. doi: 10.1167/17.3.4

Sweller, J., Van Merrienboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72:287. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *J. Educ. Meas.* 46, 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., and Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychol. Rev.* 118, 339–356. doi: 10.1037/a0022749

Vernon, P. (1965). Ability factors and environmental influences. *Am. Psychol.* 20, 723–733. doi: 10.1037/h0021472

Vernon, P. A., Nador, S., and Kantor, L. (1985). Reaction times and speed of information processing: their relationship to timed and untimed measures of intelligence. *Intelligence* 9, 357–374. doi: 10.1016/0160-2896(85)90020-0

Vigneau, F., Caissie, A. F., and Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence* 34, 261–272. doi: 10.1016/j.intell.2005.11.003

Walczyk, J. J. (1995). Testing a compensatory-encoding model. *Read. Res. Q.* 30, 396–408. doi: 10.2307/747623

Weeks, J. P., von Davier, M., and Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychol. Test Assess. Model.* 58, 671–701.

Wise, S. L., and DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educ. Assess.* 10, 1–17. doi: 10.1207/s15326977ea1001_1

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Zabal, A., Martin, S., and Rammstedt, B. (2016). *PIAAC-L Data Collection 2014*. Technical report. GESIS Papers 2016/17. Cologne: GESIS.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Engelhardt and Goldhammer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

TABLE A1 | Measurement models reading (standardized results).

	Reading ability		Reading speed	
	Loading (SE)	**Threshold (SE)**	**Loading (SE)**	**Intercept (SE)**
Item 1	0.44 (0.09)	−1.47 (0.10)	0.63 (0.03)	20.48 (0.66)
Item 2	0.71 (0.05)	−1.04 (0.07)	0.60 (0.03)	22.59 (0.73)
Item 3	0.64 (0.06)	−0.98 (0.07)	0.51 (0.04)	16.73 (0.56)
Item 4	0.48 (0.06)	−0.67 (0.06)	0.56 (0.03)	27.89 (0.92)
Item 5	0.71 (0.03)	−0.65 (0.04)	0.71 (0.02)	18.61 (0.42)
Item 6	0.57 (0.04)	−0.02 (0.04)	0.56 (0.02)	19.11 (0.45)
Item 7	0.59 (0.04)	−1.07 (0.05)	0.64 (0.02)	17.16 (0.39)
Item 8	0.45 (0.04)	0.07 (0.04)	0.66 (0.02)	21.13 (0.50)
Item 9	0.64 (0.04)	−0.57 (0.04)	0.61 (0.02)	22.57 (0.53)
Item 10	0.55 (0.04)	0.10 (0.04)	0.54 (0.03)	20.13 (0.46)
Item 11	0.82 (0.02)	−0.29 (0.04)	0.69 (0.02)	18.62 (0.44)
Item 12	0.49 (0.04)	0.67 (0.04)	0.63 (0.02)	17.51 (0.41)
Item 13	0.52 (0.04)	0.07 (0.04)	0.52 (0.03)	18.83 (0.43)
Item 14	0.64 (0.05)	−0.47 (0.06)	0.60 (0.03)	17.06 (0.55)
Item 15	0.50 (0.05)	−0.15 (0.05)	0.62 (0.03)	20.86 (0.68)
Item 16	0.52 (0.05)	0.25 (0.05)	0.49 (0.04)	17.05 (0.57)
Item 17	0.60 (0.06)	0.71 (0.06)	0.55 (0.04)	18.18 (0.60)
Item 18	0.41 (0.06)	0.22 (0.06)	0.71 (0.03)	21.11 (0.70)
Item 19	0.63 (0.07)	−0.98 (0.09)	0.46 (0.05)	16.26 (0.70)
Item 20	0.73 (0.07)	−1.65 (0.11)	0.50 (0.05)	19.07 (0.81)
Item 21	0.62 (0.12)	−2.32 (0.20)	0.55 (0.04)	20.83 (0.85)
Item 22	0.70 (0.05)	−0.31 (0.07)	0.62 (0.04)	19.97 (0.80)
Item 23	0.46 (0.07)	−0.55 (0.08)	0.49 (0.05)	16.64 (0.71)
Item 24	0.51 (0.08)	−0.67 (0.08)	0.53 (0.04)	19.26 (0.83)
Item 25	0.63 (0.04)	−0.54 (0.05)	0.75 (0.02)	23.75 (0.63)
Item 26	0.37 (0.05)	0.30 (0.05)	0.63 (0.03)	18.30 (0.51)
Item 27	0.58 (0.04)	−0.15 (0.05)	0.69 (0.02)	21.04 (0.55)
Item 28	0.49 (0.05)	−0.86 (0.05)	0.49 (0.03)	16.25 (0.45)
Item 29	0.74 (0.04)	−0.94 (0.06)	0.52 (0.03)	17.40 (0.48)
Item 30	0.79 (0.04)	−0.50 (0.06)	0.64 (0.03)	21.68 (0.75)
Item 31	0.70 (0.05)	−0.32 (0.06)	0.60 (0.03)	15.72 (0.56)
Item 32	0.67 (0.05)	−0.30 (0.06)	0.62 (0.03)	16.77 (0.58)
Item 33	0.64 (0.05)	−0.89 (0.06)	0.61 (0.03)	20.59 (0.56)
Item 34	0.62 (0.04)	−0.07 (0.04)	0.54 (0.03)	17.34 (0.48)
Item 35	0.64 (0.04)	−0.52 (0.05)	0.59 (0.03)	19.52 (0.52)
Item 36	0.57 (0.06)	−0.16 (0.06)	0.49 (0.04)	16.97 (0.66)
Item 37	0.71 (0.06)	0.46 (0.06)	0.56 (0.04)	17.75 (0.70)
Item 38	0.75 (0.05)	−0.47 (0.07)	0.47 (0.05)	19.14 (0.75)
Item 39	0.58 (0.06)	−1.47 (0.08)	0.63 (0.02)	15.95 (0.42)
Item 40	0.42 (0.05)	0.45 (0.05)	0.56 (0.03)	19.16 (0.51)
Item 41	0.74 (0.04)	0.33 (0.04)	0.58 (0.03)	20.26 (0.55)
Item 42	0.77 (0.03)	0.33 (0.04)	0.56 (0.03)	23.43 (0.64)
Item 43	0.58 (0.04)	−0.36 (0.05)	0.46 (0.03)	16.51 (0.45)
Item 44	0.77 (0.04)	−0.31 (0.06)	0.65 (0.03)	20.35 (0.72)
Item 45	0.66 (0.05)	−0.08 (0.06)	0.64 (0.03)	23.06 (0.82)
Item 46	0.56 (0.07)	0.74 (0.06)	0.72 (0.03)	18.51 (0.67)
Item 47	0.50 (0.07)	0.85 (0.07)	0.55 (0.04)	18.24 (0.67)
Item 48	0.49 (0.06)	0.35 (0.06)	0.69 (0.03)	17.93 (0.64)
Item 49	0.68 (0.07)	1.21 (0.07)	0.57 (0.04)	17.69 (0.65)

TABLE A2 | Measurement models reasoning (standardized results).

	Reasoning ability		Reasoning speed	
	Loading (SE)	**Threshold (SE)**	**Loading (SE)**	**Intercept (SE)**
Item 1	0.87 (0.05)	−2.95 (0.28)	0.36 (0.03)	4.01 (0.10)
Item 2	0.77 (0.10)	−2.61 (0.18)	0.47 (0.03)	4.33 (0.11)
Item 3	0.56 (0.07)	−1.64 (0.08)	0.55 (0.03)	4.79 (0.12)
Item 4	0.77 (0.07)	−2.41 (0.14)	0.59 (0.03)	4.11 (0.10)
Item 5	0.60 (0.05)	−1.24 (0.06)	0.62 (0.02)	4.58 (0.11)
Item 6	0.58 (0.04)	−0.46 (0.04)	0.66 (0.02)	5.17 (0.13)
Item 7	0.70 (0.09)	−2.05 (0.11)	0.66 (0.02)	4.40 (0.11)
Item 8	0.66 (0.05)	−0.94 (0.05)	0.68 (0.02)	4.93 (0.12)
Item 9	0.72 (0.04)	−0.73 (0.05)	0.66 (0.02)	4.86 (0.12)
Item 10	0.52 (0.04)	−0.37 (0.04)	0.43 (0.03)	4.79 (0.12)
Item 11	0.60 (0.04)	−0.37 (0.04)	0.44 (0.03)	4.44 (0.11)
Item 12	0.77 (0.03)	0.00 (0.04)	0.50 (0.03)	4.82 (0.12)
Item 13	0.86 (0.03)	−0.29 (0.04)	0.48 (0.03)	5.03 (0.13)
Item 14	0.71 (0.03)	−0.31 (0.04)	0.37 (0.03)	4.67 (0.12)
Item 15	0.70 (0.04)	0.60 (0.05)	0.16 (0.04)	5.42 (0.14)

TABLE A3 | Measurement models perceptual speed and word meaning activation (standardized results).

	Loading (SE)	**Intercept (SE)**
Perceptual speed		
Interval 1 (0–30 s)	0.85 (0.01)	3.09 (0.06)
Interval 2 (30–60 s)	0.89 (0.01)	4.23 (0.08)
Interval 3 (60–90 s)	0.82 (0.01)	4.45 (0.09)
Word meaning activation		
Number of correct answers	0.87 (0.02)	0.00 (0.03)

TABLE A4 | Residual correlations (and standard errors) in the measurement model for reasoning speed (standardized results).

	Item1	Item2	Item3	Item4	Item5	Item 6	Item 7	Item 8	Item 9	Item10	Item11	Item12	Item13	Item14
Item1														
Item2	0.25 (0.03)													
Item3	0.18 (0.03)	0.28 (0.03)												
Item4	0.19 (0.03)	0.35 (0.03)	0.28 (0.03)											
Item 5														
Item 6														
Item 7														
Item 8														
Item 9														
Item10														
Item11										0.24 (0.03)				
Item12										0.32 (0.03)	0.29 (0.03)			
Item13										0.32 (0.03)	0.38 (0.03)	0.41 (0.03)		
Item14										0.34 (0.03)	0.39 (0.03)	0.39 (0.03)	0.55 (0.03)	
Item15										0.41 (0.03)	0.41 (0.03)	0.38 (0.03)	0.44 (0.03)	0.48 (0.03)

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 May 2019
doi: 10.3389/fpsyg.2019.00777

Taking a Closer Look: An Exploratory Analysis of Successful and Unsuccessful Strategy Use in Complex Problems

Matthias Stadler[1]*, *Frank Fischer*[1] *and Samuel Greiff*[2]

[1] Psychologie und Pädagogik, Ludwig-Maximilians-University Munich, Munich, Germany, [2] Computer Based Assessment, University of Luxembourg, Luxembourg, Luxembourg City, Luxembourg

Influencing students' educational achievements first requires understanding the underlying processes that lead to variation in students' performance. Researchers are therefore increasingly interested in analyzing the differences in behavior displayed in educational assessments rather than merely assessing their outcomes. Such analyses provide valuable information on the differences between successful and unsuccessful students and help to design appropriate interventions. Complex problem-solving (CPS) tasks have proven to provide particularly rich process data as they allow for a multitude of behaviors several of which can lead to a successful performance. So far, this data has often been analyzed on a rather aggregated level looking at an average number of actions or predefined strategies with only a few articles investigating the specific actions performed. In this paper, we report the results of an exploratory analysis of CPS log-files that is aimed at distinguishing between students that applied the correct strategy to a problem but failed to solve it and those applying the strategy successfully. In that, the sequence of behavior displayed is reduced to interpretable parts (n-grams) that allow searching for meaningful differences between the two groups of students. This level of analysis allows finding previously undefined or unknown patterns within the data and increases our understanding of the processes underlying successful problem-solving behavior even further.

Keywords: log-file, problem-solving, n-grams, process data, educational assessment

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Hok Kan Ling,
Columbia University, United States
Ulf Kroehne,
German Institute for International Educational Research (LG), Germany

***Correspondence:**
Matthias Stadler
Matthias.Stadler@lmu.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *14 September 2018*
Accepted: *21 March 2019*
Published: *07 May 2019*

Citation:
Stadler M, Fischer F and Greiff S (2019) Taking a Closer Look: An Exploratory Analysis of Successful and Unsuccessful Strategy Use in Complex Problems. Front. Psychol. 10:777. doi: 10.3389/fpsyg.2019.00777

INTRODUCTION

The advent of computers to psychological and educational assessment has made it possible to analyze behavioral processes and sequences of actions through information captured in computer-generated log-files (records of all actions taken while working on a computerized assessment; Bunderson et al., 1989). Researchers are no longer limited to measuring the final outcome of an assessment (e.g., solved vs. not solved) but can also investigate the steps and actions resulting in the specific outcome through analyzes of test-taking behaviors. In other words, analyzing log-files allows researchers to make inferences about the latent cognitive processes involved in solving tasks from overt behavior (Greiff et al., 2015b). Log-files may, for example, inform researchers of specific mistakes made while working on a problem that may be indicative of a misunderstanding of the

problem at hand (Ifenthaler et al., 2012). Identifying specific test-taking behaviors that lead to successful and unsuccessful performance has proven to be a treasure chest for the improvement of interventions and teaching enabling the differentiation of instructions and scaffolding and providing students with avenues for learning individually tailored to their needs.

A field that has made much use of log-file analysis in the last years is the field of complex problem-solving (CPS; e.g., Goldhammer et al., 2014; Greiff et al., 2016). Analyzing students' behavior through log-files, it was shown that the application of the vary-one-thing-at-a-time strategy (VOTAT; Tschirgi, 1980), also referred to as "control of variables strategy" (Chen and Klahr, 1999), could explain a great deal of students' performance in solving complex problems (Greiff et al., 2015b). Others noted, however, that simply identifying those students that applied the VOTAT strategy is not sufficient to fully explain why some students successfully solve a task whereas others do not (Kuhn and Dean, 2005). There must be other differences in metastrategic behavior that distinguish students that apply the VOTAT strategy and successfully solve a problem and those students that apply the strategy but fail. The aim of this paper is to use data mining techniques to analyze CPS log-files to find differences in behavior that indicate successful and unsuccessful behavior beyond the already established strategies.

LOG-FILES IN COMPLEX PROBLEM-SOLVING TASKS

Traces of behavior have been gathered in psychology studies since the 1930s (Skinner, 1938). Today, modern computer-based applications of psychological assessment make it very easy to capture a variety of interaction behaviors and save them to log files for later analysis. These interaction data have been referred to virtually synonymously as "log-file data" (Arroyo and Woolf, 2005), "discrete action protocols" (Fu, 2001), or "process data" (Zoanetti, 2010), only listing the most common names. Behavioral log-files are indicators of human behavior as observed by automatic sensors that capture and record actions displayed while interacting with the assessment. They may include behavior as diverse as rich audio and video recordings or low-level keystrokes.

Complex tasks, allowing for multiple behaviors that lead to a correct solution, produce valuable log-files with sufficient variation among participants for a meaningful interpretation. The study of how individuals engage with such complex tasks is therefore synonymous with problem-solving (Vista et al., 2016). Exploration of the processes employed in problem-solving or in engaging with complex tasks can provide information about the cognitive skills that underlie successful resolution of the problems or tasks (O'Neil et al., 2003; Griffin and Care, 2015). Indicators of these cognitive skills can be deduced from behaviors, which are captured in the form of attempted or completed processes in problem-solving tasks.

Problem-solving tasks that are particularly rich in log-file data are CPS tasks. Throughout this paper, CPS is understood as "(. . .) the successful interaction with task environments that are dynamic (i.e., change as a function of the user's interventions and/or as a function of time) and in which some, if not all, of the environment's regularities, can only be revealed by successful exploration and integration of the information gained in that process" (Buchner in Frensch and Funke, 1995, p. 14). CPS tasks thus differ from static problem-solving tasks in that they require active interaction between the problem solver and the problem resulting in very meaningful log-file data (Greiff et al., 2015a).

ANALYZING LOG-FILE DATA

A priori Established Sequences of Behavior

Log-file data can be analyzed in two different ways: Based on *a priori* established sequences of behaviors (top–down) or bottom–up in an exploratory analysis that searches for patterns within the behavior displayed (Vista et al., 2016). Regarding CPS, various studies provided valuable findings by searching test-taking behavior for instances of specific, theoretically defined exploration strategies (e.g., Kröner et al., 2005; Wüstenberg et al., 2014). One of the strategies investigated most often in CPS research is the application of the VOTAT; Tschirgi (1980), also referred to as "control of variables strategy" (Chen and Klahr, 1999). When applying the VOTAT strategy, all variables of a problem are manipulated individually while the remaining variables are held constant to determine the effect of the varied independent variables on the dependent outcomes. VOTAT thus describes the principle of isolated variation of variables, which is the core component of scientific experimentation (Kuhn and Dean, 2005) and has been the almost exclusive focus of psychologists investigating the development of scientific reasoning (Zimmerman, 2000).

Empirically, multiple studies (e.g., Kröner et al., 2005; Wüstenberg et al., 2012, 2014) showed that application of VOTAT is strongly related to CPS performance (see also Funke, 2010). Most prominently, Greiff et al. (2015b) demonstrated the usefulness of the VOTAT strategy to explain performance differences within a problem-solving task that was part of the 2012 cycle of the Programme for International Student Assessment (PISA), one of the most widely recognized educational large-scale assessments (Turner and Adams, 2007). Their analysis of the *Climate Control* task showed, that applying the VOTAT strategy was strongly related to overall performance. This relation was observed both on the individual level and on the country level. However, not all students applying the VOTAT strategy solved the task leading researchers to search for other behaviors separating successful and unsuccessful problem-solvers (Kuhn and Dean, 2005).

As the empirical approach of searching for predefined behavioral patterns cannot explain why some students fail to solve tasks even though they apparently apply the correct strategy, it is necessary to take a closer look and conduct exploratory analyses searching for differences within the behaviors of students that apply the correct strategy and succeed

and those that apply the correct strategy but fail, which is what we will attempt in this paper.

Exploratory Approaches

Especially if long sequences of behavior need to be analyzed, exploratory approaches provide a helpful description of the underlying patterns. An approach proposed to explorative search for repetitive patterns within long sequences is the n-gram method (Damashek, 1995). The n-gram method summarizes a long string of entries (e.g., letters in words or separate instances of behaviors) as sequences of n consecutive elements. While this method was originally developed to classify and mine text data (Damashek, 1995), data scientists quickly noticed that it was also useful to classify behavior (mostly in the domain of web data mining; Mobasher, 2007). In this paper, we will use the n-gram approach to exploratively search for differences in behavior displayed by students that applied the VOTAT strategy to solve a complex problem and succeeded in solving it and those who applied the VOTAT strategy but failed to solve the problem.

To illustrate the n-gram approach take a problem that only allows for two different behaviors (A and B). A potential string of behaviors for the problem-solving process of a problem solver could look like this:

AABBBABABBBABBBAAABBABBBA

Table 1 illustrates how this sequence could be summarized by n-grams of the lengths $n = 2$ (bigrams), $n = 3$ (trigrams), and $n = 4$ (four-grams), each representing an increasingly more complex but less frequently appearing set of consecutive actions. In that way, the behavior of each problem solver could be described based on a set of sequences, which could then be used to either classify problem-solvers or predict future behavior (Liu and Kešelj, 2007). Due to this flexibility, n-grams form the basis of many data mining techniques (Borges and Levene, 2000).

TABLE 1 | Example of n-grams of different length with respective frequencies.

n = 2		*n* = 3		*n* = 4	
Sequence	**Frequency**	**Sequence**	**Frequency**	**Sequence**	**Frequency**
AA	2	AAA	1	AAAA	0
AB	6	AAB	2	AAAB	1
BA	6	ABB	5	AABA	0
BB	5	ABA	1	AABB	2
		BAA	1	ABAA	0
		BAB	3	ABAB	1
		BBA	5	ABBA	1
		BBB	4	ABBB	4
				BAAA	1
				BAAB	0
				BABA	1
				BABB	3
				BBAA	1
				BBAB	3
				BBBA	4
				BBBB	0

THIS STUDY

The aim of this study is to use exploratory educational data mining techniques in explaining CPS behavior. We go beyond the already established VOTAT strategy, exploring differences in behavior between students that applied the VOTAT strategy to a complex problem and successfully solved it and those that applied the strategy but failed to solve the complex problem. To analyze students' behavior, we chose the n-gram approach (Damashek, 1995) introduced above to classify students that applied the VOTAT strategy into successful and unsuccessful problem-solvers based on their behavior. Applying the n-gram approach, we summarize the participating students' behavior while solving the complex problem into a set of short sequences that can be used to find behaviors that are indicative of whether a student that applied the VOTAT strategy will also solve the complex problem. Next to presenting the empirical example, we will illustrate the methodological steps necessary to apply the n-gram approach to log-file data of CPS behavior.

EMPIRICAL EXAMPLE

Sample

For the empirical example, we relied on a large sample ($N = 1399$) of students attending the ninth grade in a Finish municipality. The data were drawn from the Vantaa panel study for the development of learning to learn competencies in basic education. This panel is sampled to be representative for the Finish population based on several demographic and socioeconomic indicators (see Vainikainen, 2014 for more information) and the findings gained are likely to be generalizable to other samples. The mean age of the students at the time of data collection was 15.8 years ($SD = 0.43$). 48% of the students were girls and 50% boys (2% missing information). The data used for this study can be found in an anonymized form on the open science framework repository created for this paper[1]. The research design and the scales were approved by the local Education Department. The same scales and design have been used also in national educational evaluations commissioned by the Ministry of Education and Culture, and by the Finnish National Board of Education, based on the Basic Education Act (1999). The measures and design have been approved, in relation with another study, also by the Ethical Committee of the Finnish National Institute for Health and Welfare. Both the students and their parents were asked to provide their informed consent in writing.

Task

Over the course of the assessment, students solved multiple CPS tasks based on the MicroDYN approach (Greiff et al., 2015a). The MicroDYN approach is based on linear structural equations (Funke, 2001) in which (in this study) three input variables were related to three output variables (see specific example below). The underlying relations were opaque to students at the onset

[1] osf.io/jycku

of the task and needed to be determined by applying adequate strategies (i.e., the VOTAT strategy) to acquire knowledge about the problems' structure and to apply that knowledge to achieve certain goals.

The example task used for this paper was the item "Handball training," which is illustrated in **Figure 1**. It illustrates problems based on the MicroDYN approach very well and is of sufficient difficulty to allow for variation in both behavior and successful solutions (Stadler et al., 2016). In this task's scenario, participants take over the role of the coach of a handball team trying to figure out how different types of training (labeled Training A, Training B, and Training C; left part of **Figure 1**) influence certain attributes of the players (i.e., Motivation, Power of the throw, Exhaustion; right part of **Figure 1**). The best strategy to solve such tasks based on the MicroDYN approach is to apply the VOTAT strategy; that is, to manipulate each variable individually (e.g., to put Training C on "++"), while keeping all other input variables constant, and to click on "apply" (in the center of **Figure 1**). The resulting changes in the outcome variables indicate the relations between the input and the output variables. After working on the scenario, the resulting knowledge (i.e., the relation between the three training strategies and the three outcomes needed to be plotted in the model underneath the task; see the lower part of **Figure 1**).

Extraction and Scoring of Log-Files

The task was implemented in the CBA item builder, a generic assessment platform, which has been designed to meet these requirements (for an overview see Rölke, 2012). This tool is provided by the German Institute for International Educational Research (DIPF) that organizes the development of the software and collects and coordinates new requirements. It allows users without programming experience to develop and deploy computer-based assessment tasks using a graphical user interface. After testing, log-files containing the response data can be downloaded in an XML format[2] from the test computer or server for further analysis. A detailed description of the embedding and scoring of CPS tasks implemented in the CBA item builder can be found in Greiff et al. (2013). An exemplary XML file can be found on the open science framework repository see text footenote[1].

For this study, we used two different scripts to extract the data from the log-files. To extract students' scores, time on task, and the use of VOTAT, we used an SPSS script already used

[2]http://www.w3.org/XML

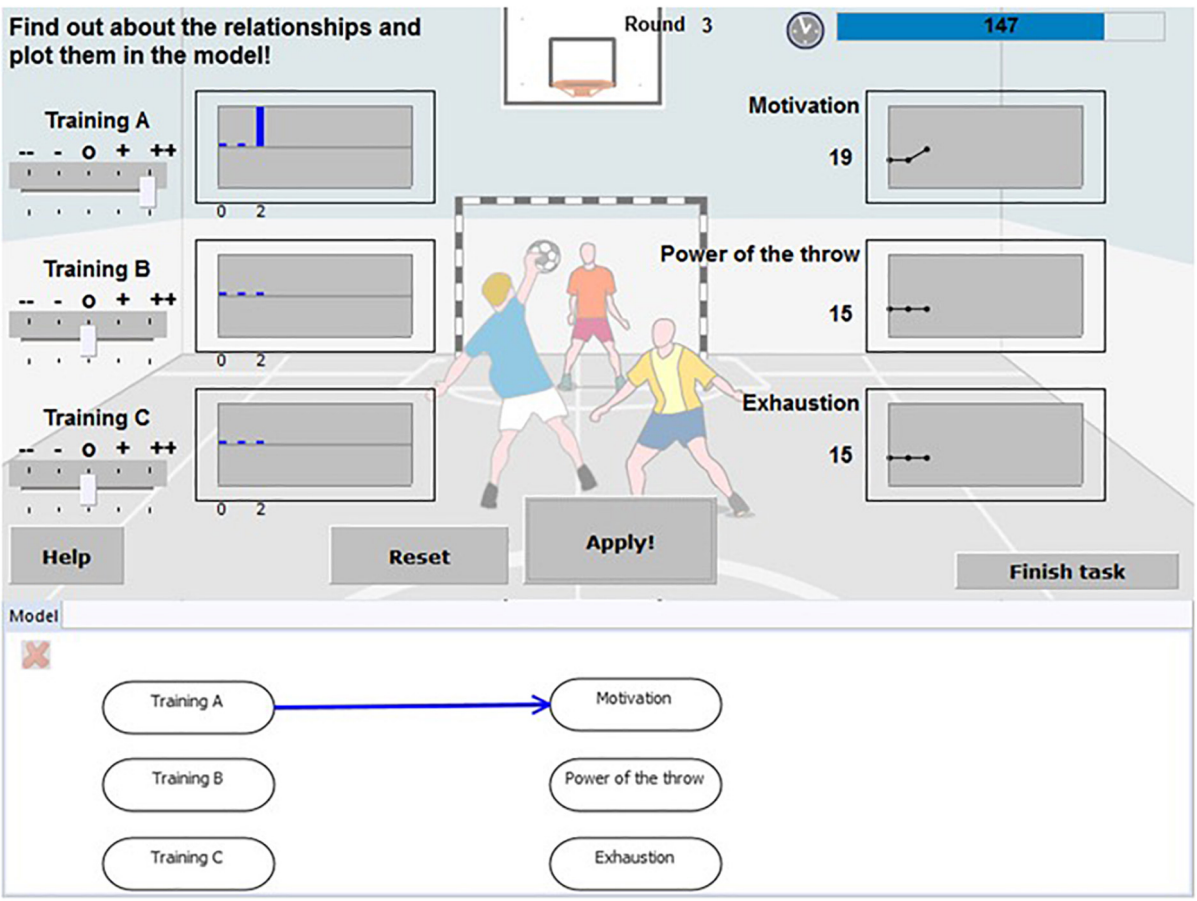


FIGURE 1 | Scenario (top) and model (bottom) of the "Handball training" task.

in various previous studies (e.g., Greiff et al., 2013). To score the application of the VOTAT strategy, log-files of students' behavior were analyzed. Full credit was given if participants manipulated each input variable at least once while keeping all other variables constant; otherwise, no credit was assigned (Wüstenberg et al., 2014). For the explorative n-gram analyses the complete string of behaviors was extracted by a customized python script[3], using the built-in xml.etree.ElementTree package.

Table 2 shows some exemplary data from a log file of the "Handball training" task. Students could make and apply changes to the input variables (i.e., the rounds in the assessment). One round was recorded every time the "Apply" button was pressed applying changes from none to all of the input variables (i.e., working on the scenario, "S") to the output variables. In the example, the participant applied the VOTAT strategy in manipulating each variable individually. The nature of the XML files does, however, not allow discerning the order in which variables were manipulated within each round. Changes in the model were recorded every time a line was drawn or removed between an input and an output variable to plot findings (i.e., working on the model, "M"). Our python script extracted the string of behavior as a vector of "M" or "S" for each participant allowing for an easy interpretation. In the example, the extracted string would be "SSS" as all three recorded behaviors were changes in the scenario. In addition, we extracted the total number of behaviors (length of the vectors) and the time spent working on the task (without reading the problem description, which was presented separately from the actual problem scenario). Use of the "reset" and "help" button was ignored as these do not provide any additional information on the solution process and were used by a marginal number of participants (N_{reset} = 53; N_{help} = 19). The python script can be found on the open science framework repository see text footnote[1].

Statistical Analysis

In order to find behavioral differences between students that applied the VOTAT strategy and successfully solved the complex problem and those students that applied the VOTAT strategy but failed, we first identified the respective students, assigning a dummy coded variable. This variable separated successful and unsuccessful students that applied the VOTAT strategy and assigned missing values to all students that did not apply the VOTAT strategy.

[3]www.python.org

To find the sequences of behaviors that led to success or failure in the problem-solving process, we applied the chi-square feature selection model, which is frequently used in natural language processing or other data mining contexts (Oakes et al., 2001). Recent publications have demonstrated how to apply this approach to problem-solving data, though (He and von Davier, 2015, 2016). The chi-square feature selection model tests whether occurrence and non-occurrence of behaviors are independent for two groups. Under the null hypothesis, the behaviors would be equally likely for both groups. Based on the observed distribution of behaviors, a chi-square value can thus be computed to evaluate the departure from this null hypothesis. A problem with this approach is potentially over-interpreting the relevance of extremely common behaviors that have little or no discriminating power while under-estimating the relevance of rather infrequent behaviors. Moreover, the added relevance of a behavior is not linear. More occurrences of a behavior indicate higher importance, but not as much relative importance as an undamped count would suggest (Manning and Schütze, 2005). To solve this problem, a weight is assigned to the observed frequency of each sequence of behaviors based on the number of participants displaying the sequence of behavior, the sequence's total frequency, and the total number of behaviors observed for a more detailed description of the chi-square feature selection model see (He and von Davier, 2016). The weight function for sequence of behavior i in total behavior j (1) was defined as:

$$\text{weigth}\left(i,j\right) = \begin{cases} \left[1+\log\left(\mathrm{f}_{i,j}\right)\right]\log\left(\frac{N}{sf_i}\right) & \text{if } \mathrm{f}_{i,j} \geq 1 \\ 0 & \text{if } \mathrm{f}_{i,j} = 0 \end{cases} \quad (1)$$

where N is the total number of sequences, f is the sequence's frequency and sf is the number of behaviors where the sequence i appears. The first clause applies to sequences occurring in the same behavior, whereas for sequences that do not appear ($f_{i,j}$ = 0), we use weight (i,j) = 0.

The scripts for all analyses can be found on the open science framework repository see text footnote[1]. **Table 3** provides the raw and weighted frequencies for all sequences of behavior of students applying the VOTAT strategy.

RESULTS

As can be seen from **Table 4**, the task was relatively difficult with only 544 (38.9%) of the students solving the task

TABLE 2 | Example data from a log file of the "Handball training" task (adapted from Greiff et al., 2013).

General information		Input variables			Output variables		
Timestamp	**Button pressed**	**Training A**	**Training B**	**Training C**	**Motivation**	**Power**	**Exhaustion**
15:13:21	Apply	0	1	0	17	15	15
15:13:23	Apply	0	0	1	17	17	15
15:13:26	Apply	1	0	0	17	19	17

TABLE 3 | The raw and weighted frequency for all sequences of behavior.

Behavior sequence	Frequency of sequences	Frequency of actions	Weight	Freq. in correct		Freq. in incorrect	
				Raw	Wgt	Raw	Wgt
Bigrams							
MM	663	5063	0.01	4073	37.57	990	9.13
MS	159	221	2.08	166	345.36	55	114.43
SM	224	327	1.66	249	414.13	78	129.73
SS	81	226	3.07	156	478.76	70	214.83
Trigrams							
MMM	628	4275	0.12	3454	408.11	821	97.01
MMS	125	159	2.33	120	279.12	39	90.71
MSM	118	147	2.38	113	269.00	34	80.94
MSS	54	66	3.08	47	144.59	19	58.45
SMM	206	267	1.75	208	363.20	59	103.02
SMS	28	33	3.47	25	86.66	8	27.73
SSM	79	96	2.76	72	198.80	24	66.27
SSS	50	128	3.49	82	286.51	46	160.73
Four-grams							
MMMM	495	3574	0.59	2895	1698.67	679	398.41
MMMS	80	101	2.77	79	218.45	22	60.83
MMSM	90	103	2.62	80	209.51	23	60.23
MMSS	41	48	3.25	34	110.37	14	45.45
MSMM	97	114	2.56	90	230.19	24	61.39
MSMS	14	15	3.65	11	40.15	4	14.60
MSSM	22	23	3.50	17	59.46	6	20.99
MSSS	36	41	3.31	28	92.70	13	43.04
SMMM	171	210	1.96	167	327.60	43	84.35
SMMS	24	28	3.53	22	77.70	6	21.19
SMSM	20	21	3.54	16	56.57	5	17.68
SMSS	12	12	3.63	9	32.64	3	10.88
SSMM	67	76	2.87	60	172.40	16	45.97
SSMS	11	11	3.64	8	29.10	3	10.91
SSSM	50	57	3.10	40	123.96	17	52.68
SSSS	25	71	4.06	42	170.71	29	117.87

Freq., frequency; Wgt, weight; S, working on the scenario; M, changing the model.

TABLE 4 | Distribution of students based on whether they solved the problem and applied the VOTAT strategy.

		Applied the VOTATstrategy		
		No	Yes	Total
Solved the problem	No	712	143	855
	Yes	21	523	544
Total		733	666	1399

correctly. Moreover, 666 (47.6%) of the students applied the VOTAT strategy. Applying the VOTAT strategy, generally, lead to a substantially higher likelihood of solving the problem ($\chi^2 = 401.10$; $df = 1$; $p < 0.001$). However, 143 (21.5%) of the students that applied the VOTAT strategy did not solve the problem.

In the exploratory analysis, we attempt to understand this observation by finding behavioral differences among the students that applied the VOTAT strategy by using the n-gram approach. There was no significant difference between the absolute number of behaviors observed for either group of students [$t(664) = 0.52$; $p = 0.601$; $d = 0.05$] nor the time spent working on the task [$t(664) = 0.27$; $p = 0.790$; $d = 0.03$]. **Table 5** displays the results of the chi-square feature selection model analyzing differences in likelihoods of specific n-grams for students that applied the VOTAT strategy and solved the problem and those that did not. Note that the possible behaviors were reduced to working on the scenario (S) and changing the model (M). N-grams with higher chi-square values are more discriminative between the two groups. Moreover, **Table 5** indicates whether the n-grams were more typical of students that solved the problem or of those that did not.

As can be seen from **Table 5**, the informational value of the n-grams increases with their length, while the general pattern does not change. The most discriminative sequence of behavior was consistently the one indicating working maximally

TABLE 5 | Summary of the chi-square feature selection model for bigrams, trigrams, and four-grams.

n = 2				*n* = 3				*n* = 4			
Sequence	χ^2	***p***	**Dir.**	**Sequence**	χ^2	**Dir.**	***p***	**Sequence**	χ^2	**Dir.**	***p***
SS	31.98	<0.001	−	SSS	59.98	−	<0.001	SSSS	76.34	−	<0.001
MM	0.92	0.337	+	MMM	12.16	+	<0.001	MMMM	67.09	+	<0.001
MS	0.23	0.632	−	MSS	4.08	−	0.043	MSSS	7.21	−	0.007
SM	0.05	0.823	+	SMM	1.95	+	0.163	SMMM	5.43	+	0.020
				MSM	0.37	+	0.543	SSSM	5.30	−	0.021
				SSM	0.17	−	0.680	MMSS	3.64	−	0.056
				MMS	0.04	−	0.841	MSMM	2.70	+	0.100
				SMS	0.00	−	1.00	SSMM	2.01	+	0.156
								MMMS	1.52	+	0.218
								MMSM	0.87	+	0.351
								SMMS	0.69	+	0.406
								SSMS	0.36	−	0.549
								MSMS	0.32	−	0.572
								MSSM	0.27	−	0.603
								SMSS	0.03	−	0.862
								SMSM	0.01	+	0.920

N-grams with higher chi-square values are more discriminative; Dir., Direction of the difference between groups. "+" represents behaviors that were more typical of students that solved the task "−" represents behaviors that were more typical of students that did not solve the task; S, working on the scenario; M, changing the model.

long in the scenario (SS, SSS, and SSSS), which was always more typical of the students that did not solve the task. This was followed by the sequence of behavior indicating working maximally long in the model (only statistically significant for MMM and MMMM), which was always more typical of students solving the task. Generally, the sequences indicating repeated changes in the scenario were associated with failing to solve the problem (statistically significant for MSS, MSSS, and SSSM), whereas the sequences indicating repeated changes in the model were associated with solving the problem (statistically significant for SMMM). The discriminative value (high chi-square values) was highest for the sequences with the longest uninterrupted sequences of one specific behavior (M or S) and least for those that indicated frequent changes between working on the scenario and working on the model (e.g., SMSM).

DISCUSSION

The aim of this study was to use exploratory educational data mining techniques in explaining problem-solving behavior. We chose one of the most established types of CPS tasks (based on the MicroDYN approach; Greiff et al., 2015a), for which the optimal strategy is well known (i.e., the VOTAT strategy; Tschirgi, 1980). However, not all students applying the VOTAT strategy also solved the tasks correctly implying that simply observing whether or not the strategy was applied is not sufficient to understand why some students succeed in solving CPS tasks while others do not (Kuhn and Dean, 2005).

Describing the whole string of behaviors observed for each individual student as a set of n-grams of different length (Damashek, 1995) allowed us to exploratively search for differences in the behavior observed within those students that applied the VOTAT strategy and successfully solved the task and those that applied the strategy but still failed to solve the task. The empirical example illustrates that given enough complexity, there are substantial differences in the frequencies of observed n-grams between the two groups. Interpreting those differences, however, requires some understanding of the task and what it takes to solve it (Banovic et al., 2016).

Correctly applied, the VOTAT strategy requires problem-solvers to make only minimal changes in the scenario, register the effects and then immediately plot the findings in the model (Wüstenberg et al., 2012). Any deviations from this algorithm will increase the cognitive load (Sweller, 2011) on the problem-solver as important information (i.e., either changes made in the scenario or findings resulting from these changes) need to be stored in working memory (Sweller, 1988). Inspecting the differences in behaviors between students that applied the VOTAT strategy and successfully solved the task and those that applied the strategy but still failed, the general pattern seemed to be that the students that solved the task spent fewer rounds continuously working on the scenario (e.g., SS, SSS, or SSSS) but more rounds working on the model (e.g., MMM and MMMM). Students that did not solve the task, thus, did not immediately plot their findings, thereby increasing their cognitive load and, in turn, the task's difficulty (Kirschner, 2002). Our findings thus highlight the importance of metastrategic competencies that enable a person to not only apply the correct strategy to solve a problem but to make use of the information gained in the process. Metastrategic competencies encompass awareness, understanding, monitoring, and

management of one's strategic performance of many kinds of cognitive tasks (Kuhn and Pearsall, 1998). As becomes obvious from our analyses, students that did not solve the problems correctly either lacked understanding of the VOTAT strategy or were not able to manage their use of the strategy. Due to the exploratory nature of our analyses, our interpretations are *post hoc* though and should be corroborated by experimental studies.

There are other limitations to be considered. In focusing only on the students applying the VOTAT strategy we reduced our sample to $N = 666$, excluding almost half of the initial sample from our analyses. However, since the aim of our paper was to find behavior differences between students that applied the VOTAT strategy and successfully solved the problem and those students that applied the VOTAT strategy but failed, students that did not apply the VOTAT strategy at all were irrelevant to our analyses. Future studies should extend our analyses to explore differences in behavior across all students not selected by *a priori* defined strategies.

Moreover, coding of the log-files into changes in the scenario and changes in the model does not allow differentiating between different changes applied to the input variables within one round of changes to the scenario (e.g., manipulations of only one variable vs. manipulations of multiple variables). However, this simplification allows for a relatively straight-forward interpretation of the resulting n-grams. A more detailed coding of changes to the scenario, on the other hand, would lead to an exponentially higher number of potential behavior sequences most of which would most likely have very little information value due to their specificity. The potential variance in changes in the input variables between successful and unsuccessful students is further reduced by the fact that all participants included in our analyses applied the VOTAT strategy (i.e., manipulated all input variables at least once individually while keeping the others constant). Since manipulating all input variables individually once is sufficient to solve the task, all further manipulations, regardless of whether single or multiple variables, will result in unnecessary additional information increasing cognitive load. Testing these assumptions will, however, require additional information to be logged (for more on the completeness of log data see Kroehne and Goldhammer, 2018).

Finally, the n-gram approach showcased in this paper is not the only explorative educational data mining approach applicable to CPS log-files, of course. Other studies have applied analyses of the interaction of behavior displayed while solving tasks such as Network Analysis (Wooldridge et al., 2018), or included the temporal order of behaviors in their analyses by displaying them as complex directed networks (Vista et al., 2016). All of these approaches share the aim of understanding problem-solving behavior on a very detailed level and the difficulties that come with that aim. Most importantly, any increase in task specificity (e.g., longer n-grams) comes, necessarily, with a decrease in generalizability. In that, perfect understanding of students' behavior in one task may be meaningless to understand performance in another task unless the structural similarities between these tasks are well understood and theoretically described. Future studies should, therefore, investigate the generalizability of behavior across different problem-solving tasks.

The findings show the potential benefit of applying explorative educational data mining approaches such as the n-gram approach in addition to searching for *a priori* defined strategies. Knowledge about how and why students that actually apply the correct strategy to solve a problem fail to actually solve it has implications for the instruction or training of CPS tasks. So far, interventions aimed at increasing CPS performance have relied on repeatedly confronting problem-solvers with problems of a similar nature (e.g., Kretzschmar and Süß, 2015). Training lead to an increase in performance and, in fact, also to an increase in strategic prowess (Lotz et al., 2017). However, no dedicated strategy training has been published to the best of our knowledge. Based on our findings, such a strategy training should consider to not only teach the VOTAT strategy but also metastrategic knowledge such as the handling of information gained through the application of VOTAT (Zohar and Peled, 2008).

CONCLUSION

In summary, our paper showcased the n-gram approach on a CPS task. The detailed description of the data provided some indication toward behavioral differences within students that apply the correct strategy toward a problem and solve it as opposed to those that apply the correct strategy and fail. We hope that the paper will help other scholars in finding ways to analyze and interpret log-file data themselves. After all, the exploitation of this rich resource through dedicated analyses is still in its infancy and we believe that it is a treasure trove worth hunting for.

ETHICS STATEMENT

Ethics approval was granted for the data collection (see Vainikainen, 2014).

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This research was supported by a project funded by the Fonds National de la Recherche Luxembourg (The Training of Complex Problem Solving; "TRIOPS").

REFERENCES

Arroyo, I., and Woolf, B. (2005). "Inferring learning and attitudes from a Bayesian network of log file data," in *Proceedings of the AIED 05, 12th International Conference on Artificial Intelligence in Education*, Chicago.

Banovic, N., Buzali, T., Chevalier, F., Mankoff, J., and Dey, A. K. (2016). "Modeling and understanding human routine behavior," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, eds J. Kaye, A. Druin, C. Lampe, D. Morris, and J. P. Hourcade (New York, NY: ACM), 248–260. doi: 10.1145/2858036.2858557

Borges, J., and Levene, M. (2000). "Data mining of user navigation patterns," in *Lecture Notes in Computer Science Lecture Notes in Artificial Intelligence*, Vol. 1836, ed. B. Masand (Berlin: Springer), 92–112. doi: 10.1007/3-540-44934-5_6

Bunderson, C. V., Inouye, D. K., and Olsen, J. B. (1989). "The four generations of computerized educational measurement," in *Educational Measurement*, ed. R. L. Linn (Washington DC: American Council on Education), 367–407.

Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi: 10.1111/1467-8624.00081

Damashek, M. (1995). Gauging similarity with n-grams: language-independent categorization of text. *Science* 267, 843–848. doi: 10.1126/science.267.5199.843

Frensch, P. A., and Funke, J. (eds) (1995). *Complex Problem Solving: The European Perspective*. London: Routledge.

Fu, W.-T. (2001). ACT-PRO action protocol analyzer: a tool for analyzing discrete action protocols. *Behav. Res. Methods Instrum. Comput.* 33, 149–158. doi: 10.3758/BF03195360

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Think. Reason.* 7, 69–89. doi: 10.1080/13546780042000046

Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cogn. Process.* 11, 133–142. doi: 10.1007/s10339-009-0345-0

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Greiff, S., Niepel, C., Scherer, R., and Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: an analysis of behavioral data from computer-generated log files. *Comput. Hum. Behav.* 61, 36–46. doi: 10.1016/j.chb.2016.02.095

Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., and Martin, R. (2015a). Sometimes less is more: comparing the validity of complex problem solving measures. *Intelligence* 50, 100–113. doi: 10.1016/j.intell.2015.02.007

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015b). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., and Funke, J. (2013). Computer-based assessment of complex problem solving: concept, implementation, and application. *Educ. Technol. Res. Dev.* 61, 407–421. doi: 10.1007/s11423-013-9301-x

Griffin, P., and Care, E. (2015). *Assessment and Teaching of 21st Century Skills*. Dordrecht: Springer.

He, Q., and von Davier, M. (2015). "Identifying feature sequences from process data in problem-solving items with n-grams," in *Springer Proceedings in Mathematics & Statistics: Volume 140. Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society, Wisconsin, 2014*, Vol. 140, eds L. A. van der Ark, D. Bolt, W.-C. Wang, J. A. Douglas, S.-M. Chow, and P. Society (Berlin: Springer), 173–190. doi: 10.1007/978-3-319-19977-1_13

He, Q., and von Davier, M. (2016). "Analyzing process data from problem-solving items with n-grams," in *Advances in Higher Education and Professional Development Book Series. Handbook of Research on Technology Tools for Real-World Skill Development*, Vol. 1, ed. Y. Rosen (Hershey, PA: Information Science Reference), 750–777. doi: 10.4018/978-1-4666-9441-5.ch029

Ifenthaler, D., Eseryel, D., and Ge, X. (2012). "Assessment for game-based learning," in *Assessment in Game-Based Learning: Foundations, Innovations and Perspectives*, eds D. Eseryel, X. Ge, and D. Ifenthaler (New York, NY: Springer), 1–8. doi: 10.1007/978-1-4614-3546-4

Kirschner, P. A. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learn. Instruct.* 12, 1–10. doi: 10.1016/S0959-4752(01)00014-7

Kretzschmar, A., and Süß, H.-M. (2015). A study on the training of complex problem solving competence. *J. Dyn. Decis. Mak.* 1:4. doi: 10.11588/jddm.2015.1.15455

Kroehne, U., and Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika* 45, 527–563. doi: 10.1007/s41237-018-0063-y

Kröner, S., Plass, J. L., and Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence* 33, 347–368. doi: 10.1016/j.intell.2005.03.002

Kuhn, D., and Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychol. Sci.* 16, 866–870. doi: 10.1111/j.1467-9280.2005.01628.x

Kuhn, D., and Pearsall, S. (1998). Relations between metastrategic knowledge and strategic performance. *Cogn. Dev.* 13, 227–247. doi: 10.1016/S0885-2014(98)90040-5

Liu, H., and Kešelj, V. (2007). Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data Knowl. Eng.* 61, 304–330. doi: 10.1016/j.datak.2006.06.001

Lotz, C., Scherer, R., Greiff, S., and Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence* 64, 98–112. doi: 10.1016/j.intell.2017.08.002

Manning, C. D., and Schütze, H. (2005). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass: MIT Press.

Mobasher, B. (2007). "Data mining for web personalization," in *The Adaptive Web. Lecture Notes in Computer Science*, Vol. 4321, eds P. Brusilovsky, A. Kobsa, and W. Nejdl (Berlin: Springer).

Oakes, M., Gaaizauskas, R., Fowkes, H., Jonsson, A., Wan, V., and Beaulieu, M. (2001). "A method based on the chi-square test for document classification," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '01*, eds D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel (New York, NY: ACM Press), 440–441. doi: 10.1145/383952.384080

O'Neil, H. F., Chuang, S.-H., and Chung, G. K. W. K. (2003). Issues in the computer-based assessment of collaborative problem solving. *Assess. Educ.* 10, 361–373. doi: 10.1080/0969594032000148190

Rölke, H. (2012). "The item builder: a graphical authoring system for complex item development," in *Proceedings of the E-Learn 2012–World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 1*, (Chesapeake, VA: AACE), 344–353.

Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis*. New York, NY: Appleton-Century-Crofts.

Stadler, M., Niepel, C., and Greiff, S. (2016). Easily too difficult: estimating item difficulty in computer simulated microworlds. *Comput. Hum. Behav.* 65, 100–106. doi: 10.1016/j.chb.2016.08.025

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202_4

Sweller, J. (2011). "Cognitive load theory," in *The Psychology of Learning and Motivation: v.55. Cognition in Education*, Vol. 55, eds J. P. Mestre and B. H. Ross (San Diego, CA: Academic Press), 37–76. doi: 10.1016/B978-0-12-387691-1.00002-8

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10. doi: 10.2307/1129583

Turner, R., and Adams, R. J. (2007). The programme for international student assessment: an overview. *J. Appl. Meas.* 8, 237–248.

Vainikainen, M.-P. (2014). *Finnish Primary School Pupils' Performance in Learning to Learn Assessments: A Longitudinal Perspective on Educational Equity*. Report No. 360, Helsinki: Unigrafia.

Vista, A., Awwal, N., and Care, E. (2016). Sequential actions as markers of behavioural and cognitive processes: extracting empirical pathways from data streams of complex tasks. *Comput. Educ.* 9, 15–36. doi: 10.1016/j.compedu.2015.10.009

Wooldridge, A. R., Carayon, P., Shaffer, D. W., and Eagan, B. (2018). Quantifying the qualitative with epistemic network analysis: a human factors case study of

task-allocation communication in a primary care team. *IISE Trans. Healthc. Syst. Eng.* 8, 72–82. doi: 10.1080/24725579.2017.1418769

Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving — More than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003

Wüstenberg, S., Stadler, M., Hautamäki, J., and Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technol. Knowl. Learn.* 19, 127–146. doi: 10.1007/s10758-014-9222-8

Zimmerman, C. (2000). The development of scientific reasoning skills. *Dev. Rev.* 20, 99–149. doi: 10.1006/drev.1999.0497

Zoanetti, N. (2010). Interactive computer based assessment tasks: how problem-solving process data can inform instruction. *Australas. J. Educ. Technol.* 26, 585–606. doi: 10.14742/ajet.1053

Zohar, A., and Peled, B. (2008). The effects of explicit teaching of metastrategic knowledge on low- and high-achieving students. *Learn. Instruct.* 18, 337–353. doi: 10.1016/j.learninstruc.2007.07.001

Conflict of Interest Statement: SG is one of two authors of the commercially available COMPRO test that is based on the multiple complex systems approach and that employs the MicroDYN approach. However, for any research and educational purposes, a free version of MicroDYN tasks is available and he receives royalties for COMPRO.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Stadler, Fischer and Greiff. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fpsyg.2019.00906

How Do Test Takers Interact With Simulation-Based Tasks? A Response-Time Perspective

Yi-Hsuan Lee[1], Jiangang Hao[1], Kaiwen Man[2] and Lu Ou[3]*

[1] Educational Testing Service, Princeton, NJ, United States, [2] Department of Human Development and Quantitative Methodology, Measurement, Statistics and Evaluation Program, University of Maryland at College Park, College Park, MD, United States, [3] ACT Inc., Iowa City, IA, United States

Many traditional educational assessments use multiple-choice items and constructed-response items to measure fundamental skills. Virtual performance assessments, such as game- or simulation-based assessments, are designed recently in the field of educational measurement to measure more integrated skills through the test takers' interactive behaviors within an assessment in a virtual environment. This paper presents a systematic timing study based on data collected from a simulation-based task designed recently at Educational Testing Service. The study is intended to understand the response times in complex simulation-based tasks so as to shed light on possible ways of leveraging response time information in designing, assembling, and scoring of simulation-based tasks. To achieve this objective, a series of five analyses were conducted to first understand the statistical properties of the timing data, and then investigate the relationship between the timing patterns and the test takers' performance on the items/task, demographics, motivation level, personality, and test-taking behaviors through use of different statistical approaches. We found that the five analyses complemented each other and revealed different useful timing aspects of this test-taker sample's behavioral features in the simulation-based task. The findings were also compared with notable existing results in the literature related to timing data.

Keywords: trialogue, response time, hierarchical modeling framework, cluster analysis, motivation, rapid-guessing behavior

OPEN ACCESS

Edited by:
Bernard Veldkamp, University of Twente, Netherlands

Reviewed by:
Chun Wang, University of Washington, United States
Carmen Llinares Millán, Universitat Politècnica de València, Spain

****Correspondence:***
Yi-Hsuan Lee
ylee@ets.org

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *01 September 2018*
Accepted: *04 April 2019*
Published: *24 April 2019*

Citation:
Lee Y-H, Hao J, Man K and Ou L (2019) How Do Test Takers Interact With Simulation-Based Tasks? A Response-Time Perspective. Front. Psychol. 10:906. doi: 10.3389/fpsyg.2019.00906

INTRODUCTION

Many traditional educational assessments use multiple-choice (MC) items and constructed-response (CR) items to measure fundamental skills, such as verbal and quantitative skills. The MC and CR items in the same form are assembled to measure the same construct but usually are not attached to a common scenario throughout the test. There is an increasing interest in the field of educational measurement in developing new capabilities for new task formats and assessment types to measure more integrated skills, such as problem-solving and critical thinking, which may not be directly assessed by those traditional educational assessments. Virtual performance assessments (VPAs), such as game- or simulation-based assessments, are often used to serve the purpose (Baker and Clarke-Midura, 2013; Mislevy et al., 2014). In a VPA, a test taker's proficiency is assessed based on his/her interactions with the virtual environment. As such, good understanding of how the test taker interacts with the virtual environment is essential for developing psychometrically sound scoring rules for VPAs,

and for designing and assembling VPAs to support the intended scoring rules. In this paper, we aim at better understanding the test taker's interactions with the virtual environment from the perspective of their response time (RT) to the items in a VPA.

There is rich literature on RT research concerning the design, assembly, and scoring aspects of traditional MC tests that are digitally based (for review papers, see, e.g., Schnipke and Scrams, 2002; Lee and Chen, 2011; Kyllonen and Zu, 2016; De Boeck and Jeon, 2019). These literature also suggests that RTs contain rich information about test takers' response processes, test-taking behaviors and strategies, and motivation. One reason is that test takers' timing behaviors reflect person-task interactions. When the major assessment outcomes to be scored are the final responses to items, test takers may adjust their timing behaviors or strategies to cope with the test conditions in order to optimize their test performance. The adjustment in behavior or strategy may occur before people take a test (during practice exams) or during a live test (Lee and Haberman, 2016). Thus, compared to item responses, their timing behaviors tend to be more sensitive to test context and content, test/item type, and test conditions. RTs have been used as ancillary information for improving precision of parameter estimation and validity of measurement beyond what is available based on item responses: For example, for tests that are intended to measure both speed and accuracy, RTs may be used to derive scores together with item responses (Maris and van der Maas, 2012; van Rijn and Ali, 2018). To have better control on test speededness, RTs may be utilized for assembling test forms in non-adaptive testing and selecting items in adaptive testing (e.g., van der Linden et al., 1999; Choe et al., 2018). In addition, RTs have been used in test security analyses and examination of general test-taking behaviors (e.g., solution behavior vs. rapid-guessing behavior, due to test speededness or low motivation).

To our knowledge, test takers' timing behaviors in VPAs have been less explored psychometrically, possibly due to limited access to large-scale empirical data from VPAs. Educational Testing Service (ETS) researchers have conducted a timing study of simulation-based tasks in the context of the National Assessment of Educational Progress (NAEP; Jia and Lee, 2018). The study focused on two simulation-based tasks, each with four items given with a time limit to around 2,000 students; the tasks assessed technology and engineering literacy of grade eight students in the United States. This study had three primary findings. First, the items that asked the students to conduct simulations or experiments (referred to as *simulation items* henceforth) required much more time to complete than the rest of the items did, but the simulation items did not appear to be especially difficult. Second, rapid-guessing behavior was not an issue for these simulation-based tasks, although the assessment was considered low-stakes to the students. Third, the correlation between the observed task time and performance was positive but almost negligible. Note that each of the two NAEP simulation-based tasks was used as part of a test form for assessing technology and engineering literacy and the scores were not reported at the task level. While RTs have also been examined in other fields, the focuses tend to be different from those in educational measurement—for example, to study varying student interactions in computer-supported collaborative learning (e.g., Jeong, 2004) or to assess learning in intelligent tutoring systems (e.g., Beck et al., 2000).

The simulation-based tasks considered in Jia and Lee (2018) were relatively short and simple. In this current work, we furthered the effort on RT analysis to study a more complex simulation-based task that has a complete storyline about how a test taker investigates volcano eruption in a virtual geology lab. This simulation-based task was developed as part of an effort to assessing collaborative problem-solving (CPS) skills in science, the ETS Collaborative Science Assessment Prototype (ECSAP; Hao et al., 2015, 2017; Liu et al., 2015). In ECSAP, there are two parallel simulation-based tasks. One is intended for individual test takers to respond, referred to as the single-user version. The other is for dyadic teams to respond collaboratively, referred to as the collaborative version. Both the individual and collaborative versions of the simulation-based tasks were modified from an earlier simulation-based task about volcano science designed to assess students' science inquiry skill (Zapata-Rivera et al., 2014). In the single-user version, each participant responded to 11 items without any time limit, and their item responses and item RTs were captured. In the collaborative version, two human participants collaborated through a chat box to interact with two virtual agents to complete the same task. In the previous research, the foci were primarily on the collaborative version of the simulation-based task to explore CPS skills and collaboration engagement through the online chats (e.g., content, frequency, and chat time) between team members and their item responses (see the CPS references above, and Halpin et al., 2017), while the single-user version was simply used as a control. No systematic timing analysis has been carried out using data collected from either version of the tasks.

In this paper, we present a systematic study on the RTs collected from the single-user version of the simulation-based task. Our goal is to understand the RTs in complex simulation-based tasks so as to shed light on possible ways of leveraging RT information in designing, assembling, and scoring of simulation-based tasks. To achieve the objective, a series of five analyses were conducted to first understand the statistical properties of the timing data, and then investigate the relationship between the timing patterns and the test takers' performance on the items/task, demographics, motivation level, personality, and test-taking behaviors through use of different statistical approaches. As will be shown, the five analyses complement each other and reveal different timing aspects of this test-taker sample's behavioral features in the simulation-based task we studied. The behavioral features observed in this simulation-based task may be quite different from those in traditional educational assessments, and the comparisons will benefit RT researchers as well as researchers who are interested in the same or similar datasets.

It is worth noting that the study concentrates on timing and response data in the simulation-based task, although in general, a simulation-based task may have many assessment metrics beyond RTs and responses that are worth exploring. Also, this study is not intended to evaluate the potential of simulation-based tasks or VPAs beyond timing and response data

for use in the field of educational measurement. For more general discussion about VPAs, please see, for example, Baker and Clarke-Midura (2013) and Mislevy et al. (2014). The rest of the paper is organized as follows. The next section provides information about the simulation-based task and the data under study. The series of five analyses are then described in detail regarding the methods and results. The Discussion section concludes the findings, addresses the implications of the results for the design, assembly, and scoring of simulation-based tasks, and discusses possible directions for future research.

DATA

As mentioned earlier, this study is based on a secondary analysis of the existing data on the simulation-based task about volcano science published in Hao et al. (2015, 2017) and Liu et al. (2015). The simulation-based task (referred to as *the task* henceforth) was designed to measure science inquiry skills on volcano science and delivered to 463 test takers on Amazon Mechanical Turk. Each test taker interacted with two virtual agents to complete 11 items embedded in a common scenario. The task began with an introduction to scientific information about volcano eruptions, followed by seven selected-response items on knowledge assessment (Items 1–7), and then four CR items on a simulation (Items 8–11). Among the four items about the simulation, the test takers were supposed to conduct a simulation on Item 8, in which they had to decide on the number of seismometers they wanted to use to monitor the volcano and then placed them in different regions around the volcano to collect data; they were then asked to explain why they chose that number of seismometers and the time duration they wanted to collect the data on Items 9 to 11. **Table 1** presents the type and format of the 11 items, with some details (e.g., number of options per MC item and what actions were required per CR item) that will be used in discussing the analysis results. In this study, all items were scored dichotomously as 0 (incorrect) or 1 (correct). For Item 8, the score was based on the correctness and completeness of the simulation. As will be shown, such items may not be difficult but are typically time-consuming. It is noteworthy that, as compared to traditional educational tests, the level of task complexity—in terms of multiple item types and formats, and the actions required to achieve a correct answer to the embedded items—is unusual. Thus, some findings in this study are likely unique to simulation-based tasks and not necessarily generalizable to traditional educational tests with MC and/or CR items.

TABLE 1 | Information about the 11 items in the task.

Item	Type	Format	Note	Chance-level proportion correct
1	MC	Single selection	1 out of 4 options	1/4
2	MC	Single selection	1 out of 4 options	1/4
3	MC	Multiple selection	2 out of 3 options	1/3
4	MC	Order	order 5 options	1/120
5	MC	Single selection	1 out of 4 options	1/4
6	MC	Single selection	1 out of 4 options	1/4
7	MC	Single selection	1 out of 4 options	1/4
8	CR		Simulation item	0
9	CR		Explain the design	0
10	CR		Explain the design	0
11	CR		Explain the design	0

The chance-level proportion correct refers to the expected probability of answering the item correctly by guessing. This information is used in Analysis 5.

The test takers could only take the items in the delivery order and were not permitted to revisit earlier items in the task. There was no time limit imposed on the task and everyone completed the task, so the data involved no missing item responses and RTs. For each test taker, the overall task time comprised two portions—one portion involving the time spent listening to scientific information about the common scenario, and the other portion involving the time spent working on the embedded items. The former portion was a fixed amount of time paced by the system, and was ignored in the rest of the study. The latter portion consisted of the item RTs under evaluation. In this study, we chose to consider the item-level RTs as the starting point to navigate the person-task interactions in the task, together with item-level responses. This choice facilitated the comparison of findings across items within the simulation-based task, and between the simulation-based task and the traditional educational tests examined in the RT literature. In this paper, for each individual, the task score refers to the sum of the responses to the 11 items, and the task time refers to the sum of the 11 item RTs.

In addition to the task, the test takers also responded to a standalone test for general science knowledge (with 37 single-selection MC items, referred to as *the MC test* henceforth), a demographic survey (including questions about their motivation level when completing the task), and a 10-item personality survey (Gosling et al., 2003). For more details about these different task/test/surveys, see Hao et al. (2017). The scores on the MC test and the responses to the survey questions were available for 445 of the 463 test takers, and this additional information was used as person covariates in the study. Thus, data from the 445 test takers were used in all analyses. Below is some information about the composition of the test-taker sample under study:

(a) About 63.6% of them were male.
(b) Their age ranged from 18 to 51, with a median of 24.
(c) They could be classified into four major ethnic groups—White (75.5%), Asian (12.8%), Black (6.1%), and others (5.6%).
(d) Regarding their career plan after college[1]—about 70.8% planned to work or worked full time, about 22.3% planned

[1]The original question and options were as follows: "What do you expect your main activity will be in the year after you leave college? A. Working full time; B. Attending graduate school; C. Serving in the military; D. Other." Given the age range of the test takers, it was assumed that the test takers selected the option that best described their situation at the time they took the task, which either had occurred or had been planned.

to attend or attended graduate school, and the rest had other plans.

(e) On the three motivation questions—did you find the task engaging? Did you find the task interesting? Did you learn something new from the task?—the fractions of the test takers answering 1, 2, or 3 (from agreeing most to least) was about 60%, 35%, and 5%, respectively.

It is worth mentioning that the test takers on Amazon Mechanical Turk were recruited to complete the task, the MC test, and the surveys. The MC test had several items designed to monitor if the test takers paid enough attention to the test and that might affect the payment. One example item was as follows: Which of the following cannot be found on earth? (a) Ocean; (b) moon (key); (c) dessert; and (d) woods. Those items were so easy that any test taker in the sample who considered them were able to answer correctly. All of the test takers included in this study answered the attention-track items correctly. Thus, it is expected that the test takers would be motivated in completing the task to some extent, although they experienced no consequences for their performance on the task and the MC test.

ANALYSES AND RESULTS

In this section, we present five analyses that were intended to investigate the following aspects of the task times and item RTs collected from the task:

(1) Statistical properties of the task times and item RTs.
(2) How did the task times relate to the test takers' performance on the task/MC test, demographics, motivation level, and personality?
(3) How did the item RTs and responses relate to each other?
(4) Did the test takers show different timing patterns across items? Did they inform differences in strategies/time allocation on the task?
(5) Did the test takers show rapid-guessing behavior on this task? Was there a clear motivation issue in this dataset?

ANALYSIS 1: STATISTICAL PROPERTIES OF THE TIMING DATA

Because the task was given without time limits, the first question to answer was how the task times and item RTs varied for different test takers. Descriptive statistics of task scores and item responses were evaluated to complement the timing analysis at different levels. In addition, how the timing variables were distributed was of interest, as later analyses involved modeling of task times and RTs.

Methods

Basic summary statistics were computed for task times and item RTs. Boxplots were made to show possible differences in the RT distributions for the 11 items. Preliminary results suggested that the histogram of task times and the histograms of item RTs had unimodal, right-skewed shapes. Thus, the distribution of task times and the distributions of item RTs were examined via QQ-plots and the Kolmogorov-Smirnov test, with respect to three theoretical models with these properties—lognormal model, gamma model, and Weibull model. These are three popular parametric models in time-to-event studies in survival analysis (Kalbfleisch and Prentice, 2002). The Kolmogorov-Smirnov test is a non-parametric test of the equality of continuous probability distributions that can be used to compare the empirical distribution function of a sample with a reference (theoretical) probability distribution. The type I error rate was set at 0.05 for evaluating the Kolmogorov-Smirnov test results.

Results

Regarding the task-level data, the task times were typically short, ranging from 1.5 to 18.3 min. The first quartile, the median, and the third quartile of the task times were equal to 3.3, 4.2, and 5.3, respectively. The task scores ranged from 0 to 11, with the first quartile, the median, and the third quartile of the task scores equal to 6, 8, and 9, respectively. Overall, the test takers had decent performance on the items without spending much time. These test takers also performed well on the MC test, with the middle 50% of test takers scoring between 25 and 32 on a 0–37 scale.

Regarding the item-level data, **Figure 1** (left panel) shows the boxplots of RTs by item (with 23 observations with RTs greater than 150 s excluded from the plot to make the RT patterns clearer to see). It is clear that the RT distributions varied across items in terms of both central location and dispersion, although the majority of the RTs were below 50 s for all items except Item 8 (this item took more time relative to other items). These RTs were generally short, as compared to those in the traditional educational assessments discussed in the RT literature. As depicted in **Figure 1** (right panel), the items were easy for the test takers. All of the items, except the last one, had a proportion correct greater than 0.5 (four were above 0.85). Items 7 and 8 present a clear contrast concerning time-consumption and difficulty—both items were very easy; but for the majority of the test takers, Item 7 could be answered in 10 s, while Item 8 took about 30 to 62 s. As shown in **Table 1**, these two items are very different in terms of item type: Item 7 is a single-selection MC item, while Item 8 is a simulation item.

Regarding the distribution of the timing data, the empirical timing distributions were compared to three theoretical models—lognormal model, gamma model, and Weibull model. **Figure 2** presents three QQ plots that compared the empirical distribution function of the task times with the best fitting distribution of the three models. Among the three QQ plots, the lognormal model approximated the task times very well and outperformed the other two models—all of the points lay on the reference line except for 8 outliers (<2%) at the right tail. Results of the Kolmogorov-Smirnov test also suggest that the lognormal model supported the observed task times. The Kolmogorov-Smirnov test statistics for the best fitting lognormal model and gamma model were equal to 0.04 (p-value = 0.13) and 0.06 (p-value < 0.001). Similarly, the lognormal model generally supported the RTs per item, although different central locations and dispersion levels should be considered for different items. Overall, results from this analysis indicate that simple statistical

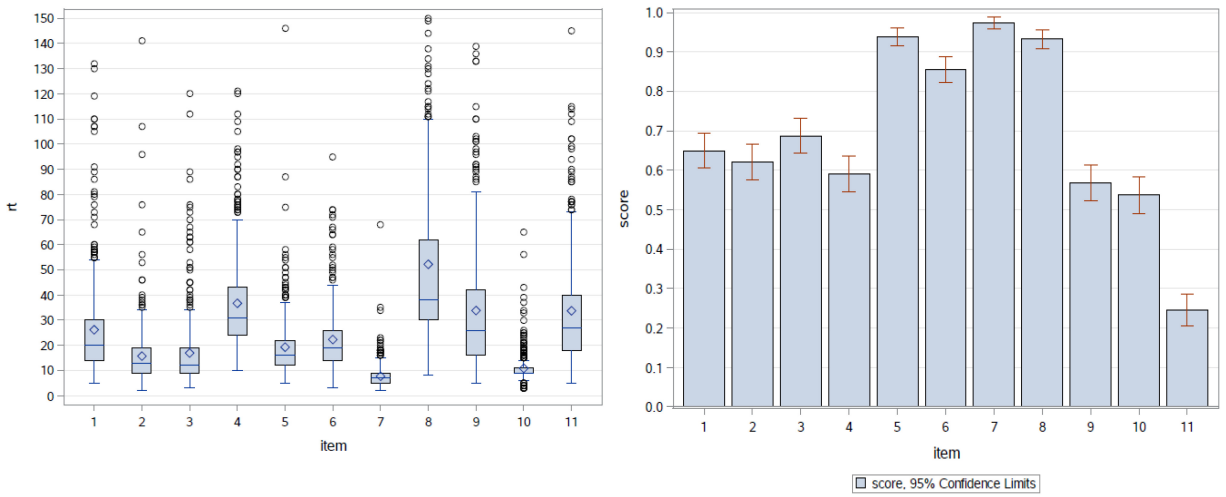


FIGURE 1 | Boxplots of RTs by item (left) and proportion correct by item (right).

models, such as lognormal regression, are appropriate for modeling the task times (Analysis 2) and for modeling the item RTs (Analysis 3) in this task.

ANALYSIS 2: HOW DID TASK TIMES RELATE TO PERFORMANCE AND OTHER INFORMATION AVAILABLE FOR THE TEST TAKERS?

As noted in the Data section, additional data were available for the test takers. Because the focus of this study was on the test takers' timing data, variables derived from the additional information, including task score, were used as person covariates (i.e., predictors) in this analysis to investigate their relationship with the task times. The research question was to what extent the variations in the task times can be explained by these person covariates.

Methods

To examine the effects of the person covariates on the task times, normal linear regression was employed to fit the log-transformed task times with different sets of predictors[2]. There were 27 possible covariates for the test takers:

- Two scores, one on the task and the other on the MC test. The correlation between these two scores was equal to 0.43 (p-value < 0.0001).
- Twelve demographic variables, including age, gender, ethnicity, high school type, experience in science, career plan after college, and home environment (related to science learning). All but "age" were treated as nominal variables.
- Three variables from the motivation questions—did you find the task engaging? Did you find the task interesting? Did you learn something new from the task? All were treated as nominal variables, each with three categories[3].
- Ten personality variables from the personality survey. All were treated as nominal variables, each with five categories[4].

Three models were considered. There was a base model that only included an intercept and no predictor. Model 1 included an intercept and eight predictors that were chosen subjectively from the 27 possible covariates. The eight predictors were the task score, age, gender, ethnicity, career plan after college, and the three motivation variables. Compared to

[2]Fitting normal linear regression with the log-transformed task times or fitting lognormal regression with the task times did not yield noticeable differences in the results, so only the former case was discussed.

[3]The responses (1–3) to the three motivation questions were ordinal in nature. Treating each motivation question as an ordinal variable had the advantage of estimating 1 fewer parameter, but it assumed that the successive response categories were equally spaced (Long and Freese, 2006, p. 421) and had monotonic effects on the log-transformed task times. To assess the effect of this variable treatment, in addition to the models presented in **Table 2**, Models 1 and 2 with the three motivation variables treated as ordinal were also considered. The corresponding adjusted R^2 were equal to 0.02 (11 degrees of freedom) and 0.05 (9 degrees of freedom), respectively. This version of Model 2 selected the same final predictors except for "Did you find the task interesting?" which did not enter the model, and this version performed slightly worse than the Model 2 in **Table 2**. An alternative approach was considered that replaced the three separate ordinal motivation variables with their sum score in Models 1 and 2, but this replacement also did not improve the model fit. Thus, treating the motivation questions as separate nominal variables was preferred with this dataset.

[4]The test takers were asked to rate the extent to which they agreed or disagreed with the statement in each personality question from 1 (disagree strongly) to 5 (agree strongly). When treating them as ordinal variables, one could compute the Pearson correlation between the responses to a pair of questions. It was found that the Pearson correlations between any two of the 10 personality variables were below 0.4 in absolute value (i.e., weak correlation), except for two cases with correlations of −0.48 and −0.66. There was no difference in handling each of the 10 personality variables as original or nominal in Model 2 in terms of adjusted R^2. However, replacing the 10 separate personality variables with their sum score in Model 2 led to an adjusted R^2 equal to 0.02 (3 degrees of freedom), with only one predictor selected (How many books at home?). Thus, this version of Model 2 was not discussed further.

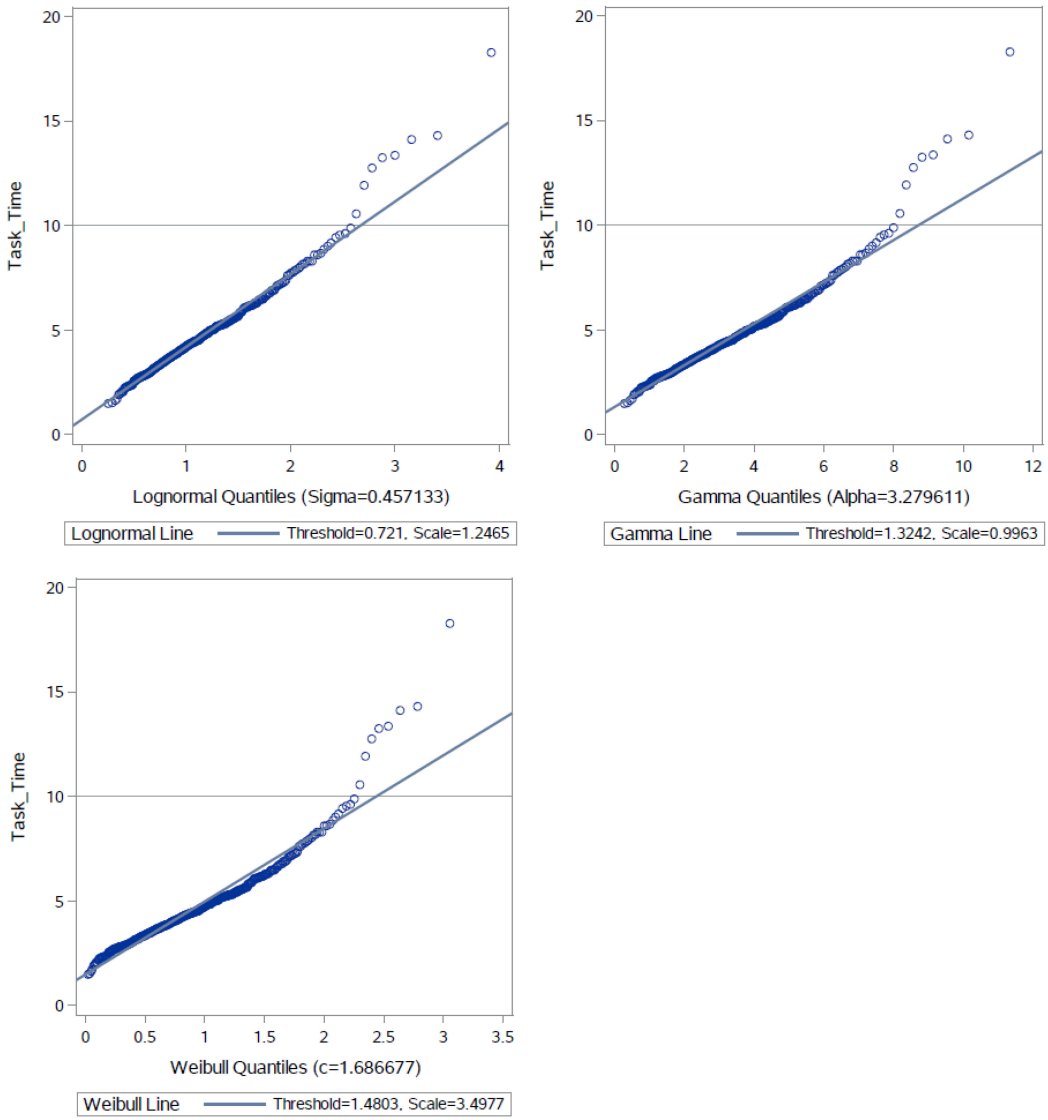


FIGURE 2 | QQ plots for task times against three theoretical models.

the rest of the person covariates, these eight predictors are more commonly available in different large-scale educational assessments, so their effects on the task times were of interest and assessed in Model 1. The second model concerns a stepwise regression (Draper and Smith, 1998, ch. 15) that identified useful predictors from all 27 possible person covariates. The predictors were added one by one to the model only if the F statistic for a predictor was significant at the 0.05 level, which is recommended by Draper and Smith (1998), p. 342) for stepwise linear regression. The same criterion was used for removal of predictors. The final model is referred to as Model 2. The residual root mean squared error (RMSE), the estimated coefficient of determinationR^2, and the estimated adjusted R^2 were reported for each model. The RMSE represents the variability of the log-transformed task times once all useful predictors are included. Adjusted R^2 was considered, because it combines information about model fit with number of parameters. Other measures, such as information criteria (Akaike, 1974; Schwarz, 1978), might be employed for the same purpose.

Results

Table 2 summaries the model-fitting results. The stepwise regression approach selected 4 predictors out of 27 and outperformed Model 1, in which the 8 predictors were chosen subjectively. The final 4 predictors in Model 2 and the estimated effects on the (log-transformed) task times are as follows:

- Career plan after college? Test takers who worked full time or attended graduate school tended to have shorter task times than those with other plans.

TABLE 2 | Model-fitting results for task times.

Model	Number of predictors	Model degrees of freedom	RMSE	R^2	Adjusted R^2
Base	0	0	0.38	0.00	0.00
Model 1	8	14	0.37	0.05	0.02
Model 2	4	11	0.36	0.08	0.06

The model degrees of freedom refers to the number of coefficients associated with the predictors and does not include the intercept.

- Personality variable—disorganized/careless? Test takers who strongly agreed that they were disorganized/careless spent less time than those who did not agree strongly.
- How many books at home? Test takers with enough books to fill one shelf, 11–25, tended to spend longer task times than did those with fewer or a lot more books.
- Did you find the task interesting? Test takers who chose 2 spent slightly less time than those who chose 1 (agreed most) or 3 (agreed least) did.

Although interesting, Model 2 explained only about 6% of the variability in the log-transformed task times and did not substantially reduce the RMSE relative to the base model. It was therefore concluded that none of the person covariates available in the dataset had clear effects on the test takers' time on task, and further details about the parameter estimates in Model 2 are omitted.

ANALYSIS 3: HOW DID ITEM RTS AND RESPONSES RELATE TO EACH OTHER?

There are many ways to examine the relationship between the observed RTs and item responses. If one assumes that the task may measure two latent traits per test taker, ability and speed, then a possible approach is the hierarchical framework for joint modeling item responses and RTs (van der Linden, 2007). This framework assumes that each test taker operates at fixed levels of speed and ability in a test. It tends to be adequate for tests with generous time limits (van der Linden, 2007, p. 292) or without any time limits—that is, the task under study.

Methods

The hierarchical framework assumes that the task measures two latent traits for each test taker *j*, one for ability θ_j and the other for speed τ_j which may be correlated among a group of test takers of size $J = 445$. It also assumes that each item *i*, $1 \leq i \leq I = 11$, can be characterized by such parameters as difficulty b_i, time-intensity β_i, time-discrimination α_i, and so on, some of which may be correlated among items in a test. Let Y_{ji} and T_{ji} be test taker *j*'s response and RT on item *i*, respectively. The hierarchical framework assumes that, conditioning on the parameters for test takers and for items, item responses Y_{ji} and RTs T_{ji} on the task items are independent and can be modeled separately at Level 1 of the framework by an IRT model (for item responses Y_{ji}) and a timing model (for RTs T_{ji}). At Level-2 of the framework, the correlation between person parameters (i.e., ability θ_j and speed τ_j) across test takers and the correlations between item parameters across items are captured in the multidimensional prior distributions and can be estimated from the data.

Due to the small sample size and short test length, the Rasch model was employed to model item responses Y_{ji}, with the conditional probability equal to

$$p\left(Y_{ji}|\theta_j, b_i\right) = \frac{1}{1 + \exp\left[-\left(\theta_j - b_i\right)\right]}.$$

According to the results in Analysis 1, the item RTs supported a lognormal model reasonably well but tended to have different central locations and levels of dispersion in the distributions for different items. Thus, a lognormal regression model with two item parameters, one for time-intensity β_i (to describe possible differences in the central location) and the other for time-discrimination α_i (to describe possible differences in the dispersion), was chosen to model the RTs. More specifically, the regression of the logarithm of T_{ji} on test taker *j*'s speed τ_j and item *i*'s time-intensity β_i may be expressed as

$$\log\left(T_{ji}\right) = \beta_i - \tau_j + \varepsilon_{ji}$$

where the random error $\varepsilon_{ji} \sim N\left(0, \alpha_i^{-2}\right)$. Parameter τ_j indicates the speed of test taker *j*, larger τ_j for faster respondents. Parameter β_i represents the time-intensity of item *i*: the larger the β_i, the more time item *i* requires for the test takers to respond. Parameter α_i represents the discriminating power of item *i* in RTs, and larger α_i corresponds to less variable T_{ji} across test takers. The probability density function (PDF) of T_{ji} is equal to

$$f\left(t|\tau_j, \beta_i, \alpha_i\right) = \frac{\alpha_i}{t\sqrt{2\pi}} \exp\left\{-\frac{\alpha_i^2}{2}\left[\log\left(t\right) - \left(\beta_i - \tau_j\right)\right]^2\right\}.$$

Level-2 of the framework involves joint models of the person parameters and of the item parameters. The joint distribution of the test taker's ability θ_j and speed τ_j, $1 \leq j \leq J$, was assumed to follow a bivariate normal distribution,

$$\begin{pmatrix}\theta_j \\ \tau_j\end{pmatrix} \sim N_2\left(\mu_p, \Sigma_p\right)$$

with the mean vector $\mu_p = (0, 0)'$ and covariance matrix

$$\Sigma_p = \begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2\end{pmatrix}.$$

Let $\rho_{\theta t} = \sigma_{\theta t}/\left(\sigma_\theta \sigma_t\right)$ be the correlation between ability θ_j and speed τ_j across *j*. Similarly, for item parameters, a bivariate normal distribution was assumed for item difficulty b_i and time-intensity β_i, $1 \leq i \leq I$,

$$\begin{pmatrix}b_i \\ \beta_i\end{pmatrix} \sim N_2\left(\mu_I, \Sigma_I\right)$$

with the mean vector $\mu_I = \left(\mu_b, \mu_\beta\right)'$ and covariance matrix

$$\Sigma_I = \begin{pmatrix}\sigma_b^2 & \sigma_{b\beta} \\ \sigma_{b\beta} & \sigma_\beta^2\end{pmatrix}.$$

Let $\rho_{b\beta} = \sigma_{b\beta}/(\sigma_b\sigma_\beta)$ be the correlation between difficulty b_i and time-intensity β_i across i. The item parameter moments were constrained from the general case, which includes time-discrimination α_i in the item parameter vector. In this study, time-discrimination α_i was estimated separately. The α_i was assumed to be independent of b_i and β_i for two reasons. First, previous studies (e.g., Bolt and Lall, 2003; Fox et al., 2014) indicate that the correlations between the time-discrimination α_i and the other item parameters (b_i and β_i) provide negligible information about the item quality or person latent traits, especially the relationship between speed and accuracy among test takers. Thus, by following the convention of jointly estimating an RT model and an IRT model, the covariances related to time-discrimination α_i were ignored. Second, forcedly estimating the covariances related to time-discrimination α_i might cause an over-fitting issue with complex hierarchical modeling, which might yield untrustworthy person parameter estimates. Thus, the mentioned constraints were applied.

A software program that implements a Bayesian MCMC approach with Just Another Gibbs Sampler (JAGS; Plummer, 2015) was employed to estimate the model parameters (Man et al., 2019). The prior distributions for estimating the mean vector and the covariance structure of the item difficulty and time-intensity were specified as follows:

$$\mu_b \sim N(0, 2), \mu_\beta \sim N(4.5, 2), \Sigma_I \sim IW(I_{I0}, \nu_{I0}),$$
$$\text{and } \alpha_i \sim InvGamma(1, 1),$$

where IW denotes the inverse-Wishart distribution, $InvGamma$ denotes the inverse-gamma distribution, I_{I0} is a 2×2 identity matrix, and ν_{I0} indicates the degree of freedom, which in this case is 1. Likewise, the prior distribution for estimating the covariance structure of the person parameters is defined as $\Sigma_p \sim IW(I_{I0}, \nu_{I0})$, the same distribution as Σ_I given above. Model parameters were estimated by the posterior mean, or the expected *a posteriori* (EAP) estimate, through the algorithm.

The *R2jags* package (Su and Yajima, 2015) was utilized to run JAGS in R (R Core Team, 2016). The potential scale reduction (PSR) factor was used for evaluating the model parameter convergence (Gelman et al., 2003).

Results

For parameter estimation with this dataset, the MCMC approach involved two chains, each with thinning of 5 using 15,000 total iterations with a 5,000 burn-in. In this study, a PSR value of a parameter estimate lower than 1.1 indicates satisfying convergence (Gelman and Rubin, 1992a,b). **Figure 3** shows that the estimation of all of the parameters converged, as all the PSR values were lower than 1.1. The current choice of hyperpriors $N(0, 2)$ and $N(4.5, 2)$ for μ_b and μ_β seemed suitable for the dataset with the use of the Rasch model and the two-parameter lognormal RT model as the two chains reached their convergence. Also, the current setting of priors follows the convention of fitting IRT and RT models with Bayesian estimation (e.g., van der Linden et al., 2010; Natesan et al., 2016; Luo and Jiao, 2018). However, whether such hyperpriors generally work for jointly modeling RTs and responses in the hierarchical framework needs to be addressed by additional sensitivity analysis.

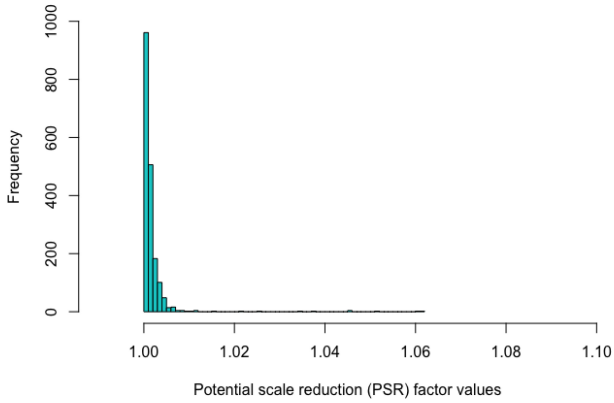


FIGURE 3 | Histogram of the values of the potential scale reduction factor based on the fitted joint model.

The histogram of the EAP estimates of the ability parameters (**Figure 4**, left panel) was skewed to the left, while the histogram of the EAP estimates of the speed parameters (**Figure 4**, right panel) was roughly symmetric. Both histograms had mean equal to 0 due to the imposed constraints for identifiability of the model parameters, but the EAP estimates of the ability parameters were much more variable than were the EAP estimates of the speed parameters.

On the other hand, there was a tiny, but statistically significant, positive correlation between the ability and speed parameters among the test takers. Based on the estimated Level-2 model parameters in **Table 3**, the estimated correlation $\hat{\rho}_{\theta t} = \hat{\sigma}_{\theta t}/(\hat{\sigma}_\theta\hat{\sigma}_t) = 0.04/(0.85 \cdot 0.08)^{1/2} = 0.17$, with a 95% credible interval (0.034, 0.302). A positive correlation between the ability parameter and the speed parameter for a test-taker sample implies that more proficient test takers tended to work faster on the task. This level of correlation is very weak compared to many reported studies based on the same hierarchical framework. For instance, Klein Entink et al. (2009) reported an estimated correlation of −0.76 for a low-stakes assessment and an estimated correlation of 0.3 for a personality questionnaire; Wang et al. (2013) found an estimated correlation of 0.71 for a high-stakes adaptive test; Zu et al. (2016) showed estimated correlations of 0.59 for a high-stakes Listening test and of 0.86 for a high-stakes quantitative reasoning test. The authors noted that the correlation between ability and speed probably depends on the test context and content, type of test, type of item, and the test conditions. There are many possible reasons for the finding of a weak positive correlation observed in this dataset, such as different item types among the 11 items (especially simulation items vs. others), no time limit on the task, and not a challenging task to the test takers so that spending more or less time did not affect the accuracy of their responses substantially.

Based on the estimated Level-2 model parameters in **Table 3**, the estimated correlation between the items' difficulty and time-intensity, $\hat{\rho}_{b\beta} = \hat{\sigma}_{b\beta}/(\hat{\sigma}_b\hat{\sigma}_\beta) = 0.24/(2.77 \cdot 0.43)^{1/2} = 0.22$

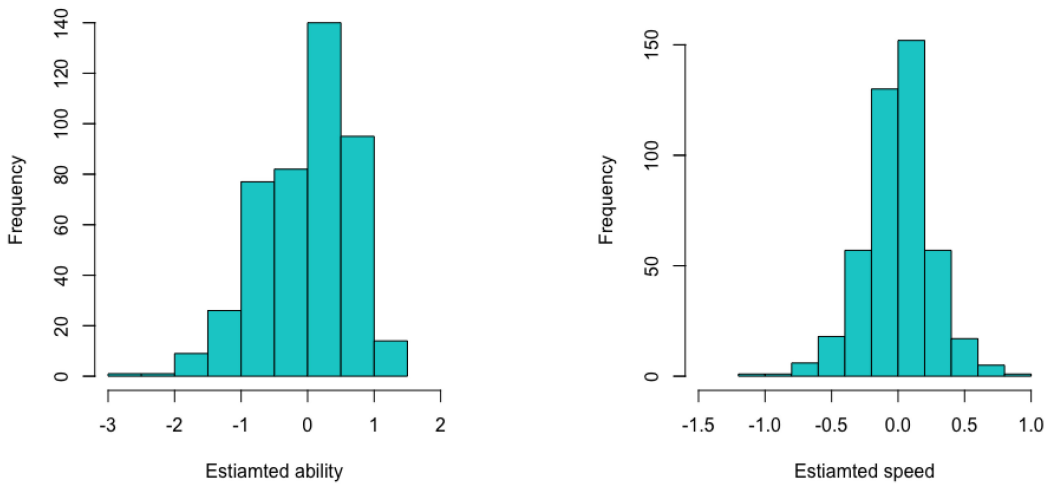


FIGURE 4 | Histograms of the EAP estimates of the ability parameters (left) and of the speed parameters (right) for the test takers in the dataset.

TABLE 3 | Estimates of item parameters and level-2 model parameters.

	RT					Rasch	
	Time-intensity			Time-discrimination		Difficulty	
	EAP	SE		EAP	SE	EAP	SE
β_1	3.06	0.02	α_1	4.02	0.30	b_1	−0.72	0.12
β_2	2.60	0.03	α_2	5.12	0.37	b_2	−0.58	0.11
β_3	2.59	0.03	α_3	3.68	0.27	b_3	−0.93	0.12
β_4	3.49	0.02	α_4	7.58	0.57	b_4	−0.43	0.12
β_5	2.81	0.02	α_5	6.28	0.46	b_5	−3.07	0.22
β_6	2.96	0.02	α_6	5.59	0.40	b_6	−2.05	0.16
β_7	1.94	0.02	α_7	7.45	0.57	b_7	−3.90	0.30
β_8	3.76	0.02	α_8	3.80	0.27	b_8	−2.94	0.22
β_9	3.27	0.03	α_9	2.65	0.19	b_9	−0.32	0.11
β_{10}	2.31	0.02	α_{10}	10.03	0.83	b_{10}	−0.18	0.11
β_{11}	3.31	0.03	α_{11}	3.16	0.23	b_{11}	1.30	0.13
Covariance matrix of item parameters								
σ_b^2	2.77							
$\sigma_{b\beta}$	0.24							
σ_β^2	0.43							
Covariance matrix of person parameters								
σ_θ^2	0.85							
$\sigma_{\theta\tau}$	0.04							
σ_τ^2	0.08							

For each item parameter, the EAP and SE are the posterior mean and posterior standard deviation, respectively.

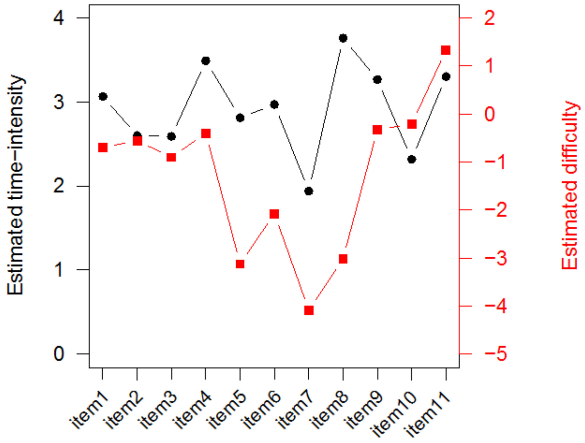


FIGURE 5 | EAP estimates of item difficulty and time-intensity by item.

with a 95% credible interval (−0.376, 0.710). Thus, there was no clear relationship between the items' difficulty and time-intensity. Item type is likely a key factor for this finding. In addition to the Level-2 model parameters, **Table 3** also summarizes the estimates of all item parameters. To better associate the combinations of the estimated item difficulty and time-intensity with the 11 items, **Figure 5** depicts their EAP estimates by item. For example, the least time-consuming item (Item 7, a single-selection MC item) was the easiest item in the task, but the most time-intensive item (Item 8, a simulation item) was also very easy. It is common for simulation-based tasks to include simulation items, which ask the test takers to follow specific instructions to conduct an experiment or a simulation, and such items are usually scored based on the completeness of the experiment/simulation. Relative to other item types, simulation items may not be difficult, but they are typically time-consuming. In the task under study, the most time-intensive but very easy item was indeed one such item, which asked the test takers to decide on the number and locations of seismometers to be placed around a volcano in order to collect proper data for later analyses. The simulation items in the two NAEP simulation-based tasks revealed the same pattern of time-intensive but easy (Jia and Lee, 2018).

Figure 6 presents the item characteristic curve based on the fitted Rasch model with the observed proportion correct for the 11 items. To evaluate the observed proportion correct, the test

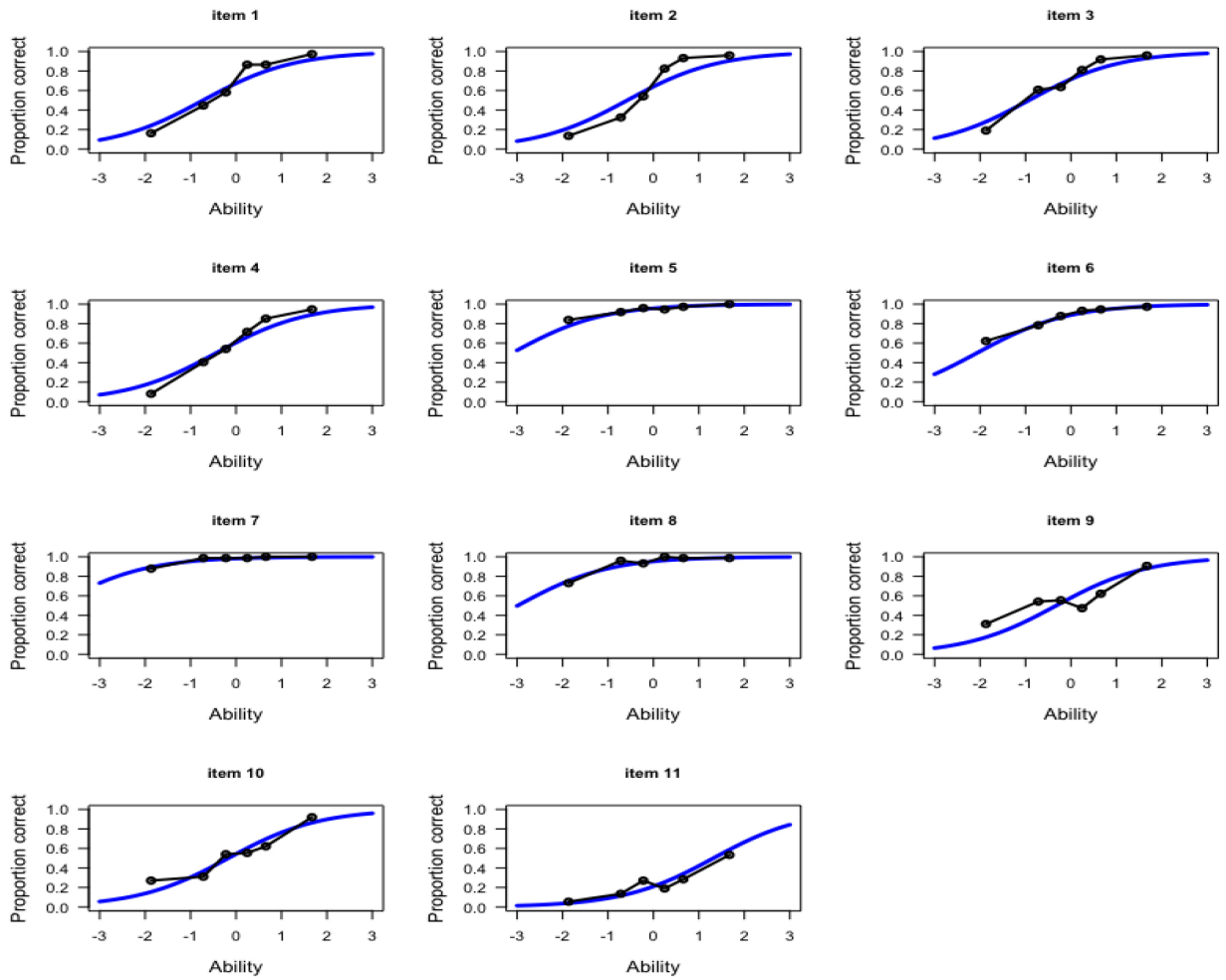


FIGURE 6 | Item characteristic curve (solid blue line) with observed proportion correct (black dots line) for the 11 items.

takers were classified into 6 equal-size groups based on their EAP estimates, and then the fraction of correct responses was computed per group for each item.

ANALYSIS 4: DID THE TEST TAKERS SHOW DIFFERENT TIMING PATTERNS ON THE TASK?

The preceding section considers a parametric approach to jointly modeling item responses and RTs. The hierarchical framework in van der Linden (2007) makes assumptions that each test taker operates at fixed levels of speed and ability, and is not designed to detect different test-taking behaviors/strategies or potential latent classes. In practice, test takers may employ different strategies to allocate their time across items. Cluster analysis is a useful approach to studying different patterns of the trend and variation in RTs across items among a test-taker sample. Test takers showing similar RT patterns would be identified as a cluster. Through examination of the identified clusters, the analysis may suggest differences in strategies/behaviors across test takers and changes in strategies/behaviors across items.

Methods

This analysis examined the RT patterns across the 11 items to look into possible trends and variations of the test takers' response processes. Each test taker's RT pattern spanned an 11-dimensional space, and a hierarchical cluster analysis was applied to the RT patterns of all test takers to find out how they clustered in the 11-dimensional space. After experimenting using a number of clustering methods and distance metrics, it was found that a hierarchical clustering approach with the Euclidean distance calculated from the RTs and the Ward linkage (Ward, 1963) led to the most interpretable clustering of test takers. By using the Ward linkage, a pair of clusters being chosen to be merged at each step of the hierarchical clustering process will minimally increase the total within-cluster variance. We determined the final number of big clusters based on the elbow point of the inter-cluster distances. After the clusters were identified, given a cluster, the mean of RTs was computed for each item, and the 11-dimensional

mean RT vector was graphed to depict the trend and variation of the general RT pattern for the cluster. To evaluate if the clusters had different overall performance in terms of accuracy and timing or item performance, the test takers' task times, task scores, and item responses were compared by cluster. The person covariates used in Analysis 2 were also considered for further investigation of the clusters.

Results

Figure 7 shows the cluster dendrogram. Based on the elbow point of the linkage (**Figure 8**), three clusters were identified. For each cluster, the average time spent on each item is shown in **Figure 9**. One may observe that cluster 1 (with 12 test takers) corresponded to a "slow" response pattern, as those test takers spent more time on average on almost all items. Cluster 2 (with 222 test takers) corresponded to a "fast" response pattern, as the test takers spent less time on average on every item. Cluster 3 (with 211 test takers) corresponded to a "moderate" response pattern, as their average RTs lie between the average RTs of those in cluster 1 and cluster 2 on most of the items. All three clusters shared a somewhat similar timing trend on most items but deviated from the trend on specific items. The common timing trend generally follows the patterns of item time-intensity observed in the boxplots of items in Analysis 1 (**Figure 1**) and estimated in the hierarchical framework in Analysis 3 (**Figure 5**). The differences among the test takers' RT patterns translated into different estimated speed. The existence of the three clusters with different RT patterns did not distort the RT distributions for individual items: the RT distributions of clusters 2 and 3 overlapped and did not appear as distinct peaks; cluster 1 only had 12 test takers and their RTs tended to appear as outliers in the overall RT distribution per item rather than a second mode. Thus, there was no evidence against using a lognormal distribution in modeling RTs (see section "Results" in Analysis 1), and the fact that the estimation of the model parameters converged successfully in the hierarchical modeling (Analysis 3) provided a sign of reasonable fit.

Figure 10 shows the boxplot of task times by cluster (left panel) and the mean task score by cluster with the associated 95% confidence limits (right panel). The task-level timing differences among the three clusters agreed with the findings regarding the item-level timing patterns discussed above (**Figure 9**). One may find that, despite the different timing patterns, the accuracy (as reflected by the task scores) was comparable across the clusters—this result suggests that, although the test takers in different clusters might have approached the items in different ways and that resulted in differences in RTs, their performances were not much affected. This finding is consistent with the observed weak positive correlation between the test taker's speed and ability estimated in the hierarchical framework. More importantly, results from the cluster analysis revealed variations in different clusters' RT patterns across items, especially between cluster 1 and the rest of the test takers (**Figure 9**). The test takers in cluster 1 spent a lot more time to figure out what to do with Item 8, which is the simulation item that asked the participants to decide on the number of seismometers they want to use to monitor the volcano and then place them in different regions around the volcano. Besides the longer RTs on average, the test takers in cluster 1 did not do as well on Item 8 as those in clusters 2 and 3—the proportions correct for clusters 1, 2, and 3 were 0.67, 0.92, and 0.96, respectively. The 12 test takers in cluster 1 also tended to spend more time on two follow-up CR questions about the simulation (Items 9 and 11) and perform slightly worse on these items. In general, such information may be leveraged to supply valuable formative feedback to students, teachers, and assessment developers to help identify potential learning gaps or design issues. With respect to the person covariates, the only more noticeable difference among the three clusters was their gender decomposition: only one-third of cluster 1 (4 out

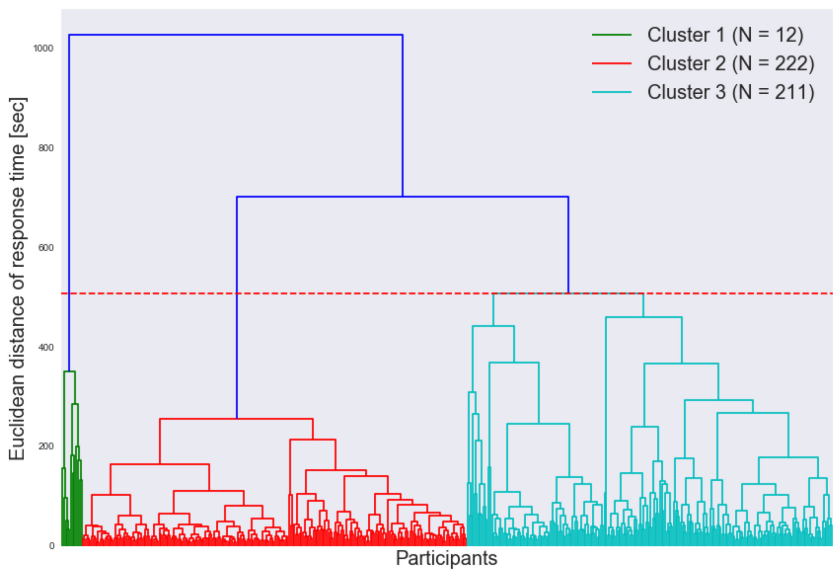


FIGURE 7 | Dendrogram of the clustering of test takers based on RTs.

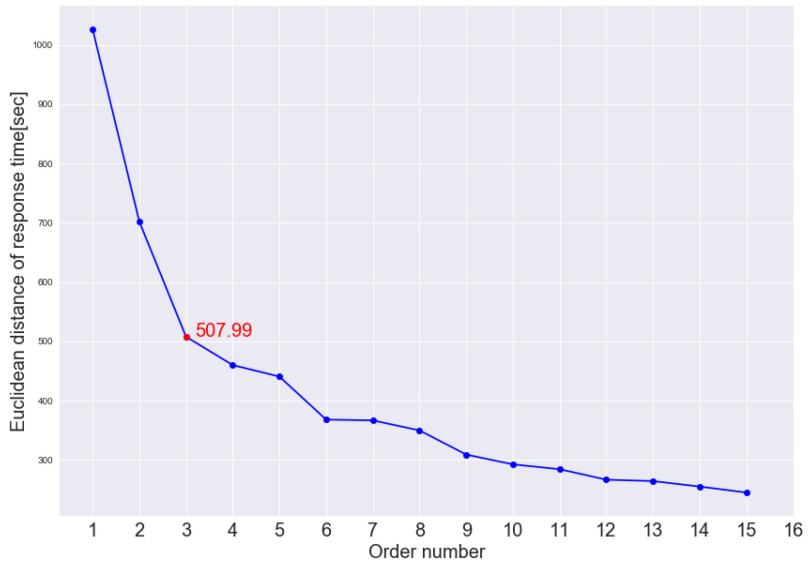


FIGURE 8 | Elbow plot of the inter-cluster distance. The red dot indicates where the elbow point is located.

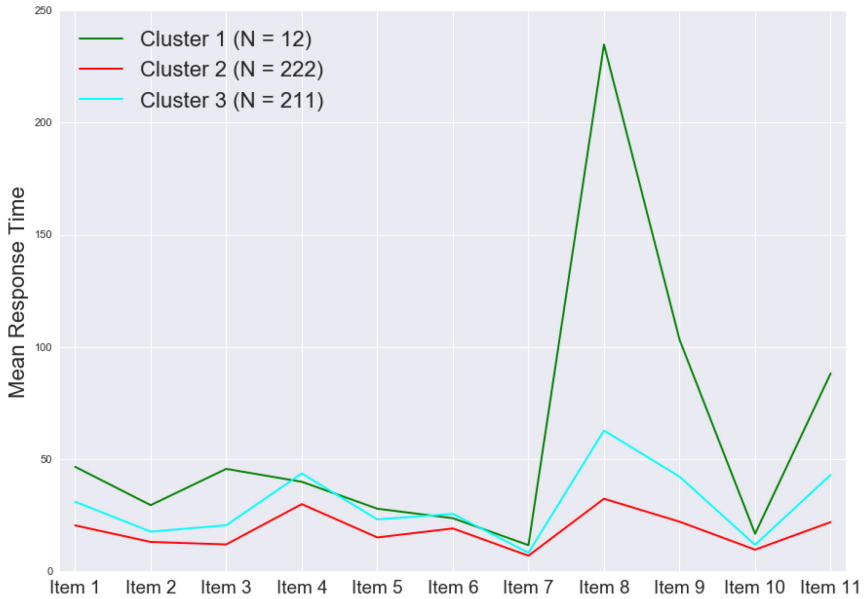


FIGURE 9 | Mean RT by item for each of the three clusters.

of 12) were male, while almost two-thirds of either cluster 2 or 3 were male (which tracked the gender decomposition in the overall sample well).

ANALYSIS 5: DID THE TEST TAKERS SHOW RAPID-GUESSING BEHAVIOR ON THE TASK?

Analysis 4 employed cluster analysis to identify clusters with different timing patterns. The analysis in this section focuses on two specific test-taking behaviors, solution behavior and rapid-guessing behavior. As noted in the Introduction section, RTs have been used to differentiate rapid-guessing behavior from solution behavior. Test takers exhibiting rapid-guessing behavior on an MC item typically spend little time relative to the majority of the test takers, and their probability of answering the item correctly is likely close to the chance-level proportion correct (i.e., the expected probability of answering an item correctly by guessing). Thus, more effective approaches to identifying rapid-guessing behavior consider both item responses and RTs (e.g., Ma et al., 2011; Lee and Jia, 2014; Wang and Xu, 2015;

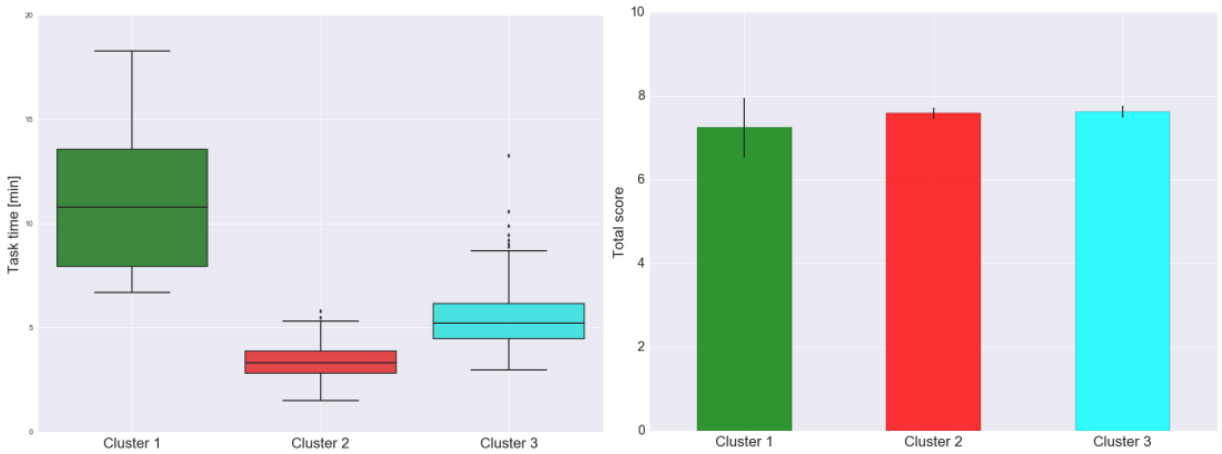


FIGURE 10 | Boxplot of task times by cluster (left) and the mean task score by cluster with the associated 95% confidence limits (right).

Guo et al., 2016). There are many reasons that may lead to the presence of rapid-guessing behavior on a test: a common issue for high-stakes assessments is test speededness, whereas a common concern for low-stakes assessments is motivation. The analysis in this section is intended to assess the extent of rapid-guessing behavior in the task. Because the task was given without time limits, clear presence of rapid-guessing behavior is more likely to indicate motivation issues. If rapid guessing is negligible or not present in a dataset, then motivation is unlikely a concern.

Methods

The non-model-based procedure in Lee and Jia (2014) was originally developed for MC tests. It was adapted by Jia and Lee (2018) to examine rapid-guessing behavior and motivation issues in the two NAEP simulation-based tasks. This procedure examines the items on a test one by one. For each item, it defines a time threshold through visual inspection of the RT distribution with the information of proportion correct evaluated at every observed RT (i.e., conditional proportion correct). For MC items, an identified time threshold for an item should classify the test takers into two groups: One group, which is assumed to exhibit solution behavior, has RTs greater than the time threshold and their proportion correct should be clearly greater than the chance level (i.e., for a 4-option single-selection MC item, the chance-level proportion correct is about 0.25). The other group, which is assumed to exhibit rapid-guessing behavior, has RTs shorter than the time threshold and conditional proportion correct close to the chance level. For items that are unlikely to be answered correctly by guessing (e.g., CR items), the chance level may be set at 0, and the rest of the procedure remains applicable (Jia and Lee, 2018).

Data with larger fractions of RTs falling below the corresponding time thresholds indicate more substantial levels of rapid guessing on the test. If no item involves the patterns of short RTs and chance-level proportion correct, or if the fraction of identified rapid guesses is negligible, then rapid guessing is considered not a concern for the test.

Results

The procedure was applied to each of the 11 items to identify possible time thresholds based on the item-level RT histograms and the associated results of conditional proportion correct. **Figure 11** presents the RT distributions of all 11 items overlaid with the conditional proportion correct represented in red points. As the identification of rapid guesses focuses on shorter RTs, the RT distributions were truncated at the 90th percentile for each item. According to **Table 1**, the chance-level proportions correct for the MC items are as follows: 0.25 for Items 1, 2, 5, 6, and 7; 1/3 for Item 3; and 1/120 for Item 4. Items 8–11 were CR items, so their chance-level proportions correct were set at 0.

Based on **Figure 11**, one could, in a strict sense, identify time thresholds of 6 and 8 (seconds) for Items 9 and 11, respectively, which classified the test takers into the two behaviors—solution behavior vs. rapid-guessing behavior. However, the size of the respective resulting group for rapid-guessing behavior was almost ignorable, that is, 1 (0%) for Item 9 and 8 (<2%) for Item 11. None of the other items had an identifiable time threshold that clearly separates the two behaviors. In fact, most of the items had decent proportions correct for pretty short RTs. Thus, it was concluded that no clear rapid-guessing behavior was detected in this dataset using timing and response data, and motivation is unlikely an issue.

DISCUSSION

This paper presents a systematic RT study on the simulation-based task about volcano science, and investigates different timing aspects of this test-taker sample's behavioral features at the task level and the item level. The goal is to understand the RTs in complex simulation-based tasks so as to gain insights into possible ways of leveraging RT information in designing, assembling, and scoring of simulation-based tasks.

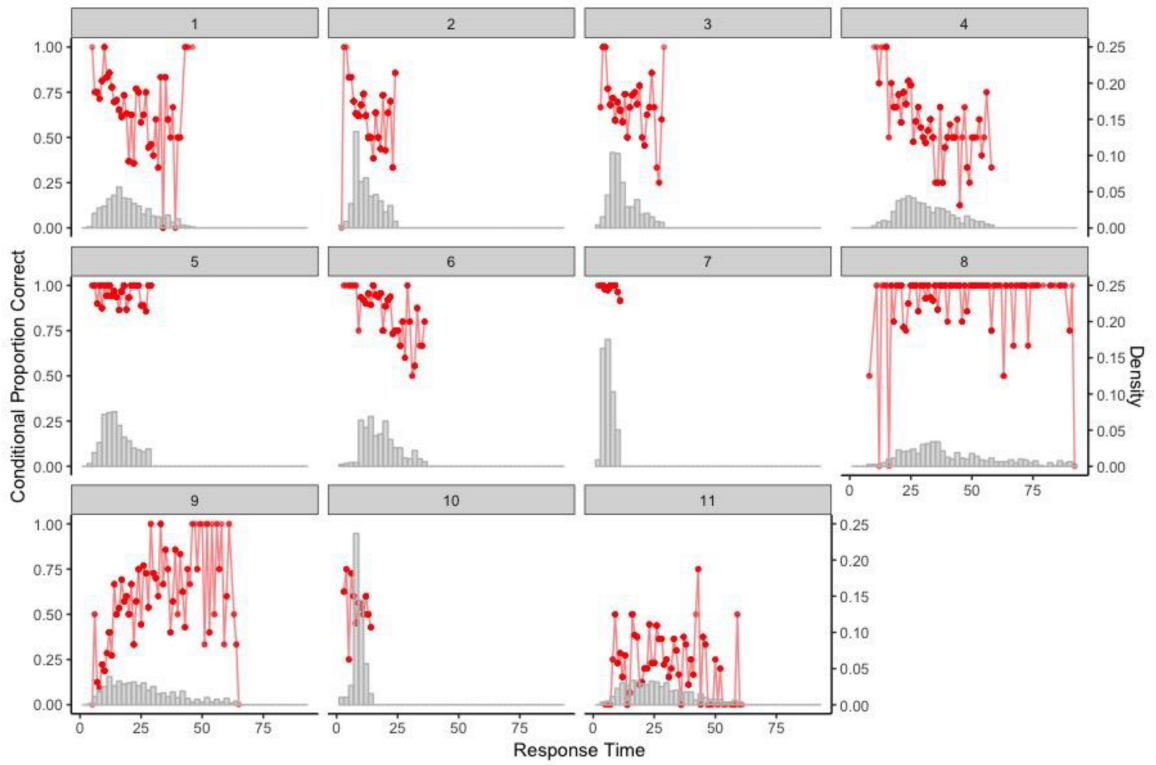


FIGURE 11 | Item-level RT distribution with conditional proportion correct for the 11 items (RTs truncated at the corresponding 90th percentile).

Information about the test takers' performance on the items/task, demographics, motivation level, and personality was also considered. The task involved 11 items of various types associated with a common scenario, and was delivered without time limits. The majority of the test takers spent 6 min or less on the 11 items and performed well.

The five timing analyses revealed the following interesting findings. First, the timing data at both the task level and the item level showed good distributional properties, which made it possible to employ relatively simple statistical models that are unimodal and right-skewed, such as lognormal regression, to analyze the relationship between the timing data and other data available for the test takers. Second, the number of observations being identified as associated with rapid guessing was negligible. Thus, it was concluded that no clear rapid-guessing behavior was observed in this dataset, and motivation was not an issue for this sample-task combination. Third, the items were not time-consuming for this sample, and there was little variability in the task times for this sample. None of the available person covariate (i.e., task performance, demographics, self-reported motivation levels, and responses to personality questions) was useful in explaining the variability in the task times, so there was no notable difference in the task times among any demographic subgroups. The two major clusters identified in the cluster analysis also did not present differences in the RT patterns among the demographic subgroups. Fourth, the results of the hierarchical modeling framework indicated a weak positive correlation estimated between the test takers' ability and speed. The three clusters identified in the cluster analysis also exhibited different RT patterns across the 11 items but comparable task scores. All three clusters shared a somewhat similar timing trend on most items but deviated from the trend on specific items. Last but not least, the hierarchical modeling framework revealed no clear association between the items' time intensity and difficulty. The simulation item had a very different combination of difficulty and time-intensity (easy but very time-consuming) compared to the other items in the task.

There are several implications of the results concerning the design, assembly, and scoring of simulation-based tasks. First, the good distributional property of the timing variables may be attributed to the "no time limit" condition, which implies no constraint on the timing variables and that results in no missing data due to lack of time in both timing and responses. Thus, censoring, a common issue in time-to-event studies in survival analysis (see, e.g., Kalbfleisch and Prentice, 2002; Lee and Ying, 2015), is not a concern in this dataset. Imposing no time limit to a simulation-based task may allow test takers to choose their own pace in working on the items. In contrast, for tasks/tests with an overall time limit, as is the case for typical educational assessments discussed in the RT literature, the presence of time limits may lead to missing item responses and RTs, some extent of speededness, truncated times at the test level and even at the item level, or may introduce between-item dependencies among each test taker's RTs. As a result, more sophisticated statistical models

may better describe RTs and responses in time-limit tests (e.g., Ranger and Ortner, 2012; Lee and Ying, 2015; Bolsinova et al., 2017; Molenaar et al., 2018).

The finding of no clear association between the items' time intensity and difficulty was interesting but not surprising. Among many possible factors, test type and item type may play an important role in this finding, as the simulation item had a very different combination of difficulty and time-intensity (easy but very time-consuming) compared to more traditional MC and CR items. The contrast between simulation items and more traditional MC and CR items in the time spent and difficulty was also discovered in the two NAEP simulation-based tasks studied in Jia and Lee (2018). Thus, this finding is possibly unique to simulation-based tasks, and is not necessarily generalizable to traditional educational tests with MC and/or CR items.

On the other hand, the weak positive correlation estimated between the test takers' ability and speed in this sample suggests that task scores (or item responses) and task times (or RTs) may reveal different useful information about the test takers on the task. The cluster analysis resulted in similar conclusions. Perhaps, two scores may be reported, one about accuracy and the other about speed/efficiency, to describe a test taker's performance on a simulation-based task. The finding of no notable difference in the task times, or in the RT patterns of the two major clusters, among any demographic subgroups indicates that fairness in terms of timing was not an issue for this sample-task combination. However, it is unclear how test takers would change their behaviors when they were told that all process data would be examined and scored. Further research is needed to evaluate such impact on person-task interactions. As already mentioned in Analysis 3, this level of correlation is unusual as compared to existing findings in the RT literature. There are many possible factors for this observation. For example, test type and item type (especially simulation items vs. others) are likely relevant. Test design and condition may be another factor—the task was delivered without time limit and was not high-stakes, so the test takers were not urged to complete accurately and quickly. Range restriction (e.g., Raju and Brand, 2003) is another possibility. The dataset under study came from a test-taker sample that seemed proficient in the task. This factor may also explain the lack of association between the test takers' time spent on the task and the available person covariates. Further empirical studies should focus on different simulation-based tasks and/or different test-taker populations to assess the generalizability of the weak positive association observed in this study.

One potential issue with scoring the current simulation-based task is that the task length may be too short to produce reliable scores on any aspects of the task performance. The sample size was also limited in this dataset. The short task length probably results from practical constraints on the overall task time, which not only includes the time spent responding to the embedded items but also includes the time spent listening to information about the common scenario. Our study indicated that the 11 items were generally not time-consuming for this test-taker sample. Thus, it may be adequate to include a few more non-simulation items to better assess what the test takers know and can do while not making the overall task time overly excessive. Designing the simulation-based task with more items, together with a larger sample, would also open up the possibility of using more complicated statistical models to capture the more complex person-task interactions. For instance, the simulation in the task may introduce additional dependencies among the associated items, or the test takers may change their behaviors across items of different types. Extensions of the hierarchical framework (van der Linden, 2007) with more complex IRT models may better describe the additional dependencies among the associated items. Mixture models may be used to detect heterogeneous behaviors with multiple classes underlying the responses and RTs (e.g., Molenaar et al., 2018), or to detect the test takers' shifting between solution behavior and rapid-guessing behavior with two underlying classes (e.g., Wang and Xu, 2015). Further work in this direction is worth considering.

Analysis 5 in the study concluded no notable rapid-guessing behavior or motivation issue in this dataset. Possible explanations include that this task was more engaging to this sample of test takers, the task was not too challenging to the test takers so they were willing to work on the items, and so on. Jia and Lee (2018) also found no issue with rapid-guessing behavior in the two NAEP simulation-based tasks. It is likely that simulation-based tasks are more interesting and engaging to test takers, but more research with different datasets—in terms of different tasks, different test-taker populations, different test conditions, and so on—is needed to further investigate the benefit of using simulation-based tasks in various settings. It may also be useful to retrieve more fine-grained data at the action level, including timing, processes, and others, to look further into person-task interactions (see, e.g., Ercikan and Pellegrino, 2017; Man and Harring, 2019). In any case, it will be valuable to conduct a systematic RT study similar to the one presented herein to assess different timing aspects of a test-taker sample's behavioral features in the simulation-based task of interest. Findings from such an RT study will lead to a better understanding of the person-task interactions and therefore offer insights into possible ways to leverage RT information in designing, assembling, and scoring of the simulation-based task of interest.

AUTHOR CONTRIBUTIONS

Y-HL conducted analyses 1 and 2, and led the writing and revisions of the manuscript. JH supplied information about the data, conducted analysis 4, and drafted the corresponding section. KM conducted analysis 3 and drafted the corresponding section. LO conducted analysis 5 and drafted the corresponding section. All authors reviewed and revised the manuscript.

ACKNOWLEDGMENTS

The authors thank Chun Wang, Carmen Llinares Millán, Bernard Veldkamp, Matt Johnson, Yue Jia, and Jim Carlson for their advice. Any opinions expressed in this article are those of the authors and not necessarily of ETS and ACT. The work presented in this manuscript was performed during the ETS summer internship of KM and LO.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transac. Automat. Control* 19, 716–723. doi: 10.1109/tac.1974.1100705

Baker, R. S. J. D., and Clarke-Midura, J. (2013). "Predicting successful inquiry learning in a virtual performance assessment for science," in *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization*, (Rome: User Modeling Inc.), 203–214. doi: 10.1007/978-3-642-38844-6_17

Beck, J., Woolf, B. P., and Beal, C. R. (2000). "ADVISOR: a machine learning architecture for intelligent tutor construction," in *Proceedings of the 7th National Conference on Artificial Intelligence*, (Cambridge, MA: The MIT Press), 552–557.

Bolsinova, M., Tijmstra, J., Molenaar, D., and De Boeck, P. (2017). Conditional independence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. *Front. Psychol.* 8:202. doi: 10.3389/fpsyg.2017.00202

Bolt, D. M., and Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Appl. Psychol. Meas.* 27, 395–414. doi: 10.1177/0146621603258350

Choe, E. M., Kern, J. L., and Chang, H.-H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *J. Educ. Behav. Stat.* 43, 135–158. doi: 10.3102/1076998617723642

De Boeck, P., and Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Front. Psychol.* 10:102. doi: 10.3389/fpsyg.2019.00102

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis*, 3rd Edn. New York, NY: John Wiley.

Ercikan, K., and Pellegrino, J. W. (2017). *Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes*. New York, NY: Taylor & Francis.

Fox, J. P., Klein Entink, R. H., and Avetisyan, M. (2014). Compensatory and noncompensatory multidimensional randomized item response models. *Br. J. Math. Stat. Psychol.* 67, 133–152. doi: 10.1111/bmsp.12012

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. New York, NY: Chapman & Hall.

Gelman, A., and Rubin, D. B. (1992a). A single series from the Gibbs sampler provides a false sense of security. *Bayesian Stat.* 4, 625–631.

Gelman, A., and Rubin, D. B. (1992b). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the big-five personality domains. *J. Res. Pers.* 37, 504–528. doi: 10.1177/1359105317720819

Guo, H., Rios, J. A., Haberman, S. J., Liu, O. L., Wang, J., and Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Appl. Meas. Educ.* 29, 173–183. doi: 10.1080/08957347.2016.1171766

Halpin, P. F., von Davier, A. A., Hao, J., and Liu, L. (2017). Measuring student engagement during collaboration. *J. Educ. Meas.* 54, 70–84. doi: 10.1111/jedm.12133

Hao, J., Liu, L., von Davier, A. A., and Kyllonen, P. C. (2015). "Assessing collaborative problem solving with simulation based tasks," in *Proceedings of 11th International Conference on Computer Supported Collaborative Learning*, Gothenburg.

Hao, J., Liu, L., von Davier, A. A., and Kyllonen, P. C. (2017). "Initial steps towards a standardized assessment for collaborative problem solving (CPS): practical challenges and strategies," in *Innovative Assessment of Collaboration*, eds A. A. von Davier, M. Zhu, and P. C. Kyllonen (New York, NY: Springer), 135–156. doi: 10.1007/978-3-319-33261-1_9

Jeong, A. (2004). The combined effects of response time and message content on growth patterns of discussion threads in computer-supported collaborative argumentation. *J. Dis. Educ.* 19:36.

Jia, Y., and Lee, Y.-H. (2018). "Using timing data to investigate test-taking behaviors for scenario-based tasks—A case study in the context of national assessment of educational progress (NAEP)," in *Paper Presented at the International Test Commission Conference in Montreal*, Canada.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edn. New York, NY: Wiley.

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Kyllonen, P. C., and Zu, J. (2016). Use of response time for measuring cognitive ability. *J. Intell.* 4:14. doi: 10.3390/jintelligence4040014

Lee, Y.-H., and Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychol. Test Assess. Model.* 53, 359–379.

Lee, Y.-H., and Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *Int. J. Test.* 16, 240–267. doi: 10.1080/15305058.2015.1085385

Lee, Y. H., and Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large Scale Assess. Educ.* 2:1.

Lee, Y.-H., and Ying, Z. (2015). A mixture cure-rate model for responses and response times in time-limit tests. *Psychometrika* 80, 748–775. doi: 10.1007/s11336-014-9419-8

Liu, L., Hao, J., von Davier, A., Kyllonen, P. C., and Zapata-Rivera, D. (2015). "A tough nut to crack: measuring collaborative problem solving," in *Handbook of Research on Computational Tools for Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf (Hershey, PA: IGI-Global).

Long, J. S., and Freese, J. (2006). *Regression Models for Categorical Dependent Variables Using Stata*, 2nd Edn. New York, NY: StataCorp LP.

Luo, Y., and Jiao, H. (2018). Using the stan program for Bayesian item response theory. *Educ. Psychol. Meas.* 78, 384–408. doi: 10.1177/0013164417693666

Ma, L., Wise, S. L., Thum, Y. M., and Kingsbury, G. (2011). Detecting response time threshold under the computer adaptive testing environment. *Paper presented at the annual meeting of the National Council of Measurement in Education*, New Orleans, LA.

Man, K., and Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educ. Psychol. Meas.* doi: 10.1177/0013164418824148

Man, K., Harring, J. R., Jiao, H., and Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Appl. Psychol. Meas.* doi: 10.1177/0146621618824853

Maris, G., and van der Maas, H. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika* 77, 615–633. doi: 10.1007/s11336-012-9288-y

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., et al. (2014). *Psychometric Considerations in Game-Based Assessment*. New York, NY: Institute of Play.

Molenaar, D., Bolsinova, M., and Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *Br. J. Math. Stat. Psychol.* 71, 205–228. doi: 10.1111/bmsp.12117

Natesan, P., Nandakumar, R., Minka, T., and Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational bayes. *Front. Psychol.* 7:1422.

Plummer, M. (2015). *JAGS: Just Another Gibbs Sampler Version 4.0.0.* .

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raju, N. S., and Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Appl. Psychol. Meas.* 27, 52–71. doi: 10.1177/0146621602239476

Ranger, J., and Ortner, T. (2012). The case of dependency of responses and response times: a modeling approach based on standard latent trait models. *Psychol. Test Assess. Model.* 54, 128–148.

Schnipke, D. L., and Scrams, D. J. (2002). "Exploring issues of examinee behavior: insights gained from response-time analyses," in *Computer-Based Testing: Building the Foundation for Future Assessments*, eds C. N. Mills, M. Potenza, J. J. Fremer, and W. Ward (Hillsdale, NJ: Lawrence Erlbaum Associates), 237–266.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Su, Y. S., and Yajima, M. (2015). *R2jags: Using R to Run JAGS. (Version 0.5)*.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J., Klein Entink, R. H., and Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Appl. Psychol. Meas.* 34, 327–347. doi: 10.1177/0146621609349800

van der Linden, W. J., Scrams, D. J., and Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Appl. Psychol. Meas.* 23, 195–210. doi: 10.1177/01466219922031329

van Rijn, P. W., and Ali, U. S. (2018). A generalized speed-accuracy response model for dichotomous items. *Psychometrika* 83, 109–131. doi: 10.1007/s11336-017-9590-9

Wang, C., Fan, Z., Chang, H. H., and Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J. Educ. Behav. Stat.* 38, 381–417. doi: 10.3102/1076998612461831

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi: 10.1080/01621459.1963.10500845

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., and Katz, I. R. (2014). "Assessing science inquiry skills using trialogues," in *Intelligent Tutoring Systems. ITS 2014. Lecture Notes in Computer Science*, Vol. 8474, eds S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia (Cham: Springer), doi: 10.1007/978-3-319-07221-0_84

Zu, J., Robin, F., and Bontya, A. (2016). Examining the practical gain of joint modeling responses and response times. *Paper Presented at the International Meeting of the Psychometric Society*, Asheville, NC. doi: NODOI

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Lee, Hao, Man and Ou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Education

ORIGINAL RESEARCH
published: 04 June 2019
doi: 10.3389/feduc.2019.00049

Detecting Examinees With Pre-knowledge in Experimental Data Using Conditional Scaling of Response Times

Sarah L. Toton and Dennis D. Maynes*

Caveon, Data Forensics, American Fork, UT, United States

The detection of examinees who have previously accessed proprietary test content is a primary concern in the context of test security. Researchers have proposed using item response times to detect examinee pre-knowledge, but progress in this area has been limited by a lack of real data containing credible information about pre-knowledge and by strict statistical assumptions. In this work, an innovative, but simple, method is proposed for detecting examinees with pre-knowledge. The proposed method represents a conditional scaling that assesses an examinee's response time to a particular item, compared to a group of examinees who did not have pre-knowledge, conditioned on whether or not the item was answered correctly. The proposed method was investigated in empirical data from 93 undergraduate students, who were randomly assigned to have pre-knowledge or not. Participants took a computerized GRE Quantitative Reasoning test and were given no items, half the items, or half the items with correct answers to study before the test, depending on their condition. Exploratory analysis techniques were used to investigate the resulting values at both the item and person-level, including factor analyses and cluster analyses. The proposed method achieved impressive accuracy of separation between disclosed and undisclosed items and examinees with and without pre-knowledge (96 and 97% accuracy for cluster analyses, respectively), demonstrating detection power for item disclosure and examinee pre-knowledge. The methodology requires minimal assumptions about the data and can be used for a variety of modern test designs that preclude other types of data forensic analyses.

Keywords: test security, latencies, cheating, pre-knowledge, data forensics, response times, experiment

OPEN ACCESS

Edited by:
Douglas F. Kauffman, Medical University of the Americas – Nevis, United States

Reviewed by:
Kathy Ellen Green, University of Denver, United States
Mohamed A. Ali, Grand Canyon University, United States

****Correspondence:***
Sarah L. Toton
sarah.toton@caveon.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *01 September 2018*
Accepted: *16 May 2019*
Published: *04 June 2019*

Citation:
Toton SL and Maynes DD (2019) Detecting Examinees With Pre-knowledge in Experimental Data Using Conditional Scaling of Response Times. Front. Educ. 4:49. doi: 10.3389/feduc.2019.00049

INTRODUCTION

Greater access to technology and the rise of standardized testing have led to an increase in threats of cheating on tests. Pre-knowledge of exam content occurs when exam items, options, and/or answers (presumed or actual) are harvested and shared with future examinees. Examinees may memorize or record exam content and divulge that information to future examinees via conversations, forums or online groups, review courses, shared files, or even by selling exam content online. The result is examinees who have accessed exam content prior to testing, gaining an unfair advantage through pre-knowledge of the content. When items are disclosed and pre-knowledge is gained, test security is violated and the validity of test scores should be questioned. In this paper, a new method for

analyzing response time data is proposed that makes minimal assumptions about the nature of the data and provides meaningful information about extreme response times, aiding in the detection of disclosed items and examinees with pre-knowledge.

There are several known cases in which widespread pre-knowledge was observed after exam content was disclosed. In one case, the certifications of 139 physicians were suspended for soliciting or sharing exam content through exam preparation courses (American Board of Internal Medicine, 2010). In another case, more than 250 examinees who allegedly accessed or shared exam content on social media sites were identified and penalized (Federation of State Boards of Physical Therapy, 2015). In yet another case, the average GRE score for several countries increased by 50–100 points on test sections with scaled scores ranging from 200 to 800 (Kyle, 2002). It was discovered that exam content was being posted and widely accessed by examinees in these countries. Information on the prevalence of pre-knowledge is understandably difficult to obtain, but cases like these can provide information about the potential magnitude of the issue when the stakes for exams are high.

Pre-knowledge is difficult to detect because there are not usually external signs of this type of cheating (Bliss, 2012). Typically, examinees with pre-knowledge have accessed and memorized exam content before testing. These factors mean that the detection of pre-knowledge must be accomplished through data forensics analyses rather than through surveillance by test proctors or video cameras.

Expected Patterns of Examinees With Pre-Knowledge

Examinees with pre-knowledge presumably do not answer test questions by conventional independent and intellectual means. Given that standardized tests have many respondents providing predictable and valid response patterns, examinees with pre-knowledge likely have different, identifiable response patterns in their data. In order to detect examinee pre-knowledge, it is important to understand how pre-knowledge influences data patterns. Differences are likely to manifest in both the item scores and the response time patterns. Additionally, examinees with pre-knowledge are likely respond to disclosed and undisclosed items differently, even after item difficulty and complexity are taken into account.

Examinees with pre-knowledge are likely to receive a higher score than they would have given their own ability, although their exact score depends on: the accuracy of the source of the pre-knowledge, their capability to memorize and recall the test content, and their ability level prior to obtaining pre-knowledge. Thus, score is not likely to be a powerful predictor of pre-knowledge unless it is analyzed in combination with other variables, such as item response times (RTs) or latencies.

There are many possible patterns that could represent pre-knowledge, but at a basic level, examinees with pre-knowledge are expected to:

- score higher on disclosed items than other items, because they likely accessed answers to these items while preparing for the exam (van der Linden and van Krimpen-Stoop, 2003; Belov, 2016a; Toton et al., in preparation);
- avoid or neglect studying for the test, thus receiving lower scores on undisclosed items than other items; and
- respond more quickly to disclosed items than to other items, because they were previously exposed to these items, and thus may spend less time reading the item content and selecting a response (van der Linden and van Krimpen-Stoop, 2003; Toton et al., in preparation).

Although we have separated items into disclosed and undisclosed for the purpose of this research, it is very common that the status of a group of items is unknown. Thus, it is important to note that detecting an item as undisclosed does not necessarily indicate that the item has not been disclosed. In addition, different subsets of examinees may have pre-knowledge of different items.

The above patterns in item scores and response times are generally expected to be true when pre-knowledge is present, although variability in these patterns is to be expected. The statistical methods to detect pre-knowledge use some or all of these patterns in order to identify examinees who are suspected of having pre-knowledge and/or items that may have been disclosed.

Statistics to Detect Examinee Pre-Knowledge

A wide variety of statistical approaches for detecting pre-knowledge have been explored in the literature and it is beyond the scope of this paper to review all of them. We will briefly describe categories of methods to detect pre-knowledge and a few methods to represent each category. For more comprehensive reviews on the wide variety of statistics used to detect pre-knowledge, see Bliss (2012), Eckerly (2017), or Scott (2018). Categories of methods to detect pre-knowledge include analyses of person-fit, similarity, score differences, and response times. Some methods may span multiple categories. For example, the method proposed in this paper applies the logic of person-fit and score-differencing statistics to response time data.

Person-Fit Statistics

Person-fit statistics only require item scores to compute and are a part of a typical psychometric analysis. Thus, computing person-fit statistics to detect pre-knowledge is a standard approach for psychometricians. Examinees without pre-knowledge are expected to respond to test items in Guttman patterns, such that examinees are expected to answer easier test items correctly up until a specific difficulty level and then answer all test items harder than that level incorrectly (Guttman, 1944). Person-fit statistics quantify the degree of misfit between an examinee's responses and the expected Guttman pattern.

One person-fit statistic that is a common baseline for detecting pre-knowledge is l_z (Drasgow et al., 1985). This statistic is the standardized log likelihood of a test response, based on an Item

Response Theory (IRT) model. It is assumed that l_z is normally distributed, but that is often not the case in live data. Karabatsos (2003) assessed the performance of 36 person-fit indices in detecting unusual response patterns (e.g., a high ability examinee answering easy items incorrectly, but difficult items correctly), including l_z. The best statistic, H_t (Sijtsma and Meijer, 1992), was a non-parametric statistic that compared an examinee's response pattern to the response patterns of all other examinees. Generally, non-parametric (i.e., those not based on IRT models) person-fit indices performed better than parametric indices. Another study, which applied IRT models to simulated data to detect cheating found that the *lco* difference (Ferrando, 2007), a sum of squared, standardized residuals across items, performed better than other methods, including H_t (Clark et al., 2014). All of the tested person-fit methods performed poorly on data with low rates of cheating.

The main limitations of using person-fit statistics to detect pre-knowledge are that they have relatively lower power than other statistics (Belov, 2016a) and that all types of misfit to the model is identified and flagged, so drawing inferences about what may have caused the misfit to the model (e.g., examinee pre-knowledge) is extremely difficult. Additionally, if pre-knowledge is widespread, it may be that examinees *without* pre-knowledge exhibit misfit.

Similarity Statistics

Similarity statistics quantify the agreement between examinees' responses. Answer-copying statistics are a type of similarity statistics that are typically computed for a single pair of examinees. Similarity statistics are a broader category than answer-copying statistics since they do not require labeling of source and copier examinees and are generally designed to detect groups larger than two who may have shared a source of pre-knowledge.

Angoff (1974) developed and researched the performance of eight answer-copying statistics. The two best statistics were those that (1) identified anomalously large numbers of identical incorrect responses in a pair of examinees, compared to other pairs of examinees with similar products of incorrect and identical incorrect responses, and (2) represented the maximum identical incorrect responses, or omitted responses, in a string of identical responses, compared to other pairs of examinees with similar scores. Similarly, Frary et al. (1977) developed the g_2 statistic to compare the number of identical responses in a pair to the expected number of identical responses. To compute the expected number of identical responses, a source and a copier are labeled in the pair and then the probability that the copier selected the same response as the source is calculated and these probabilities are summed over all items. This difference between the expected identical responses and observed identical incorrect responses is then standardized. This statistic is based in classical test theory (CTT). Wollack (1997) expanded on this work, developing the omega (ω) statistic, which is computed in a similar manner, but is based in the framework of IRT and uses the nominal response model (Bock, 1972) to obtain the probabilities that each examinee will select a particular response option.

van der Linden and Sotaridona (2006) and Maynes (2017) also used the nominal response model to estimate probabilities of an examinee selecting a particular response. The generalized binomial test developed by van der Linden and Sotaridona (2006) counts the total identical answers between examinees and compares it to the expected number. The M4 similarity statistic compares the observed identical incorrect, identical correct, and non-matching responses to the expected counts, based on the examinees' scores, using a generalized trinomial distribution (Maynes, 2017). Both of these methods provide a way to assess mismatch between the expected level of similarity between examinees and the observed level of similarity.

The main limitation of using similarity statistics to detect pre-knowledge is that their performance is known to be affected by the examinees' scores. High-scoring examinees should have strong agreement in their responses, since they provide mostly correct responses. Additionally, similarity statistics were developed for use on fixed form tests, where all examinees receive the same items, but often cannot achieve sufficient power in modern test designs such as computerized adaptive testing (CAT) or linear-on-the-fly testing (LOFT) because the number of items in common across examinees is typically very small. Thus, similarity analyses are powerful but often cannot be conducted with confidence in data obtained from modern test designs, due to the small number of common items.

Score-Differencing Statistics

If information about the items is known, different subsets of items can be scored and then the scores compared across subsets. For example, if a subset of items is known to be undisclosed (e.g., a group of new items is added to an exam), scores on the undisclosed items can be compared to scores on the remainder of the items. Examinees with much lower scores on the undisclosed subset and much higher scores on the remainder of the items should be detected as anomalous. These analyses are often referred to as score-differencing or differential person functioning analyses.

The Deterministic, Gated Item Response Theory Model proposed by Shu et al. (2013) detects examinees with pre-knowledge by separating probably disclosed items and probably undisclosed items. Score differences between the two item types are computed and examinees are split into those suspected of having pre-knowledge and those who are not based on those differences. Depending on the classification of the examinee and the item type, the ability level on the measured construct and the ability level due to cheating are both estimated. In this way, the model can identify a "true" ability level, untainted by the influence of probably disclosed items. Eckerly et al. (2015) expanded on this work, proposing a process of purifying the scale of item and person parameters by removing detected examinees from the computation of item difficulty parameters and then using the purified item parameters to re-estimate ability estimates. This modification reduced false-positives, while maintaining detection power.

Belov (2017) developed the posterior shift statistic, which is a Bayesian analysis that compares posterior distributions of ability between subsets of items. For example, comparing known

disclosed items to remaining items, or known undisclosed items to the remaining items. An expansion of this work injects a posterior shift statistic into a specially organized Markov Chain Monte Carlo to simultaneously detect disclosed items and examinees with pre-knowledge in a situation where a subset of undisclosed items is known (Belov, 2016b). The results suggest that the method performs well, even when the known undisclosed subset actually contains up to 15% disclosed items, but not when the known undisclosed subset contains 30% disclosed items. Thus, the method is robust to some error in specifying the undisclosed subset, but requires that the majority of the items in the undisclosed subset be accurately identified.

The main limitation of score-differencing statistics is that the additional information on items they require to compute is often unavailable or inaccurate. This can limit the application of score-differencing statistics to live data.

Flawed key analyses

Flawed key analyses are a subset of score-differencing statistics, although they are also strongly related to similarity analyses. Items are often disclosed without official answer keys and examinees who create their own keys often make errors. If a disclosed answer key contains errors and is known, a flawed key analysis can provide valuable information about pre-knowledge. Flawed keys are commonly discovered online or by finding keys used by rogue review courses, who sometimes create practice tests out of live items and provide an answer key to score the test. To conduct a flawed key analysis when the disclosed key is known, each test is scored using the actual key and the disclosed key, and examinees with significantly higher scores on the disclosed keys are identified (Scott, 2018).

Some research has focused on estimating latent sources, which includes flawed keys, from response data. To estimate a flawed key, similarity statistics can be utilized. Scott (2018) proposed a method of estimating disclosed keys that involved computing (Wollack, 1997) omega for all possible pairs of examinees. Four methods were compared to estimate the disclosed key, selecting each item's key as the response from the (1) source in the most source-copier pairs that exceeded a threshold of omega, (2) source in the source-copier pair with the largest omega value, (3) examinee's response pattern that was in the most source-copier pairs where omega exceeded a threshold, and (4) the source that was associated with the most copiers with maximum omega values. The results showed that the fourth method performed with very high accuracy in live data and the third performed best in simulated data, indicating that the estimation of flawed keys may require different methods in different contexts.

Maynes and Thomas (2017) analyzed clusters of examinees to estimate disclosed keys. This analysis assumes that a similar cluster of examinees has been detected using the Wollack and Maynes (2016) method, which is similar to nearest neighbors clustering. In the Wollack and Maynes (2016) method, pairwise similarity is computed for all examinees. Then, similarity values and test responses are plotted against each other in a dense graph of edges. Edges that fall below a selected threshold are removed and then clusters are created by labeling the groups of connected edges. Maynes and Thomas (2017) analyzed such clusters to extract the disclosed key using four methods, selecting each item's key as the response that (1) was most commonly chosen, (2) maximized the corrected item-total correlation, (3) contributed most to a chi-square, comparing the expected responses from the nominal response model to the actual responses, and (4) had the highest Kullback-Leibler divergence (Kullback and Leibler, 1951). Simulations varying the latent ability of the disclosed key and the amount of answer copying were performed. The results showed that the method that maximized the corrected-item total correlation was the best at accurately estimating the disclosed key, particularly with high amounts of answer-copying or large cluster sizes.

Haberman and Lee (2017) expanded this research to identify multiple disclosed keys, which were estimated using multidimensional IRT models. Once the multiple disclosed keys were estimated, examinees whose responses were identical or nearly identical to those disclosed keys were identified. This method appears to have good power and low false-positive rates, but it was tested on live data, so it is impossible to determine the power of the method by comparing the detected examinees to the examinees who actually used disclosed keys.

The main limitation of flawed key analyses is that they require additional information about disclosed keys to be known or estimated. However, once disclosed keys are obtained, a flawed key analysis can provide very powerful and compelling evidence that an examinee had pre-knowledge.

Response Time Statistics

The change from paper-and-pencil based testing to computer-based testing allowed the automatic collection of RT data, which inspired researchers to develop a variety of methodologies to detect pre-knowledge using RTs. van der Linden and van Krimpen-Stoop (2003) noted that due to the continuous nature of RTs, they contain more information, variability, and granularity than item score data.

van der Linden and van Krimpen-Stoop (2003) proposed a model to identify unusual RTs, particularly those resulting from pre-knowledge and item harvesting, on computerized adaptive tests. Item scores were analyzed using a three-parameter logistic IRT model (Birnbaum, 1968) and the RTs were modeled using a log-normal model. The log-normal RT model, originally published by Thissen (1983), estimates the time required for an item, the speed of the examinee, and the average RT for the population. The model by van der Linden and van Krimpen-Stoop (2003) assumes that the model for item scores is independent of the model for response times. Expected RTs for each item were estimated, using both maximum likelihood and Bayesian estimates. Extremely unusual RTs were detected by investigating the residual differences between the expected RTs and the observed RTs. The Bayesian residuals had better detection rates than the others, but also had higher false-positive rates. This method assumes that RTs have the same variance across items and examinees and that IRT model parameters for the data are accurate and consistent with known parameters. van der Linden and Guo (2008) expanded upon this research, proposing a combination of the RT and IRT models, allowing the examinees' RTs to be adjusted for their speed to investigate

the correspondence of their RT patterns with the estimated time required for each item.

Meijer and Sotaridona (2006) introduced the effective response time, which is an estimate of the time necessary to respond to an item correctly. Like van der Linden and van Krimpen-Stoop (2003), they proposed a three-parameter logistic model with a log-normal model for the RTs, estimating the speed of the examinee, the time required for each item, and the average RT for the population. However, only RTs of examinees with correct responses and with probabilities of answering correctly greater than the estimated pseudo-guessing parameter were used in the computation of the effective response time. This eliminated the effects of RTs caused by random guesses and other test taker behaviors that may yield uninformative RTs. This method has reasonable Type I error rates and has been used to detect pre-knowledge in K-12 data (Liu et al., 2013).

These methods require strict assumptions be met to model the RT data. They assume that each examinee has a constant working speed, the majority of examinees do not have pre-knowledge, item parameters are known, and that difficult items take longer than easier items. Additionally, item complexity, which can be described as the number of steps that must be completed to respond to an item, appears to be largely ignored. It is possible to have a very easy item that requires a large number of steps. It is possible that the difficulty of the item could be low, while the complexity of the item is high, which should affect the RTs of the item. Most of the available models for RTs ignore important factors that influence the data and require strict assumptions about the nature of the data.

In this paper, a simple, but innovative, method for using response times to detect disclosed items and examinees with pre-knowledge is proposed that makes minimal assumptions about the data and is appropriate for a wide variety of test designs.

Computing Conditional zRTs

We propose a measure of the extremeness of a response time that compares the RT for each person on each item to a group of examinees without pre-knowledge who received the same score on that item. Comparing a sample of data tainted by pre-knowledge to another tainted sample and expecting to detect extreme examinees is unlikely to yield the desired results. In the proposed method, the group of examinees without pre-knowledge (i.e., the uncontaminated comparison group) was the control condition of an experiment on pre-knowledge. However, in practice there is often no such group. Data from pilot testing or from the first day of testing for a particular exam form could be used and assumed to be uncontaminated by pre-knowledge, since exams are unlikely to be compromised prior to the first day of administrations (barring the help of a program insider or hacking the server).

The RTs for each examinee on each item can be transformed using the natural log transformation to approximate normality. After computing means and standard deviations of the log RTs for each item for the uncontaminated comparison group, the individual log RTs can be converted into conditional zRTs, so named because the log RTs are conditioned on item score, compared to the uncontaminated comparison group, and the statistics of interest are computed as *z*-scores. The conditional zRT for person j on item i with an item score of s can be computed using Equation (1).

$$zRT_{ijs} = \frac{RT_{ijs} - \bar{x}_{Cis}}{\sigma_{Cis}} \tag{1}$$

In Equation (1), RT_{ijs} represents the log-transformed response time for person j on item i with an item score of s (0 or 1), $\bar{x}_{Cis}$ represents the average log RT for the uncontaminated comparison group on item i with an item score of s, and σ_{Cis} represents the standard deviation of the log RTs for the uncontaminated comparison group on item i with an item score of s. Thus, conditional zRTs can be computed by taking the difference between an examinee's RT on an item and the average RT on that item for an uncontaminated comparison group with the same item score, divided by the standard deviation of the RTs on the item for the uncontaminated comparison group with the same item score.

After they are computed, conditional zRTs can be analyzed using exploratory grouping techniques, such as cluster analysis, to identify groups of items (disclosed and undisclosed) and examinees (with and without pre-knowledge). In the following sections, each particular component of the conditional zRTs that captures important information is discussed.

Comparing RTs Based on Item Score

Very fast response times can be produced by fast examinees, by testing strategies such as rapid guessing, or by examinees with pre-knowledge. To distinguish between these response patterns, item score should be taken into account. Response time data are messy and depend upon item difficulty, item complexity, personal testing style, testing strategies, and the speededness of a test. RTs can encompass a huge range and still be representative of normal test-taking behavior. Underlying multidimensionality coupled with high variability makes it difficult to determine which RTs should be considered extreme. However, conditioning RTs based on item score can help to distinguish between the potential behavioral causes for extreme RTs. For example, random guesses may yield fast RTs with mostly incorrect responses, but responses made with pre-knowledge may yield fast RTs with mostly correct responses.

Examinees with pre-knowledge who attempt to mask their response patterns would also likely be detected by the proposed method, as their pattern of conditional zRTs is likely to differ from examinees without pre-knowledge. For example, an examinee with pre-knowledge may complete the test quickly and then spend a large amount of time on a single item to lengthen the test time, but conditional zRTs should indicate that the examinee was anomalously fast and responded correctly on a large number of items, then anomalously slow on a single item.

Using an Uncontaminated Comparison Group to Obtain Comparison RTs

With data contaminated by cheating, it can be difficult to find meaningful separation in groups of items or persons, especially when information about the categories is unknown. One risk when analyzing contaminated data is that the results may not actually represent pre-knowledge. Thus, inferences made from contaminated data may penalize examinees with potentially life-changing consequences. Because many normal test-taking behaviors and strategies are not well-understood, model bias is a distinct possibility. It is important that examinees are not detected simply because they have employed a different test-taking strategy or exhibited an unusual test-taking style.

When computing conditional zRTs, the log RT for each examinee on each item is compared to the log RTs for that item obtained from an uncontaminated comparison group. If the uncontaminated comparison group is sufficiently large, it can be assumed to include a wide variety of test-taking strategies and other normal variation in test-taking behaviors.

Obtaining A Dataset With Pre-Knowledge

Pre-knowledge is difficult to study because high-quality data are scarce. Data containing pre-knowledge are generally gathered from one of three sources: simulations, using measures that approximate pre-knowledge in real data, or experiments that empirically manipulate pre-knowledge. Each of these methods has strengths and weaknesses, which are comprehensively discussed in Thomas and Maynes (2018).

Simulations

Simulations are often used to test methodologies for detecting pre-knowledge, but they do not contribute to understanding the natural patterns caused by pre-knowledge and are unlikely to capture the full variability of normal test-taking behaviors and strategies used by real examinees. Using simulated data that is too clean and does not capture the noise of normal test-taking may artificially inflate the performance of detection methods that are tested on such data.

Real Data

In real data, the examinees with pre-knowledge are typically unknown. Examinees who are suspected of having pre-knowledge may be identified, but the credibility of this information varies widely between testing programs and exams. Thus, using this information as a dependent variable can introduce a significant element of uncertainty.

Experiments

Using experiments to manipulate pre-knowledge allows for the creation of known examinees with pre-knowledge for a known subset of items. Tiemann et al. (2014) conducted an experimental study of pre-knowledge by asking 20 participants to write down test content they remembered after taking a test. The participants were told the next participant would be able to use this content while taking the test. Some participants were informed they would be asked to remember the content after the test and some were not. The results showed that many participants remembered imprecise information about the test content, but few remembered specific information. Ten participants who were informed in advance that they would be asked to remember test content demonstrated poor recall of the content, with only seven items and two answers remembered specifically and correctly. Surprisingly, 10 participants who were not informed they would be asked to remember test content in advance demonstrated better recall of the content, with nine items and 11 answers remembered specifically and correctly. The next part of the study investigated if access to the test content provided by previous examinees prior to testing raised scores. Two cheat sheets were created, one based on content remembered by participants who were informed they would be asked to recall test content in advance and one from those who were not. Students were randomly assigned to receive one of these cheat sheets to study and then took the test. The results showed no significant differences in test scores between students who received cheat sheets and students who did not.

A major advantage of laboratory experiments is that the data encompass the complexity of test-taking behavior and the identities of examinees with pre-knowledge can be known. However, previous studies attempting to mimic pre-knowledge in the laboratory have shown null results (Tiemann et al., 2014), possibly because participants were unmotivated or because the pre-knowledge provided was too weak to find effects. In the current study, experimental data were analyzed because of the benefits of having known groups of items and examinees while capturing normal test-taking variability.

Experimental Design

The data utilized in this study were collected with a 3 (Pre-Knowledge: Control vs. ItemOnly vs. Item+Answer) × 2 (Item Disclosure: Disclosed vs. Undisclosed) within and between-subjects design. All participants took a computerized GRE Quantitative Reasoning test. Pre-knowledge was manipulated by allowing some participants access to some test items (with or without accompanying answers) prior to the test. In the control condition, participants took the test but were not exposed to any of the test items in advance. In the experimental conditions, participants were allowed to study 12 of the 25 test items for 20 min before the test. In the ItemOnly condition, only the items themselves were provided to the participants (without their corresponding answers). In the Item+Answer condition, both the items and the correct answers were provided to the participants.

The design of this study was different from previous experimental studies in a few key ways. First, perfectly accurate disclosed test items (and answers in the Item+Answer condition) were provided to participants, rather than attempting to have potentially unmotivated participants harvest the items themselves. Second, the effect of the examinees' motivation to harvest items or cheat was removed by simply assigning some examinees to study provided disclosed materials and some not to. The task was not identified as relevant to cheating in any way, as the researchers who conducted the study simply requested that examinees study the materials.

The conditions only differed in whether or not they received test content in advance, and, if they did, the nature of that test content. Thus, pre-knowledge was investigated regardless of the examinees' ability to harvest items or motivation to cheat.

It was expected that participants in the study who were in the experimental conditions and given pre-knowledge of some items would select more correct answers on disclosed items than undisclosed items and respond more quickly to disclosed items than undisclosed items, as discussed in the Expected Patterns of Examinees with Pre-Knowledge section of this paper.

The Goals of the Current Study

The goal of the current study is to implement the proposed method of computing conditional zRTs, described above, in experimental data to create datapoints that capture complex information about the response patterns of the examinees. Particular attention will be paid to the feasibility of calculating the conditional zRTs and the assessing which components of the statistic provide the most important information. The conditional zRTs will then be analyzed using exploratory grouping methods, such as cluster and factor analyses, to attempt to identify groups of items and groups of examinees. The performance of the method in identifying groups of items (disclosed and undisclosed) and examinees (with and without pre-knowledge) will be assessed using the known groups contained in the experimental data. The purpose of the study is to assess a new method for detecting examinees with pre-knowledge and disclosed items in experimental data. If the method performs well, it may be able to be used to detect disclosed items and examinee pre-knowledge in a variety of live testing data, including those with modern test designs that preclude many data forensics analyses.

METHODS

Participants

Participants were 93 undergraduate psychology students (28 men, 64 women, and one gender non-conforming) at the University of Virginia. The sample was composed of primarily first year students (61% first year, 18% second year, 17% third year, 3% fifth year, and 1% exchange students). Participants took part in a 90 min laboratory session through the psychology department participant pool and were compensated with 1.5 h credit.

Procedure

This study was conducted according to the recommendations of, and was approved by, the Institutional Review Board for Social and Behavioral Sciences at the University of Virginia. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Participants were randomly assigned to condition immediately after signing up for a timeslot to participate in the study. When participants arrived at the lab, they were asked to leave their belongings with the researcher and enter a small room with a desk and computer. All participants were given a yellow sheet of laminated paper that they could put under the door if they wanted to ask questions or contact the researcher. This prevented the participants from discovering that other participants were receiving different treatment than they were (i.e., studying test items in advance) and completing the test more quickly. After gathering informed consent and giving participants instructions, the researcher left the participant alone.

Participants in the experimental conditions were first given a packet of materials that they were told to study for 20 min before taking the test. The packet contained instructions on how to use the study materials, a test form with 12 of the 25 test items, and two pieces of scratch paper. The 12 test items in the ItemOnly condition contained only the test items and possible answer choices; the same 12 items were provided in the Item+Answer condition, including red circles indicating the correct answers.

After the pre-knowledge stage, participants in the experimental conditions took a computerized Qualtrics version of an out-of-circulation, paper-and-pencil GRE Quantitative Reasoning test (Educational Testing Service, 2017) that the researchers were granted permission to use. Computerizing the test allowed for the collection of response time data for each item.

Participants in the control condition proceeded immediately to the computerized test, without first completing the pre-knowledge stage. All participants were given 40 min to complete the test. Participants who finished any stage of the experiment in less than the allotted time could notify the researchers by slipping a yellow paper under the door (to avoid participants from overhearing others) and advance to the next stage of the experiment.

After the test, participants were given 20 min to complete a battery of individual difference measures to assess factors that might impact their performance on the test or their general willingness to cheat. This battery included self-reported effort and time spent studying the packet of 12 items (for experimental conditions only), information about the testing experience of the participant during the study, math proficiency, previous exposure to GRE study materials, demographics, test anxiety (Westside Test Anxiety Scale, Driscoll, 2007), the Big Five personality traits (TIPI, Gosling et al., 2003), Moral Foundations (MFQ30, Graham et al., 2011), and religiosity/spiritualism. The order of the scales after demographics was randomized across participants.

For more comprehensive information about the study design, participants, or measures, see Toton et al. (in preparation).

RESULTS

The data used in this paper was obtained by disclosing a perfectly accurate key of 12 of the 25 items to a subset of the examinees. Examinees with pre-knowledge exhibited higher scores and faster RTs on disclosed items than examinees without pre-knowledge (Toton et al., in preparation). In contrast, the examinees with and without pre-knowledge did not differ

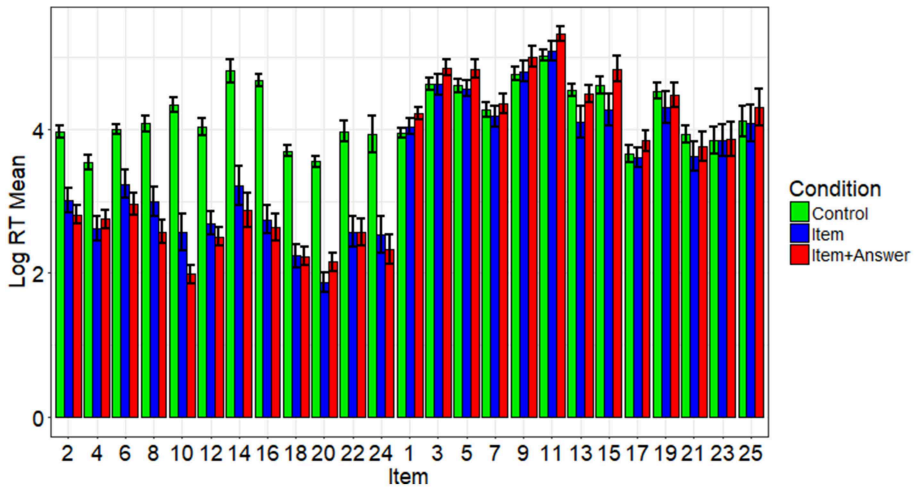


FIGURE 1 | Average log RTs by item and condition. This figure shows the average log RT by condition (see color) and item (see x-axis). Even-numbered items were disclosed and odd-numbered items were not. The error bars represent the standard errors.

significantly in item scores or log RTs for undisclosed items. Log RTs for each item for participants in all conditions are presented in **Figure 1**. For more information about how the three conditions differed from one another on all measures, see Toton et al. (in preparation).

Three different transformations for approximating normality in the RT data were investigated: the square root, logistic, and inverse transformations. The logistic, or natural log transformation, was the best transformation for achieving an approximately normal distribution for the majority of the items (14/25) in the data, as assessed by the Shapiro-Wilk test (Shapiro and Wilk, 1965). The results suggested that the square root transformation could also have been used as it was the best transformation for approximating normality for 10 of the 25 items.

Descriptive statistics for each condition are presented in **Table 1**. **Figures 2**, **3** show the response patterns of item scores and log RTs for an example examinee with and without pre-knowledge. These response patterns are a graphical representation of the information that is captured in the conditional zRTs. The examinees whose data are shown in **Figures 2**, **3** were selected because their data mimicked patterns that were expected based on theoretical ideas of taking the test with and without pre-knowledge. However, there is large variability in the patterns for examinees, particularly in the patterns of examinees who appear to be guessing for much of the test or examinees who appear to be very high in ability on the tested construct. The selected data shown in **Figures 2**, **3** demonstrate support for the theoretical ideas concerning the response patterns exhibited by examinees who took the test with and without pre-knowledge.

To compute the conditional zRTs, the means and standard deviations of the log RTs for each item were computed for control participants ($N = 33$) for both correct and incorrect responses. The conditional zRTs for each person on each item were computed using Equation (1). No conditional zRTs were computed for incorrect responses to Item 4 because no examinees responded to this item incorrectly in the control or other conditions. No conditional zRTs were computed for incorrect responses to Item 1, because there was only one participant in the control condition who responded to the item incorrectly, so the standard deviation could not be computed. Conditional zRTs were not necessary in this case because no participants in the experimental conditions had incorrect responses to this item that needed to be compared to the control group. Information on the sample size, mean, and standard deviations of the log RTs for the control condition that were used to compute the conditional zRTs are presented in **Table 2**.

TABLE 1 | Descriptive statistics by condition.

	Control	ItemOnly	Item+Answer
N	33	30	30
Disclosed item scores	7.67 (2.41)	8.57 (2.91)	11.33 (1.12)
Undisclosed item scores	6.91 (2.48)	7.13 (2.66)	7.43 (2.71)
Disclosed item RT	76.59 (64.12)	32.74 (59.35)	20.46 (27.38)
Undisclosed item RT	98.09 (63.60)	103.76 (88.36)	123.95 (94.52)
Disclosed item LN RT	4.05 (0.78)	2.73 (1.15)	2.53 (0.93)
Undisclosed item LN RT	4.36 (0.75)	4.25 (1.03)	4.48 (1.00)

This table presents the means and standard deviations (in parentheses) of item scores, raw response times (RTs) across examinees and items in seconds (Item RTs), and log RTs across examinees and items (Item LN RTs) for disclosed and undisclosed items for participants in each of the three conditions.

Assessing Components of the Conditional zRTs

The computation of conditional zRTs based on item score and using an uncontaminated comparison group of examinees

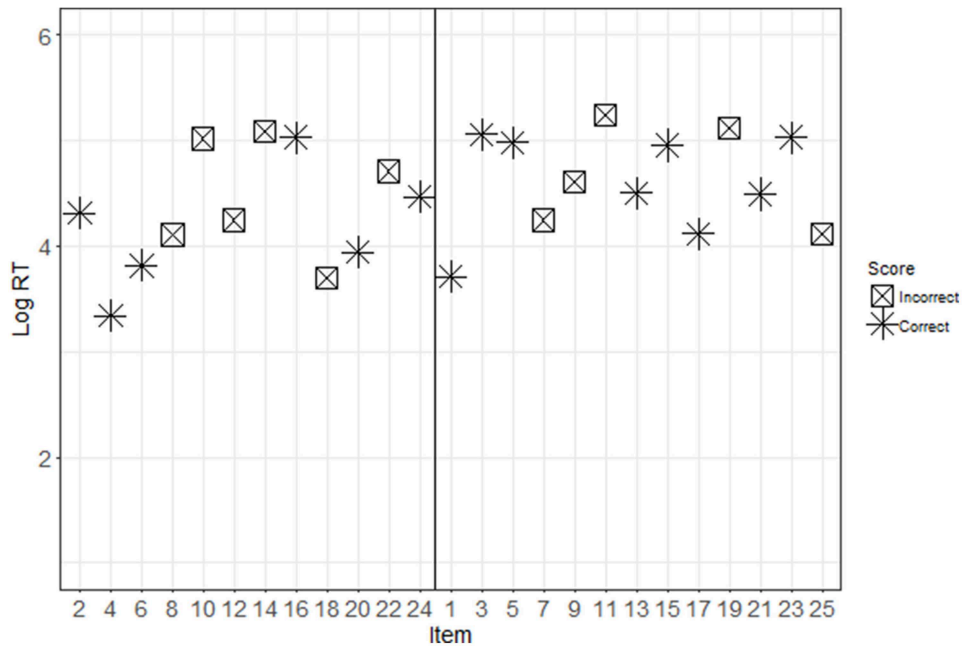


FIGURE 2 | Response patterns for a control participant. This figure shows the item scores (see marker shape) and log RTs (y-axis) for disclosed (even, presented on the left) and undisclosed (odd, presented on the right) items. Note that for this control participant, there is no discernable difference between the response patterns for the disclosed and undisclosed items.

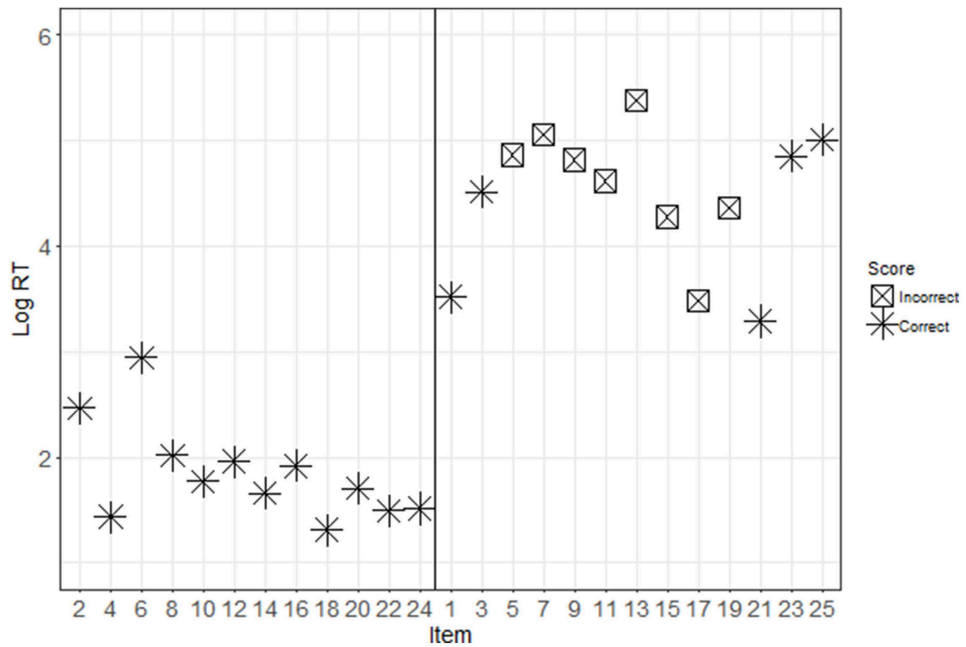


FIGURE 3 | Response patterns for an ItemOnly participant. This figure shows the item scores (see marker shape) and log RTs (y-axis) for disclosed (even, presented on the left) and undisclosed (odd, presented on the right) items. Note that for this ItemOnly participant, there is a discernable difference between the response patterns for the disclosed and undisclosed items, such that disclosed items are more likely to be answered correctly and quickly.

TABLE 2 | Descriptive statistics of control group for computing conditional zRTs.

Item	*N* correct control	Average correct control	*SD* correct control	*N* incorrect control	Average incorrect control	*SD* incorrect control
1	32	3.95	0.38	1	3.92	NA
2	17	3.96	0.51	16	3.96	0.44
3	24	4.63	0.52	9	4.65	0.30
4	33	3.54	0.61	0	NA	NA
5	21	4.60	0.47	12	4.61	0.57
6	31	4.00	0.41	2	3.89	0.39
7	18	4.13	0.58	15	4.44	0.59
8	16	4.09	0.62	17	4.06	0.73
9	6	4.99	0.30	26	4.70	0.52
10	19	4.27	0.52	14	4.44	0.63
11	7	5.17	0.39	24	4.97	0.46
12	14	3.97	0.49	18	4.03	0.74
13	23	4.51	0.49	10	4.65	0.52
14	13	4.91	0.99	19	4.92	0.60
15	14	4.62	0.46	17	4.58	0.86
16	28	4.71	0.43	4	4.44	0.95
17	21	3.79	0.56	11	3.41	0.79
18	22	3.72	0.32	9	3.65	0.62
19	13	4.78	0.42	16	4.31	0.72
20	26	3.46	0.31	4	4.14	0.44
21	23	4.05	0.41	6	3.50	1.12
22	18	4.11	0.53	10	3.73	0.99
23	16	4.01	0.57	11	3.56	1.44
24	16	4.19	1.23	12	3.58	1.45
25	10	4.39	0.67	18	3.97	1.32

This table provides the sample size, average, and standard deviation of the log RTs from the control condition that were used to compute the conditional zRTs for both correct and incorrect responses.

without pre-knowledge were decisions based in theoretical considerations. To test if these factors impacted the performance of the conditional zRTs, zRTs were also computed:

- without conditioning on score, using the full sample as a comparison group,
- conditioning on score, using the full sample as a comparison group, and
- without conditioning on score, using the control condition as a comparison group.

The comparison groups were used to compute the means and standard deviations in Equation (1). Note that in the current data, the majority of the participants had item pre-knowledge, so using the full sample as the comparison group should be dramatically different from using just the control condition. There were 60 examinees with pre-knowledge of 12 items, which means that there were 720 total responses of examinees with pre-knowledge to disclosed items (ignoring the possibility of missing data and assuming examinees with pre-knowledge display the expected patterns consistently). If all of these were answered anomalously quickly, then one would expect approximately 720 extremely fast zRTs. When the full sample was used as a comparison group, 69 zRTs were detected when the zRTs were not conditioned on score and 57 were detected when the zRTs were conditioned on score (see **Table 3**). However, when the control condition was used as the comparison group, more than 500 zRTs were detected; 504 when the zRTs were not conditioned on score and 521 when zRTs were conditioned on score.

TABLE 3 | Comparing extreme results using different methods of computing zRTs.

Label	Conditioned on score	Comparison group	Extreme zRT count	Fast, extreme zRT count
zRTs- Full	No	Full sample	91	69
Conditional zRTs-Full	Yes	Full sample	85	57
zRTs-Control	No	Control condition	568	504
Conditional zRTs	Yes	Control condition	589	521

This table presents the number of zRTs more extreme than a z of ±1.96, the critical value for a z-distribution at $\alpha = 0.05$, and the number of those representing fast zRTs, with z-scores more extreme than −1.96 for several methods of computing zRTs.

The zRTs computed using the full sample were very different than those computed using the control participants as the comparison group, particularly for participants in the ItemOnly and Item+Answer conditions (see **Figure 4**). Participants in the control condition had similar zRTs no matter which

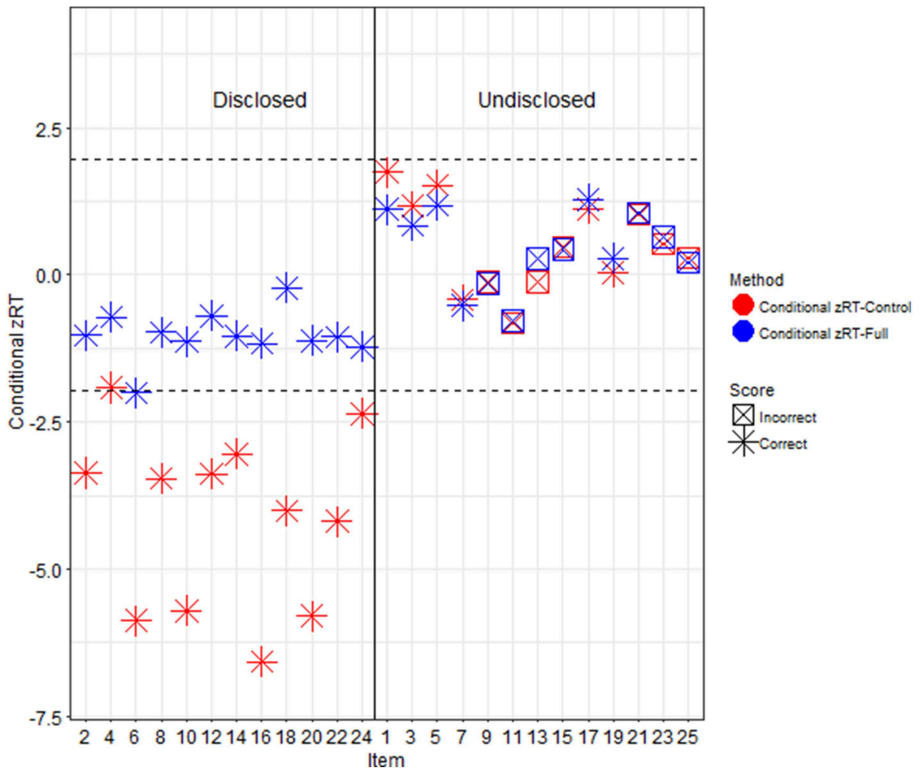


FIGURE 4 | Conditional zRTs computed with different comparison groups. The conditional zRTs for one Item+Answer participant computed with all participants as the comparison group are presented in blue and with only control participants as the comparison group in red. The dashed lines represent *z*-scores of ±1.96; it is expected that 95% of the distribution of conditional zRTs will fall in between these lines. When all participants were used as the comparison group, the disclosed and undisclosed items appeared similar in terms of conditional zRTs, but when only control participants were used as the comparison group, the disclosed items for this participant appeared to be much more extreme.

computation was used, but participants in the ItemOnly or Item+Answer conditions were much more likely to have extreme zRTs detected when the control condition was used as a comparison group.

Separating Disclosed and Undisclosed Items

The goal of the item-level analyses was to use exploratory data techniques to create groups of items and then to assess those groups to investigate if they represented disclosed and undisclosed items. Cluster and factor analyses were conducted and are presented below. Note that correlations and Kolmogorov-Smirnov tests were also investigated, and achieved good separation of item groups, but are not presented here for the sake of brevity.

Cluster Analyses

K-means cluster analysis is an exploratory data analysis technique used to group datapoints into a number of clusters, where the number of clusters to create is specified by the researcher (Lloyd, 1957). First, *k* number of cluster center are randomly placed among the data. Second, the points closest to each center are assigned to that cluster. Third, the mean of the points assigned to each cluster is computed and the center is moved to this point. The second and third steps continue until a stable solution is obtained. If group memberships of the data are known, the estimated group memberships obtained in the cluster analysis can be compared with the known group memberships. Cluster solutions were obtained using the "kmeans" function in the "stats" package of R (R Core Team, 2016), using the Hartigan and Wong (1979) algorithm. This algorithm is an efficient *k*-means analysis that prevents re-analysis of data points that were not assigned to a different cluster in the last step. Comparisons of each datapoint to the cluster centers were based on Euclidean distance and iterations continued until the total within sum of squares was minimized. Cases with missing conditional zRTs were omitted from these analyses, leaving 70 complete cases for analysis. There were nine control participants with missing data, eight ItemOnly participants, and six Item+Answer participants. Two and three-cluster solutions were computed to investigate grouping accuracy for disclosed and undisclosed items and the robustness of the grouping accuracy when different numbers of clusters were specified.

Two-cluster solution for items

The two-cluster solution grouped the items into two groups containing 14 and 11 items, respectively (see **Figure 5**). The

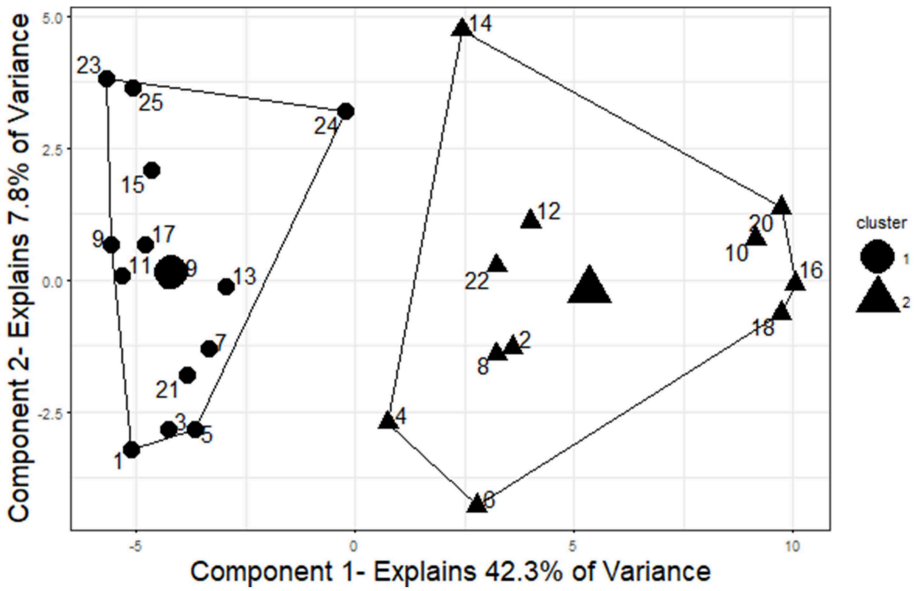


FIGURE 5 | Two-cluster solution for items. The two clusters represent good separation between disclosed (even-numbered) and undisclosed (odd-numbered) items. The cluster centers and cluster membership are indicated by shape (triangles represent the cluster of size 11 and circles represent the cluster of size 14). To visualize the cluster results, components were obtained using principal component analysis, and the clusters are plotted on those components. This plot was created using the "fviz_cluster" function in the "factoextra" package of R (Kassambara and Mundt, 2017).

group of 14 items contained all 13 undisclosed items as well as one disclosed item that was incorrectly grouped (Item 24). The group of 11 items contained all of the remaining disclosed items. Overall, 24 of the 25 items were grouped correctly into disclosed or undisclosed items. Thus, this cluster analysis produced groupings that were 96% accurate in separating disclosed and undisclosed items.

Three-cluster solution for items

The three-cluster solution closely mirrored the two-factor solution, again grouping all of the disclosed items except item 24 into a cluster of size 11. The cluster of size 14 observed in the two-cluster solution was split into two clusters in the three-cluster solution, one of size eight and one of size six. The cluster of size eight included only undisclosed items and the cluster of size six included five undisclosed items and one disclosed item (Item 24). This cluster solution was as accurate as the two-cluster solution in grouping items based on their disclosure status.

Factor Analyses

Three sets of data were analyzed by factor analysis to explore latent components that underly similarities in data: conditional zRTs, item scores, and item log RTs. Factor solutions were assessed for the three sets of data using the "fa" function in the "psych" package of R, excluding missing data in a pairwise fashion (Revelle, 2016). The factor analyses were conducted using principal axis factoring and promax, or oblique, rotations to allow correlations between the factors. For the purposes of this analysis, factor loadings of <0.30 were considered low. One and two-factor solutions were computed to analyze differences in grouping accuracy for disclosed and undisclosed items.

Conditional zRTs

The one-factor solution for conditional zRTs explained 26% of the variance and showed that the 12 disclosed items had factor loadings that were positive and strong on the factor, as well as three undisclosed items (see **Table 4**). The remaining undisclosed items had weak loadings on the factor, ranging from $\lambda = -0.12$ to $\lambda = 0.25$. Thus, 22 of the 25 items were grouped with items of the same disclosure status (88% accuracy).

The two-factor results for the conditional zRTs explained 42% of the variance and showed good simple structure, with all 12 disclosed items loading strongly on the first factor and 12 of the 13 undisclosed items loading strongly on the second factor (see **Table 4**). There were no items with cross-loadings of 0.30 or greater. One undisclosed item, Item 1, did not load strongly onto either factor, but had a negative loading on factor one and a positive loading on factor two. Thus, 24 of the 25 items were grouped with items of the same disclosure status (96% accuracy). When all cases with missing data were excluded, leaving 70 examinees for analysis, the one and two-factor solutions exhibited perfect simple structure with clear separation between disclosed and undisclosed items and no cross loadings of 0.30 or greater.

Item scores

The factor analyses for item scores excluded item four, since no participants responded to that item incorrectly. The one-factor solution for item scores explained 16% of the variance and showed that 10 disclosed items and seven undisclosed items had factor loadings of 0.30 or greater on the factor. The remaining one disclosed item and six undisclosed items had factor loadings

TABLE 4 | Factor loadings of conditional zRTs for items.

		One factor solution	Two factor solution	
	Item	**Factor 1 loading**	**Factor 1 loadings**	**Factor 2 loadings**
Disclosed items	2	**0.70**	**0.69**	0.06
	4	**0.63**	**0.58**	0.13
	6	**0.57**	**0.56**	0.06
	8	**0.68**	**0.66**	0.08
	10	**0.69**	**0.80**	−0.15
	12	**0.74**	**0.75**	0.03
	14	**0.69**	**0.61**	0.20
	16	**0.75**	**0.80**	−0.04
	18	**0.72**	**0.85**	−0.19
	20	**0.75**	**0.84**	−0.11
	22	**0.72**	**0.73**	0.03
	24	**0.64**	**0.62**	0.08
Undisclosed items	1	−0.12	−0.19	0.14
	3	0.24	−0.05	**0.62**
	5	0.18	−0.09	**0.56**
	7	0.24	−0.03	**0.59**
	9	0.25	−0.09	**0.76**
	11	0.17	−0.14	**0.65**
	13	**0.40**	0.09	**0.68**
	15	0.27	−0.04	**0.67**
	17	0.12	−0.10	**0.47**
	19	**0.32**	0.08	**0.52**
	21	**0.36**	0.16	**0.43**
	23	0.16	−0.08	**0.51**

Factor loadings for one and two-factor solutions for conditional zRTs of items. Factor loadings above 0.30 are bolded.

of <0.30 on the factor. Thus, 16 of the 25 items were grouped with items of the same disclosure status (64% accuracy).

The two-factor solution for item scores explained 22% of the variance and showed that nine disclosed items had factor loadings of 0.30 or greater on the first factor and that seven undisclosed items and one disclosed item had factor loadings of 0.30 or greater on the second factor, with no cross loadings of 0.30 or greater. The remaining one disclosed item and six undisclosed items did not load strongly onto either factor. Thus, 16 of the 25 items were grouped with items of the same disclosure status (64% accuracy).

Log RTs

The one-factor solution for log RTs explained 26% of the variance and showed that 12 disclosed items and six undisclosed items had factor loadings of 0.30 or greater on the factor. The remaining seven undisclosed items had factor loadings of <0.30 on the factor. Thus, 19 of the 25 items were grouped with items of the same disclosure status (76% accuracy).

The two-factor solution for log RTs explained 40% of the variance and showed that all 12 disclosed items had factor loadings of 0.30 or greater on the first factor and that 12 of the 13 undisclosed items had factor loadings of 0.30 or greater on the second factor, with no cross loadings of 0.30 or greater. The one remaining undisclosed item did not load strongly onto either factor but had a negative loading on factor one and a positive loading on factor two. Thus, 24 of the 25 items were grouped with items of the same disclosure status (96% accuracy).

Separating Examinees With and Without Pre-Knowledge

The goal of the person-level analyses was to use cluster analyses to create groups of examinees. Cluster solutions were obtained using the "kmeans" function in the "stats" package of R (R Core Team, 2016). Cases with missing conditional zRTs were omitted from these analyses, leaving 70 complete cases for analysis. Two and three-cluster solutions were computed to investigate grouping accuracy for examinees with and without pre-knowledge, the robustness of the grouping accuracy when different numbers of clusters were specified, and to investigate if the three-cluster solution distinguished between control, Item, and Item+Answer participants. Correlational results were also investigated, and showed separation of examinees with and without pre-knowledge, but are not presented here for the sake of brevity.

Two-Cluster Solution for Examinees

The two-cluster solution grouped the examinees into two groups containing 26 and 44 participants, respectively (see **Figure 6**). The group of 26 examinees contained 24 control participants, one ItemOnly participant, and one Item+Answer participant. The group of 44 examinees contained 22 ItemOnly participants and 23 Item+Answer participants. Overall, 68 of the 70 examinees with no missing conditional zRTs were grouped correctly into examinees with or without pre-knowledge (97% accuracy).

Three-Cluster Solution for Examinees

The three-cluster solution grouped the items into groups containing 26, 27, and 17 examinees, respectively (see **Figure 7**). The results closely mirrored those obtained in the two-factor solution, as the cluster of size 26 contained the same examinees as in the two-cluster analysis (24 control, one Item, and one Item+Answer). The cluster of 27 examinees contained only examinees with pre-knowledge, 15 ItemOnly participants and 12 Item+Answer participants. Similarly, the cluster of size 17 contained only examinees with pre-knowledge, six ItemOnly participants, and 11 Item+Answer participants. This cluster solution was as accurate as the two-cluster solution in grouping examinees based on whether or not they had pre-knowledge.

DISCUSSION

Response times were transformed with the logistic transformation, conditioned on item score, and compared to log RTs from a group of examinees without pre-knowledge to compute conditional zRTs. There were two item score combinations that conditional zRTs could not be computed

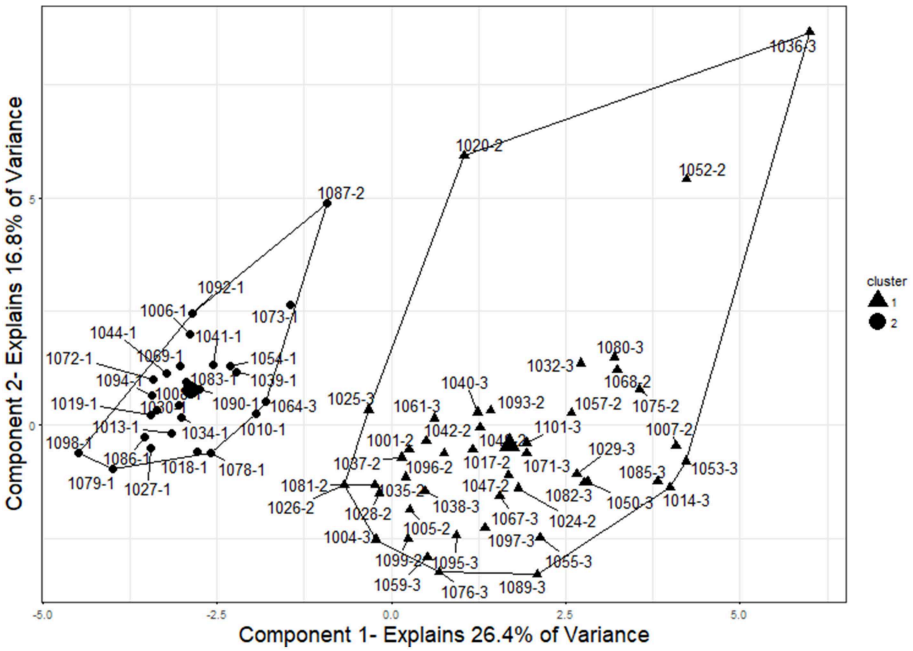


FIGURE 6 | Two-cluster solution for examinees. The two clusters represent good separation between control (IDs that end in−1), ItemOnly (IDs that end in−2), and Item+Answer (IDs that end in−3) participants. Two examinees were grouped incorrectly by this analysis (1064-3 and 1087-2). The cluster centers and cluster membership are indicated by shape (circles represent the cluster of size 26 and triangles represent the cluster of size 44). To visualize the cluster results, components were obtained using principal component analysis, and the clusters are plotted on those components. This plot was created using the "fviz_cluster" function in the "factoextra" package of R (Kassambara and Mundt, 2017).

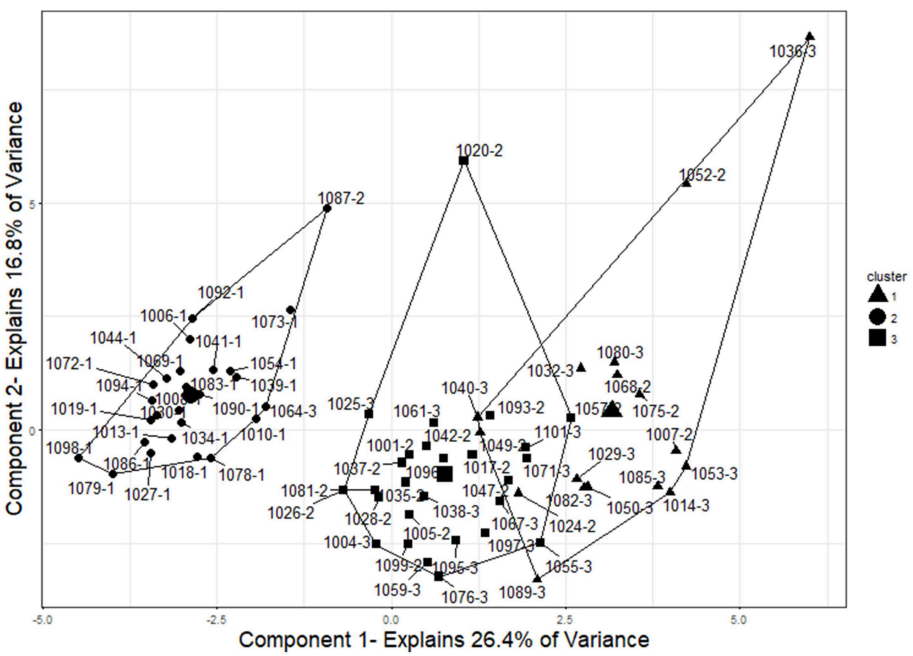


FIGURE 7 | Three-cluster solution for examinees. The three clusters represent good separation between control (IDs that end in−1), ItemOnly (IDs that end in−2), and Item+Answer (IDs that end in−3) participants. The cluster centers and cluster membership are indicated by shape (circles represent the cluster of size 26, squares represent the cluster of size 27, and triangles represent the cluster of size 17). Two examinees were grouped incorrectly by this analysis (1064-3 and 1087-2). To visualize the cluster results, components were obtained using principal component analysis, and the clusters are plotted on those components. This plot was created using the "fviz_cluster" function in the "factoextra" package of R (Kassambara and Mundt, 2017).

for. Item 1 had one incorrect response and Item 4 had no incorrect responses, so the control condition means and/or standard deviations could not be computed. The participant who responded to Item 1 incorrectly was a control condition participant. Thus, conditional zRTs would not have been computed for these item score combinations even if the mean and standard deviation for the control group had been available, because there were no response times for these item and score combinations from the experimental conditions that needed to be compared to the control condition.

Computing the zRTs using the control condition as the comparison yielded more extremely fast zRTs than using the full sample as the comparison group. Given that there were 60 examinees with pre-knowledge of 12 items, we expected around 720 extremely fast responses, assuming no missing data and that examinees with pre-knowledge consistently exhibit expected patterns. Thus, the detection of around 500 zRTs using the control sample as a comparison is much more consistent with the expectation of around 720 than the detection of 57–69 zRTs when using the full sample as the comparison group. Conditioning on score appeared to be less important to the performance of the conditional zRTs, than the choice of comparison group. The number of extremely fast zRTs computed using the same comparison groups were similar, with and without conditioning on score (for the full sample comparison group 57 and 69 and for the control condition comparison group 521 and 504, respectively). However, the number of extremely fast zRTs computed using different comparison groups were very different, even when they were matched by if they were conditioned on score (for those conditioned on score 57 and 521 and for those not conditioned on score 69 and 504, respectively). Thus, the use of a comparison group without compromise appears to be a key element in obtaining useful information from conditional zRTs.

Cluster and factor analyses found distinct groups of items, which showed very good correspondence to the disclosed and undisclosed item groups. Person-level analyses of correlations and cluster analyses identified separations between examinees with and without pre-knowledge. The cluster analyses were 96% accurate when grouping items (24 of 25 correctly grouped) and 97% accurate when grouping examinees (68 of 70 correctly grouped). The one item that was incorrectly grouped was second to last in the disclosed materials that participants received and thus may have exhibited a weaker effect of pre-knowledge than some of the other items.

Factor analyses of item scores showed poor performance in separating disclosed and undisclosed items, but factor analyses on the log RTs were about as effective at grouping items as the conditional zRTs. The two-factor solutions separated items better than the one-factor solutions. Thus, if two factors are observed in the conditional zRTs or log RTs in what was expected to be unidimensional data, item disclosure and pre-knowledge may be present. These findings indicate that disclosed items can be identified in strongly contaminated data simply by performing factor analyses on the log RTs. However, it is possible that this result is due to the large proportion of examinees with pre-knowledge in the current data and that factor analysis of conditional zRTs would outperform factor analysis of log RTs if lower rates of item disclosure or pre-knowledge were present.

The two and three-cluster solutions for grouping examinees resulted in the same accuracy. There were three conditions in the experimental study representing examinees with no pre-knowledge, examinees with pre-knowledge of items, and examinees with pre-knowledge of items and answers. The results for both cluster solutions indicate that examinees were grouped by whether they had any pre-knowledge rather than the nature of that pre-knowledge, such that examinees with pre-knowledge of items and examinees with pre-knowledge of items and answers were identified as a single group. These findings suggest that it does not matter if answers were provided, as examinees with pre-knowledge of the items exhibited similar patterns to examinees with pre-knowledge of the items and answers. It is possible that participants in the ItemOnly condition were able to solve the items they had pre-knowledge of, creating their own answer key, and thus obtaining a similar amount of pre-knowledge to those who had the answer key provided.

Many techniques for analyzing potential pre-knowledge require that examinees are administered the same items or forms, limiting their utility in practice. Conditional zRTs were developed and selected for research because they can be applied to tests administered using fixed-forms *and* to tests administered using other modern test designs, such as CAT, LOFT, or multi-stage adaptive testing (MSAT). This is a very important advantage of using the proposed method.

Conditional zRTs are easy to compute, analyze, and explain to exam stakeholders. The computation of conditional zRTs only requires item scores, response times, and a group of examinees to serve as an uncontaminated comparison group. Strict assumptions regarding the nature of the data are not required. The assumptions of z-scores are that the distribution is normal, hence the logistic transformation of RTs. The computation of the conditional zRTs assumes that the comparison group means and standard deviations are representative of the population of examinees without pre-knowledge who received the same score on the item. The statistical methodology is intuitive and interpreting the results does not require formidable statistical expertise. Conditional zRTs are datapoints that take into account an examinee's score on an item and the amount of time a typical examinee without pre-knowledge would take to complete the item to achieve that same score. These values could be used by testing programs of all sizes and can be used in MSAT, CAT, and LOFT test designs. These factors indicate that conditional zRTs may provide a useful method to conduct data forensics for non-traditional test designs. Thus, the main strength of conditional zRTs is the flexibility they offer for detecting item disclosure and examinee pre-knowledge across a wide variety of situations and sample sizes.

Limitations

In the current study, the majority of the participants had pre-knowledge. Thus, it is possible that analyses of the conditional zRTs were able to separate disclosed from undisclosed items and

examinees with pre-knowledge from those without because of the high rates of item disclosure (48%) and examinee pre-knowledge (65%) in the data. The computation of conditional zRTs should not be influenced by the proportion of item compromise or pre-knowledge, as each examinee's RTs are directly compared to a control sample of RTs with the same item scores. However, the performance of the grouping techniques, such as cluster analysis, in detecting groups of items and/or persons may be impacted by the proportion of compromise and pre-knowledge. Presumably, accuracy would decrease with smaller proportions of item compromise and examinee pre-knowledge. If the approximate baseline rates of item compromise or examinee pre-knowledge are known, the cluster analyses can be weighted with this information. If such information is unknown, other grouping techniques may perform better. We hope this research will serve to inspire future research on conditional zRTs in data with various rates of item disclosure and pre-knowledge. Additional research should compare conditional zRTs to other methods, such as the IRT-based l_z person-fit statistic, score-differencing statistics, and similarity analyses. Investigating conditional zRTs in a broader range of data will illuminate when this method is most appropriate and effective as well as identifying best practices for analyzing them.

One issue that may limit the utility of the proposed method is obtaining a sufficient sample size of RTs uncontaminated by pre-knowledge for comparison. Ideally, the uncontaminated comparison group would be composed of 30 or more examinees for each possible item score. In practice, the feasibility of such a sample size is likely dependent on the characteristics of a test, such as the design. For example, it may be more difficult to achieve sample sizes of 30 or more for a CAT exam where not every examinee receives the same items, because of low administration rates for some items. Additionally, an easy item should yield a sufficient sample size for comparison of examinees with an item score of one, but may not yield a sufficient sample size of examinees with an item score of zero. In the current research, the naturally occurring available sample size was used for comparison, but future research may benefit from investigating methods for improving such comparisons with small sample sizes.

Although it is expected that examinees with pre-knowledge have identifiable response patterns, it is important to note that disclosed test content varies in accuracy and may be utilized imperfectly because of human error or other factors. The scores of examinees with pre-knowledge depend on the accuracy of the disclosed test items or key. Even in a case where the correct answer for every item on a test is disclosed, examinees may have different abilities or tendencies to memorize and recall that information accurately. Some examinees with pre-knowledge may intentionally use disclosed content imperfectly; for example, by only accessing difficult items or answering some disclosed test items incorrectly to avoid detection. Some examinees may be aware that their RTs are being monitored by testing companies and engage in behaviors to make their test-taking behaviors look less suspicious. Thus, although it is expected that examinees with pre-knowledge will display distinct response patterns in comparison to examinees without pre-knowledge, there may be considerable individual variability. It is also likely that examinees with pre-knowledge will invent novel ways of responding to avoid detection, which makes the detection of pre-knowledge an evolving problem.

Future Research

Future research on conditional zRTs should attempt to improve the quality of information in the uncontaminated comparison group of log RTs. The purpose of the comparison group is to accurately capture typical response times of examinees who do not have pre-knowledge and who put an appropriate level of effort into responding to the items. In the current study, all log RTs for the examinees in the control group were used to compute the log RT means and standard deviations. Extreme outliers or rapid guesses were not excluded when computing these comparison group statistics because of the small sample size of the control condition and because such data are typical in testing data. However, the information provided by conditional zRTs could be improved by identifying and removing rapid guesses and other extreme data points prior to the computation of the comparison group statistics. This should cause increased separation in conditional zRTs between normal responses and rapid guesses or responses of examinees with pre-knowledge.

Another possible way to improve the quality of the information in the comparison group would be to use a statistic other than the mean in the calculation of the conditional zRTs. *Z*-scores use the mean and standard deviation to scale data and are thus most appropriate for normal distributions with sufficient sample sizes. In the current study, log transformations of response times were used to approximate normality, but it is possible that some item RT distributions remained skewed after the transformation and the mean did not accurately represent the center of the distribution because of the presence of outliers. In such cases, the median is more appropriate as an indicator of the center of the distribution. Using the median would prevent the skewness of the distribution from causing some log RTs to appear more or less extreme than they should, based on the comparison group. With sufficient sample size, it may also be possible to compute conditional zRTs matching examinees with a comparison group of examinees without pre-knowledge with a similar ability level. Although examinees with pre-knowledge likely have inflated ability estimates, this matching would compare their log RTs to the log RTs of examinees without pre-knowledge with a similar ability level. Such a comparison would likely detect the log RTs of the examinees with pre-knowledge as anomalous, even when the examinees have high ability levels. These adjustments could improve the quality of information in the comparison group, leading to improved separation of extreme log RTs from typical log RTs.

One potentially fruitful future direction for the use of conditional zRTs would be to identify groups of examinees based on their patterns. Latent profile analysis, or similar grouping methods could be used to distinguish types of responding. For example, differentiating high ability examinees from examinees

with pre-knowledge or low ability examinees from random guessers. This would advance the ultimate goal of this type of research, which is to develop statistical models to identify likely examinee behaviors. Investigating how the examinee behavior patterns for an exam shift over time may provide valuable information, such as when test content may have been disclosed or the population of examinees has changed.

Summary

In this study, a new method was proposed for analyzing response times to detect pre-knowledge, computing a conditional scaling by comparing each examinee's response time on each item to a sample of response times on the same item from examinees who did not have pre-knowledge and who received the same score on that item. The comparisons were conducted using logistically transformed response times and computed as *z*-scores, hence the name of the resulting values, conditional zRTs. These conditional zRTs were computed and analyzed in an experimental data set obtained by randomly selecting some examinees to have pre-knowledge of half of the test items. Some examinees received only test items and some also received correct answers. The results showed that the computation of conditional zRTs was feasible, even with a small uncontaminated comparison group, and that using an uncontaminated comparison group was more important to the performance of the zRTs than conditioning on score. Exploratory analyses of the conditional zRTs found strong separation between disclosed and undisclosed items and between examinees with and without pre-knowledge.

ETHICS STATEMENT

This study was conducted according to the recommendations of, and was approved by, the Institutional Review Board for Social and Behavioral Sciences at the University of Virginia. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

ST and DM developed the idea for the proposed method, both contributing critical pieces. ST provided the experimental data and drafted the manuscript. DM provided critical revisions.

ACKNOWLEDGMENTS

The authors would like to thank Karen Schmidt and Yasmeen Nabhani for helping to design and conduct the experiment that yielded the data analyzed in this research. Also, thanks to Marsha Line and David Silverman for volunteering their time to administer the experiment to participants.

REFERENCES

American Board of Internal Medicine (2010, June 9). Unprecedented action reflects ongoing commitment to protect the integrity of the medical board certification process. Available online at: https://www.abim.org/news/abim-sanctions-physicians-for-ethical-violations.aspx (accessed November 19, 2018).

Angoff, W. (1974). The development of statistical indices for detecting cheaters. *J. Am. Stat. Assoc.* 69, 44–49. doi: 10.1080/01621459.1974.10480126

Belov, D. I. (2016a). Comparing the performance of eight item preknowledge detection statistics. *Appl. Psychol. Meas.* 40, 83–97. doi: 10.1177/0146621615603327

Belov, D. I. (2016b). *A New Approach to Detecting Cluster Aberrancy*. LSAC Research Report (16-05), Newtown, PA.

Belov, D. I. (2017). On the optimality of the detection of examinees with aberrant answer changes. *Appl. Psychol. Meas.* 41, 338–352. doi: 10.1177/0146621617692077

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores (Chapters 17–20)*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley), 397–479.

Bliss, T. J. (2012). *Statistical Methods to Detect Cheating on Tests: A Review of the Literature* (Unpublished doctoral dissertation). Brigham Young University, Provo, UT.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Clark, M., Skorupski, W., Jirka, S., McBride, M., Wang, C., and Murphy, S. (2014). *An Investigation into Statistical Methods to Identify Aberrant Response Patterns*. Available online at: http://researchnetwork.pearson.com/wp-content/uploads/Data-Forensics-RD-Research-Report-Person-Fit.pdf (accessed August 8, 2018).

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *Br. J. Math. Stat. Psychol.* 38, 67–86. doi: 10.1111/j.2044-8317.1985.tb00817.x

Driscoll, R. (2007). *Westside Test Anxiety Scale Validation*. American Test Anxiety Association. Available online at: https://files.eric.ed.gov/fulltext/ED495968.pdf

Eckerly, C. A. (2017). "Detecting preknowledge and item compromise: understanding the status quo," in *Handbook of Quantitative Methods for Detecting Cheating on Tests,* eds G. J. Cizek and J. A. Wollack (New York, NY: Routledge), 101–123.

Eckerly, C. A., Babcock, B. G., and Wollack, J. A. (2015). "Preknowledge detection using a scale-purified deterministic gated IRT model," *Paper Presented at the Annual Meeting of the National Council of Measurement in Education* (Chicago, IL).

Educational Testing Service (2017). *Practice Book for the Paper-based GRE® revised General Test (2nd ed.).* Princeton, NJ. Retrieved from: https://www.ets.org/s/gre/pdf/practice_book_GRE_pb_revised_general_test.pdf

Federation of State Boards of Physical Therapy (2015, August 7). PEAT Investigation [Press release]. Available online at: https://www.fsbpt.org/NewsEvents/News/PEATInvestigation.aspx

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behav. Res.* 42, 481–507. doi: 10.1080/00273170701382583

Frary, R. B., Tideman, T. N., and Watts, T. M. (1977). Indices of cheating on multiple choice tests. *J. Educ. Stat.* 2, 235–256. doi: 10.3102/10769986002004235

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. Jr. (2003). A very brief measure of the big five personality domains. *J. Res. Pers.* 37, 504–528. doi: 10.1016/S0092-6566(03)00046-1

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the moral domain. *J. Pers. Soc. Psychol.* 101, 366–385. doi: 10.1037/a0021847

Guttman, L. A. (1944). A basis for scaling qualitative data. *Am. Sociol. Rev.* 91, 139–150. doi: 10.2307/2086306

Haberman, S. J., and Lee, Y.-H. (2017). *A Statistical Procedure for Testing Unusually Frequent Exactly Matching Responses and Nearly Matching Responses.* Research Report No. RR-17-23,Educational Testing Service, Princeton, NJ.

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: a K-means clustering algorithm. *Appl. Stat.* 28, 100–108. doi: 10.2307/2346830

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Appl. Meas. Educ.* 16, 277–298. doi: 10.1207/S15324818AME1604_2

Kassambara, A., and Mundt, F. (2017). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.5. Available online at: https://CRAN.R-project.org/package=factoextra

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86. doi: 10.1214/aoms/1177729694

Kyle, T. (2002). *Cheating Scandal Rocks GRE, ETS*. Retrieved from http://www.thedartmouth.com/article/2002/08/cheating-scandal-rocks-gre-ets/ (accessed November 1, 2018).

Liu, X. L., Primoli, V., and Plackner, C. (2013). "Utilization of response time in data forensics of K-12 computer-based assessment," *Presented at the Conference on the Statistical Detection of Test Fraud* (Madison, WI).

Lloyd, S. P. (1957). Least squares quantization in PCM. *IEEE Trans. Information Theor.* 2, 129–137. doi: 10.1109/TIT.1982.1056489

Maynes, D. D. (2017). "Detecting potential collusion among individual examines using similarity analysis," in *Handbook of Quantitative Methods for Detecting Cheating on Tests,* eds G. J. Cizek and J.A. Wollack (New York, NY: Routledge), 47–69.

Maynes, D. D., and Thomas, S. L. (2017). "A method for estimating the latent source used by answer copiers," *Paper Presented at the Meeting of the National Council on Measurement in Education* (New York, NY).

Meijer, R. R., and Sotaridona, L. (2006). *Detection of Advance Item Knowledge Using Response Times in Computer Adaptive Testing*. LSAC research report series; No. CT 03-03, Law School Admission Council, Newton, PA.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Revelle, W. (2016). *psych: Procedures for Personality and Psychological Research*. Evanston, IL: Northwestern University. Available online at : http://CRAN.R-project.org/package=psych Version = 1.6.4.

Scott, M. W. (2018). *A Series of Papers on Detecting Examinees Who Used a Flawed Answer Key* (Unpublished doctoral dissertation). Utah State University, Logan, UT.

Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591–611. doi: 10.1093/biomet/52.3-4.591

Shu, Z., Henson, R., and Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika* 78, 481–497. doi: 10.1007/s11336-012-9311-3

Sijtsma, K., and Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's non-parametric IRT model. *Appl. Psychol. Meas.* 16, 149–157. doi: 10.1177/014662169201600204

Thissen, D. (1983). "Timed testing: an approach using item response theory," in *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing,* ed D. J. Weiss (New York, NY: Academic Press), 179–203.

Thomas, S. L., and Maynes, D. D. (2018). "The use of data mining techniques to detect cheating," in *Data Analytics and Psychometrics: Informing Assessment Practices,* eds H. Jiao, R.W. Lissitz, and A. Van Wie (Charlotte, NC: Information Age Publishing), 77–108.

Tiemann, G. C., Miller, H. L., Kingston, N. M., and Foster, D. (2014). "Protecting item content via the discrete-option multiple-choice item type," *Presented at the Conference on Test Security*.

van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

van der Linden, W. J., and Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *J. Educ. Behav. Stat.* 31, 283–304. doi: 10.3102/10769986031003283

van der Linden, W. J., and van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika* 68, 251–265. doi: 10.1007/BF02294800

Wollack, J. A. (1997). A nominal response model approach for detecting answer copying. *Appl. Psychol. Meas.* 21, 307–320. doi: 10.1177/014662169702 14002

Wollack, J. A., and Maynes, D. M. (2016). "Detection of test collusion using cluster analysis," in *Handbook of Quantitative Methods for Detecting Cheating on Tests,* eds G. J. Cizek and J. A. Wollack (New York, NY: Routledge), 124–150.

Conflict of Interest Statement: ST and DM were employed by Caveon, a company that conducts data forensics analyses for a variety of clients, at the time this research was conducted.

Copyright © 2019 Toton and Maynes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 February 2019
doi: 10.3389/fpsyg.2019.00083

Understanding Test Takers' Choices in a Self-Adapted Test: A Hidden Markov Modeling of Process Data

Meirav Arieli-Attali[1,2]*, *Lu Ou*[2] *and Vanessa R. Simmering*[2]

[1] Department of Psychology, Fordham University, New York, NY, United States, [2] ACTNext, ACT Inc., Iowa City, IA, United States

With the rise of more interactive assessments, such as simulation- and game-based assessment, process data are available to learn about students' cognitive processes as well as motivational aspects. Since process data can be complicated due to interdependencies in time, our traditional psychometric models may not necessarily fit, and we need to look for additional ways to analyze such data. In this study, we draw process data from a study on self-adapted test under different goal conditions (Arieli-Attali, 2016) and use hidden Markov models to learn about test takers' choice making behavior. Self-adapted test is designed to allow test takers to choose the level of difficulty of the items they receive. The data includes test results from two conditions of goal orientation (performance goal and learning goal), as well as confidence ratings on each question. We show that using HMM we can learn about transition probabilities from one state to another as dependent on the goal orientation, the accumulated score and accumulated confidence, and the interactions therein. The implications of such insights are discussed.

Keywords: hidden Markov model, self-adapted test, likelihood ratio test, goal orientation, confidence

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Shiyu Wang,
University System of Georgia,
United States
Denise Reis Costa,
University of Oslo, Norway

***Correspondence:**
Meirav Arieli-Attali
meirav.attali@act.org;
mattali@fordham.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *02 September 2018*
Accepted: *11 January 2019*
Published: *06 February 2019*

Citation:
Arieli-Attali M, Ou L and Simmering VR (2019) Understanding Test Takers' Choices in a Self-Adapted Test: A Hidden Markov Modeling of Process Data. Front. Psychol. 10:83. doi: 10.3389/fpsyg.2019.00083

1. INTRODUCTION

With the rise of interactive assessment and learning programs, process data become available to infer about students' cognitive and motivational aspects. Process data can help us learn about students' strategies, preferences, and attitudes. In the context of problem solving, detecting strategies may reveal the cognitive processes needed to perform the task, and may even be considered as a factor in ability estimating (DiCerbo and Behrens, 2012; Liu et al., 2018). However, interactive assessments such as simulation- and game-based assessments often afford opportunities to make choices about the course of game/simulation (e.g., which variables to try in the simulation, which path to take in the game) that are not directly connected to ability albeit may influence its assessment. Such choices may be a result of or reflect metacognitive or motivational aspects of task performance. For example, students' self-estimated knowledge and belief in their ability, students' tendency toward challenge, or whether students are motivated to do their best or just perform at minimum effort are just a few of the factors that may play a role in choices made in interactive assessment.

Metacognition of task performance is rarely assessed as part of educational or academic assessments, yet it is acknowledged as important in student performance (Camara et al., 2015). One aspect of metacognition is the Feeling of Knowledge (FOK; Koriat, 1993) that is evoked naturally when attempting to answer a question. The cognitive process of attempting to answer

a question evokes the FOK based on the implicit and explicit accessibility cues (the easiness of accessing the answer, the vividness of the clues, the amount of information activated, etc.), and the content of that knowledge, its coherence, and the inferences that can be made from various clues retrieved (cf. Koriat, 1993, 2000). The more information activated and the easier it is accessed, the more confident a person is in his or her answer. Asking people to evaluate their level of confidence in answering a question is the most common way to eliciting their FOK estimation and is a moderately valid predictor of actual knowledge (Koriat, 1993, 2000; Wright and Ayton, 1994).

Feeling of knowing and estimation of one's own ability relate to and affect a student's engagement or motivation when performing a task, which is called the "expectancy component" in the Expectancy-Value Model of motivation by Pintrich and colleagues (Pintrich, 1988; Pintrich and De Groot, 1990; Pintrich and Schunk, 2002). Another component of the Expectancy-Value Model is the perceived value of the task. One aspect of perceived value is the goal orientation toward the task. Research on goal orientation of task performance yields a primary distinction between "performance" and "learning" goals (Dweck and Leggett, 1988). Individuals with a performance goal strive to perform at their best to demonstrate their skills to themselves or others, while individuals with a learning goal toward a task strive to learn from the task caring less about demonstrating their skills. Although individuals often exhibit these attitudes in general (Dweck et al., 1995), studies have shown that the orientation goal can be changed via psychological intervention given prior to performing a task and even only by the instructions of the task (Dweck, 2006). One of the pervasive findings regarding this distinction is that students with a learning goal are more motivated and seek more challenges (Dweck, 2006; Blackwell et al., 2007; Yeager and Dweck, 2012).

In this study we tap into motivational and metacognitive aspects of task performance via modeling process data. We are analyzing data from a previous study (Arieli-Attali, 2016) that applied the goal-orientation manipulation in a self-adapted test, while collecting also confidence ratings. Self-adapted testing is designed to allow test takers to choose the level of the difficulty of the items they receive. In her study, Arieli-Attali (2016) instructed participants in one condition to perform at their best on the test, with incentive of a reward; participants in the second condition were instructed to use the self-adapted test as a learning tool for a test the following day. Main findings showed that participants in the learning goal condition chose overall more difficult items (about half a level on average out of seven possible levels) compared to the performance goal condition, after controlling for pre-test performance, manifested both in the start of the test (the first choice) and the mean choices across all items. In addition, participants in the learning goal condition reverted to a strategy of choosing only the easiest level for all items significantly less frequently than those in the performance goal condition did (3.4% compared to 11.5%, respectively), and showed more exploratory behavior by choosing a wider range of difficulty levels (range of 3 levels compared to 2.5 levels in the performance goal condition).These results support the general theory and converge with previous findings by Dweck and colleagues about the higher motivation and tendency to seek more challenges when one is holding a learning goal orientation. Regarding confidence ratings, Arieli-Attali found that those in the learning goal condition showed under-confidence while those in the performance goal condition showed over-confidence (−1.4 vs. +1.9% respectively), similar to a recent study by Dweck and colleagues (Ehrlinger et al., 2016). Using the process data from Arieli-attali's study will allow us to tap deeper into the dynamics of choices as changing over time and depending on goal orientation and confidence rating. Before we describe the details of the current study, we provide a brief summary of research on self-adapted testing.

Self-adapted tests are designed to allow test takers to choose the level of difficulty of the items they receive (Rocklin and O'Donnell, 1987; Wise et al., 1992; Hontangas et al., 2004; Arieli-Attali, 2016). Such tests provide both product data—which items were answered correctly—as well as process data—what difficulty levels were chosen across time. Using an item response theory modeling approach, each test taker's ability can be estimated using the product data regardless of the item difficulty levels chosen. However, the difficulty preferences (the process data) may also be useful as an indication of the test taker's metacognitive and/or motivational state.

Previous studies on self-adapted tests were primarily concerned with the product data and its reliability and validity. However, there were also studies that looked into the process data particularly to examine the strategies of test takers in choosing the difficulty levels (Rocklin, 1989; Johnson et al., 1991; Ponsoda et al., 1997; Hontangas et al., 2000; Revuelta, 2004). In these studies, strategies were examined with regards to correct or incorrect responses to the adjacent preceding item, based on the assumption that the "results" on a previous item, whether correct or incorrect, would affect the next choice. Researchers were interested in uncovering the "rules," if existed, in examinees' choices, mostly adopting the approach of defining predetermined rules and looking in the data to find them. For example, Rocklin (1989) defined a "flexible strategy" as a selection of an easier level after an incorrect response, and a more difficult level after a correct response. This strategy is intuitive and in fact simulates the sequence of item difficulty produced by a Computer Adaptive Test (CAT) algorithm that maximizes test accuracy, where test takers often receive an easier item after incorrect response, and a harder item after a correct response, based on item response theory (Hambleton and Swaminathan, 1985). Defining such a strategy is based on the intuition that this would also be the most "rational" strategy people are using in their choices. In addition to the flexible strategy, Rocklin (1989) defined two variations: the "failure tolerant" and "failure intolerant." In the former, selections do not change after incorrect response (thus, showing tolerance to incorrect/failure), and in the latter, selections do not change after correct responses. Findings from this study and another study that followed (Johnson et al., 1991) showed that few test takers adhere to one of the three clear-cut categories, while most people exhibit more of a mixed strategy (or what Johnson et al., 1991 termed as "sluggishly flexible") where test takers selected a harder level after one or a string of several correct responses, and selected an easier level after one or a string

of several incorrect responses. In other studies (e.g., Hontangas et al., 2000; Revuelta, 2004) authors made somewhat different distinctions (such as totally rigid, partly flexible, and partly rigid); however, the findings were still very similar, showing that the majority of test takers are in the "partly rigid partly flexible" category, supporting previous findings. In Revuelta (2004)'s study, the author also reported that a majority of selections (about 60%) had the same difficulty level as the previous item.

In the current study, we take a different approach to look at the sequences of difficulty choices. Although we still look at transitions, we adopt a hidden latent approach rather than direct analysis of the observed choices. In addition, due to the inter-dependencies among difficulty choices, we apply a hidden Markov model (HMM). Under an HMM we assume independence between the observed choices conditional on respective latent states, which follow a first-order Markov process such that the current state only depends on the previous state. We explain initial states and state transitions in terms of probabilities and the effects of covariates on these probabilities. The HMM approach, as well as other variations of Markov models, are becoming increasingly popular among the educational measurement community for cognitive modeling (Yudelson et al., 2013; Li et al., 2016; LaMar, 2018; Wang et al., 2018) and analyses involving serially dependent process data (Vermunt et al., 1999; Dutilh et al., 2010; Bergner et al., 2017; Shu et al., 2017). We add to the literature an application of the HMM approach in characterizing test takers' behavior in self-adapted tests. The advantages of using this approach in our context are three-fold: (1) the introduction of the latent state as the metacognitive and/or motivational state that drives the observed difficulty choices can separate the stochasticity in the underlying metacognitive process from measurement errors; (2) it allows the same observed difficulty level to be a reflection of different latent states depending on the choices before and after (see **Figure 5** below for a specific example); (3) the estimation is robust against some design decisions such as the number of difficulty levels offered in different applications of self-adapted testing (whether 5, 7, or 9 difficulty levels are offered may change the observed sequence).

2. THE CURRENT STUDY

In this paper we conduct a secondary analysis of the data from Arieli-Attali (2016). The original study evaluated how the goal orientation conditions affected test takers' item difficulty choices, as well as the influence of different feedback conditions that will not be considered here. The aim of the current analysis is to model test takers' choices of item difficulty under the two orientation goal conditions, while taking into account the correctness and confidence ratings of previous items. We applied a first order Markov process, that looked at the change of the current state/class as dependent on the previous one. However, we used accumulated correctness and confidence as predictors. That is, we assumed that accumulated prior results of overall success (accumulated correct answers) and overall state of FOK (accumulated confidence) would affect the latent state and hence the next observed choice.

Using HMM we obtained the transition probabilities between the latent classes. Transition from a class with lower difficulty level to one with a higher difficulty level (i.e., an upward transition) represents a scenario where a test taker was willing to take on higher difficulty levels presumably due to increase in motivation, openness to challenge and exploration and/or increase in self-perceived ability due to evidence of success. On the contrary, a transition from choosing higher to lower difficulty items (i.e., a downward transition) illustrates the case where a test taker preferred to lower the difficulty, presumably due to a decrease in motivation or to alleviate stress, and/or as a strategy to get a better score/feedback (get more items correct).

Our first research question concerned modeling the transitions between latent states given the current state in the two goal conditions. Based on Arieli-Attali (2016)' results we anticipated that participants in the performance goal condition would not only have higher probability of choosing the lower difficulty state initially but also transition less from this state.

Our second research question addressed transitions in difficulty as dependent on correctness of and confidence on past items responses. We hypothesized that overall accumulated correctness and confidence would interact such that being correct and confident would generally enhance upward transitions while being incorrect and unconfident would enhance downward transitions. Regarding transitions in the mis-match cases of being correct with low confidence (under-confident) or being incorrect with high confidence (over-confidence), we hypothesized overall more transitions in both directions resulting from the conflict between confidence and feedback about correctness.

The paper is organized as follows: we first describe the data and the modeling approach. Next we provide some insights into the data using visualization of the raw data, the most common sequences and the patterns observed. We then report the results of the HMM analysis addressing specifically the two research questions. Lastly, we discuss these results in relation to their contribution to the emerging field of analyzing process data in assessment.

3. METHODS

3.1. Participants, Design, and Procedure

Arieli-Attali (2016) reported a final sample of 583 adult participants (age range = 18–74 years, M = 33.09; 45% women), recruited through Amazon Mechanical Turk (limited to native English speakers and residents of the US or Canada), who participated in a task over 2 days. Ethics approval for the study was obtained from Fordham University Institutional Review Board and a written informed consent was obtained from all participants (for the IRB approval and informed consent form see appendix E in Arieli-Attali, 2016). Our analysis includes data only from Day 1 of the experiment. On Day 1, participants completed a 24-item non-adaptive pre-test and a 40-item self-adapted test, both comprising open-ended general knowledge items. We used the pre-test scores that were obtained in the form of percentage of correct responses (ranged from 0.22 to 1, with a mean of

0.75, and standard deviation of 0.16). Following completion of the pre-test, participants were randomly assigned to one of two goal conditions: 286 participants were in the performance goal condition (condition = 1), instructed to maximize their score on the test, and 297 were in the learning goal condition (condition = 0), instructed to use the test as a learning tool for the test the next day. During the self-adapted test, participants chose a difficulty level for each item out of seven difficulty levels offered. After responding to each question, participants rated their confidence in their answer on a scale from 0 to 100 with 10-point intervals. After submission of the answer and the confidence ratings, participants received feedback whether their answer was correct or not and were provided with the correct answer. Coding of correctness was 0 for incorrect and 1 for correct. The observed item difficulty levels were integers from 1 to 7, which we divided by 7 to arrive at a range comparable with other variables used in the model fitting. Confidence reporting was converted proportionally to a scale from 0 to 1.

3.2. Modeling

We modeled test takers' choices of item difficulty using a hidden Markov model (HMM; Vermunt et al., 1999; Böckenholt, 2005; Visser and Speekenbrink, 2010; Visser, 2011) that assumed the manifest variables (i.e., item difficulty choices) are conditionally independent given an underlying latent Markov chain with a finite number of latent states or classes of the general difficulty preferences. We assumed that there are M states in the Markov chain. In the following text, we use "state" and "class" interchangeably to refer to the latent state of the M-state Markov chain, which is denoted as $S_{i,j}$, where integers i and j, respectively index participants and items. The categorical variable $S_{i,j}$ was an integer element from the finite set $\{1, 2, \cdots, M\}$ and varies across people and items. In our measurement model (as shown in the upper panel of **Figure 1**), we assumed that the conditional distribution of the manifest choices of item difficulty, $y_{i,j}$, given $S_{i,j}$, was univariate normal with mean $\mu_{S_{i,j}}$ and variance of $\sigma^2_{S_{i,j}}$. Although $y_{i,j}$ was ordinal in our current study, we treated it as continuous because we conceptualized the 7 manifest difficulty levels as a continuum representing participants' preferences of item difficulty and the intervals between any two points were approximately equal. The seven-level difficulty structure corresponded to the seven categories of a categorized item difficulty continuous scale (−3, −2, −1, 0, 1, 2, 3). The average difficulties of items at each difficulty level are: −3.3, −1.8, −0.9. −0.2, 0.5, 1.0, and 1.8 for level 1 through 7 respectively (corresponding to 92, 80, 68, 55, 41, 30, and 16% average probability of correct answer at each level) (Arieli-Attali, 2016). So the data were an ordinal approximation of a continuous variable. Practically, the rule of thumb is that ordinal variables with five or more categories can often be used as continuous without substantial harm to the analysis (Johnson and Creech, 1983; Norman, 2010; Rhemtulla et al., 2012). There were 7 categories in our study. We preferred to treat the data as continuous rather than as categorical for ease of interpretation. Depending on the magnitude of $\mu_{S_{i,j}}$, each class thus represented a more general item difficulty level that the participants feel comfortable choosing but may stochastically end at different manifest choices according to the measurement model.

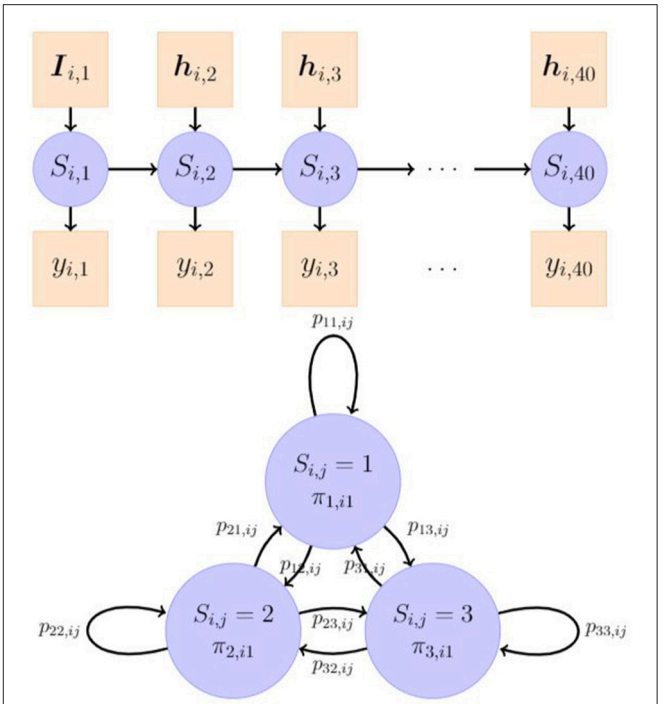


FIGURE 1 | An illustration of a 3-state hidden Markov model. The latent categorical $S_{i,j}$ is linked to the observed variable $y_{i,j}, j = 1, 2, \cdots, 40$ through a measurement model. $\pi_{m,i1}, m = 1, 2, 3$ is the probability of individual i's being initially in the class m and is explained by observed covariates $\boldsymbol{I}_{i,j}$. $p_{lm,ij}$ is the probability of individual i's transitioning from class l at item $j-1$ to class m at item j, and is explained by observed covariates $\boldsymbol{h}_{i,j}$.

In the latent model (as shown in **Figure 1**), we assumed that the change process of $S_{i,j}$ followed a first-order Markov chain process, where the current state only depended on the previous state. We described the dynamics of $S_{i,j}$ through its initial state and transitions between the states. The former depends on a $M \times 1$ vector of initial state probabilities, $\boldsymbol{\pi}_{i1} = [\pi_{m,i1}]$, and the latter is characterized by a $M \times M$ matrix of transition probabilities of moving from a state l to a state m, $\boldsymbol{P}_{ij} = [p_{lm,ij}]$, whose k-th row is denoted as $\boldsymbol{P}_{ij,k}$. Individual differences in the dynamic processes of $S_{i,j}$ were assumed to lie in the initial state probabilities and the transition probabilities, represented by two multinomial logistic regression models as follows:

$$\Pr\left(S_{i,1} = m|\boldsymbol{I}_{i,1}\right) \triangleq \pi_{m,i1} = \frac{\exp(a_m + \boldsymbol{b}_m^T \boldsymbol{I}_{i,1})}{\sum_{k=1}^{M} \exp(a_k + \boldsymbol{b}_k^T \boldsymbol{I}_{i,1})}, \quad (1)$$

$$\Pr\left(S_{i,j} = m|S_{i,j-1} = l, \boldsymbol{h}_{i,j}\right) \triangleq p_{lm,ij} = \frac{\exp(c_{lm} + \boldsymbol{d}_{lm}^T \boldsymbol{h}_{i,j})}{\sum_{k=1}^{M} \exp(c_{lk} + \boldsymbol{d}_{lk}^T \boldsymbol{h}_{i,j})}, \quad (2)$$

where $m = 1, 2, \cdots, M$ denotes the latent classes, $\boldsymbol{I}_{i,1}$, $\boldsymbol{h}_{i,j}$ are vectors of covariates used for prediction in the logistic regressions, a_m and c_{lm} denote the logit intercepts, and $\boldsymbol{b}_m$,

and $\boldsymbol{d}_{lm}$ denote the regression coefficients of the covariates in the associated log-odds (LO) relative to a specified reference class. In the current study, we predicted the initial class probabilities, $\pi_{m,i1}$, using the goal condition (abbreviated as *d*), pre-test score (abbreviated as *p*), and their interactions, and explain the transition probabilities, $p_{lm,ij}$, using the goal condition, accumulated correctness (abbreviated as *r*), accumulated confidence (abbreviated as *f*), and the interactions therein. The accumulated correctness and confidence at item *j* were calculated as the percentage of correctness or average confidence among items from the beginning to item *j*.

For identification purposes, both Equations (1) and (2) require specification of a reference class where all parameters in the regression equation are zero, which ensures that the initial class probabilities across all classes and the probability of moving into any class from a single class sum to 1.0. $\pi_{m,i1}$ is the probability of individual *i*'s being initially in the class *m*, and the regression coefficients $\boldsymbol{b}_m$ denote the effects of the covariates in the LO of being initially in the class *m* relative to the reference class. $p_{lm,ij}$ is the probability of individual *i*'s transitioning from class *l* at item $j-1$ to class *m* at item *j*, and the slopes in $\boldsymbol{d}_{lm}$ represent the effects of the covariates on the LO of transitioning from the *l*th class into the *m*th class relative to transitioning into the reference class. The choice of the reference class will only affect the logit regression parameters to be estimated, but will not influence the fit indices, the other parameter estimates, and the transformed estimated probabilities by a notable significant amount. Theoretically, the probability of being in the reference class cannot be zero in the model. Practically, it is recommended to choose a class that is presumably large enough and can make interpretation of results easier, for example, the normative class, the largest class, or the intermediate class. In this study, we used the default latent reference class of the R package depmixS4 (i.e., the first class), which turned out to be the medium class based on its mean estimate, but the findings should not be sensitive to this choice.

We can summarize Equations (1) and (2) into vector forms of $\boldsymbol{\pi}_{i1} = g([a_m + \boldsymbol{b}_m^T \boldsymbol{I}_{i,1}])$ and $\boldsymbol{P}_{ij} = g([c_{lm} + \boldsymbol{d}_{lm}^T \boldsymbol{h}_{i,j}])$, where $g(\cdot)$ is the softmax (normalized exponential) function. In our full model (also shown in **Table 1**), $\boldsymbol{I}_{i,1}$ is a 3×1 vector of the covariates *d*, *p*, and their interaction *dp*, and $\boldsymbol{h}_{i,j}$ is a 7×1 vector of the covariates including *d*, *r*, *f*, three two-way interactions (*df*, *dr*, and *fr*), and one three-way interaction (*dfr*). Accordingly, there are 3 parameters in $\boldsymbol{b}_m$ and 7 parameters in $\boldsymbol{d}_{lm}$. Altogether, there are $2M + 4(M-1) + 8M(M-1)$ parameters in the model, consisting of $2M$ parameters in the measurement model, (3+1) parameters each for $M-1$ regressions of initial class probabilities, and (7+1) parameters each (i.e., c_{lm} and $\boldsymbol{d}_{lm}$) for $M(M-1)$ regressions of $M(M-1)$ transition probabilities.

Parameters of the model can be estimated using the expectation-maximization (EM) algorithm, where the expectation of the complete log-likelihood function of the parameters given the observations $y_{i,j}$ and states $S_{i,j}$ are iteratively maximized to yield parameter estimates. In the R package depmixS4 (Visser and Speekenbrink, 2010), the EM algorithm has been implemented for unconstrained models, using the standard glm routine and the nnet.default routine in the nnet package (Venables and Ripley, 2002) in the maximization step for maximizing different parts of the expectations obtained in the expectation step. For more information on the estimation, we direct the readers to check the Visser and Speekenbrink (2010) paper.

Model fit of hidden Markov models can be compared using Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Konishi et al., 2004). The lower the AIC or BIC, the better the model fits the data. The fit of nested models can also be examined using likelihood ratio tests (LRT; Vermunt et al., 1999; Giudici et al., 2000). If $p < 0.05$, the more general model shows significant improvements in fit than the constrained model at the .05 level.

Additionally, given a sequence of observations $\{y_{i,j}\}$ and a hidden Markov model, we could get the most probable sequence of the state estimates of $\{S_{i,j}\}$, using the Viterbi algorithm (Viterbi, 1967; Forney, 1973; Rabiner, 1989). In the depmixS4 package, one can use the posterior() function to obtain the Viterbi most probable states, as well as the highest probabilities of a state sequence ending in a certain state at item *j* with all observations up to the item *j* taken into account.

4. RESULTS

In this section, we first provide a description and visualization of the data, along with the HMM general results about state classifications and initial state modeling, followed by two sets of our transitions modeling questions: (1) modeling transitions between states in the two goal conditions; (2) modeling transitions based on accumulated correctness and confidence and their interactions.

4.1. Description of Data

Here we summarize the most relevant characteristics of the data. First we present the choice sequences and the visualization of the data: **Figure 2** was created using the R package TraMineR (Gabadinho et al., 2011), and shows all the difficulty choice sequences and the ten most frequent sequences for the performance (P) and learning (L) goal conditions. The most frequent sequences are those with no transitions, where participants chose a level and stayed with it for the entire 40-item test, most frequently the extreme levels (level 1 and 7). Although there was not a clear difference between the conditions in the number or proportion of participants choosing to start and stay at the highest difficulty level (level 7; 3 participants in the performance goal condition, constituting 1.05%, and 5 in the learning goal condition, taking up 1.68%), substantially more participants chose to start at the lowest difficulty level (level 1) and stay there in the performance goal condition (33 or 11.54%) than in the learning goal condition (10 or 3.37%). In the learning goal condition there were also frequent sequences of starting and staying at level 2, 4, and 5 (as can be seen in the right-most panel), while in the performance goal condition these sequences were not frequent. Generally, there were also more switches in difficulty levels in the learning goal condition than in the performance goal condition. The average number of upward (i.e., from a lower manifest difficulty level to a higher one) and downward (i.e.,

TABLE 1 | Fit indices and parameter estimates across fitted models.

Model	A	B	B1	B2a	B2b	B3
M	2	3	3	3	3	3
$[1, \boldsymbol{I}_{i,j}]$			[1, d, p, dp]	[1, d, p, dp]	[1, d, p, dp]	[1, d, p, dp]
$[1, \boldsymbol{h}_{i,j}]$				[1, d]	[1, f, r, fr]	[1, d, f, r, fr, df, dr, dfr]
AIC	−16658.95	−27223.54	−27282.80	−27306.69	−27466.63	−27487.31
BIC	−16602.55	−27110.74	−27121.66	−27097.20	−27160.46	−26987.77
df	7	14	20	26	38	62
logLik	8336.474	13625.771	13661.399	13679.344	13771.316	13805.653
LRT			B ⊂ B1	B1 ⊂ B2a	B1 ⊂ B2b	B2b ⊂B3
$\Delta_{\chi^2}[\Delta_{df}]$			71.26*[6]	35.89*[6]	219.83*[18]	68.67*[24]
$\mu_1(\sigma_1)$	0.19 (0.07)	0.19 (0.07)	0.19 (0.07)	0.19 (0.07)	0.19 (0.07)	0.19 (0.07)
$\mu_2(\sigma_2)$	0.62 (0.22)	0.51 (0.12)	0.51 (0.12)	0.51(0.12)	0.51(0.11)	0.51(0.11)
$\mu_3(\sigma_3)$		0.86 (0.13)	0.86 (0.13)	0.86 (0.13)	0.86 (0.13)	0.86 (0.13)
$\boldsymbol{\pi}_{i1}$	$\begin{bmatrix}0.33\\0.67\end{bmatrix}$	$\begin{bmatrix}0.36\\0.45\\0.19\end{bmatrix}$	$g\left(\begin{bmatrix}1\\d\\p\\dp\end{bmatrix}^{\top}\begin{bmatrix}1.74&0&-3.21\\0.05&0&-2.29\\-3.06&0&3.15\\0.6&0&2.56\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\d\\p\\dp\end{bmatrix}^{\top}\begin{bmatrix}1.77&0&-3.16\\-0.01&0&-2.39\\-3.07&0&3.1\\0.62&0&2.68\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\d\\p\\dp\end{bmatrix}^{\top}\begin{bmatrix}1.75&0&-3.13\\0.03&0&-2.16\\-3.11&0&3.04\\0.62&0&2.35\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\d\\p\\dp\end{bmatrix}^{\top}\begin{bmatrix}1.75&0&-3.14\\0.08&0&-1.99\\-3.07&0&3.07\\0.49&0&2.1\end{bmatrix}\right)$
$\boldsymbol{P}_{ij,1}$	$\begin{bmatrix}0.96\\0.04\end{bmatrix}$	$\begin{bmatrix}0.93\\0.05\\0.02\end{bmatrix}$	$\begin{bmatrix}0.93\\0.05\\0.02\end{bmatrix}$	$g\left(\begin{bmatrix}1\\d\end{bmatrix}^{\top}\begin{bmatrix}2.54&0&-1.13\\0.6&0&0.21\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\f\\r\\fr\end{bmatrix}^{\top}\begin{bmatrix}2.96&0&-1.35\\-2.95&0&1.33\\-0.52&0&-0.66\\4.12&0&-0.33\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\d\\f\\r\\fr\\df\\dr\\dfr\end{bmatrix}^{\top}\begin{bmatrix}1.77&0&-1.33\\3.67&0&-1.04\\-1.32&0&1.98\\1.6&0&-0.42\\1.01&0&-1.87\\-4.65&0&-0.07\\-5.9&0&0.89\\7.92&0&1.28\end{bmatrix}\right)$
$\boldsymbol{P}_{ij,2}$	$\begin{bmatrix}0.02\\0.98\end{bmatrix}$	$\begin{bmatrix}0.04\\0.92\\0.04\end{bmatrix}$	$\begin{bmatrix}0.04\\0.92\\0.04\end{bmatrix}$	$g\left(\begin{bmatrix}1\\d\end{bmatrix}^{\top}\begin{bmatrix}-3.17&0&-2.96\\-0.05&0&-0.19\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\f\\r\\fr\end{bmatrix}^{\top}\begin{bmatrix}-1.21&0&-3.3\\-1.9&0&0.84\\-4.12&0&-1.37\\4.41&0&1.42\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\d\\f\\r\\fr\\df\\dr\\dfr\end{bmatrix}^{\top}\begin{bmatrix}-0.91&0&-3.32\\-0.94&0&0.06\\-1.9&0&1.77\\-5.33&0&-1.71\\5.38&0&0.83\\0.58&0&-2.59\\2.81&0&0.66\\-2.63&0&2.1\end{bmatrix}\right)$
$\boldsymbol{P}_{ij,3}$		$\begin{bmatrix}0.04\\0.06\\0.9\end{bmatrix}$	$\begin{bmatrix}0.04\\0.06\\0.9\end{bmatrix}$	$g\left(\begin{bmatrix}1\\d\end{bmatrix}^{\top}\begin{bmatrix}-0.39&0&2.74\\-0.16&0&0\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\f\\r\\fr\end{bmatrix}^{\top}\begin{bmatrix}0.24&0&2.93\\-0.96&0&-1.59\\-2.36&0&-0.04\\3.31&0&2.03\end{bmatrix}\right)$	$g\left(\begin{bmatrix}1\\d\\f\\r\\fr\\df\\dr\\dfr\end{bmatrix}^{\top}\begin{bmatrix}0.91&0&3.43\\-2.56&0&-1.58\\-2.13&0&-3.36\\-3.65&0&-0.11\\5.48&0&3.78\\4.08&0&4.82\\4.3&0&0.71\\-6.77&0&-4.79\end{bmatrix}\right)$

*$p < 0.05$; *d, condition; p, pre-test score; f, accumulated mean confidence; r, accumulated mean correctness.*

from a higher manifest difficulty level to a lower one) transitions in the learning condition were 7.43 and 6.51, respectively, both slightly higher than in the performance condition (6.07 and 5.40, respectively).

Regarding the distribution of choices, among all chosen item difficulty levels (i.e., a total of 583 × 40 choices), 22.85% were at level 1, ranked as the highest proportion and followed by 19.67% at level 4, 16.13% at level 3, 14.22% at level 2, 10.73% at level 5, 8.87% at level 7, and 7.52% at level 6. The distribution of the manifest choices is displayed in **Figure 3**, which suggests that the marginal distribution of the data should follow a mixture distribution. The chosen item difficulty levels were negatively correlated with answer correctness (point-biserial correlation r_{pb} = −0.30, p <0.001) and perceived confidence (r = −0.28, p <0.001), while the latter two variables were positively correlated (r_{pb} = 0.60, p <0.001).

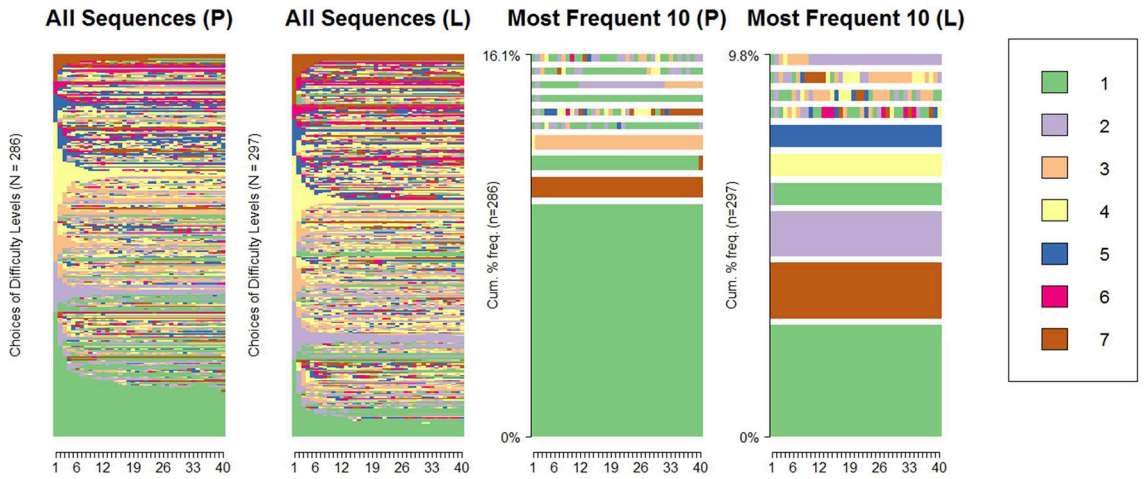


FIGURE 2 | Sequences of difficulty choices and the most frequent 10 sequences across the performance (P) and learning (L) conditions.

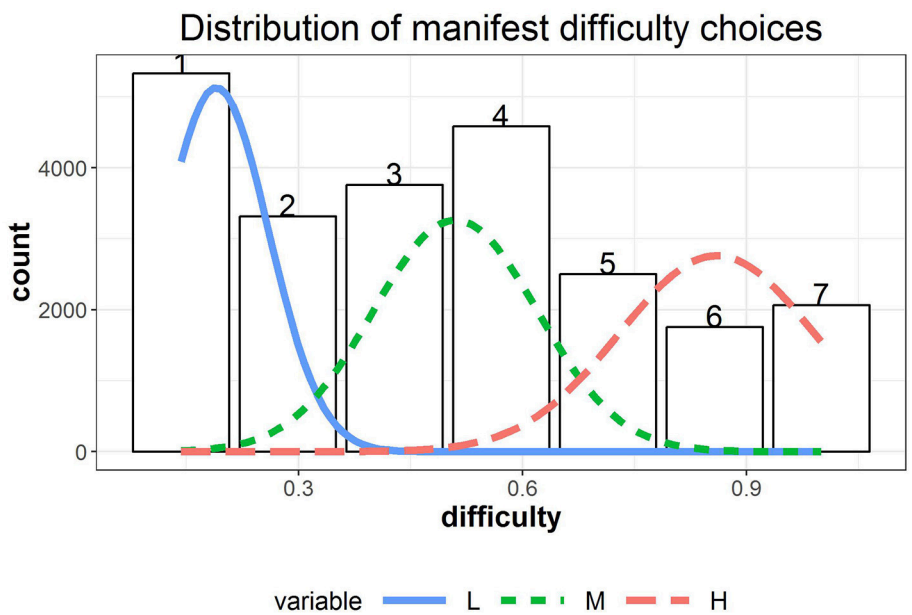


FIGURE 3 | The distribution of manifest difficulty choices overlaid with the normal densities from the fitted 3-state HMM model.

To examine the item dependencies in the difficulty choices, we obtained the residuals of the manifest difficulty data after removing the participant and item effects in a generalized additive mixed model using the R package mgcv (Wood, 2006). The autocorrelation functions (ACFs) of the residuals are plotted in **Figure 4** using the R package itsadug (van Rij et al., 2017), where the first panel displays the average ACF across participants, and the rest five are the ACFs for 5 randomly selected individuals. Although there were individual differences in the ACFs, on average the lag-1 autocorrelation was relatively high, around 0.44, suggesting the need of a first-order Markov model.

4.2. Hidden Markov Modeling Results

We used R package depmixS4 (Visser and Speekenbrink, 2010) to fit a series of HHM models to the data, which are summarized in **Table 1**. Comparison analyses indicated that a 3-state HMM (Model B; AIC = −27223.54, BIC = −27110.74) provided a better fit to the data than a 2-state HMM (Model A; AIC = −16658.95, BIC = −16602.55) based on the AIC and BIC (see **Table 1**). We hence present the results from 3-state HMMs. The parameter estimates of $\mu_{S_{i,j}}$ and $\sigma_{S_{i,j}}$ in the measurement model are summarized in **Table 1**. Based on **Table 1**, the three latent states respectively represent low [L; $\mu_1(\sigma_1)$ = 0.19 (0.07)], medium [M; $\mu_2(\sigma_2)$ = 0.51 (0.12)], and high [H; $\mu_3(\sigma_3)$ = 0.86

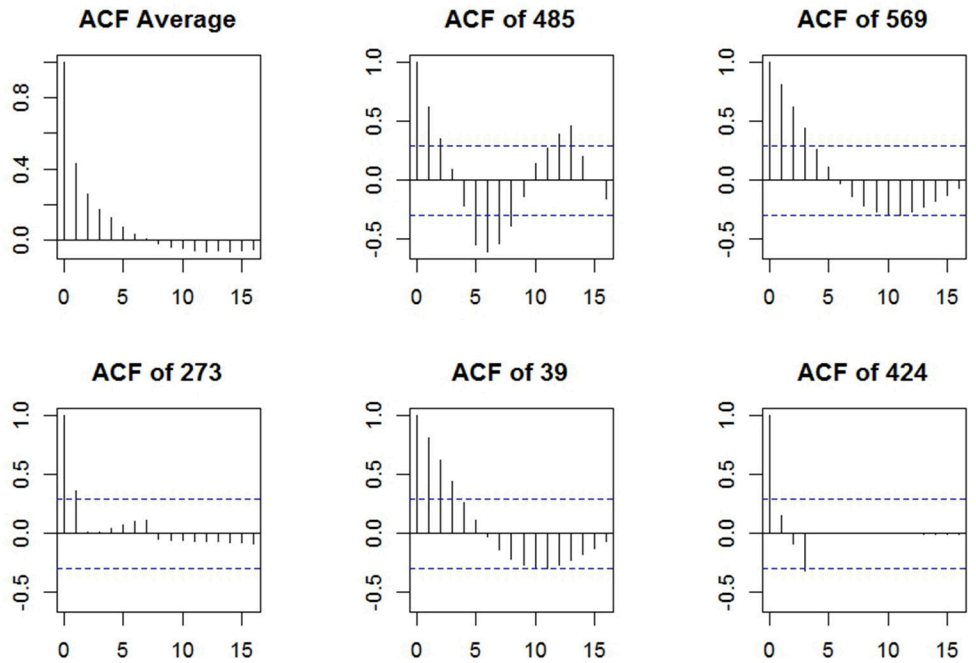


FIGURE 4 | The autocorrelation functions in the residuals of manifest difficulty choices after removing participant and item effects; the dashed blue lines represent the 95% confidence limits.

(0.13)] item difficulty levels. The estimated normal densities are shown as overlaid on the manifest distribution in **Figure 3**. The fitted mixture distribution of the hidden Markov models was still able to capture the manifest distribution of the chosen difficulty levels.

Figure 5 shows four representative participants' trajectories of item difficulty choices, accumulated confidence, and accumulated correctness, accompanied by the estimated most probable state at each item colored differently in the background. For example, participant 27 in the learning goal condition stayed at the low-level difficulty across time (switching between level 1 and 2) and the most probable latent state throughout was the L latent class (background colored blue). The accumulated correctness was generally high (above 70%) and the accumulated confidence was relatively low (mostly below 50%), yet they co-varied across time. Participant 347 in the performance goal condition, on the other hand, chose high-difficulty items across time (levels 5, 6, and 7) and the most probable latent state was the H latent class (background colored pink). The levels of confidence and correctness for this participant were almost identical, with a decline at approximately item 8. Participants 374 and 468 showed more transitions in their choices of difficulty levels. Participant 468 showed a gradual increase in item difficulty choices reflected in the transition of the most probable latent state from L to M to H latent states (blue → green → pink) with a steady high accumulated correctness albeit moderately low accumulated confidence. Lastly, participant 374 showed many transitions upwards and downwards, while correctness and confidence were moderately low. Note that participant 374 provides an illustration of how the same manifest/observed difficulty level can be associated with different most probable latent states: level 4 (just above .5 on the y-axis) was linked to the H state when the surrounding difficulty choices were higher (between item 10 and 20), but linked to the M state when the preceding choices were lower (between item 25 and 30) (see arrows on the figure).

Similar to Arieli-Attali (2016) in predicting choices, we used pre-test score (i.e., percentage of correctness), goal condition, and their interaction as predictors of initial difficulty latent state; the resulting Model is Model B1. As noted above Arieli-Attali (2016) reported that test takers' selection of difficulty on the first item differed across goal conditions, with lower difficulty chosen in the performance group, after controlling for pre-test performance. Our model analysis adds to this finding by using the three latent states rather than manifest difficulty levels. Parameter estimates and fit indices are shown in **Table 1**. Model B1 fits significantly better than Model B based on the LRT ($\Delta_{\chi^2} = 71.26$, $\Delta_{df} = 6$, $p < 0.05$). As it is not intuitive for us to draw conclusions from the parameter estimates in the LO sense, we illustrate the logistic regression results in terms of expected probabilities evaluated at certain values of the predictors in stacked bar figures. **Figure 6** indicate that when participants' pre-test scores are controlled, the expected probability of starting the test in a low-difficulty state compared to medium- or high-difficulty, is higher in the performance goal condition. Also it is evident from **Figure 6**, that within a condition, the higher the pre-test score, the higher the probability that the participant would initially be in a medium- or high-difficulty state. In particular, participants who answer fewer than half of the pre-test items correctly are more likely (above the

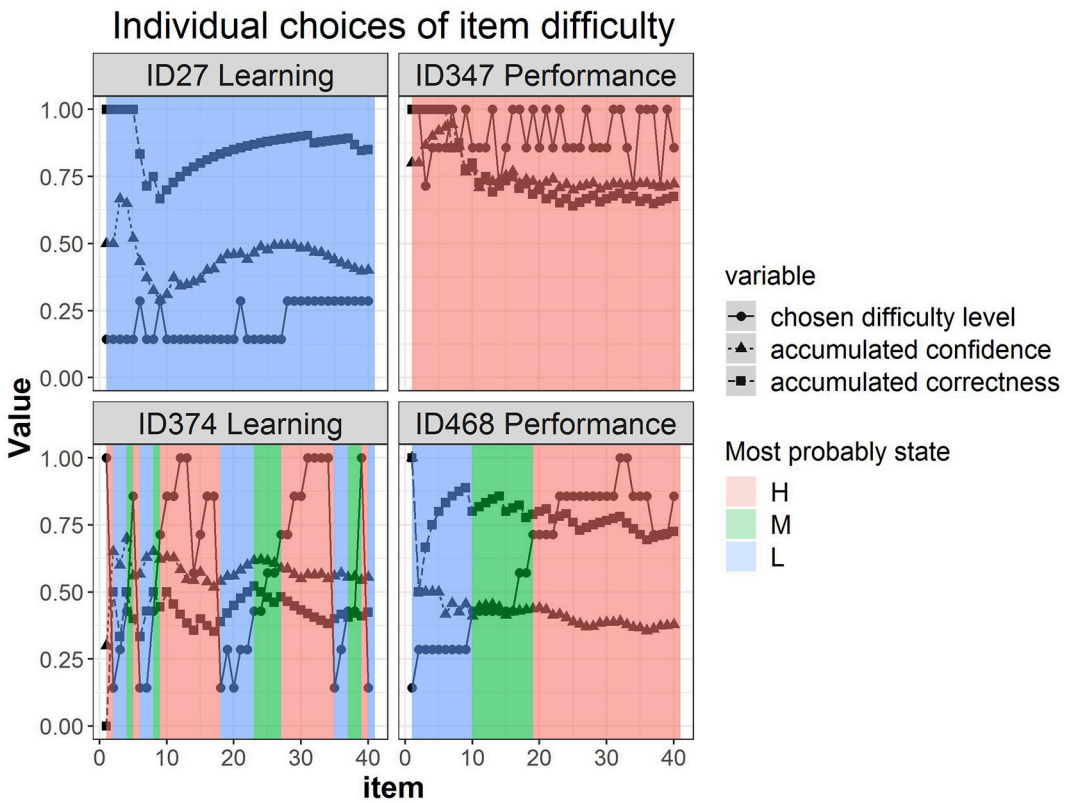


FIGURE 5 | Four representative individuals' trajectories of item difficulty choices, accumulated confidence, and accumulated correctness, with the estimated most probable state at each item as identified by the 3-state hidden Markov model colored differently in the background.

50%) to be in the low-difficulty initial state. Participants who have higher or full pre-test scores are more likely to be in initial state of medium- or high-difficulty. Now we turn to model transitions.

4.3. Research Question 1: Modeling Transitions in the Two Goal Conditions

Our first research question addressed modeling transitions between states in the two goal conditions. We added a multinomial logistic regression of the transition probabilities with condition as predictor to Model B1 (i.e., Model B2a), which significantly improves the fit of Model B1 ($\Delta_{\chi^2} = 35.89$, $\Delta_{df} = 6$, $p < 0.05$) and has a lower AIC value[1]. Fitting results of Model B2a are presented in **Table 1** and **Figure 7**. **Figure 7** shows the expected probability of transitions to and from each of the three latent states separately for each condition. As this figure shows, in both conditions the most probable choice behavior is staying in the same latent difficulty state with probabilities of over 90% (recall that different manifest difficulty levels were included in each latent state). However, when looking at the transitions between conditions, the model predicts a higher likelihood of staying at low difficulty and a lower likelihood of upward transitions from low to medium difficulty in the performance goal condition. In other words, participants in the performance goal condition are expected to transition less from the low state, confirming and adding to the results reported by Arieli-Attali (2016) that test takers in the performance goal condition tended to choose the lower level more frequently than in the learning goal condition, shown here also when considering latent states and transitions between states. Note that transitions from the medium or high state (either upwards or downwards) were similar between the two goal conditions.

4.4. Research Question 2: Modeling Transitions Based on Correctness and Confidence

We next fitted a more general model than Model B1, with accumulated correctness and confidence across items as predictors without condition (i.e., Model B2b), to evaluate the influence of these characteristics on transitions. Parameter estimates and fit indices are presented in **Table 1** and expected probabilities are displayed in **Figure 8**. Compared to Model B1, B2b fits the data significantly better ($\Delta_{\chi^2} = 219.83$, $\Delta_{df} = 18$, $p < 0.05$) and has lower AIC and BIC values. Note that the figure presents the four extreme quadrants of the two continuous scales. The horizontal line represents the accumulated correctness showing the extreme ends of the scale as "all incorrect" and "all correct" (from left to right), while the vertical line represents the accumulated confidence, showing the extremes of lowest and

[1]Please note that the BIC of B2a is larger than that of B1.

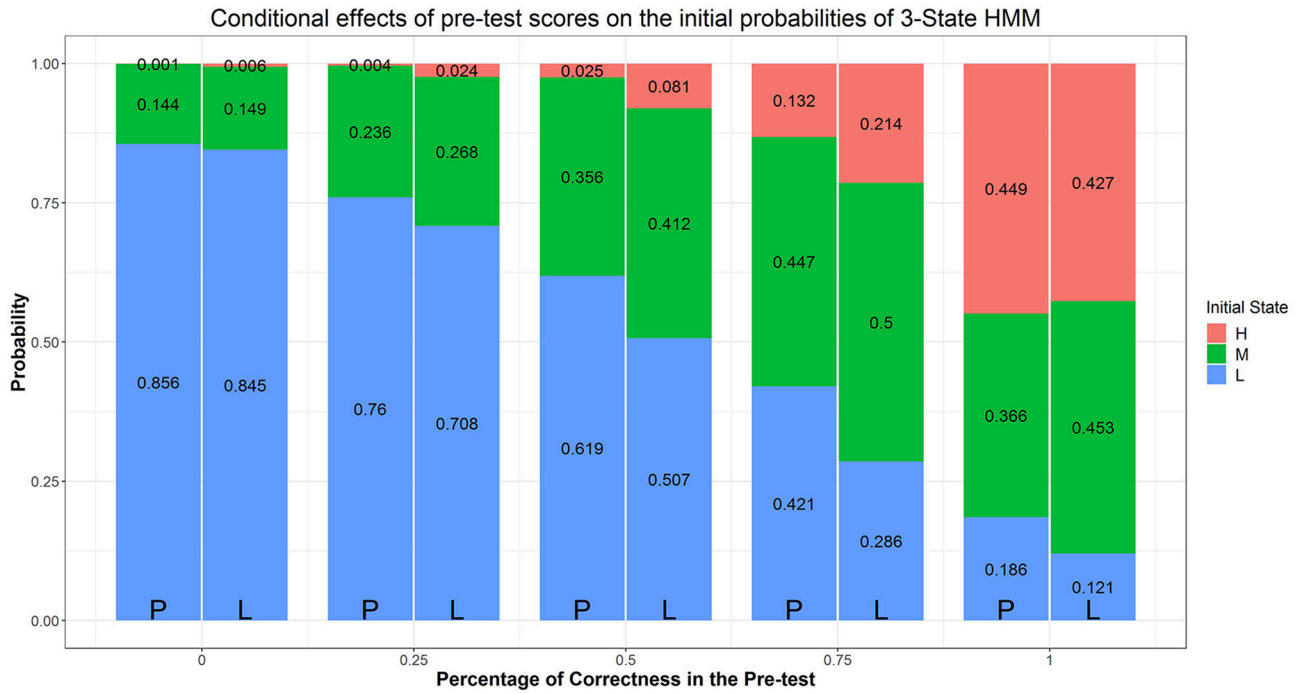


FIGURE 6 | Effects of condition and prescore on initial class probabilities. The numbers in the stacked bars are expected probabilities evaluated at certain values of the predictors based on the model fitting results.

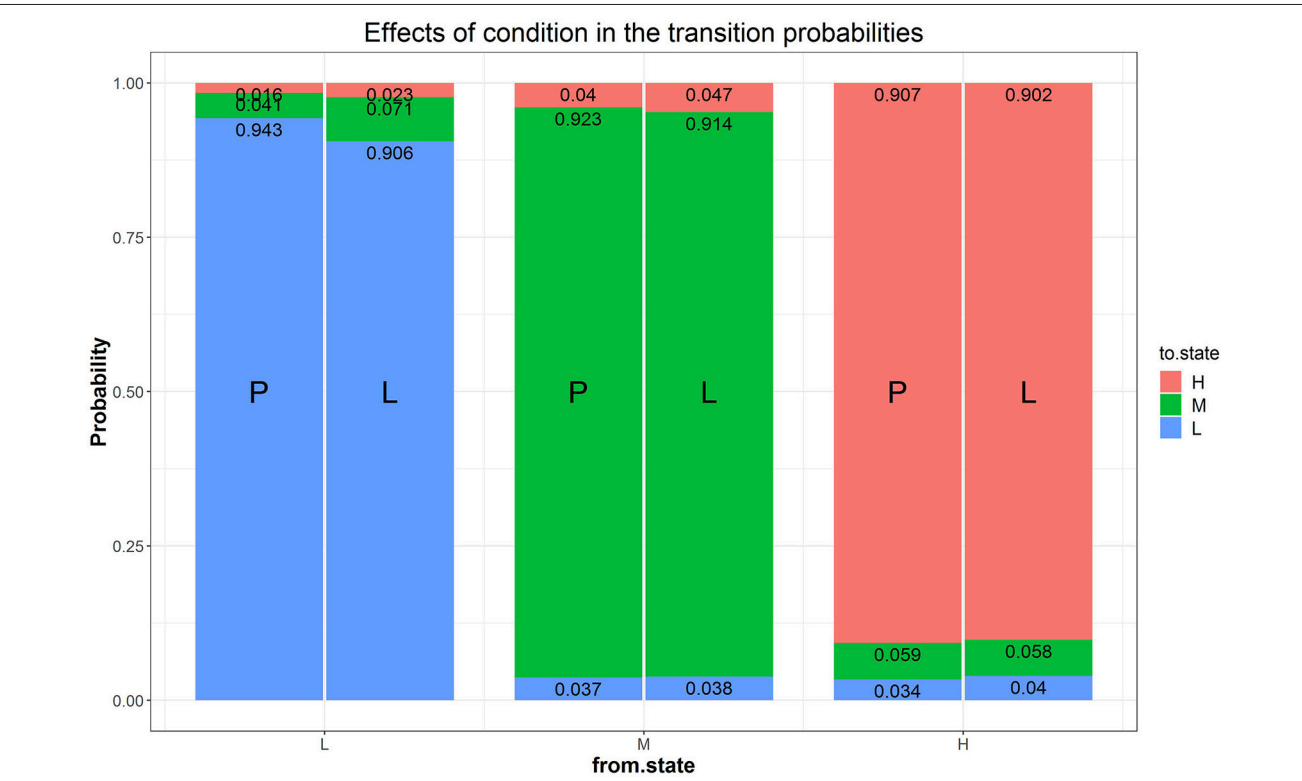


FIGURE 7 | Effects of condition on the transition probabilities. The numbers in the stacked bars are expected probabilities evaluated at certain values of the predictors based on the model fitting results.

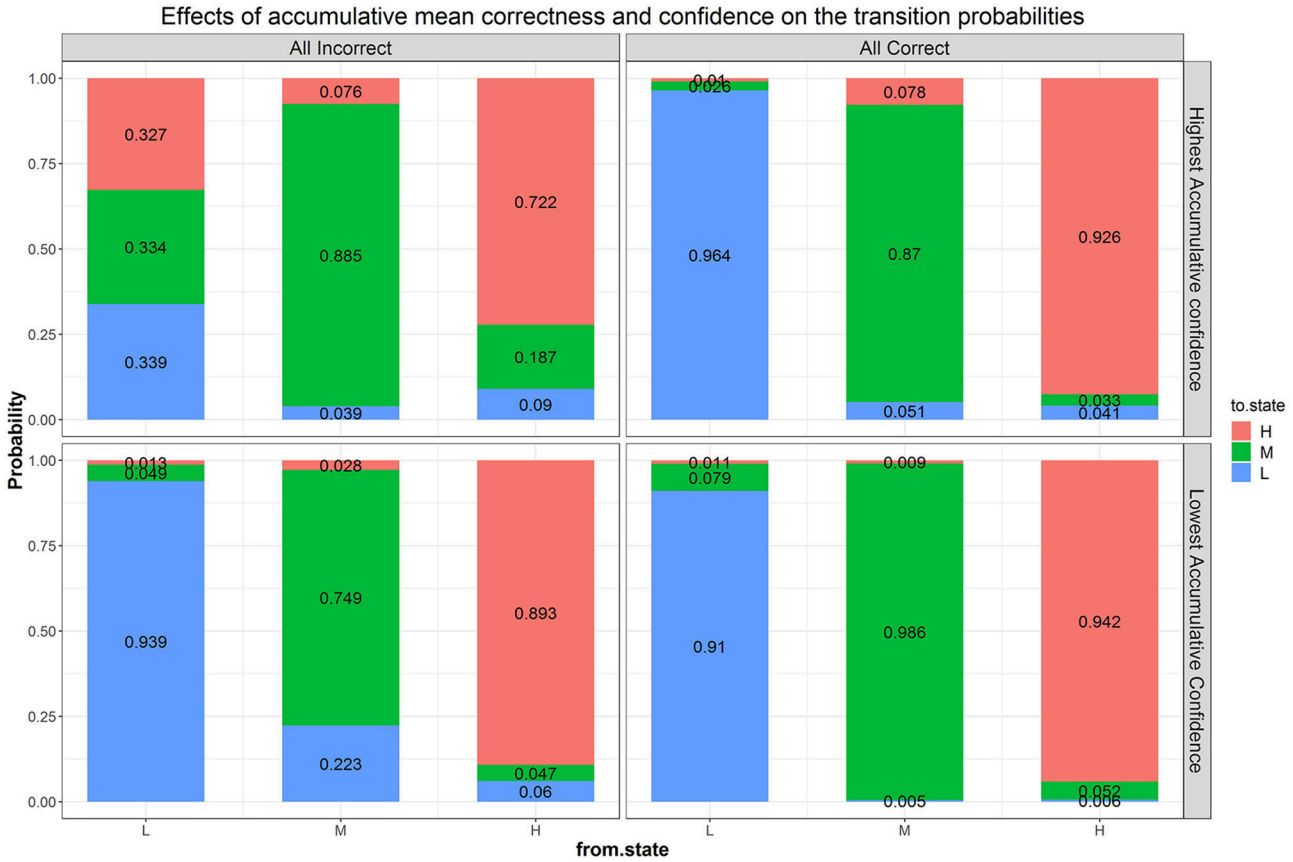


FIGURE 8 | Effects of accumulated correctness, confidence, and their interactions on the transition probabilities. The numbers in the stacked bars are expected probabilities evaluated at certain values of the predictors based on the model fitting results.

highest confidence (from bottom to top). As this figure shows, with high accumulated correctness (top and bottom right-side panels), expected probability of transitions is low and staying at the same difficulty state has the highest likelihood across the confidence scale. However, when accumulated correctness decreases (toward the quadrants in the top and bottom left-side panels) there is higher likelihood for transitions in both directions, and the likelihood of transitions increases as the confidence increases (i.e., illustrating the interaction between these factors). In particular, we can see expected downward transitions from the medium state when confidence is low (22.3%; bottom left-side panel), and from the high-state when confidence is high (27.7%; top left-side panel), as expected. However, we can also see that when the accumulated confidence is highest (top left-side panel; indicating over-confidence) participants are more likely to transition upwards from the low state (66.1%) equally to either the medium- or high-state. In other words, staying at the same state is the least probable in this case relative to other quadrants and states (recall that this quadrant is the extreme end of the confidence scale, and transition upwards from the low state are expected to increase as confidence increases). To get a sense of the frequency of participants with different relations between accumulated correctness and confidence, in particular considering the representation within each of the four quadrants illustrated in **Figure 8**, we show in **Figure 9** the relation between accumulated correctness and confidence after 10, 20, 30, and 40 items. As can be seen, the data cluster along the diagonal increasingly as the number of items increased, with sparse representation in the quadrants with mismatches between correctness and confidence. This suggests that test takers were overall well-calibrated in their confidence, with little representation of over- and under-confidence.

We then further added back goal condition as a predictor of the transition probabilities to Model B2b (i.e., Model B3), which significantly improved the fit of Model B2b ($\Delta_{\chi^2} = 68.67$, $\Delta_{df} = 24$, $p < 0.05$) and has a lower AIC value[2]. **Figure 10** shows the same transition probabilities as in **Figure 8** split by goal condition. The downward transitions when accumulated correctness decreases are also evident when split into the goal condition and are more so in the learning goal condition. The findings about higher likelihood of upward transitions in the over-confident quadrant are still evident when split into the goal conditions, with somewhat more transitions in the performance goal compared to learning goal condition (73.8 and 65.2%,

[2]Please note that the BIC of B3 is larger than that of B2b.

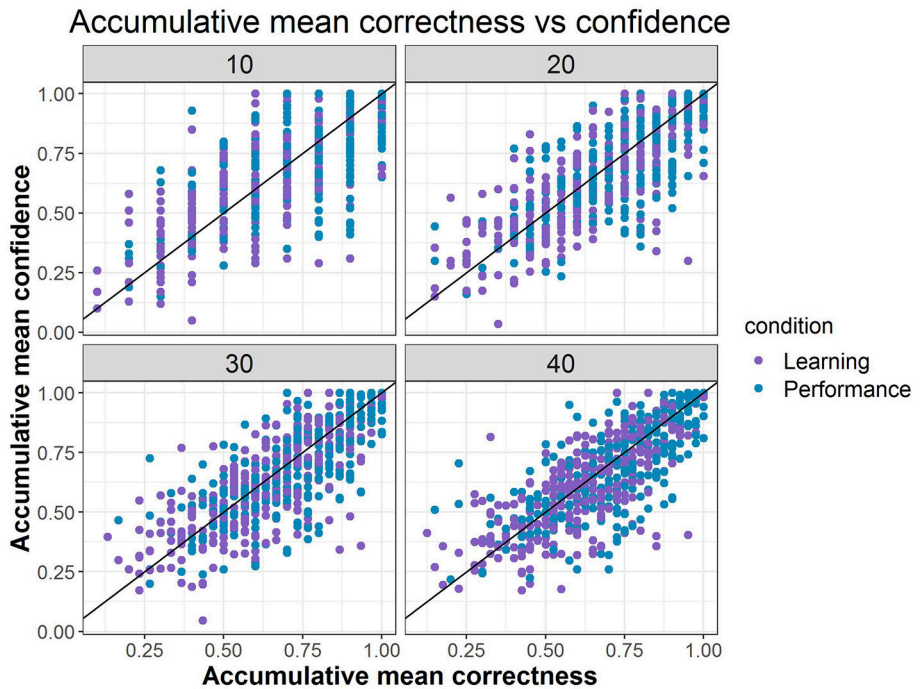


FIGURE 9 | The distribution of accumulated confidence and correctness.

respectively at the extreme quadrant of the confidence scale). A new finding from this split analysis is that there are also more transitions in the performance goal condition when accumulated correctness is high but confident is low (27%; bottom right-side panel, the quadrant indicating under-confidence).

5. DISCUSSION

The purpose of our secondary data analysis from Arieli-Attali (2016) was to apply a hidden Markov model to test takers' choices of item difficulty in a self-adapted test. We investigated whether those choices could be modeled by the goal condition (learning vs. performance), as well as the test takers' correctness and confidence across items. Analysis of the data using the hidden Markov model identified three latent states of difficulty from the seven manifest levels. These three latent states correspond to low, medium and high difficulty levels, and may be an indication of a low, medium or high self-estimated ability and/or motivation. We first modeled test takers' initial difficulty state based on their pre-test scores and goal condition, confirming past results (Arieli-Attali, 2016) about preference of lower difficulty in the performance goal condition, showing it here also as a higher expected probability of starting in the low state in the performance goal condition after controlling for pre-test scores. The results here add to the understanding that this is not just the single first choice influenced by the goal orientation (in addition to the self-perceived ability), but rather it is the participant's latent state that is influenced and therefore drives the choices accordingly. This result further confirms that when the goal orientation is to excel at a task individuals may avoid taking on challenges (Dweck, 2006).

We then used the model to predict transitions across items, and found the highest likelihood was to remain at the same difficulty state across items. This is the main contribution of applying a latent state approach in this context, because manifested choices may show transitions attributable to random variability while actual latent states are less likely to change. When using only goal condition as a predictor, there was no difference in transitions from the middle- or high- states between the two goal conditions, however there was a slightly lower likelihood of upward transitions from the low state in the performance goal condition relative to the learning goal condition, confirming the overall finding that test takers in the performance goal condition applied a strategy of the "easy way out," keeping low effort (Arieli-Attali, 2016).

The main contribution of this analysis is in the application of the HMM to model the interaction between answer correctness and confidence. We have shown that the likelihood of transitions increased when the accumulated correctness decreases. This result is intuitive as it means that participants were attentive to the correctness feedback and when they were overall wrong they tended to transition or change their metacognitive/motivational state. We found that downward transitions were more likely across the confidence scale as expected, but upward transitions were more likely when confidence increased for those who were in the low state, that is, we found that when confidence was highest, it reached the highest likelihood of about 2/3

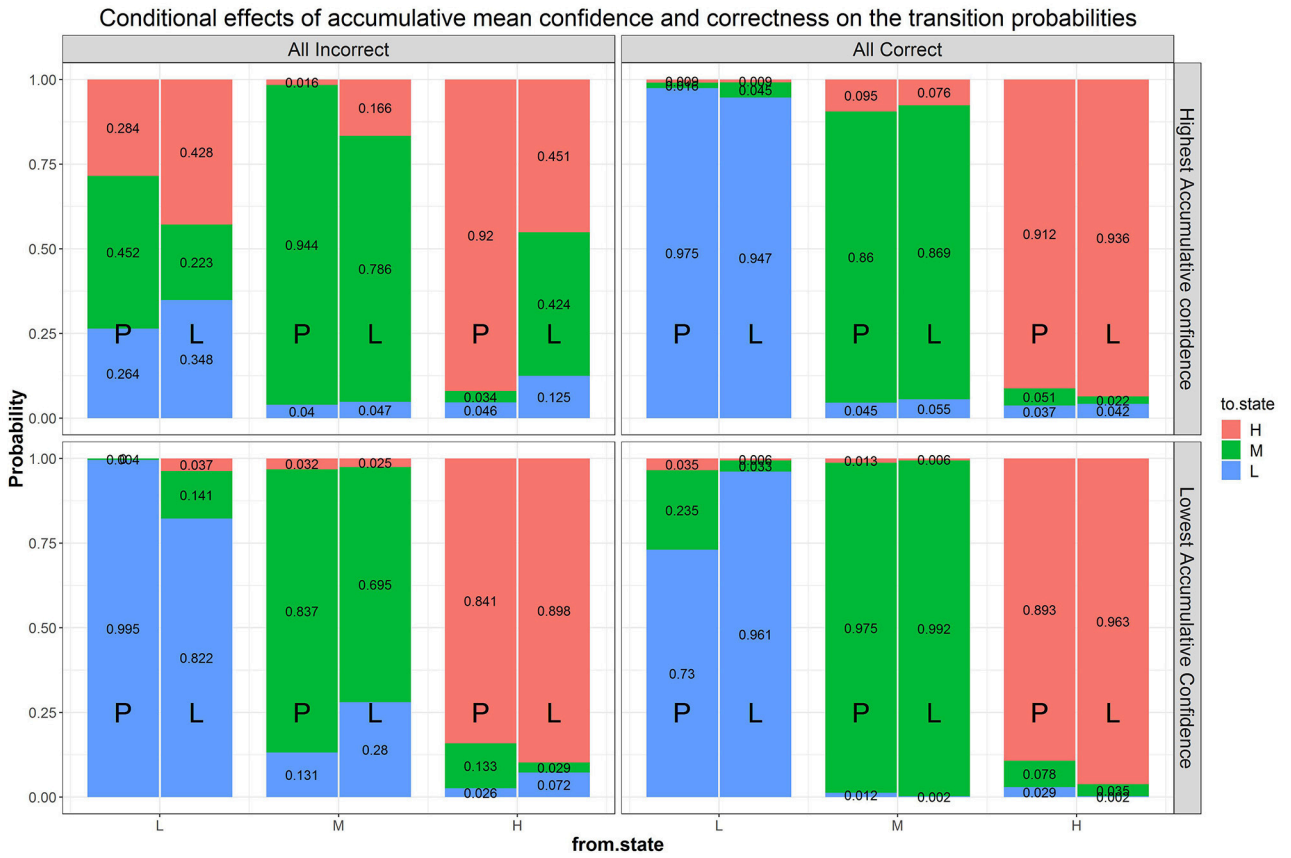


FIGURE 10 | Effects of condition, correctness, confidence, and their interactions on the transition probabilities. The numbers in the stacked bars are expected probabilities evaluated at certain values of the predictors based on the model fitting results.

upward transitions in the over-confidence end of the scale. This finding can be related to the literature on confidence and learning from errors by Metcalfe and colleagues (Butterfield and Metcalfe, 2001; Metcalfe and Xu, 2018). This line of research generally showed that people who made an error with high confidence were more likely to correct their mistake compared to a situation when the error was made with low confidence (the hypercorrection phenomenon). One of the explanations of this phenomenon is the surprise/attention explanation, which says that individuals experience surprise at being wrong when they were sure they were right, and as a consequence they rally their attentional resources (Butterfield and Metcalfe, 2006; Metcalfe et al., 2012). In our study we showed that individuals with high confidence who were proven incorrect were more likely to change difficulty state as reflected in more transitions upwards. The transitions upwards may be a reflection of being more attentive or putting forth more effort, similar to what occur under the hypercorrection phenomenon.

We also found that accumulated correctness and confidence interacted with goal condition in predicting transitions. The transitions when accumulated correctness decreases were also likely when split into the goal conditions but the downward transitions have higher likelihood in the learning goal condition, while the upward transitions in the over-confidence case have higher likelihood in the performance goal condition. This analysis also revealed a new finding of higher likelihood of upward transitions in the performance goal condition when accumulated correctness was high but confident was low, i.e., in the under-confidence end of the scale. These two findings together, that in the performance goal condition test takers were more likely to transition upwards from the low state in both mis-matched conditions (over- and under- confidence), indicate the specific interaction of the goal with correctness and confidence, and may suggest that when (1) participants are instructed to do their best, (2) they experience mis-match between what they think they know and what they actually know (feedback of correctness), and (3) they are in the low state without possible downward transition, they try to "find their luck" someplace else or decide to put more effort. This finding may suggest that mis-calibration between confidence and correctness could serve as a motivating factor, as being in the low state in the performance goal condition has been shown to stem from low motivation (Arieli-Attali, 2016). This combined pattern was not found for the learning goal condition, suggesting that evidence about mis-calibration when one is striving to learn has less of an effect (i.e., it had an effect in over-confidence, but not in under-confidence).

These results are consistent with the literature on goal orientation, showing that participants who are encouraged to use the test for learning rather than focusing on performance are more likely to seek challenges and show resilience amid difficulties (Yeager and Dweck, 2012). However, our additional findings about the interaction between correctness, confidence, and goal orientation further shed light on the complexity of the choices made in self-adapted test. The interactions we found suggest that the test takers' goal (i.e., whether the participant needs to maximize one's score, as the goal of the test), confidence across items (as a reflection of one's internal states), and correctness (as an outside feedback) together may form a recursive feedback loop that results in the changes of an individual's motivational and/or metacognitive state and further affects choice behavior.

To summarize, in this study we explored ways to learn about the motivation and feeling of knowledge of test takers and its affect on their actions while engaging in an interactive self-adapted test, via analyzing process data. Motivation and engagement is particularly crucial in low stakes assessment programs (such as the National Assessment of Educational Progress program, or the Trends in International Mathematics and Science Study), where test scores have no personal consequences for individuals, potentially resulting in low motivation to do one's best, and subsequently threatening the validity of the test scores. While low stakes programs make attempts to make their tests more interactive and appealing to participants in order to increase their engagement, we offer insights on how goal orientation, correctness and confidence influence choices that determine the course of the test. More research is needed to learn about how complex choice making in simulation- and game-based assessment can be modeled by factors inherent to the simulation or the game (such as curiosity, challenge seeking, sense of satisfaction, and the like).

AUTHOR CONTRIBUTIONS

MA-A: research design, data interpretation, and paper writing; LO: data analysis and interpretation, and paper writing; VS: data interpretation and paper writing.

FUNDING

The study was supported and funded by ACTNext, by ACT, Inc. ACTNext is the center of innovation for ACT, Inc., a testing company.

REFERENCES

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, eds B. N. Petrov and F. Csaki (Budapest: Akademiai Kiado), 267–281.

Arieli-Attali, M. (2016). *Self Adapted Testing as Formative Assessment: Effects of Feedback and Scoring on Engagement and Performance*. Ph.D. thesis, Fordham University, ProQuest LLC.

Bergner, Y., Walker, E., and Ogan, A. (2017). "Dynamic bayesian network models for peer tutoring interactions, in Innovative Assessment of Collaboration, Methodology of Educational Measurement and Assessment, eds A. A. von Davier, M. Zhu, and P. C. Kyllonen (Cham: Springer International Publishing), 249–268.

Blackwell, L. S., Trzesniewski, K. H., and Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. *Child Dev.* 78, 246–263. doi: 10.1111/j.1467-8624.2007.00995.x

Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions. *Psychol. Methods* 10, 65–83. doi: 10.1037/1082-989X.10.1.65

Butterfield, B., and Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *J. Exp. Psychol.* 27, 1491–1494. doi: 10.1037/0278-7393.27.6.1491

Butterfield, B., and Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacogn. Learn.* 1, 69–84. doi: 10.1007/s11409-006-6894-z

Camara, W., O'Connor, R., Mattern, K., and Hanson, M. A. (2015). *Beyond Academics: A Holistic Framework for Enhancing Education and Workplace Success.* ACT Research Report Series (Iowa City, IA: ACT), 1–91.

DiCerbo, K. E., and Behrens, J. T. (2012). "Implications of the digital ocean on current and future assessment," in *Computers and Their Impact on State Assessment: Recent History and Predictions for the Future*, eds R. W. Lissitz and H. Jiao (Charlotte, NC: Information Age Publishing), 273–306.

Dutilh, G., Wagenmakers, E., Visser, I., and van der Maas, H. L. J. (2010). A phase transition model for the speed-accuracy trade-off in response time experiments. *Cogn. Sci.* 35, 211–250. doi: 10.1111/j.1551-6709.2010.01147.x

Dweck, C. S. (2006). *Mindset: The New Psychology of Success*. New York, NY: Random House Incorporated.

Dweck, C. S., and Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychol. Rev.* 95, 256–273. doi: 10.1037/0033-295X.95.2.256

Dweck, C. S., yue Chiu, C., and yi Hong, Y. (1995). Implicit theories and their role in judgments and reactions: a world from two perspectives. *Psychol. Inq.* 6, 267–285. doi: 10.1207/s15327965pli0604_1

Ehrlinger, J., Mitchum, A. L., and Dweck, C. S. (2016). Understanding overconfidence: theories of intelligence, preferential attention, and distorted self-assessment. *J. Exp. Soc. Psychol.* 63, 94–100. doi: 10.1016/j.jesp.2015.11.001

Forney, G. D. (1973). The viterbi algorithm. *Proc. IEEE* 61, 268–278. doi: 10.1109/PROC.1973.9030

Gabadinho, A., Ritschard, G., Müller, N. S., and Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *J. Stat. Softw.* 40, 1–37. doi: 10.18637/jss.v040.i04

Giudici, P., Rydén, T., and Vandekerkhove, P. (2000). Likelihood-ratio tests for hidden Markov models. *Biometrics* 56, 742–747. doi: 10.1111/j.0006-341X.2000.00742.x

Hambleton, R. K., and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Springer Science & Business Media.

Hontangas, P., Olea, J., Ponsoda, V., Revuelta, J., and Wise, S. L. (2004). Assisted self-adapted testing: a comparative study. *Eur. J. Psychol. Assess.* 20, 2–9. doi: 10.1027/1015-5759.20.1.2

Hontangas, P., Ponsoda, V., Olea, J., and Wise, S. L. (2000). The choice of item difficulty in self adapted testing. *Eur. J. Psychol. Assess.* 16, 3–12. doi: 10.1027//1015-5759.16.1.3

Johnson, D. R., and Creech, J. C. (1983). Ordinal measures in multiple indicator models: a simulation study of categorization error. *Am. Sociol. Rev.* 48, 398–407. doi: 10.2307/2095231

Johnson, P. L., Roos, L. L., Wise, S. L., and Plake, B. S. (1991). Correlates of examinee item choice behavior in self-adapted testing. *Mid-Western Educ. Res.* 4, 25–28.

Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika* 91, 27–43. doi: 10.1093/biomet/91.1.27

Koriat, A. (1993). How do we know that we know? The assessibility model of the feeling of knowing. *Psychol. Rev.* 100, 609–639. doi: 10.1037/0033-295X.100.4.609

Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Conscious. Cogn.* 9, 149–171. doi: 10.1006/ccog.2000.0433

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika* 83, 67–88. doi: 10.1007/s11336-017-9570-0

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

Liu, H., Liu, Y., and Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: application of the modified Multilevel Mixture IRT model. *Front. Psychol.* 9:1372. doi: 10.3389/fpsyg.2018.01372

Metcalfe, J., Butterfield, B., Habeck, C., and Stern, Y. (2012). Neural correlates of people's hypercorrection of their false beliefs. *J. Cogn. Neurosci.* 24, 1571–1583. doi: 10.1162/jocn_a_00228

Metcalfe, J., and Xu, J. (2018). Learning from one's own errors and those of others. *Psychon. Bull. Rev.* 25, 402–408. doi: 10.3758/s13423-017-1287-7

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Adv. Health Sci. Educ.* 15, 625–632. doi: 10.1007/s10459-010-9222-y

Pintrich, P., and Schunk, D. (2002). *Motivation in Education: Theory, Research, and Applications*. Upper Saddle River, NJ: Merrill.

Pintrich, P. R,. and De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *J. Educ. Psychol.* 82, 33–40. doi: 10.1037/0022-0663.82.1.33

Pintrich, P. R. (1988). "A process-oriented view of student motivation and cognition," in *Improving Teaching and Learning Through Research: New Directions for institutional Research*, eds L. A. Mets and J. S. Stark (San Francisco, CA: Jossey-Bass; Spring), 65–79.

Ponsoda, V., Wise, S. L., Olea, J., and Revuelta, J. (1997). An investigation of self-adapted testing in a Spanish high school population. *Educ. Psychol. Meas.* 57, 210–221. doi: 10.1177/0013164497057002002

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Revuelta, J. (2004). Estimating ability and item-selection strategy in self-adapted testing: a latent class approach. *J. Educ. Behav. Stat.* 29, 379–396. doi: 10.3102/10769986029004379

Rhemtulla, M., Brosseau-Liard, P. É., and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychol. Methods* 17, 354–373. doi: 10.1037/a0029315

Rocklin, T. (1989). "Individual differences in item selection in computerized self adapted testing," in *Paper presented at the Annual Meeting of the American Educational Research Association* (San Francisco, CA), 1–15.

Rocklin, T., and O'Donnell, A. M. (1987). Self-adapted testing: a performance-improving variant of computerized adaptive testing. *J. Educ. Psychol.* 79, 315–319. doi: 10.1037/0022-0663.79.3.315

Shu, Z., Bergner, Y., Zhu, M., Hao, J., and von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychol. Test Assess. Model.* 59, 109–131. Available online at: https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2017_20170323/07_Shu.pdf

van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2017). *itsadug: Interpreting time series and autocorrelated data using GAMMs.* R package version 2.3. Available online at: https://cran.r-project.org/web/packages/itsadug/itsadug.pdf

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics With S, 4th Edn*. New York, NY: Springer.

Vermunt, J. K., Langeheine, R., and Böckenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *J. Educ. Behav. Stat.* 24, 179–207. doi: 10.3102/10769986024002179

Visser, I. (2011). Seven things to remember about hidden markov models: a tutorial on Markovian models for time series. *J. Math. Psychol.* 55, 403–415. doi: 10.1016/j.jmp.2011.08.002

Visser, I., and Speekenbrink, M. (2010). depmixS4: An R package for hidden Markov models. *J. Stat. Softw.* 36, 1–21. doi: 10.18637/jss.v036.i07

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* 13, 260–269. doi: 10.1109/TIT.1967.1054010

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

Wise, S. L., Plake, B. S., Johnson, P. L., and Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *J. Educ. Meas.* 29, 329–339. doi: 10.1111/j.1745-3984.1992.tb00381.x

Wood, S. N. (2006). *Generalized Additive Models: An Introduction With R*. Boca Raton, FL: CRC press.

Wright, G., and Ayton, P. (1994). *Subjective Probability*. Oxford, UK: John Wiley & Sons.

Yeager, D. S., and Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educ. Psychol.* 47, 302–314. doi: 10.1080/00461520.2012.722805

Yudelson, M. V., Koedinger, K. R., and Gordon, G. J. (2013). "Individualized bayesian knowledge tracing models," in *International Conference on Artificial Intelligence in Education* (Springer), 171–180.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Arieli-Attali, Ou and Simmering. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 November 2018
doi: 10.3389/fpsyg.2018.02231

Data Mining Techniques in Analyzing Process Data: A Didactic

Xin Qiao and Hong Jiao*

University of Maryland, College Park, College Park, MD, United States

Due to increasing use of technology-enhanced educational assessment, data mining methods have been explored to analyse process data in log files from such assessment. However, most studies were limited to one data mining technique under one specific scenario. The current study demonstrates the usage of four frequently used supervised techniques, including Classification and Regression Trees (CART), gradient boosting, random forest, support vector machine (SVM), and two unsupervised methods, Self-organizing Map (SOM) and *k*-means, fitted to one assessment data. The USA sample ($N = 426$) from the 2012 Program for International Student Assessment (PISA) responding to problem-solving items is extracted to demonstrate the methods. After concrete feature generation and feature selection, classifier development procedures are implemented using the illustrated techniques. Results show satisfactory classification accuracy for all the techniques. Suggestions for the selection of classifiers are presented based on the research questions, the interpretability and the simplicity of the classifiers. Interpretations for the results from both supervised and unsupervised learning methods are provided.

Keywords: data mining, log file, process data, educational assessment, psychometric

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Daniel Bolt,
University of Wisconsin-Madison, United States
Hongyun Liu,
Beijing Normal University, China

****Correspondence:***
Xin Qiao
xin.qiao56@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *14 March 2018*
Accepted: *29 October 2018*
Published: *23 November 2018*

Citation:
Qiao X and Jiao H (2018) Data Mining Techniques in Analyzing Process Data: A Didactic. Front. Psychol. 9:2231. doi: 10.3389/fpsyg.2018.02231

INTRODUCTION

With the advance of technology incorporated in educational assessment, researchers have been intrigued by a new type of data, process data, generated from computer-based assessment, or new sources of data, such as keystroke or eye tracking data. Most often, such data, often referred to as "data ocean," is of very large volume and with few ready-to-use features. How to explore, discover and extract useful information from such an ocean has been challenging.

What analyses should be performed on such process data? Even though specific analytic methods are to be used for different data sources with specific features, some common analysis methods can be performed based on the generic characteristics of log files. Hao et al. (2016) have summarized several common analytic actions when introducing the package in Python, glassPy. These include summary information about the log file, such as the number of sessions, the time duration of each session, and the frequency of each event, can be obtained through a summary function. In addition, event n-grams, or event sequences of different lengths, can be formed for further utilization of similarity measures to classify and compare persons' performances. To take the temporal information into account, hierarchical vectorization of the rank ordered time intervals and the time interval distribution of event pairs were also introduced. In addition to these common analytic techniques, other existing data analytic methods for process data are Social Network Analysis (SNA; Zhu et al., 2016), Bayesian Networks/Bayes nets (BNs; Levy, 2014), Hidden Markov Model (Jeong et al., 2010), Markov Item Response Theory (Shu et al., 2017), diagraphs (DiCerbo et al., 2011) and process mining (Howard et al., 2010). Further, modern data mining techniques,

including cluster analysis, decision trees, and artificial neural networks, have been used to reveal useful information about students' problem-solving strategies in various technology-enhanced assessments (e.g., Soller and Stevens, 2007; Kerr et al., 2011; Gobert et al., 2012).

The focus of the current study is about data mining techniques and this paragraph provides a brief review of related techniques that have been frequently utilized and lessons that have been learned related to analyzing process data in technology-enhanced educational assessment. Two major classes of data mining techniques are supervised and unsupervised learning methods (Fu et al., 2014; Sinharay, 2016). Supervised methods are used when subjects' memberships are known and the purpose is to train a classifier that can precisely classify the subjects into their own category (e.g., score) and then be efficiently generalized to new datasets. Unsupervised methods are utilized when subjects' memberships are unknown and the goal is to categorize the subjects into clearly separate groups based on features that can distinguish them apart. Decision trees, as a supervised data classification method, has been used very often in analysing process data in educational assessment. DiCerbo and Kidwai (2013) used Classification and Regression Tree (CART) methodology to create the classifier to detect a player's goal in a gaming environment. The authors demonstrated the building of the classifier, including feature generation, pruning process, and evaluated the results using precision, recall, Cohen's Kappa and A' (Hanley and McNeil, 1982). This study proved that the CART could be a reliable automated detector and illustrated the process of how to build such a detector with a relative small sample size ($n = 527$). On the other hand, cluster analysis and Self-Organizing Maps (SOMs; Kohonen, 1997) are two well-established unsupervised techniques that categorize students' problem-solving strategies. Kerr et al. (2011) showed that cluster analysis can consistently identify key features in 155 students' performances in log files extracted from an educational gaming and simulation environment called *Save Patch* (Chung et al., 2010), which measures mathematical competence. The authors described how they manipulated the data for the application of clustering algorithms and showed evidence that fuzzy cluster analysis is more appropriate than hard cluster analysis in analyzing log file process data from game/simulation environment. Most importantly, the authors demonstrated how cluster analysis can identify both effective strategies and misconceptions students have with respect to the related construct. Soller and Stevens (2007) showed the power of SOM in terms of pattern recognition. They used SOM to categorize 5284 individual problem-solving performances into 36 different problem-solving strategies, each exhibiting different solution frequencies. The authors noted that the 36 strategy classifications can be used as input to a test-level scoring process or externally validated by associating them with other measures. Such detailed classifications can also serve as valuable feedback to students and instructors. Chapters in Williamson et al. (2006) also discussed extensively the promising future of using data mining techniques, like SOM, as an automated scoring method. Fossey (2017) has evaluated three unsupervised methods, including *k*-means, SOM and Robust Clustering Using Links (ROCK) on analyzing process data in log files from a game-based assessment scenario.

To date, however, no study has demonstrated the utilization of both supervised and unsupervised data mining techniques for the analysis of the same process data. This study aims at filling this gap and provides a didactic of analyzing process data from the 2012 PISA log files retrieved from one of the problem-solving items using both types of data mining methods. This log file is well-structured and representative of what researchers may encounter in complex assessments, thus, suitable for demonstration purposes. The goal of the current study is 3-fold: (1) to demonstrate the use of data mining methods on process data in a systematic way; (2) to evaluate the consistency of the classification results from different data mining techniques, either supervised or unsupervised, with one data file; (3) to illustrate how the results from supervised and unsupervised data mining techniques can be used to deal with psychometric issues and challenges.

The subsequent sections are organized as follows. First, the PISA 2012 public dataset, including participants and the problem-solving item analyzed, is introduced. Second, the data analytic methods used in the current study are elaborated and the concrete classifier development processes are illustrated. Third, the results from data analyses are reported. Lastly, the interpretations of the results, limitations of the current study and future research directions are discussed.

METHODS

Participants

The USA sample ($N = 429$) was extracted from the 2012 PISA public dataset. Students were from 15 years 3 months old to 16 years 2 months old, representing 15-year-olds in USA (Organisation for Economic Co-operation Development, 2014). Three students with missing student IDs and school IDs were deleted, yielding a sample of 426 students. There were no missing responses. The dataset was randomly partitioned into a training dataset ($n = 320$, 75.12%) and a test dataset ($n = 106$, 24.88%). The size of the training dataset is usually about 2 to 3 times of the size of the test dataset to increase the precision in prediction (e.g., Sinharay, 2016; Fossey, 2017).

Instrumentation

There are 42 problem-solving questions in 16 units in 2012 PISA. These items assess cognitive process in solving real-life problems in computer-based simulated scenarios (Organisation for Economic Co-operation Development, 2014). The problem-solving item, TICKETS task2 (CP038Q01), was analyzed in the current study. It is a level-5 question (there were six levels in total) that requires a higher level of exploring and understanding ability in solving this complex problem (Organisation for Economic Co-operation Development, 2014). This interactive question requires students explore and collect necessary information to make a decision. The main cognitive processes involved in this task are planning and executing. Given the problem-solving scenario, students need to come up with a plan and test it and modify it if needed. The item asks students to use their concession fare to

find and buy the cheapest ticket that allows them to take 4 trips around the city on the subway within 1 day. One possible solution is to choose 4 individual concession tickets for city subway, which costs 8 zeds while the other is to choose one daily concession ticket for city subway, which costs 9 zeds. **Figure 1** includes these two options. Students can always use "CANCEL" button before "BUY" to make changes. Correctly completing this task requires students to consider these two alternative solutions, then make comparisons in terms of the costs and end up choosing the cheaper one.

This item is scored polytomously with three score points, 0, 1, or 2. Students who derive only one solution and fail to compare with the other get partial credits. Students who do not come up with either of the two solutions, but rather buy the wrong ticket, get no credit on this item. For example, the last picture in **Figure 1** illustrates the tickets for four individual full fare for country trains, which cost 72 zeds. "COUNTRY TRAINS" and "FULL FARE" are considered as unrelated actions because they are not the necessary actions to accomplish the task this item requires. In terms of scoring, unrelated actions are allowed as long as the students buy the correct ticket in the end and make comparisons during the action process.

Data Description

The PISA 2012 log file dataset for the problem-solving item was downloaded at http://www.oecd.org/pisa/pisaproducts/

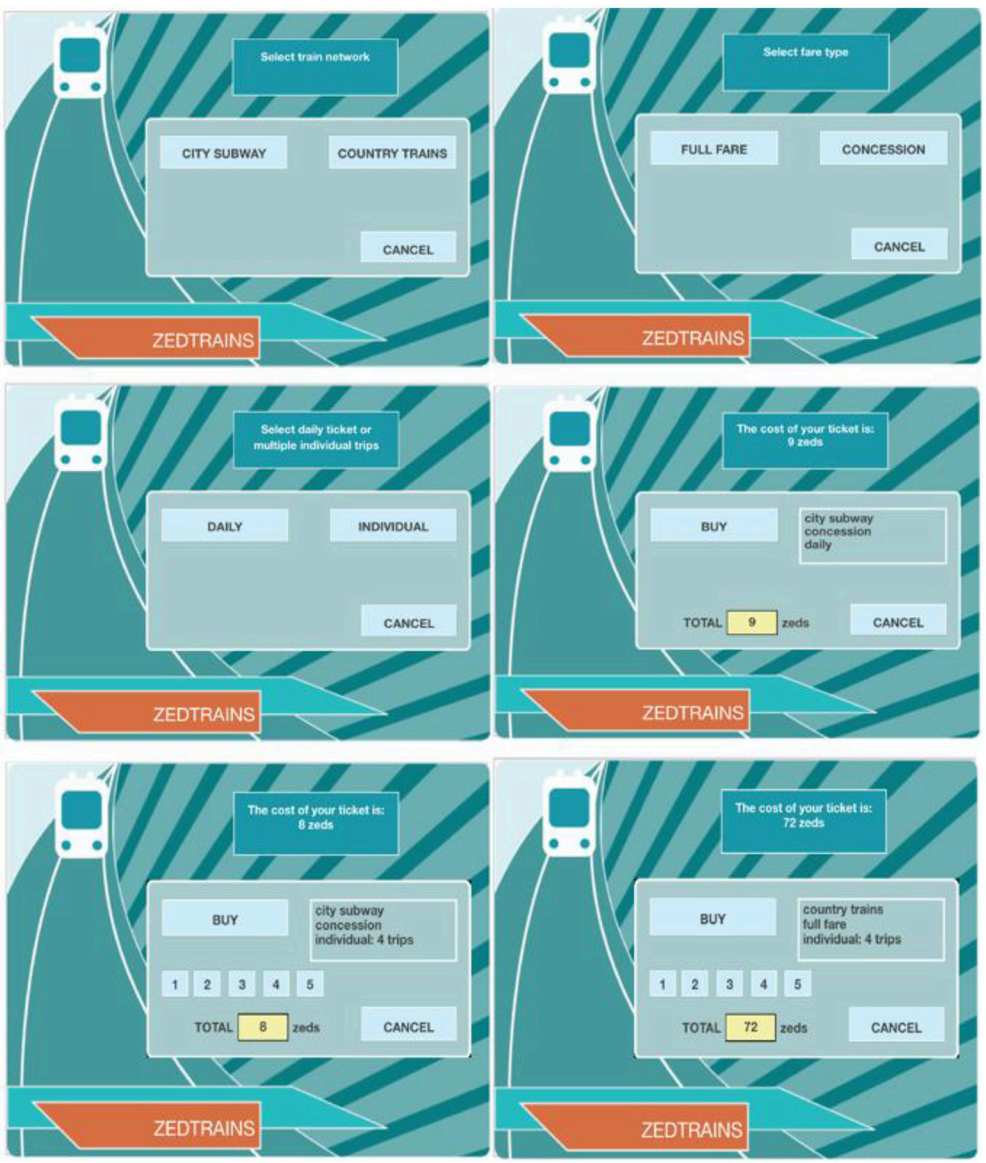


FIGURE 1 | PISA 2012 problem-solving question TICKETS task2 (CP038Q01) screenshots. (For more clear view, please see http://www.oecd.org/pisa/test-2012/testquestions/question5/).

database-cbapisa2012.htm. The dataset consists of 4722 actions from 426 students as rows and 11 variables as columns. Eleven variables (see **Figure 2**) include: *cnt* indicates country, which is USA in the present study; *schoolid* and *StIDStd* indicate the unique school and student IDs, respectively; *event_number* (ranging from 1 to 47) indicates the cumulative number of actions the student took; *event_value* (see raw event_values presented in **Table 1**) tells the specific action the student took at one time stamp and *time* indicates the exact time stamp (in seconds) corresponding to the *event_value*. Event notifies the nature of the action (start item, end item, or actions in process). Lastly, *network*, *fare_type*, *ticket_type*, and *number_trips* all describe the current choice the student had made. The variables used were *schoolid*, *StIDStd*, *event_value* and *time*. ID variables helped to identify students, while *event_value* and *time* variables were used to generate features. The scores for all students were not provided in the log file, thus, hand coded and carefully double checked based on the scoring rule. Among the 426 students, 121 (28.4%) got full credit, 224 (52.6%) got partial credit and 81 (19.0%) did not get any credit. Full, partial, and no credit were coded as 2, 1, and 0, respectively.

Feature Generation And Selection

Feature Generation

Features generated can be categorized into time features and action features, as summarized in **Table 1**. Four Time features were created: T_time, A_time, S_time, and E_time, indicating total response time, action time spent in process, starting time spent on first action, and ending time spent on last action, respectively. It was assumed that students with different ability levels may differ in the time they read the question (starting time spent on first action), the time they spent during the response (action time spent in process), and the time they used to make final decision (ending time spent on last action). Different researchers have proposed various joint modeling approaches for both response accuracy and response times, which explain the relationship between the two (e.g., van der Linden, 2007; Bolsinova et al., 2017). Thus, the total response times are expected to differ as well.

However, in this study, action features were created by coding different lengths of adjacent action sequences together. Thus, this study generated 12 action features consisting of only one action (unigrams), 18 action features containing two ordered adjacent actions (bigrams), and 2 action features created from four sequential actions (four-grams). Further, all action sequences generated were assumed to have equal importance and no weights were assigned to each action sequence. In **Table 1**, "concession" is a unigram, consisting of only one action, that is, the student bought the concession fare; on the other hand, "S_city" is a bigram, consisting of two actions, which are "Start" and "city subway," representing the student selected the city subway ticket after starting the item.

Sao Pedro et al. (2012) showed that features generated should be theoretically important to the construct to achieve better interpretability and efficiency. Following their suggestion, features were generated as the indicators of the problem-solving ability measured by this item, which is supported by the scoring rubric. For example, one action sequence consisted of four actions, which was coded as "city_con_daily_cancel," is crucial to scoring. If the student first chose "city_subway" to tour the city, then used the student's concession fare ("concession"), looked at the price of daily pass ("daily") next and lastly, he/she clicked "Cancel" to see the other option, this action sequence is necessary but not sufficient for a full credit.

The final recoded dataset for analysis is made up of 426 students as rows and 36 features (including 32 action sequence features and 4 time features) as columns. Scores for each student served as known labels when applying supervised learning methods. The frequency of each generated action feature was calculated for each student.

cnt	schoolid	StIDStd	event	time	event_number	event_value	network	fare_type	ticket_type	number_trips
USA	157	4846	START_ITEM	652.4	1	Start	NULL	NULL	NULL	NULL
USA	157	4846	ACER_EVENT	659.5	2	city_subway	city_subway	NULL	NULL	0
USA	157	4846	ACER_EVENT	660.2	3	concession	city_subway	concession	NULL	0
USA	157	4846	ACER_EVENT	675.5	4	daily	city_subway	concession	daily	0
USA	157	4846	ACER_EVENT	677.7	5	Cancel	NULL	NULL	NULL	0
USA	157	4846	ACER_EVENT	678.2	6	country_trains	country_trains	NULL	NULL	0
USA	157	4846	ACER_EVENT	679.8	7	Cancel	NULL	NULL	NULL	0
USA	157	4846	ACER_EVENT	680.4	8	city_subway	city_subway	NULL	NULL	0
USA	157	4846	ACER_EVENT	680.9	9	concession	city_subway	concession	NULL	0
USA	157	4846	ACER_EVENT	681.8	10	individual	city_subway	concession	individual	0
USA	157	4846	ACER_EVENT	682.8	11	trip_4	city_subway	concession	individual	4
USA	157	4846	ACER_EVENT	688.5	12	Buy	city_subway	concession	individual	4
USA	157	4846	END_ITEM	692.6	13	End	NULL	NULL	NULL	NULL

FIGURE 2 | The screenshot of the log file for one student.

TABLE 1 | 15 raw event values and 36 generated features.

Event_value (15)	Start, End, city_subway, concession, full_fare, daily, Cancel, country_trains, individual, Buy, trip_1, trip_2, trip_3, trip_4, trip_5
Time features (4)	T_time, A_time, S_time, E_time
Single actions (12)	All in raw event_values except for Start, End and Buy
Two actions coded together (18)	S_city (Start ⟶ city_subway) S_country (Start ⟶ country_trains) city_full (city_subway ⟶ full_fare) city_concession (city_subway ⟶ concession) country_full (country_trains ⟶ full_fare) country_concession (country_trains ⟶ concession) concession_daily (concession ⟶ daily) concession_individual (concession ⟶ individual) full_daily (full_fare ⟶ daily) full_individual (full_fare ⟶ individual) individual_trip4 (individual ⟶ trip_4) other_cancel (other ⟶ Cancel) daily_cancel (daily ⟶ Cancel) trip4_cancel (trip_4 ⟶ Cancel) daily_buy (daily ⟶ Buy) trip4_buy (trip_4 ⟶ Buy) individual_other (individual ⟶ other) other_buy (other ⟶ Buy)
Four actions coded together (2)	city_con_ind_4 (city_subway ⟶ concession ⟶ individual ⟶ trip_4) city_con_daily_cancel (city_subway⟶ concession ⟶ daily ⟶ Cancel)

Feature Selection

The selection of features should base on both theoretical framework and the algorithms used. As features were generated from a purely theoretical perspective in this study, no such consideration is needed in feature selection.

Two other issues that need consideration are redundant variables and variables with little variance. Tree-based methods handle these two issues well and have built-in mechanisms for feature selection. The feature importance indicated by tree-based methods are shown in **Figure 3**. In both random forest and gradient boosting, the most important one is "city_con_daily_cancel." The next important one is "other_buy," which means the student did not choose trip_4 before the action "Buy." The feature importance indicated by tree-based methods is especially helpful when selection has to be made among hundreds of features. It can help to narrow down the number of features to track, analyze, and interpret. The classification accuracy of the support vector machine (SVM) is reduced due to redundant variables. However, given the number of features (36) is relatively small in the current study, deleting highly correlated variables ($\rho \geq 0.8$) did not improve classification accuracy for SVM.

Clustering algorithms are affected by variables with near zero variance. Fossey (2017) and Kerr et al. (2011) discarded variables with 5 or fewer attempts in their studies. However, their data were binary and no clear-cut criterion exists for feature elimination when using cluster algorithms in the analysis of process data. In the current study, 5 features with variance no >0.09 in both training and test dataset were removed to achieve optimal classification results. Descriptive statistics for all 36 features can be found in **Table A1** in Appendix A.

In summary, a full set of features (36) were retained in the tree-based methods and SVM while 31 features were selected for SOM and *k*-means after the deletion of features with little variance.

Data Mining Techniques

This study demonstrates how to utilize data mining techniques to map the selected features (both action and time) to students' item performance on this problem-solving item in 2012 PISA. Given students' item scores are available in the data file, supervised learning algorithms can be trained to help classify students based on their known item performance (i.e., score category) in the training dataset while unsupervised learning algorithms categorize students into groups based on input variables without knowing their item performance. No assumptions about the data distribution are made on these data mining techniques.

Four supervised learning methods: Classification and Regression Tree (CART), gradient boosting, random forest, and SVM are explored to develop classifiers while, two unsupervised learning methods, Self-organizing Map (SOM) and *k*-means, are utilized to further examine different strategies used by students in both the same and different score categories. CART was chosen because it worked effectively in a previous study (DiCerbo and Kidwai, 2013) and is known for its quick computation and simple interpretation. However, it might not have the optimal performance compared with other methods. Furthermore, small changes in the data can change the tree structure dramatically (Kuhn, 2013). Thus, gradient boosting and random forest, which can improve the performance of trees via ensemble methods, were also used for comparison. Though SVM has not been used much in the analysis of process data yet, it has been applied as one of the most popular and flexible supervised learning techniques for other psychometric analysis such as automatic scoring (Vapnik, 1995). The two clustering algorithms, SOM and *k*-means, have been applied in the analysis of process data in log files (Stevens and Casillas, 2006; Fossey, 2017). Researchers have suggested to use more than one clustering methods to validate the clustering solutions (Xu et al., 2013). All the analyses were conducted in the software program Rstudio (RStudio Team, 2017).

Classifier Development

The general classifier building process for the supervised learning methods consists of three steps: (1) train the classifier through estimating model parameters; (2) determine the values of tuning parameters to avoid issues such as "overfitting" (i.e., the statistical model fits too closely to one dataset but fails to generalize to other datasets) and finalize the classifier; (3) calculate the accuracy of the classifier based on the test dataset. In general, training and tuning are often conducted based on the same training dataset. However, some studies may further split the training dataset into two parts, one for training while the other for tuning. Though tree-based methods are not affected by the scaling issue, training and test datasets are scaled for SVM, SOM, and *k*-means.

Given the relatively small sample size of the current dataset, training, and tuning processes were both conducted on the training dataset. Classification accuracy was evaluated with the test dataset. For the CART technique, the cost-complexity

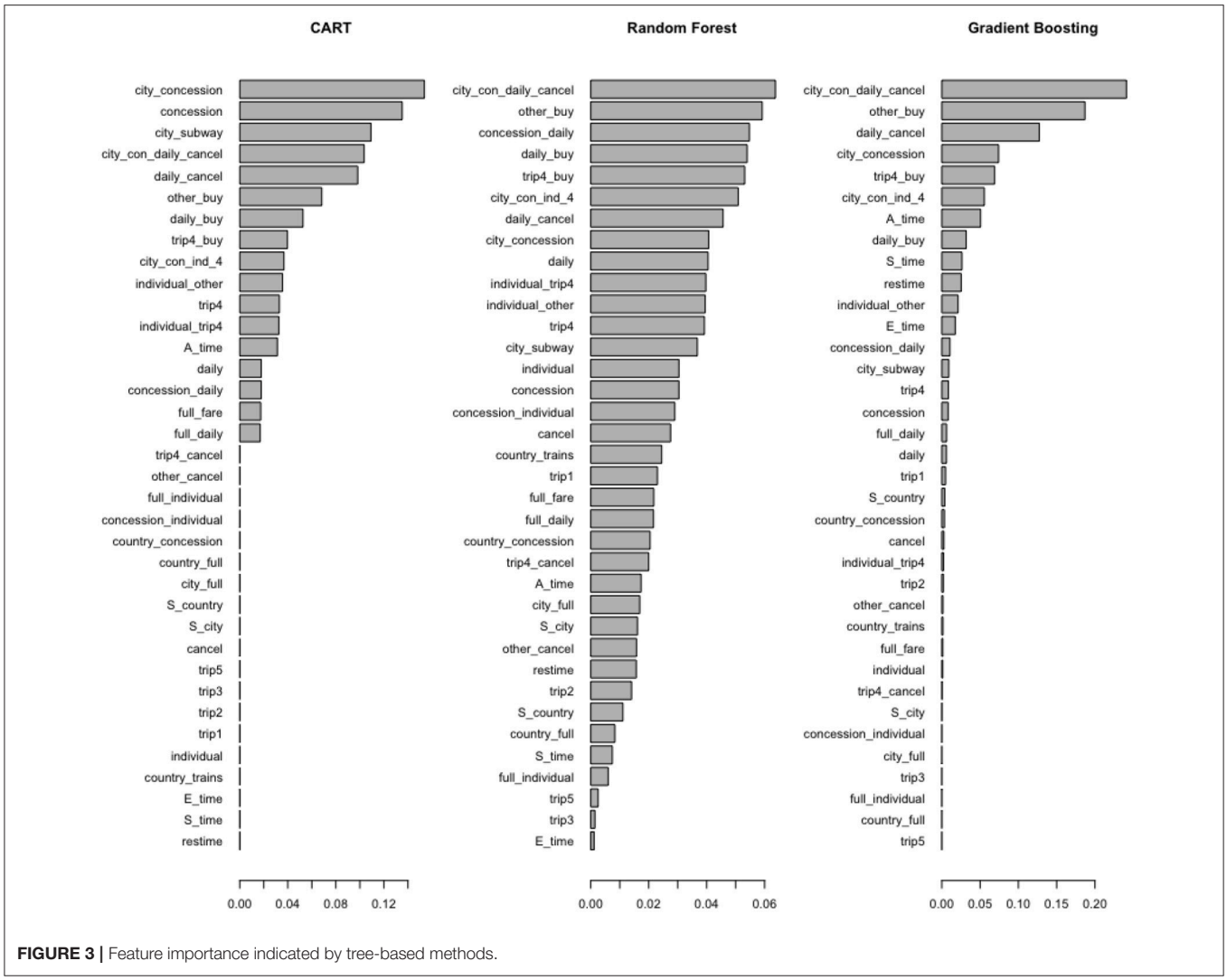


FIGURE 3 | Feature importance indicated by tree-based methods.

parameter (*cp*) was tuned to find the optimal tree depth using R package *rpart*. Gradient boosting was carried out using R package *gbm*. The tuning parameters for gradient boosting were the number of trees, the complexity of trees, the learning rate and the minimum number of observations in the tree's terminal nodes. Random forest was tuned over its number of predictors sampled for splitting at each node (m_{try}) using R package *randomForest*. A radial basis function kernel SVM, carried out in R package *kernlab*, was tuned through two parameters: scale function σ and the cost value C, which determine the complexity of the decision boundary. After the parameters were tuned, the classifiers were trained fitting to the training dataset. 10-fold-validation was conducted for supervised learning methods in the training processes. Cross-validation is not necessary for random forest when estimating test error due to its statistical properties (Sinharay, 2016).

For the unsupervised learning methods, SOM was carried out in the R package *kohonen*. Learning rate declined from 0.05 to 0.01 over the updates from 2000 iterations. *k*-means was carried out using the *kmeans* function in the *stats* R package with 2000 iterations. Euclidian distance was used as a distance measure for both methods. The number of clusters ranged from 3 to 10. The lower bound was set to be 3 due to the three score categories in this dataset. The upper bound was set to be 10 given the relative small number of features and small sample size in the current study. The R code for the usage of both supervised and unsupervised methods can be found in **Appendix B**.

Evaluation Criterion

For the supervised methods, students in the test dataset are classified based on the classifier developed based on the training dataset. The performance of supervised learning techniques was evaluated in terms of classification accuracy. Outcome measures include overall accuracy, balanced accuracy, sensitivity, specificity, and Kappa. Since item scores are three categories, 0, 1, and 2, sensitivity, specificity and balanced accuracy were calculated as follows.

$$Sensitivity = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}, \quad (1)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives\ +\ False\ Positives}, \quad (2)$$

$$Balanced\ Accuracy\ =\ \frac{Sensitivity\ +\ Specificity}{2} \quad (3)$$

where sensitivity measures the ability to predict positive cases, specificity measures the ability to predict negative cases and balanced accuracy is the average of the two. Overall accuracy and Kappa were calculated for each method based on the following formula:

$$Overall\ Accuracy\ =\ \frac{True\ Positives\ +\ True\ Negatives}{Total\ Cases} \quad (4)$$

$$Kappa\ =\ \frac{p_o - p_e}{1 - p_e} \quad (5)$$

where overall accuracy measures the proportion of all correct predictions. Kappa statistic is a measure of concordance for categorical data. In its formula, p_o is the observed proportion of agreement, p_e is the proportion of agreement expected by chance. The larger these five statistics are, the better classification decisions.

For the two unsupervised learning methods, the better fitting method and the number of clusters were determined for the training dataset by the following criteria:

1. Davies-Bouldin Index (DBI; Davies and Bouldin, 1979) calculated as in Equation 6, can be applied to compare the performance of multiple clustering algorithms (Fossey, 2017). The algorithm with the lower DBI is considered the better fitting one which has the higher between-cluster variance and smaller within-cluster variance.

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k} max_{i\neq j}\frac{S_i + S_j}{M_{ij}} \quad (6)$$

   where k is the number of clusters, S_i and S_j are the average distances from the cluster center to each case in cluster i and cluster j. M_{ij} is the distance between the centers of cluster i and cluster j. Cluster j has the smallest between-cluster distance with cluster i or has the highest within-cluster variance, or both (Davies and Bouldin, 1979).
2. Kappa value (see Equation 5) is a measure of classification consistency between these two unsupervised algorithms. It is usually expected not smaller than 0.8 (Landis and Koch, 1977).

To check the classification stability and consistency in the training dataset, the methods were repeated in the test dataset, DBI and Kappa values were computed.

RESULTS

The tuning and training results for the four supervised learning techniques are first reported and then the evaluation of their performance on the test datasets. Lastly, the results for the unsupervised learning methods are presented.

Supervised Learning Methods

The tuning processes for all the classifiers reached satisfactory results. For the CART, *cp* was set to 0.02 to achieve minimum error and the simplest tree structure (error < 0.2, number of trees < 6), as shown in **Figure 4**. The final tuning parameters for gradient boosting: the number of trees = 250, the depth of trees = 10, the learning rate = 0.01 and the minimum number of observations in the trees terminal nodes = 10. **Figure 5** shows that when the maximum tree depth equaled 10, the RMSE was minimum as iteration reached 250 with the simplest tree structure. The number of predictors sampled for splitting at each node (m_{try}) in the random forest was set to 4 to achieve the largest accuracy, as shown in **Figure 6**. In the SVM, the scale function σ was set to 1 and the cost value C set to 4 to reach the smallest training error 0.038.

The performance of the four supervised techniques was summarized in **Table 2**. All four methods performed satisfactorily, with almost all values larger than 0.90. The gradient boosting showed the best classification accuracy overall, exhibiting the highest Kappa and overall accuracy (Kappa = 0.94, overall accuracy = 0.96). Most of their subclass specificity and balanced accuracy values also ranked top, with only sensitivity

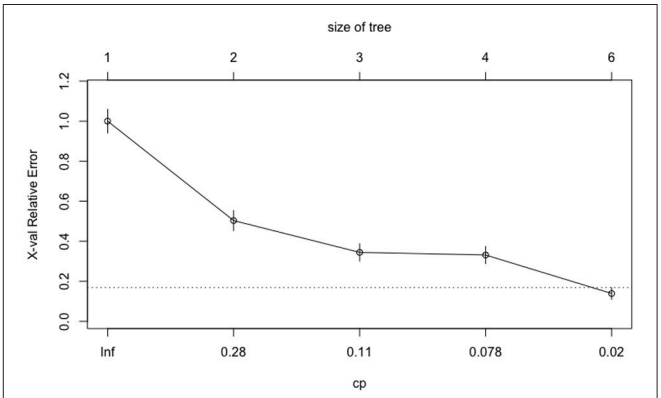


FIGURE 4 | The CART tuning results for cost-complexity parameter (*cp*).

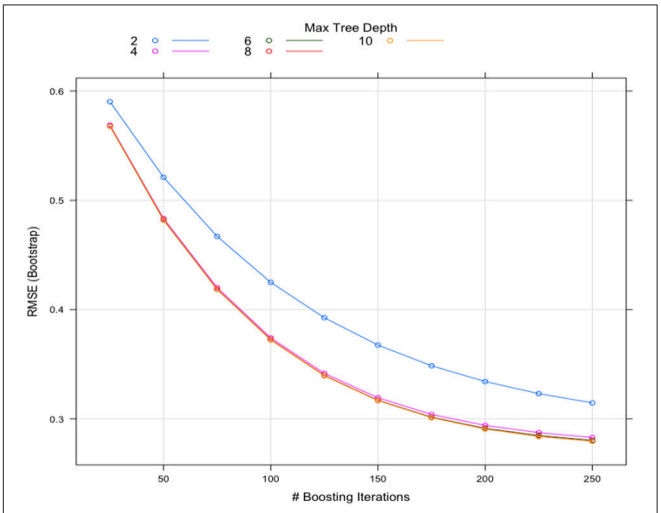


FIGURE 5 | The Gradient Boosting tuning results.

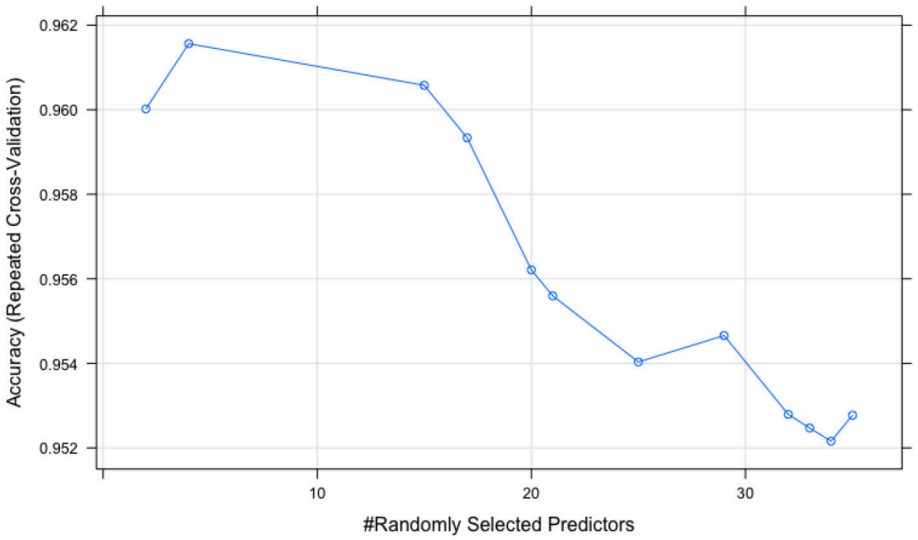


FIGURE 6 | The random forest tuning results (peak point corresponds to $m_{try} = 4$).

TABLE 2 | Average of accuracy measures of the scores.

Method	R package	Kappa	Overall accuracy	Sensitivity			Specificity			Balanced accuracy		
				0	1	2	0	1	2	0	1	2
CART	rpart	0.92	0.95	0.89	0.97	0.97	0.98	0.96	0.99	0.93	0.96	0.98
Random Forest	randomForest	0.92	0.95	0.89	0.95	10.0	0.99	0.96	0.97	0.94	0.95	0.99
Gradient Boosting	gbm	0.94	0.96	0.89	0.97	10.0	0.99	0.96	0.99	0.94	0.96	0.99
Support Vector Machine	kernlab	0.92	0.95	0.94	0.93	10.0	0.98	0.98	0.97	0.96	0.96	0.99

for score = 0, specificity for score = 1 and balanced accuracy for score = 0 smaller than those from SVM. SVM, random forest, and CART performed similarly well, all with a slightly smaller Kappa and overall accuracy values (Kappa = 0.92, overall accuracy = 0.95).

Among the four supervised methods, the single tree structure from CART built from the training dataset is the easiest to interpret and plotted in **Figure 7**. Three colors represent three score categories: red (no credit), gray (partial credit), and green (full credit). The darker the color is, the more confident the predicted score is in that node, the more precise the classification is. In each node, we can see three lines of numbers. The first line indicates the main score category in that node. The second line represents the proportions of each score category, in the order of scores of 0, 1, and 2. The third line is the percentage of students falling into that node. CART has a built-in characteristic to automatically choose useful features. As shown in **Figure 7**, only five nodes (features), "city_con_daily_cancel," "other_buy," "trip4_buy," "concession," and "daily_buy," were used in branching before the final stage. In each branch, if the student performs the action (>0.5), he/she is classified to the right, otherwise, to the left. As a result, students with a full credit were branched into one class, in which 96% truly belonged to this class and accounted for 29% of the total data points. Students who earned a partial credit were partitioned into two classes, one purely consisted of students in this group and the other consisted of 98% students who truly got partial credit. For the no credit group, students were classified into three classes, one purely consisted of students in this group and the other two classes included 10 and 18% students from other categories. One major benefit from this plot is that we can clearly tell the specific action sequences that led students into each class.

Unsupervised Learning Methods

As shown in **Table 3**, the candidates for the best clustering solution from the training dataset were *k*-means with 5 clusters (DBI = 0.19, kappa = 0.84) and SOM with 9 clusters (DBI = 0.25, kappa = 0.96), which satisfied the criterion of a smaller DBI value and kappa value $\geq$ 0.8. When validated with the test dataset, the DBI values for *k*-means and SOM all increased. It could be caused by the smaller sample size of the test dataset. Due to the low kappa value for the 5-cluster solution in the validation sample, the final decision on the clustering solution was SOM with 9 clusters. The percentage of students in each score category in each cluster is presented in **Figure 8**. The cluster analysis results obtained based on both SOM and *k*-means can be found in **Table A2** in Appendix A.

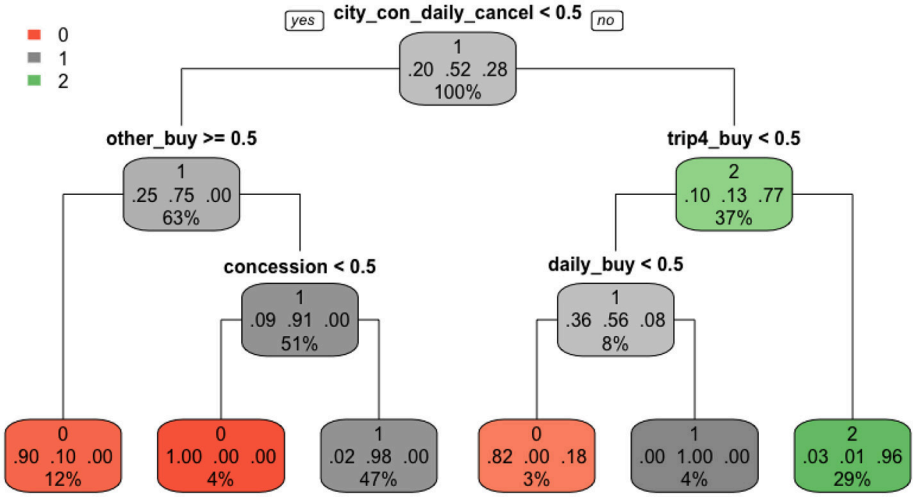


FIGURE 7 | The CART classification.

TABLE 3 | Clustering Algorithms' Fit (DBI) and Agreement (Cohen's Kappa).

Number of clusters	Training dataset (n = 320)			Test dataset (n = 106)		
	DBI		Kappa	DBI		Kappa
	k-means	SOM		*k*-means	SOM	
3	1.427	1.54	0.037	1.741	1.696	0.900
4	1.792	1.447	0.061	1.444	1.178	0.078
5	**0.188****	1.296	**0.843**	1.098	1.133	**0.320****
6	1.448	1.087	0.934	1.057	1.171	0.390
7	1.413	1.023	0.835	1.177	0.920	0.891
8	0.198	1.057	0.753	1.063	1.034	0.894
9	1.099	**0.249***	**0.959**	1.288	0.979	0.831
10	1.442	0.251	0.884	1.288	0.816	0.627

**Best fitting solution with the training dataset but lower Kappa value with the test dataset, indicating the disagreement between k-means and SOM.*
**Final chosen solution. Bold values indicate potential final clustering solution and are discussed in the text.*

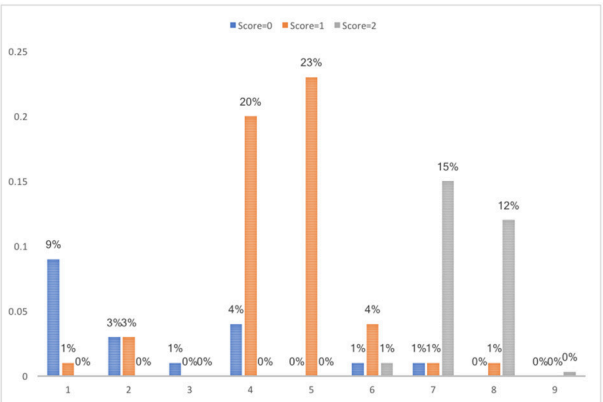


FIGURE 8 | Percentage in each score category in the final SOM clustering solution with 9 clusters from the training dataset.

To interpret, label and group the resulting clusters, it is necessary to examine and generalize the students' features and the strategy pattern in each of the cluster. In alignment with the scoring rubrics and ease of interpretation, the nine clusters identified in the training dataset are grouped into five classes and interpreted as follows.

1. Incorrect (cluster1): students bought neither individual tickets for 4 trips nor a daily ticket.
2. Partially correct (cluster 4–5): students bought either individual tickets for 4 trips or a daily ticket but did not compare the prices.
3. Correct (cluster 7 and 8): students did compare the prices between individual tickets and a daily ticket and chose to buy the cheaper one (individual tickets for 4 trips).
4. Unnecessary actions (cluster 2, 3, and 6): students tried options not required by the question, e.g., country train ticket, other number of individual ticket.
5. Outlier (cluster 9): the student made too many attempts and is identified as an outlier.

Such grouping and labeling can help researchers better understand the common strategies used by students in each score category. It also helps to identify errors students made and can be a good source of feedback to students. For those students mislabeled above, they share the major characteristics in the cluster. For example, 4% students who got no credit in cluster 4 in the training dataset bought daily ticket for the city subway without comparing the prices, but they bought the full fare instead of using student's concession fare. These students are different from those in cluster 1 who bought neither daily tickets nor individual tickets for 4 trips. Thus, students in the same score category were classified into different clusters, indicating

that they made different errors or took different actions during the problem-solving process. In summary, though students in the same score category generally share the actions they took, they can also follow distinct problem-solving processes. Students in different score categories can also share similar problem-solving process.

SUMMARY AND DISCUSSIONS

This study analyzed the process data in the log file from one of the 2012 PISA problem-solving items using data mining techniques. The data mining methods used, including CART, gradient boosting, random forest, SVM, SOM, and *k*-means, yielded satisfactory results with this dataset. The three major purposes of the current study were summarized as follows.

First, to demonstrate the analysis of process data using both supervised and unsupervised techniques, concrete steps in feature generation, feature selection, classifier development and outcome evaluation were presented in the current study. Among all steps, feature generation was the most crucial one because the quality of features determines the classification results to a large extent. Good features should be created based on a thorough understanding of the item scoring procedure and the construct. Key action sequences that can distinguish correct and incorrect answers served as features with good performance. Unexpectedly, time features, including total response time and its pieces, did not turn out to be important features for classification. This means that considerable variance of response time existed in each score group and the differences in response time distributions among the groups was not large enough to clearly distinguish the groups (see **Figure A1** in Appendix A). This study generated features based on theoretical beliefs about the construct measured and used students as the unit of analysis. The data could be structured in other ways according to different research questions. For example, instead of using students as the unit of analysis, the attempts students made can be used as rows and actions as columns, then the attempts can be classified instead of people. Fossey (2017) included a detailed tutorial on clustering algorithms with such data structure in a game-based assessment.

Second, to evaluate classification consistency of these frequently used data mining techniques, the current study compared four supervised techniques with different properties, namely, CART, gradient boosting, random forest, and SVM. All four methods achieved satisfactory classification accuracy based on various outcome measures, with gradient boosting showing slightly better overall accuracy and Kappa value. In general, easy interpretability and graphical visualization are the major advantages of trees. Trees also deal with noisy and incomplete data well (James et al., 2013). However, the trees are easily influenced by even small changes in the data due to its hierarchical splitting structure (Hastie et al., 2009). SVM, on the contrary, generalizes well because once the hyperplane is found, small changes to data cannot greatly affect the hyperplane (James et al., 2013). Given the specific dataset in the current study, even the CART method worked very well. In addition, the CART method can be easily understood and provided enough information about the detailed classifications between and within each score category. Thus, based on the results in the current study, the CART method is sufficient for future studies on similar datasets. Unsupervised learning algorithms, SOM and *k*-means, also showed convergent clustering results based on DBI and Kappa values. In the final clustering solution, students were grouped into 9 clusters, revealing specific problem-solving processes they went through.

Third, supervised and unsupervised learning methods serve to answer different research questions. Supervised learning methods can be used to train the algorithm to predict memberships in the future data, like automatic scoring. Unsupervised methods can reveal the problem-solving strategy patterns and further differentiate students in the same score category. This is especially helpful for formative purposes. Students can be provided with more detailed and individualized diagnostic reports. Teachers can better understand students' strengths and weaknesses, and adjust instructions in the classroom accordingly or provide more targeted tutoring to specific students. In addition, it is necessary to check any indication for cheating behavior in the misclassified or outlier cases from both types of data mining methods. For example, students answered the item correctly within an extremely short amount of time can imply item compromise.

This study has its own limitations. Other data mining methods, such as other decision trees algorithms and clustering algorithms, are worth of investigation. However, the procedure demonstrated in this study can be easily generalized to other algorithms. In addition, the six methods were compared based on the same set of data rather than data under various conditions. Therefore, the generalization of the current study is limited due to factors such as sample size and number of features. Future studies can use a larger sample size and extract more features from more complicated assessment scenarios. Lastly, the current study focuses on only one item for the didactic purpose. In the future study, process data for more items can be analyzed simultaneously to get a comprehensive picture of the students.

To sum up, the selection of data mining techniques for the analysis of process data in assessment depends on the purpose of the analysis and the data structure. Supervised and unsupervised techniques essentially serve different purposes for data mining with the former as a confirmatory approach while the latter as an exploratory approach.

AUTHOR CONTRIBUTIONS

XQ as the first author, conducted the major part of study design, data analysis and manuscript writing. HJ as the second author, participated in the formulation and refinement of the study design and provided crucial guidance in the statistical analysis and manuscript composition.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02231/full#supplementary-material

REFERENCES

Bolsinova, M., De Boeck, P., and Tijmstra, J. (2017). Modeling conditional dependence between response time and accuracy. *Psychometrika* 82, 1126–1148. doi: 10.1007/s11336-016-9537-6

Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., et al. (2010). "Testing instructional design variations in a prototype math game," in *Current Perspectives From Three National RandD Centers Focused on Game-based Learning: Issues in Learning, Instruction, Assessment, and Game Design*, eds R. Atkinson (Chair) (Denver, CO: Structured poster session at the annual meeting of the American Educational Research Association).

Davies, D. L., and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 224–227. doi: 10.1109/TPAMI.1979.4766909

DiCerbo, K. E., and Kidwai, K. (2013). "Detecting player goals from game log files," in *Poster presented at the Sixth International Conference on Educational Data Mining* (Memphis, TN).

DiCerbo, K. E., Liu, J., Rutstein, D. W., Choi, Y., and Behrens, J. T. (2011). "Visual analysis of sequential log data from complex performance assessments," in *Paper presented at the annual meeting of the American Educational Research Association* (New Orleans, LA).

Fossey, W. A. (2017). *An Evaluation of Clustering Algorithms for Modeling Game-Based Assessment Work Processes*. Unpublished doctoral dissertation, University of Maryland, College Park. Available online at: https://drum.lib.umd.edu/bitstream/handle/1903/20363/Fossey_umd_0117E_18587.pdf?sequence=1 (Accessed August 26, 2018).

Fu, J., Zapata-Rivera, D., and Mavronikolas, E. (2014). *Statistical Methods for Assessments in Simulations and Serious Games* (ETS Research Report Series No. RR-14-12). Princeton, NJ: Educational Testing Service.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., and Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *J. Educ. Data Min.* 4, 111–143. Available online at: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/24 (Accessed November 9, 2018)

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Hao, J., Smith, L., Mislevy, R. J., von Davier, A. A., and Bauer, M. (2016). *Taming Log Files From Game/Simulation-Based Assessments: Data Models and Data Analysis Tools* (ETS Research Report Series No. RR-16-10). Princeton, NJ: Educational Testing Service.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd edn.* New York, NY: Springer. doi: 10.1007/978-0-387-84858-7

Howard, L., Johnson, J., and Neitzel, C. (2010). "Examining learner control in a structured inquiry cycle using process mining." in *Proceedings of the 3rd International Conference on Educational Data Mining*, 71–80. Available online at: https://files.eric.ed.gov/fulltext/ED538834.pdf#page=83 (Accessed August 26, 2018).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning, Vol 112*. New York, NY: Springer.

Jeong, H., Biswas, G., Johnson, J., and Howard, L. (2010). "Analysis of productive learning behaviors in a structured inquiry cycle using hidden Markov models," in *Proceedings of the 3rd International Conference on Educational Data Mining*, 81–90. Available online at: http://educationaldatamining.org/EDM2010/uploads/proc/edm2010_submission_59.pdf (Accessed August 26, 2018).

Kerr, D., Chung, G., and Iseli, M. (2011). *The Feasibility of Using Cluster Analysis to Examine Log Data From Educational Video Games* (CRESST Report No. 790). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA. Available online at: https://files.eric.ed.gov/fulltext/ED520531.pdf (Accessed August 26, 2018).

Kohonen, T. (1997). *Self-Organizing Maps*. Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-97966-8

Kuhn, M. (2013). *Predictive Modeling With R and the Caret Package [PDF Document]*. Available online at: https://www.r-project.org/conferences/useR-2013/Tutorials/kuhn/user_caret_2up.pdf (Accessed November 9, 2018).

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Levy, R. (2014). *Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments* (CRESST Report No.837). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA. Available online at: https://files.eric.ed.gov/fulltext/ED555714.pdf (Accessed August 26, 2018).

Organisation for Economic Co-operation and Development (2014). *PISA 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-Life Problems, Vol. 5*. Paris: PISA, OECD Publishing.

RStudio Team (2017). *RStudio: Integrated development environment for R (Version 3.4.1) [Computer software]*. Available online at: http://www.rstudio.com/

Sao Pedro, M. A., Baker, R. S. J., and Gobert, J. D. (2012). "Improving construct validity yields better models of systematic inquiry, even with less information," in *User Modeling, Adaptation, and Personalization: Proceedings of the 20th UMAP Conference*, eds J. Masthoff, B. Mobasher, M. C. Desmarais, and R. Nkambou (Heidelberg: Springer-Verlag), 249–260. doi: 10.1007/978-3-642-31454-4_21

Shu, Z., Bergner, Y., Zhu, M., Hao, J., and von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychol. Test Assess. Model.* 59, 109–131. Available online at: https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2017_20170323/07_Shu.pdf (Accessed November 9, 2018)

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educ. Meas. Issues Pract.* 35, 38–54. doi: 10.1111/emip.12115

Soller, A., and Stevens, R. (2007). *Applications of Stochastic Analyses for Collaborative Learning and Cognitive Assessment (IDA Document D-3421)*. Arlington, VA: Institute for Defense Analysis.

Stevens, R. H., and Casillas, A. (2006). "Artificial neural networks," in *Automated Scoring of Complex Tasks in Computer-Based Testing*, eds D. M. Williamson, R. J. Mislevy, and I. I. Bejar (Mahwah, NJ: Lawrence Erlbaum Associates, Publishers), 259–312.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag. doi: 10.1007/978-1-4757-2440-0

Williamson, D. M., Mislevy, R. J., and Bejar, I. I. (2006). *Automated Scoring of Complex Tasks in Computer-Based Testing, eds*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers. doi: 10.4324/9780415963572

Xu, B., Recker, M., Qi, X., Flann, N., and Ye, L. (2013). Clustering educational digital library usage data: a comparison of latent class analysis and k-means algorithms. *J. Educ. Data Mining* 5, 38–68. Available online at: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/21 (Accessed November 9, 2018)

Zhu, M., Shu, Z., and von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *J. Educ. Meas.* 53, 190–211. doi: 10.1111/jedm.12107

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Qiao and Jiao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 03 August 2018
doi: 10.3389/fpsyg.2018.01372

Analysis of Process Data of PISA 2012 Computer-Based Problem Solving: Application of the Modified Multilevel Mixture IRT Model

Hongyun Liu[1,2], *Yue Liu*[1] *and Meijuan Li*[3,4]*

[1] *Faculty of Psychology, Beijing Normal University, Beijing, China,* [2] *Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, Beijing, China,* [3] *Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, Beijing, China,* [4] *Educational Supervision and Quality Assessment Research Center, Beijing Academy of Educational Sciences, Beijing, China*

Computer-based assessments provide new insights into cognitive processes related to task completion that cannot be easily observed using paper-based instruments. In particular, such new insights may be revealed by time-tamped actions, which are recorded as computer log-files in the assessments. These actions, nested in individual level, are logically interconnected. This interdependency can be modeled straightforwardly in a multi-level framework. This study draws on process data recorded in one of complex problem-solving tasks (Traffic CP007Q02) in Program for International Student Assessment (PISA) 2012 and proposes a modified Multilevel Mixture IRT model (MMixIRT) to explore the problem-solving strategies. It was found that the model can not only explore whether the latent classes differ in their response strategies at the process level, but provide ability estimates at both the process level and the student level. The two level abilities are different across latent classes, and they are related to operational variables such as the number of resets or clicks. The proposed method may allow for better exploration of students' specific strategies for solving a problem, and the strengths and weaknesses of the strategies. Such findings may be further used to design targeted instructional interventions.

Keywords: computer-based problem solving, PISA2012, process data, the modified multilevel mixture IRT model, the process level, the student level

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Yunxiao Chen,
Emory University, United States
Matthias Stadler,
University of Luxembourg,
Luxembourg

***Correspondence:**
Meijuan Li
lmj182@163.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *03 February 2018*
Accepted: *16 July 2018*
Published: *03 August 2018*

Citation:
Liu H, Liu Y and Li M (2018) Analysis of Process Data of PISA 2012 Computer-Based Problem Solving: Application of the Modified Multilevel Mixture IRT Model. Front. Psychol. 9:1372. doi: 10.3389/fpsyg.2018.01372

INTRODUCTION

The problem-solving competence is defined as the capacity to engage in cognitive processing to understand and resolve problem situations where a solution is not immediately obvious. It includes the willingness to engage in these situations in order to achieve one's potential as a constructive and reflective citizen (OECD, 2014; Kurniati and Annizar, 2017). Problem solving can be conceptualized as a sequential process where the problem solver must understand the problem, devise a plan, carry out the plan, and monitor the progress in relation to the goal (Garofalo and Lester, 1985; OECD, 2013). These problem-solving skills are key to success in all pursuits, and they can be developed in school through curricular subjects. Therefore, it is no surprise that the problem-solving competency is increasingly becoming the focus of many testing programs worldwide.

Advances in technology have expanded opportunities for educational measurement. Computer-based assessments, such as simulation-, scenario-, and game-based assessments, constantly change item design, item delivery, and data collection (DiCerbo and Behrens, 2012; Mislevy et al., 2014). These assessments usually provide an interactive environment in which students can solve a problem through choosing among a set of available actions and taking one or more steps to complete a task. All student actions are automatically recorded in system logs as coded and time-stamped strings (Kerr et al., 2011). These strings can be used for instant feedback to students, or for diagnostic and scoring purposes at a later time (DiCerbo and Behrens, 2012). And they are called process data. For example, the problem solving assessment of PISA 2012, which is computer-based, used simulated real-life problem situations, such as a malfunctioning electronic device, to analyze students' reasoning skills, problem-solving ability, and problem-solving strategies. The computer-based assessment of problem solving not only ascertains whether students produce correct responses for their items, but also records a large amount of process data on answering these items. These data make it possible to understand students' strategies to the solution. So far, to evaluate students' higher order thinking, more and more large-scale assessments of problem solving become computer-based.

Recent research has focused on characterizing and scoring process data and using them to measure individual student's abilities. Characterizing process data can be conducted via a variety of approaches, including visualization, clustering, and classification (Romero and Ventura, 2010). DiCerbo et al. (2011) used diagraphs to visualize and analyze sequential process data from assessments. Bergner et al. (2014) used cluster analysis to classify similar behaving groups. Some other researchers used decision trees, neural networks, and Bayesian belief networks (BBNs) (Romero et al., 2008; Desmarais and Baker, 2012; Zhu et al., 2016), to classify the performance of problem solvers (Zoanetti, 2010) and to predict their success (Romero et al., 2013). Compared to characterizing process data, the research of scoring process data is very limited. Hao et al. (2015) introduced "the editing distance" to score students' behavior sequences based on the process data in a scenario-based task of the National Assessment of Educational Progress (NAEP). Meanwhile, these process data have been used in psychometric studies. Researchers analyzed students' sequential response process data to estimate their ability by combining Markov model and item response theory (IRT) (Shu et al., 2017). It is noteworthy that all these practices have examined process data that describe students' sequential actions to solve a problem.

All the actions, recorded as process level data, which are nested in individual level, are logically interconnected. This interdependency allows a straightforward modeling in a multi-level framework (Goldstein, 1987; Raudenbush and Bryk, 2002; Hox, 2010). This framework is similar to those used in longitudinal studies, yet with some differences. In longitudinal studies, measurements are typically consistent to show the development pattern of certain traits. For process data, however, actions are typically different within each individual. These successive actions are used to characterizing individuals' problem solving strategies.

It is common in computer-based assessments that a nested data structure exists. To appropriately analyze process data (e.g., time series actions) within a nested structure (e.g., process within individuals), the multi-level IRT model can be modified by allowing process data to be a function of the latent traits at both process and individual levels. It is noteworthy that in the modified model, the concept of "item" in IRT changed to each action in individuals' responses, which was scored based on certain rules.

With respect to the assessment of problem solving competency, the focus of this study is the ability estimate at the student level. We were not concerned with individual's ability reflected from each action at the process level, since the task needs to be completed by taking series actions. Even for individuals with high problem solving ability, the first few actions may not accurately reflect test takers' ability. As a result, more attention was put on the development of ability at the process level because it can reveal students' problem solving strategies. Mixture item response theory (MixIRT) models have been used in describing important effects in assessment, including the differential use of response strategies (Mislevy and Verhelst, 1990; Rost, 1990; Bolt et al., 2001). The value of MixIRT models lies in that they provide a way of detecting different latent groups which are formed by the dimensionality arising directly from the process data. These groups are substantively useful because they reflect how and why students responded the way they did.

In this study, we incorporated the multilevel structure into a mixture IRT model and used the modified multilevel mixture IRT (MMixIRT) model to detect and compare the latent groups in the data that have differential problem solving strategies. The advantage of this approach is the usage of latent groups. Although they are not immediately observable, these latent groups, which are defined by certain shared response patterns, can help explain process-level performance about how members of one latent group differ from another. The approach proposed in this study was used to estimate abilities both at process and student levels, and classify students into different latent groups according to their response strategies.

The goal of this study is to illustrate steps involved in applying the modified MMixIRT model in a computer-based problem solving assessment then to further present and interpret the results. Specifically, this article focuses on (a) describing and demonstrating the modified MMixIRT model using a task of PISA 2012 problem-solving process data; (b) interpreting the different action patterns; (c) analyzing the correlation between characteristics of different strategies and task performance, as well as some other operational variables such as the number of resets or clicks. All the following analysis was based on one sample data set.

MEASUREMENT MATERIAL AND DATASET

Problem Solving Item and Log Data File

This study illustrates the use of the modified MMixIRT model in analyzing process data through one of the problem-solving

tasks in PISA 2012 (Traffic CP007Q02). The task is shown in **Figure 1**. In this task, students were given a map and the travel time on each route, and then they were asked to find the quickest route from Diamond to Einsten, which takes 31 min.

The data are from the task's log file (CBA_cp007q02_logs12_SPSS.SAV, data source: http://www.oecd.org/pisa/data/) (an example of log data file is shown in **Appendix 1**). The data file contains four variables associated with the process. The "event" variable refers to the type of event, which may be either system generated (start item, end item) or student generated (e.g., ACER_EVENT, Click, Dblclick). The "time" variable is the event time for this item, given in seconds since the beginning of the assessment, with all click and double-click events included. The "event_value" variable is recorded in two rows, as a click event involves selecting or de-selecting a route of the map. For example, in the eleventh row where the state of the entire map is given, 1 in the sequence means that the route was selected, and 0 means that it was not; the twelfth row records an event involving highlighting, or un-highlighting. A route of the map represents the same click event, and it is in the form "hit_segment name" (The notes on log file data can be downloaded from http://www.oecd.org/pisa/data/). All the "click" and "double-click" events represent that a student performs a click action that is not related to select a route. **Table 1** shows the label, the route and the correct state of the entire selected routes.

Sample

The study sample was drawn from PISA 2012 released dataset, consisting of a total of 413 students from 157 American schools who participated in the traffic problem-solving assessment (47.2% as females). The average age of students was 15.80 years ($SD = 0.29$ years), ranging from 15.33 to 16.33 years.

For the traffic item response, the total effective sample size under analysis was 406, after excluding seven incomplete responses. For the log file of the process record, there were 15,897 records in the final data file, and the average record number for each student was 39 ($SD = 33$), ranging from 1 to 183. The average response time was 672.64 s ($SD = 518.85$ s), ranging from 58.30 to 1995.20 s.

THE MODIFIED MMIXIRT MODEL FOR PROCESS DATA

Process-Level Data Coding

In this task log file, "ACER_EVENT" is associated with "click." However, in this study we only collected the information of ACER_EVENT and deleted the redundant click data. Then, we split and rearranged the data by routes, making each row represent a step in the process of individual students, and each column represent a route (0 for de-selecting, and 1 for selecting). **Table 2** shows part of the reorganized data file, indicating how individual student selected each route in each step. For example, the first line

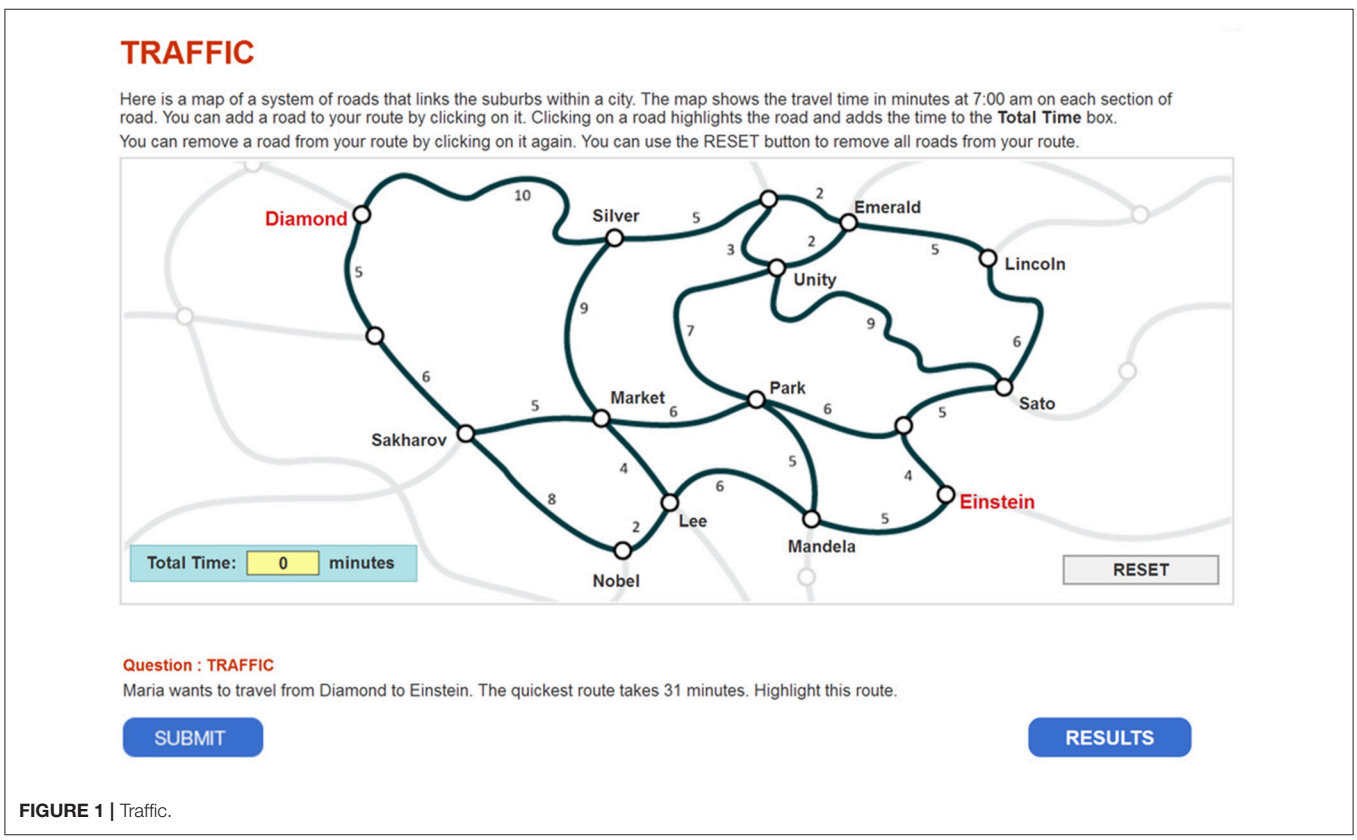


FIGURE 1 | Traffic.

TABLE 1 | The routes of the map.

Label	Route	Included or not in the correct routes
P1	Diamond-Nowhere	1
P2	Diamond-Silver	0
P3	Emerald-Lincoln	0
P4	Emerald-Unity	0
P5	Lee-Mandela	1
P6	Lincoln-Sato	0
P7	Mandela-Einstein	1
P8	Market-Lee	1
P9	Market-Park	0
P10	Nobel-Lee	0
P11	Nowhere-Einstein	0
P12	Nowhere-Emerald	0
P13	Nowhere-Sakharov	1
P14	Nowhere-Unity	0
P15	Park-Mandela	0
P16	Park-nowhere	0
P17	Sakharov-Market	1
P18	Sakharov-Nobel	0
P19	Sato-nowhere	0
P20	Silver-Market	0
P21	Silver-nowhere	0
P22	Unity-Park	0
P23	Unity-Sato	0

1, Yes; 0, No.

represents that student 00017 selected P2 in his/her first step.

Process data were first recoded for the analysis purpose. Twenty-three variables were created to represent a total number of available routes that can possibly be selected (similar to 23 items). The right way for solving this problem is to select the following six routes: Diamond–Nowhere–Sakharov–Market–Lee–Mandela–Einstein (i.e., P1, P5, P7, P8, P13, and P17). For the correct routes, the scored response was 1 if one was selected, and 0 otherwise; for the incorrect routes, the scored response was 0 if one was selected, and 1 otherwise. Each row in the data file represents an effective step (or action) a student took during the process. In each step, when a route was selected or not, the response for this route was recoded accordingly. When a student finished an item, all the steps during the process were recorded. Therefore, for the completed data set, the responses of the 23 variables were obtained and the steps were nested within students.

The Modified MMixIRT Model Specification

The MMixIRT model has mixtures of latent classes at the process level or at both process and student levels. It assumes that possible heterogeneity exists in response patterns at the process level and therefore are not to be ignored (Mislevy and Verhelst, 1990; Rost, 1990). Latent classes can capture the interactions among the responses at the process level (Vermunt, 2003). It is interesting to note that if no process-level latent classes exist, there are no student-level latent classes, either. The reason lies in that student-level units are clustered based on the likelihood of the processes belonging to one of the latent classes. For this particular consideration, the main focus in this study is to explore how to classify the process-level data, and the modified MMixIRT model only focus on latent classes at the process level.

The MMixIRT model accounts for the heterogeneity by incorporating categorical or continuous latent variables at different levels. Because mixture models have categorical latent variables and item response models have continuous latent variables, latent variables at each level may be categorical or continuous. In this study, the modified MMixIRT includes both categorical (latent class estimates) and continuous latent variables at the process level and only continuous (ability estimates) latent variables at the student level.

The modified MMixIRT model for process-level data is specified as follows:

Process-Level

$$\mathrm{P}\left(y_{jki}=1\middle|\theta_{jkg},C_{jk}=g\right)=\frac{\exp(\alpha_{ig.W}\theta_{jkg}-\beta_{ig})}{1+\exp(\alpha_{ig.W}\theta_{jkg}-\beta_{ig})} \quad (1)$$

$$P\left(y_{jk1}=\omega_1,y_{jk2}=\omega_2,\cdots,y_{jkI}=\omega_I\right)= \sum_{g=1}^{G}\gamma_{jkg}\prod_{i=1}^{I}P\left(y_{jki}=1\middle|\theta_{jkg},C_{jk}=g\right)^{\omega_i} (1-P\left(y_{jki}=1\middle|\theta_{jkg},C_{jk}=g\right))^{(1-\omega_i)} \quad (2)$$

Student-Level

$$\mathrm{P}\left(y_{ki}=1\middle|\theta_k\right)=\frac{\exp(\alpha_{i.B}\theta_k-\beta_i)}{1+\exp(\alpha_{i.B}\theta_k-\beta_i)} \quad (3)$$

For the process level, in Equation (1), $\boldsymbol{i}$ is an index for ith route ($i = 1, \ldots, I$), $\boldsymbol{k}$ is an index for a student ($k = 1,\ldots, K$), $\boldsymbol{j}$ is an index for the jth valid step of a student during the response process ($j = 1, \ldots, J_k$),(J is the total steps of the kth student) and g indexes the latent classes ($C_{jk} = 1, \ldots, g \ldots G$, where G is the number of latent classes), $\boldsymbol{C_{jk}}$ is a categorical latent variable at the process level for the jth valid step of student k, which captures the heterogeneity of the selections of routes in each step. $P\left(y_{jki}=1\middle|\theta_{jkg},C_{jk}=g\right)$ is the probability of selecting an route i in the jth step of student k, which is predicted by the two-parameter logistic (2PL) model, and $\boldsymbol{\alpha_{ig.W}}$ is the discrimination parameter of process-level in class g, W means within-level, $\boldsymbol{\beta_{ig}}$ is the location parameter in class g, and $\boldsymbol{\theta_{jkg}}$ is the latent ability of examinee $\boldsymbol{k}$ for a specific step $\boldsymbol{j}$ during the process of selecting the route, which is called the process ability in this study ($\boldsymbol{\theta_{jkg}} \sim \boldsymbol{N(\mu_{jkg}, \sigma^2_{jkg})}$). The process abilities across different latent classes are constrained to follow a normal distribution ***(θ_jk ∼N(0, 1)).*** In Equation (2), $P\left(y_{jk1}=\omega_1,y_{jk2}=\omega_2,\cdots,y_{jkI}=\omega_I\right)$ is the joint probability of the actions in the j^{th} step of student k. ω_i denotes either selected or not selected for ith route. For the correct routes, 1 represents that the route was selected, and 0

otherwise; for the incorrect routes, 0 represents that the route was selected, and 1 otherwise. γ_{jkg} is the proportion of the jth step in each latent class and $\sum_{g=1}^{G} \gamma_{jkg} = 1$. As can be seen from the Equation (2), the probability of the actions (y_{jki}) are assumed to be independent from each other given class membership, which is known as the local independence assumption for mixture models.

For the student level, in Equation (3), $\boldsymbol{\alpha_{i,B}}$ is the item discrimination parameter where $\boldsymbol{B}$ represents between-level. $\boldsymbol{\beta_i}$ is the item location parameter which is correlated with the responses of the final step of the item. $\boldsymbol{\theta_k}$ is the ability estimate at the student level based on the final step of the process, which also represents the problem-solving ability of student $\boldsymbol{k}$ in this study ($\boldsymbol{\theta_k \sim N(0, 1)}$).

Figure 2 demonstrates a modified two-level mixture item response model with within-level latent classes. The squares in the figure represent item responses, the ellipses represent latent variables, and 1 inside the triangle represents a vector of 1 s. As is shown in the figure, the response for each route of the jth step [$\boldsymbol{y_{jk1},\ldots, y_{jki},\ldots, y_{jkI}}$] is explained by both categorical and continuous latent variables ($\boldsymbol{C_{jk}}$ and $\boldsymbol{\theta_{jkg}}$, respectively) at the process level; and the final response of students for each route [$\boldsymbol{y_{k1},\ldots, y_{ki},\ldots, y_{kI}}$] is explained by a continuous latent variable ($\boldsymbol{\theta_k}$) at the student level. The arrows from the continuous latent variables to the item (route) represent item (route) discrimination parameters ($\boldsymbol{\alpha_{ig,W}}$ at the process level and $\boldsymbol{\alpha_{i,B}}$ at the student level), and the arrows from the triangle to the item responses represent item location parameters at both levels. The dotted arrows from the categorical latent variable to the other arrows indicate that all item parameters are class-specific.

It should be noted that the MMixIRT model is different from the traditional two-level mixture item response model in the definition of the latent variables at the between-level. In the standard MMixIRT model, the between-level latent variables are generally obtained from the measurement results made by within-level response variables [$\boldsymbol{y_{jk1},\ldots, y_{jki},\ldots, y_{jkI}}$] on between-level latent variables (Lee et al., 2017). In this study,

TABLE 2 | Example of the reorganized data file.

StIDStd	Time	Event_number	Event_value	P1	P2	P3	P4	P5	P6	P7	P8	...	P21	P22	P23
00017	837.6000	2.00	'01000000000000000000000	0	1	0	0	0	0	0	0	...	0	0	0
00017	839.8000	4.00	'11000000000000000000000	1	1	0	0	0	0	0	0	...	0	0	0
00017	841.1000	7.00	'11000000000010000000000	1	1	0	0	0	0	0	0	...	0	0	0
00017	841.7000	9.00	'11000000000010000100000	1	1	0	0	0	0	0	0	...	0	0	0
00017	842.7000	11.00	'11000000010010000100000	1	1	0	0	0	0	0	0	...	0	0	0
00017	844.8000	13.00	'11000000010010000101000	1	1	0	0	0	0	0	0	...	0	0	0
00017	846.4000	15.00	'11000000010000000101000	1	1	0	0	0	0	0	0	...	0	0	0
00017	847.4000	17.00	'01000000010000000101000	0	1	0	0	0	0	0	0	...	0	0	0
00017	848.4000	19.00	'01000000010000000001000	0	1	0	0	0	0	0	0	...	0	0	0
00017	850.6000	21.00	'01000000000000000001000	0	1	0	0	0	0	0	0	...	0	0	0
00017	851.6000	23.00	'01000000010000000001000	0	1	0	0	0	0	0	0	...	0	0	0
00017	852.5000	25.00	'01000000000000000001000	0	1	0	0	0	0	0	0	...	0	0	0
00017	853.4000	27.00	'01000000100000000001000	0	1	0	0	0	0	0	0	...	0	0	0
00017	853.7000	29.00	'01000000100000010001000	0	1	0	0	0	0	0	0	...	0	0	0

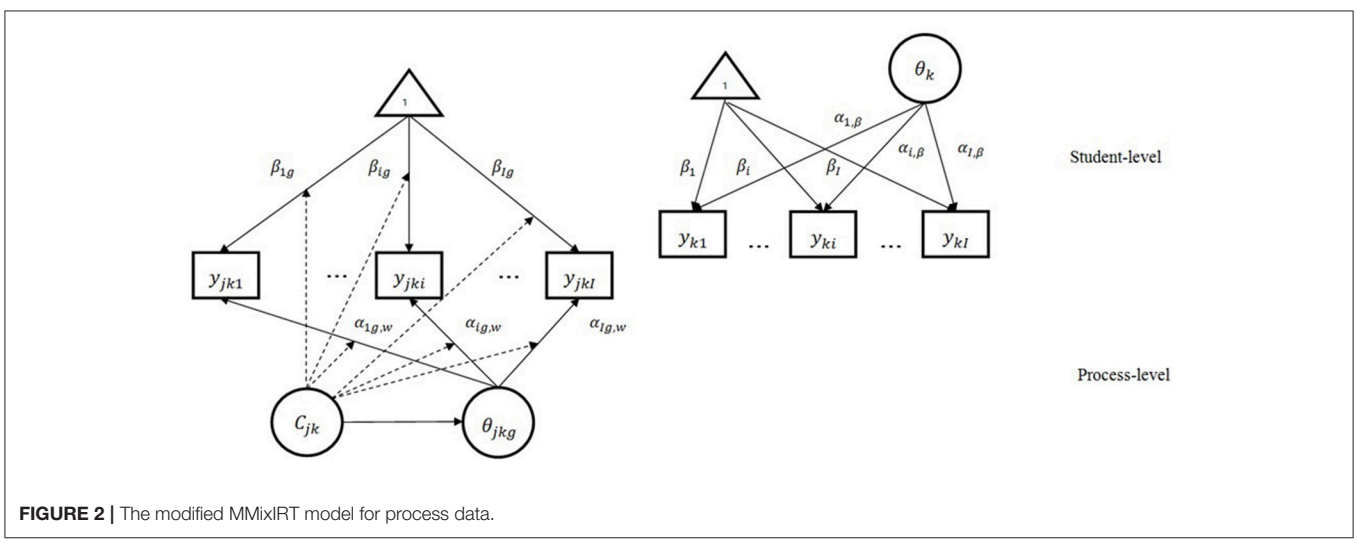


FIGURE 2 | The modified MMixIRT model for process data.

the process-level data mainly reflect the strategies for problem solving, while the responses at the last step represent students' final answers on this task. Therefore, students' final responses are used to estimate their problem-solving abilities (latent variable at the between-level, i.e., ability of the student level) in the modified MMixIRT model.

Mplus Software (Muthén and Muthén, 1998-2015) was used to estimate the parameters of the modified MMixIRT model, as specified above. In addition, the detailed syntax are presented in **Appendix 5**.

RESULTS

Results of Descriptive Statistics

Table 3 shows the proportion of each route selected by the students in the correct group and in the wrong group, respectively. The correct group consists of students who selected the right routes, and the wrong group refers to students who failed to do so. There are a total of 476 students, with 377 in the correct group and 99 in the wrong group. The results show that most of the students in the correct group selected the right routes, while a large number of students in the wrong group selected the wrong routes. To further explore the differences of the proportion of students selecting the wrong routes in the two groups, χ^2-tests were conducted. No significant differences were found between the correct group and the wrong group in terms of the proportion of students who clicked four wrong routes, including P4 [$\chi^2_{(1)} = 0.370$, $P > 0.05$], P9 [$\chi^2_{(1)} = 3.199$, $P > 0.05$], P10 [$\chi^2_{(1)} = 3.636$, $P > 0.05$], and P15 [$\chi^2_{(1)} = 2.282$, $P > 0.05$]. This further suggests that it was difficult for the correct group to avoid these routes during their response process, and even quite a number of students in the correct group experienced trial and error before eventually solving the problem.

TABLE 3 | The proportion of route selection.

Route	Selected proportion	
	Wrong group	Correct group
P1	**40.023**	**69.504**
P2	38.158	19.872
P3	3.290	1.688
P4	0.635	0.815
P5	**16.055**	**25.148**
P6	2.481	1.287
P7	**15.699**	**22.260**
P8	**4.340**	**21.953**
P9	25.379	23.435
P10	12.586	12.007
P11	16.559	10.819
P12	4.304	2.601
P13	**36.846**	**64.109**
P14	8.404	3.622
P15	5.182	6.886
P16	19.122	12.771
P17	**16.653**	**43.530**
P18	17.629	13.157
P19	4.884	1.923
P20	17.579	10.732
P21	15.369	7.211
P22	5.531	1.759
P23	4.296	1.377

The right routes are printed in bold.

Results of the Modified MMixIRT Model

Model Selection

The determination of the number of latent classes has been discussed in many studies (Tofighi and Enders, 2008; Li et al., 2009; Peugh and Fan, 2012). Several statistics of the mixture IRT models are often computed to compare relative fits of these models. Akaike's (1974) information criterion (AIC) incorporates a kind of penalty function for over-parameterization on model complexity. A criticism of AIC has been that it is not asymptotically consistent because the sample size is not directly involved in its calculation (Janssen and De Boeck, 1999; Forster, 2004). Schwarz (1978) proposed BIC as another information-based index, which attains asymptotic consistency by penalizing over-parameterization by using a logarithmic function of the sample size. For the sample size in BIC, the number of persons is used in multilevel model (Hamaker Ellen et al., 2011) and in multilevel item response model (Cohen and Cho, 2016). Most studies suggested the BIC value as the best choice because it was a sample-based index that also penalized the sophisticated model. However, Tofighi and Enders (2008) indicated in their simulation study that a sample size-adjusted BIC (aBIC) was an even better index. Smaller AIC, BIC, and aBIC values indicate a better model fit for mixture IRT models. Besides, entropy value has been used to measure how well a mixture model separates the classes; an entropy value close to 1 indicates good classification certainty (Asparouhov and Muthén, 2014).

The model selection results for the modified MMixIRT models are given in **Table 4**. The model fit indicates that *LL*, *AIC*, *BIC*, and *aBIC* decreased consistently as the class number increased to eight classes, and the nine-class model did not converge. As noted above, the best fit for *AIC*, *BIC*, and *aBIC* was determined or dictated by the smallest value in the ordered set of models from the least to the most complex. As suggested by Rosato and Baer (2012), selecting a robust latent class model is a balance between the statistical result of the model fit and the substantive meaning of the model. The model that fits best and yields meaningful classes should be retained. In this study the proportions of latent classes were examined to ensure the empirical significance, and the interpretability of each class was considered accordingly. For the 6-class model, the proportion of each class was 18.1, 30.7, 18.1, 20.1, 7.2, and 5.9%. And for the 7-class model, the proportion was 19.9, 13.4, 6.0, 12.3, 13.5, 27.4, and 7.5%. Compared to the 6-class model, in the 7-class model, the extra class of the steps was similar to class 2 of the 6-class model, while mixing class 4 at the same time. This makes the 7-class model hard to interpret. For the 8-class model, the proportion of one of the classes was too small (only 2.7%). Taking into account

both the model fit index and the interpretability of each class, the 6-class model was retained in this study.

Description of Class Characteristics

The most likely latent class membership are displayed in **Table 5**. In this matrix, steps from each class have an average probability of being in each class. Large probabilities are expected on the diagonal. The numbers on diagonal are greater than 0.9. It can be concluded from the results that the modified MMixIRT model can classify students properly based on process data.

Figure 3 presents the characteristics of route selection for each class based on the 6-class mixture IRT model, with ①, ②, ③.... indicating the order of the routes. Based on the results of the modified MMixIRT model, the number of clicks of the 23 routes (P1–P23) in each class is listed in **Appendix 2**. The characteristics of route selection can be obtained pursuant to routes that get more clicks than others in each class, as well as the relations among routes shown in **Figure 1**. For example, P17, P13, P1, P8, P5, P16, and P7 in Class 1 were clicked more than other routes; however, **Figure 1** shows that there is no obvious relationship between P16 and other routes. Therefore, the characteristic of Class 1 was defined as P1-P13-P17-P8-P5-P7 and P16 was removed. These routes were sequenced by the number of clicks they got, with the most clicked routes taking the lead. As indicated in **Figure 3**, different latent classes have typical characteristics depending on the similarity of the correct answers. For example, the route selection strategy of Class 1 best approximated the ideal route required by the item. Based on their last click, almost all the students in Class 1 gave the correct answer. Therefore, Class 1 could be regarded as the correct answer class, while the rest classes took different wrong routes.

The numbers in circles (①, ②, ③....) indicate the order of the routes.

As is illustrated in **Table 6**, different classes demonstrated different means of process-level ability. It is obvious that the mean process ability in Class 1 is the highest (0.493), followed by Class 6, Class 2, Class 4, yet Class 5 and Class 3 with the lowest process-level ability. A closer check of these classes in **Figure 3** indicates that the selected routes of Class 5 and Class 3 were incredibly far away from the correct one, and they took far more than 31 min. Therefore, it is no surprise that the mean process-level ability estimates of these two classes were the lowest and were both negative (−1.438 and −0.935, respectively). In addition, as can be seen in the number of students, almost all the students in Class 1 provided the right answer, demonstrating that different latent classes had different probabilities of the correct answer. In summary, the process-level ability is different across latent classes, which is related to different strategies of students' route selection or cognitive process.

The Sequence of Latent Classes at the Process Level

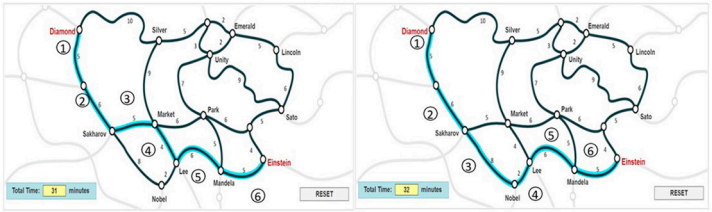
Based on the results of the modified MMixIRT model, the characteristics of the strategy shifts between step-specific classes were explored and summarized. To capture the characteristics of students' strategy shifts during the response, it is necessary to identify the typical route selection strategy of each class in the first place. In this study, if a student applied the strategy of a certain class three or more times consecutively, it was considered that the student had employed the strategy of this class at the process level. Three times was chosen as the rule of thumb because it demonstrated enough stability to classify a solution behavior. Then the strategy shifts of each student during their clicking procedure could be obtained in orders. The typical route selection strategy of different classes and the class shifts of students in the correct group are presented in **Appendixes 3, 4**, respectively. The results in **Appendix 4** provide useful and specific information about the strategy shifts used by students over time. For example, in the correct group, 58 students shifted from one class to another, including 22 from Class 2 to Class 1, 3 from Class 3 to Class 1, 30 from Class

TABLE 5 | Most likely latent class membership of each class.

	Most likely latent class membership					
	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**
Class 1	0.945	0.000	0.006	0.033	0.004	0.012
Class 2	0.001	0.936	0.002	0.033	0.013	0.015
Class 3	0.002	0.020	0.949	0.011	0.017	0.001
Class 4	0.029	0.004	0.007	0.949	0.002	0.010
Class 5	0.002	0.007	0.018	0.002	0.969	0.002
Class 6	0.016	0.014	0.001	0.025	0.002	0.942

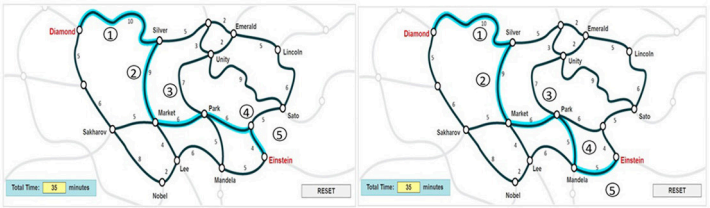
TABLE 4 | Model comparison and selection.

No of class	**No of Free parameters**	***LL Value***	***Akaike (AIC)***	***Bayesian (BIC)***	***Sample-Size Adjusted BIC***	***Entropy***
1	46	−112745.581	225583.161	225936.108	225789.923	
2	95	−99334.232	198858.463	199587.375	199285.472	0.957
3	144	−92723.338	185734.676	186839.552	186381.931	0.860
4	193	−89375.035	179134.070	180607.239	179997.077	0.920
5	242	−87186.912	174857.823	176714.629	175945.571	0.936
6	291	−85974.117	172530.234	174763.005	173838.228	0.908
7	340	−84864.882	170409.764	173018.500	171938.004	0.904
8	389	−83821.533	168421.066	171405.766	170169.552	0.893

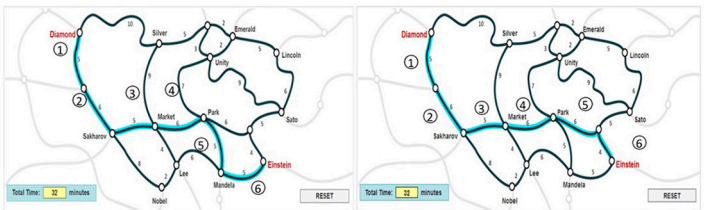


Class 1

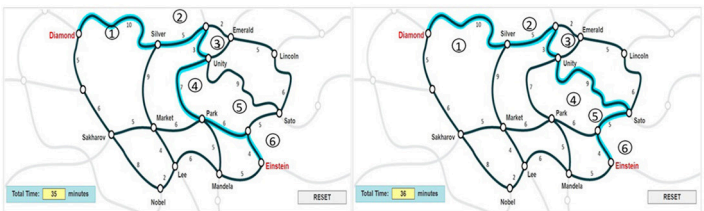
Class 2 & Class 6



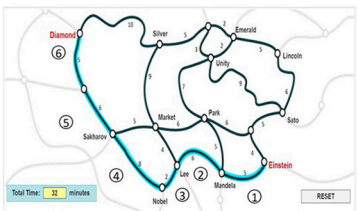
Class 3



Class 4



Class 5



Class 6

FIGURE 3 | Route selection strategy by class.

4 to Class 0, and 3 from Class 6 to Class 1. It is noteworthy that when students did not apply any strategies for more than three times consecutively, it was regarded as class 0 in this study.

The Relationship of the Two Level Ability Estimates and Operational Variables

To validate whether students with different patterns of actions will have different process-level ability, the descriptive statistics

TABLE 6 | Means and standard deviations of process level abilities.

	Latent class size for process-level		No of Students		Process-level ability	
	Count	Proportion	Right	Wrong	Mean	*SD*
Class 1	2875	18.1	307	3	0.493	0.678
Class 2	4867	30.7	0	41	0.323	0.903
Class 3	2867	18.1	0	14	−0.935	0.386
Class 4	3192	20.1	0	26	0.292	0.556
Class 5	1138	7.2	0	12	−1.438	0.404
Class 6	940	5.9	0	3	0.424	0.698
Total	15879	100	307	99	0.000	0.934

In the column of no of Students, the last step of the process within each student is classified into one of the six latent classes. Then, the numbers of students who gave the correct or wrong answer are summarized based on the latent classes.

were conducted of operational variables such as the number of route clicks and resets and their correlation with the mean ability estimate of process-level ability (See **Table 7** for details). To further explore the differences of click actions between the correct group and the wrong group, several *T*-tests were conducted. The results indicate that students in the correct group did significantly fewer resets than their counterparts in the wrong group [$t_{(404)} = 2.310$, $P < 0.05$]. No significant differences were detected of the number of routes clicked or the response time between the correct group and the wrong group [$t_{(404)} = 1.656$, $P = 0.099$; $t_{(404)} = -0.199$, $P = 0.843$]. The results in **Table 7** suggest two things. Firstly, positive correlation existed between the estimate of student-level ability and that of process-level ability. This means that the process-level ability estimate provides consistency and auxiliary diagnostic information about the process. The students with higher process-level ability had higher ability estimates of student level. Secondly, for the process-level ability, a significant negative correlation existed between the mean process-level ability estimate and variables such as the valid number of route clicks and the number of resets for students in the correct group. It is concluded that in the correct group, the less frequently a student clicks the routes and resets the whole process, the higher process-level ability he or she is likely to obtain. For students in the wrong group, however, no significant correlations were observed between the mean ability estimate and the variables discussed above. Instead, a significant negative correlation was found between the mean process-level ability estimate and the absolute time of difference from 31 min. For these students, their process-level ability decreased as the time cost by the wrong routes increased. Third, the mean process-level ability estimate for the correct group was 0.310, in contrast to −0.175 for the wrong group, which reveals a significant difference between the two groups [$t_{(404)} = 8.959$, $P < 0.001$]. In terms of student-level ability, the estimate for the correct group was significantly higher than for the wrong group [$t_{(404)} = 112.83$, $P < 0.001$].

The result in **Table 8** indicates that the sequence of latent classes are consistent with the ability estimates at both process and student levels. For students in the correct group, the mean process-level ability estimate decreased as the number of class shifts, clicks and resets increased. Students with higher process-level ability tended to select the correct route immediately or after a few attempts. Consequently, these students clicked and reset for fewer times because they had a clearer answer in mind and therefore were more certain about it. In contrast, for students in the wrong group, the mean ability estimates at both process and student levels were rather small when the number of class shifts were 0 and 1. When the number of class shifts was 0, students failed to stick with a specific strategy to solve the problem during the process. It took them a longer response time with about two resets on average; as a result, the time cost for their route selection was nearly twice the target time. When the number of class shifts was 1, these students simply stuck to a totally wrong route for the entire time, with shorter response time and fewer numbers of clicks. However, unlike the correct group, the number of class shifts in the wrong group showed a non-linear relationship with the mean ability at both process and student levels. At first, when the number of class shifts increased from 0 to 4, the ability estimates at both levels increased as well. The explanation was that because these students figured out the right routes, they should have higher abilities than the 0 shift group that sticks to the wrong route all the time. For example, students with four shifts all ended up using strategy of Class 1, which was the right strategy class (**Appendix 4**). Therefore, they were supposed to have the highest process ability in the wrong group. However, when the number of class shifts increased from 5 to 6, the process-level ability estimate dropped. This has much to do with the fact that too many shifts reflected little consideration and a lack of deep cognitive processing.

DISCUSSION

A modified MMixIRT model was described for modeling response data at process and student levels. The model developed in this study combined the features of an IRT model, a latent class model, and a multilevel model. The process-level data provide an opportunity to determine whether latent classes or class shifts differ in their response strategies to solve the problem. The student-level data can be used to account for the differences of students' problem solving abilities. The ability estimate at both process and student levels are different across latent classes. The modified MMixIRT model makes it possible to describe differential strategies based on process-level and student-level characteristics. If a student's specific strategies and their strengths and weaknesses can be described in the process of solving a problem, then the assessment of a student's proficiency in problem solving can guide instructional interventions in target areas.

As process data from various computer-based assessment or educational learning system have become common, there is an urgent call for analyzing such data in an accurate way. The psychometrical model-based approach has a great potential in this aspect. Latent classes and the characteristics of latent class shifts obtained from process data can reveal students' reasoning skills in problem-solving. The findings of characteristics of process-level latent classes make it easy to uncover meaningful

TABLE 7 | Correlation between ability estimates and operational variables in process.

Item response result	Click action variable	Mean ability of process level	Ability of student level	Mean	*SD*
Correct ($N = 307$)	No of Route Clicks	−0.657**	/	79.760	63.874
	No of Resets	−0.467**	/	0.919	1.737
	Absolute Time of Difference from 31	/	/	0.000	0.000
	Response Time	0.048	/	675.540	525.710
	Mean Ability of Process Level	/	/	0.310	0.447
	Ability of Student Level	/	/	1.371	0.000
Wrong ($N = 99$)	No of Route Clicks	−0.050	0.142	93.030	84.138
	No of Resets	−0.124	0.098	1.394	1.910
	Absolute Time of Difference from 31	−0.248*	−0.179	5.210	10.869
	Response Time	−0.087	0.022	663.620	499.466
	Mean Ability of Process Level	/	0.597**	−0.175	0.530
	Ability of Student Level	0.597**	/	−0.432	0.281
Total ($N = 406$)	No of Route Clicks	−0.439**	−0.066	83.000	69.484
	No of Resets	−0.378**	−0.103*	1.035	1.790
	Absolute Value of Difference from 31	−0.269**	−0.407***	1.300	5.802
	Response Time	0.015	0.012	672.640	518.849
	Mean Ability of Process Level	/	0.454***	0.192	0.512
	Ability of Student Level	0.454**	/	0.931	0.787
	Correct Responses	0.407**	0.985***	0.756	0.430

*"31" indicated in "absolute value of difference from 31" in Column 8 refers to the time taken in walking the right route for the item CP007Q02. *p < 0.05, **p < 0.01, ***p < 0.001.*

TABLE 8 | Ability estimates and the operational variables in the different numbers of class shifts in the correct group and wrong group.

Correct or wrong answer group	No of class shifts	No of students	Process-level ability (Mean)	Student-level ability (Mean)	Response time (Mean)	Valid No of click (Mean)	Absolute value of difference from 31 (Mean)	No of Reset (Mean)
Correct group ($N = 307$)	1	32	0.650	1.371	714.941	19.375	0	0.156
	2	58	0.692	1.371	609.116	31.655	0	0.121
	3	69	0.468	1.371	814.619	60.667	0	0.275
	4	73	0.196	1.371	601.215	93.795	0	1.192
	5	63	−0.141	1.371	649.711	134.143	0	1.937
	6	12	−0.279	1.371	679.617	212.25	0	3.5
Wrong group ($N = 99$)	0	11	−0.453	−0.548	991.7	36.909	29.182	2.091
	1	15	−0.439	−0.552	377.713	22.867	1.067	1.067
	2	12	0.139	−0.312	470.392	37.75	1.417	0.5
	3	12	0.466	−0.275	552.042	71.917	0.917	0.667
	4	20	−0.151	−0.438	784.455	94.4	1.250	0.85
	5	24	−0.343	−0.492	690.038	170.292	5.042	2.292
	6	5	−0.348	−0.162	921.02	234	1.000	2.6

and interesting action patterns from the process data, and to compare patterns from different students. These findings provide valuable information to psychometricians and test developers, help them better understand what distinguishes successful students from unsuccessful ones, and eventually lead to better test design. In addition, as shown in this study, some operational variables such as the number of resets and the number of clicks or double clicks are related to the ability estimates at both process and student levels and therefore can predict student scores on problem solving assessment. Since students' different abilities capture individual patterns in process data, it can be used to score or validate the rubrics. Williamson et al. (2006) explain that a "key to leveraging the expanded capability to collect and record data from complex assessment tasks is implementing automated scoring algorithms to interpret data of the quantity and complexity that can now be collected" (p. 2).

The extension of the modified MMixIRT approach proposed in this study can be implemented in several ways. Firstly, it can be simplified in removing the process-level ability parameters, and also be extended to include student-level latent classes instead of abilities. Secondly, one of the advantages of this proposed model is that item parameters can be constrained to be equal across the process-level and student-level. So the abilities of both levels are on the same scale and can be compared and evaluated. Lastly, the main benefits of multilevel IRT modeling lie in the possibility of estimating the latent traits (e.g., problem solving) at each level. More measurement errors can be accounted for by considering other relevant predictors such as motivations (Fox and Glas, 2003).

The psychometrical model-based approach also has its limitations. First, even though latent class shifts preserve the sequential information in action series, they do not capture all the related information. For instance, for the purpose of convenient analysis in this study, some unstable characteristics of a latent class such as random shifts were not used in our definition of class characteristics and class shifts. Fortunately, in many cases, as in this study, this missing information does not affect the results. If it becomes an issue in some cases, it can be addressed by considering more details about the latent class shifts to minimize the ambiguity. Second, this study only takes a single route as an analysis unit, yet failing to consider possible route combinations. For example, in some cases two routes are available, it makes full sense to combine these two routes into one to conduct analysis, because the link between these routes is exclusive. In the future, we may consider the transition model for different route combinations, such as Bi-Road. In terms of the generalizability of the modified MMixIRT model for solving complicated problems, if the process data for another single task can be recoded or restructured as the data file in this study, similar models can be applied to explore the latent classes and characteristics of the problem solving process. However, the difficulty during the analysis lies in how to recode the responses into dichotomous data. For multiple tasks, a three-level model can be applied, with the first level as the process level, the second as the task level and the third as the student level. If there are plenty of tasks, the ability estimates of the student will stay stable. Therefore, while the generalizability of the model may be conditional, the main logic of the MMixIRT approach can be generalized.

AUTHOR CONTRIBUTIONS

HL research design, data analysis, and paper writing. YL paper writing. ML data analysis, and paper writing.

FUNDING

Supported by National Natural Science Foundation of China (31571152); Special Found for Beijing Common Construction Project (019-105812).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01372/full#supplementary-material

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Asparouhov, T., and Muthén, B. (2014). Auxiliary variables in mixture modeling: three-step approaches using Mplus. *Struct. Equat. Model.* 21, 329–341. doi: 10.1080/10705511.2014.915181

Bergner, Y., Shu, Z., and von Davier, A. A. (2014). "Visualization and confirmatory clustering of sequence data from a simulation based assessment task," in *Proceedings of the 7th International Conference on Educational Data Mining,* 177–184. Available online at: http://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/177_EDM-2014-Full.pdf

Bolt, D. M., Cohen, A. S., and Wollack, J. A. (2001). A mixture item response for multiple choice data. *J. Educ. Behav. Stat.* 26, 381–409. doi: 10.3102/10769986026004381

Cohen, A. S., and Cho, S. J. (2016). "Information criteria," in *Handbook of Item Response Theory: Statistical Tools,* ed W. J. van der Linden (Boca Raton, FL: Chapman and Hall/CRC), 363–378.

Desmarais, M. C., and Baker, R. S. J. D. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Model. User Adapt. Interact.* 22, 9–38. doi: 10.1007/s11257-011-9106-8

DiCerbo, K. E., and Behrens, J. T. (2012). "Implications of the digital ocean on current andfuture assessment," in *Computers and their Impact on State Assessment: Recent History and Predictions for the Future,* eds R. Lissitz and H. Jiao (Charlotte, NC: Information Age Publishing), 273–306.

DiCerbo, K. E., Liu, J., Rutstein, D. W., Choi, Y., and Behrens, J. T. (2011). Visualanalysis of sequential log data from complex performance assessments. *Paper presented at the Annual Meeting of the American Educational Research Association* (New Orleans, LA).

Forster, M. R. (2004). *Simplicity and Unification in Model Selection*. University of Wisconsin–Madison. Available online at: http://philosophy.wisc.edu/forster/520/Chapter%230.pdf

Fox, J. P., and Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika* 68, 169–191. doi: 10.1007/bf02294796

Garofalo, J., and Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *J. Res. Math. Educ.* 16, 163–176. doi: 10.2307/748391

Goldstein, H. (1987). Multilevel models in education and social research. *High. Educ. Res. Dev.* 28, 664–645.

Hamaker Ellen, L., van Hattum, P., Kuiper Rebecca, M., and Hoijtink, H. (2011). "Model selection based on information criteria in multilevel modeling," in *Handbook of Advanced Multilevel Analysis,* eds J. J. Hox and J. K. Roberts (New York, NY: Routledge), 231–256.

Hao, J., Shu, Z., and von Davier, A. (2015). Analyzing process data from game/scenario-basedtasks: an edit distance approach. *J. Educ. Data Mining* 7, 33–50.

Hox, J. J. (2010). *Multilevel Analysis Methods: Techniques and Applications*. New York, NY: Routledge.

Janssen, R., and De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behav. Res.* 34, 245–268. doi:10.1207/s15327906mb340205

Kerr, D., Chung, G., and Iseli, M. (2011). *The Feasibility of Using Cluster Analysis to Examine Log Data From Educational Video Games*. CRESST Report, No. 790.

Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kurniati, D., and Annizar, A. M. (2017). The analysis of students' cognitive problem solving skill in solving PISA standard-based test item. *Adv. Sci. Lett.* 23, 776–780. doi: 10.1166/asl.2017.7466

Lee, W. Y., Cho, S. J., and Sterba, S. K. (2017). Ignoring a multilevel structure in mixture item response models: impact on parameter recovery and model selection. *Appl. Psychol. Meas.* 42, 136–154. doi: 10.1177/0146621617711999

Li, F., Cohen, A. S., Kim, S. H., and Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Appl. Psychol. Meas.* 33, 353–373. doi: 10.1177/0146621608326422

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., et al. (2014). *Psychometric Considerations in Game-Based Assessment*. GlassLab Research White Paper. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., and Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika* 55, 195–215. doi: 10.1007/bf02295283

Muthén, L. K., and Muthén, B. O. (1998-2015). *Mplus Users Guide, 7th Edn*. Los Angeles, CA: Muthén and Muthén.

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing. Available online at: http://www.oecd.org/pisa/pisaproducts/pisa2012draftframeworks-mathematicsproblemsolvingandfinancialliteracy.htm

OECD (2014). *PISA 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-LIFE Problems Vol. 5, PISA*, OECD Publishing. Available online at: http://www.oecd.org/education/pisa-2012-results-volume-v.htm

Peugh, J., and Fan, X. (2012). How well does growth mixture modeling identify heterogeneous growth trajectories? a simulation study examining GMM's performance characteristics. *Struct. Equat. Model.* 19, 204–226. doi: 10.1080/10705511.2012.659618

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn*. Newbury Park, CA: SAGE.

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., and Ventura, S. (2013). Web usage mining for predicting final marks of students that use moodle courses. *Comput. Appl. Eng. Educ.* 21, 135–146. doi: 10.1002/cae.20456

Romero, C., and Ventura, S. (2010). Educational datamining: a review of the state of the art. *IEEE Trans. Syst. Man Cybern. Part C* 40, 601–618. doi: 10.1109/tsmcc.2010.2053532

Romero, C., Ventura, S., and García, E. (2008). Data mining in course management systems: moodle case study and tutorial. *Comput. Educ.* 51, 368–384. doi: 10.1016/j.compedu.2007.05.016

Rosato, N. S., and Baer, J. C. (2012). Latent class analysis: a method for capturing heterogeneity. *Soc. Work Res.* 36, 61–69.doi: 10.1093/swr/svs006

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Appl. Psychol. Meas.* 14, 271–282. doi: 10.1177/014662169001400305

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Shu, Z., Bergner, Y., Zhu, M., Hao, J., and von Davier, A. A. (2017). An item response theory analysis of problem-solving processed in scenario-based tasks. *Psychol. Test Assess. Model.* 59, 109–131.

Tofighi, D., and Enders, C. K. (2008). *Identifying the Correct Number of Classes in Growth Mixture Models,* Vol. 13. Charlotte, NC: Information Age Publishing Inc.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociol. Methodol.* 33, 213–239. doi: 10.1111/j.0081-1750.2003.t01-1-00131.x

Williamson, D., M., Bejar, I. I., and Mislevy, R. J. (2006). "Automated scoring of complex tasks in computer-based testing: an introduction," in *Automated Scoring of Complex Tasks in Computer-Based Testing,* eds D. M. Williamson, I. I. Bejar, and R. J. Mislevy (Hillsdale, NJ: Laurence Erlbaum), 1–13.

Zhu, M., Shu, Z., and von Davier, A. A. (2016). Using networks to visualize and analyzeprocess data for educational assessment. *J. Educ. Meas.* 53, 190–211. doi: 10.1111/jedm.12107

Zoanetti, N. (2010). Interactive computer based assessment tasks: how problem-solving process data can inform instruction. *Austral. J. Educ. Technol.* 26, 585–606. doi: 10.14742/ajet.1053

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Liu, Liu and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 March 2019
doi: 10.3389/fpsyg.2019.00646

Mapping Background Variables With Sequential Patterns in Problem-Solving Environments: An Investigation of United States Adults' Employment Status in PIAAC

Dandan Liao[1], Qiwei He[2]* and Hong Jiao[1]*

[1] Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, United States, [2] Educational Testing Service, Princeton, NJ, United States

Adult assessments have evolved to keep pace with the changing nature of adult literacy and learning demands. As the importance of information and communication technologies (ICT) continues to grow, measures of ICT literacy skills, digital reading, and problem-solving in technology-rich environments (PSTRE) are increasingly important topics for exploration through computer-based assessment (CBA). This study used process data collected in log files and survey data from the Programme for the International Assessment of Adult Competencies (PIAAC), with a focus on the United States sample, to (a) identify employment-related background variables that significantly related to PSTRE skills and problem-solving behaviors, and (b) extract robust sequences of actions by subgroups categorized by significant variables. We conducted this study in two phases. First, we used regression analyses to select background variables that significantly predict the general PSTRE, literacy, and numeracy skills, as well as the response time and correctness in the example item. Second, we identified typical action sequences by different subgroups using the chi-square feature selection model to explore these sequences and differentiate the subgroups. Based on the malleable factors associated with problem-solving skills, the goal of this study is to provide information for improving competences in adult education for targeted groups.

Keywords: process data, problem solving, sequential pattern, background variables, large-scale assessment, PIAAC

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Hongyun Liu,
Beijing Normal University, China
Daniel Bolt,
University of Wisconsin–Madison, United States

****Correspondence:***
Dandan Liao
dandanl@terpmail.umd.edu
Qiwei He
qhe@ets.org

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *23 August 2018*
Accepted: *07 March 2019*
Published: *27 March 2019*

Citation:
Liao D, He Q and Jiao H (2019) Mapping Background Variables With Sequential Patterns in Problem-Solving Environments: An Investigation of United States Adults' Employment Status in PIAAC. Front. Psychol. 10:646. doi: 10.3389/fpsyg.2019.00646

INTRODUCTION

Adult assessments have evolved to keep pace with the changing nature of adult literacy and learning demands. As the importance of information and communication technologies (ICT) continues to grow, measures of ICT skills are increasingly important topics for exploration through computer-based assessment (CBA). The Programme for the International Assessment of Adult Competencies (PIAAC) is the first international household survey of adult skills predominantly collected using ICT skills. Conducted in 40 countries, this international survey measures key cognitive and workplace skills including literacy, numeracy, and problem-solving in

technology-rich environments (PSTRE). These skills are not only critical to individual prosperity but are also key drivers of economic growth and societal advancement (Organisation for Economic Co-operation and Development [OECD], 2013b, p. 3).

Specifically, PSTRE assessment focuses on the ability of "using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks" (Organisation for Economic Co-operation and Development [OECD], 2012). As digital technology has become an indispensable part of human lives, there is an increasing need for measuring the ability to solve problems in conjunction with basic computer literacy skills. PSTRE assessment renders it possible to measure how well adults process, analyze, and address problems for specific goals in a computer-based environment.

According to a recent report published by the National Center for Education Statistics (Rampey et al., 2016), United States respondents on average scored lower than respondents from other countries in the PSTRE domain (Organisation for Economic Co-operation and Development [OECD], 2013b, p. 11). In particular, the United States sample had the largest proportion of respondents scoring below Level 1, which is the minimum proficiency level required to complete simple problem-solving tasks in daily life (Organisation for Economic Co-operation and Development [OECD], 2013b, p. 21).

Some facts about specific subgroups of United States respondents are also concerning. Scores for millennials (adults born after 1980 and between ages 16–34) in the United States were among the lowest of all participating countries even though over half of them spent 35 hours per week on digital media (Organisation for Economic Co-operation and Development [OECD], 2013b, p. 21; Goodman et al., 2015). It was found that 41% of respondents with less than high school education chose to take the paper version of PIAAC, compared to 17% for high school graduates and 5% for those with a college degree or above (Organisation for Economic Co-operation and Development [OECD], 2013b, p. 21). Further, 30% of those who reported being out of the workforce took the paper-based test as opposed to 14% for adults in the labor force (Organisation for Economic Co-operation and Development [OECD], 2013b, p. 21), suggesting a correlation between skills required for completing the computerized version of the assessment and employability (Vanek, 2017).

An issue that PIAAC attempts to provide a clear picture for is the match between supply and demand for employment skills (Organisation for Economic Co-operation and Development [OECD], 2016, p. 3). There has been increasing interest in exploring the relationship between proficiency levels and subgroups by employment-related variables, such as employment status and skills used at work (e.g., Organisation for Economic Co-operation and Development [OECD], 2016, p. 102–103; Perry et al., 2016). However, assessment of skills is merely one step toward a more balanced labor market. Knowing which subgroups performed better is a good starting point, but the processes that gave rise to the final proficiency levels are more informative for providing necessary education.

To bridge the gap between supply and demand and provide targeted intervention, it is important to understand which subgroups performed at a lower level and why. Specifically, how did these respondents arrive at a specific wrong answer, and how did subgroups differ in terms of problem-solving strategies? In this regard, more fine-grained investigation on multiple sources of data is needed, which cannot be easily achieved by utilizing response data alone.

The use of computers as the delivery platform enables data collection not just on whether respondents are able to solve the tasks. It also gives information on how they solved them, which is referred to as process data. Process data has great potential for providing insight into different phases of educational learning. One key application area is allowing intelligent tutoring systems to adapt to respondents' needs in real time based on their characteristics (e.g., Baker, 2007; D'Mello et al., 2008; Scheuer and McLaren, 2011). Another area that has attracted much interest is to model changes in knowledge over time via Bayesian knowledge tracing (e.g., Corbett and Anderson, 1994; Baker et al., 2008; Pavlik et al., 2009).

More importantly, several studies have revealed the critical role of process data in understanding different problem-solving strategies (e.g., Hurst et al., 1997; Vendlinski and Stevens, 2002; He and von Davier, 2015, 2016; He et al., 2018). Vendlinski and Stevens (2002) identified three strategy levels that students adopted to solve a chemistry item: limited, prolific, and efficient. Students who used a limited strategy tried only a few options before attempting to solve the item, whereas the prolific strategy was to explore almost all options on the menu, similar to the "unfocused problem-solving strategy" found in Hurst et al. (1997). On the contrary, students with efficient strategy concentrated only on the key pieces of information, resulting in the highest probability of a correct answer. He and von Davier (2015) further pointed out that the pattern of robust sequences of actions differed significantly by respondents' performance levels by respondents' performance levels, which was found consistently were consistent across countries. Those in the higher-performing group tended to use more tools such as search and sort, had clearer understanding of the subgoals, and were able to recover from initial mistakes. The lower-performing group, however, demonstrated more hesitative behaviors, such as clicking "cancel" repeatedly, and only had a vague idea about the purpose of the item (He and von Davier, 2016). He et al. (2018) continued investigating the differences in problem-solving strategies associated with background variables on one PIAAC item across six countries. It was found that test takers with high levels of skills for using ICT at home were more likely to have higher PSTRE performance. Respondents with different genders had significant differences in digital task-solving strategies. In fact, older people, female, and those with low ICT skill use at home or work showed a need for more intervention to improve their PSTRE skills.

Based on the results from He and von Davier (2015, 2016) and He et al. (2018), the present study mainly focuses on employment-related variables and the United States sample to further identify important factors associated with problem-solving skills. Specifically, two research questions are addressed via exploring the process data from one representative PSTRE item:

(1) Which employment-related background variables are significantly related to performance in the PSTRE, literacy, and numeracy domains in the United States sample?
(2) For those subgroups that showed significantly different performance on a representative PSTRE item, what features can we extract from process data to best characterize their behaviors?

By analyzing process data in different employment situations and with different work experience, we are able to see different behavioral patterns by subgroups during the process of solving digital tasks. The rest of this paper is structured as follows. In Section "Materials and Methods" we elaborate on the data and instrument used in this study, and introduce the proposed approach (i.e., regression analysis and feature identification) to map the background variables with action sequences in process data. The results corresponding to the two research questions are presented in Section "Results", with special attention to generalizing results for the United States population. In the last section, we summarize the findings and discuss the limitations and potential future work using process data in large-scale assessments.

MATERIALS AND METHODS

Datasets and Instruments

The PSTRE assessment in PIAAC 2012 study included 14 items, with seven in each of the two booklets.[1] Respondents who responded to the PSTRE items had to have some prior computer experience and to have passed the first two stages of core computer-based assessments. The PSTRE items were generally designed in four different environments—email, web, word processor, and spreadsheet; each item involved one or two environments as summarized in **Table 1**.

Item U02, the Meeting Room Assignment item, was chosen as an example to illustrate the present study. There are three environments involved in this item: email, web, and word processor. Respondents were asked to read through a list of emails of meeting room requests in the email environment, and then try to fill out as many requests as possible in the room reservation system in a web environment.

There are four reasons why we decided to use U02 as an example:

(1) U02 was rather difficult for United States respondents: 932 (70%) respondents received no credit, 294 (22%) received partial credit, and only 114 (9%) got full credit. Such an item could potentially provide more information to identify reasons for failure when tracking respondents' process data. Researchers have found that for a moderately difficult item, respondents tend to demonstrate more heterogeneous use of strategies, aberrant response behavior, and response time (e.g., Vendlinski and Stevens, 2002; Goldhammer et al., 2014; de Klerk et al., 2015). To explore the difference between respondents who at least got part of the item correct and those who received no credit, the polytomous scores were dichotomized by collapsing partial credit and full credit in the present study.
(2) U02 had multiple environments (email, web, and word processor), which tended to have more diverse actions from which to extract information.
(3) Compared to items at the beginning or the end, items in the middle of the booklet were less likely to demonstrate position effect (e.g., Wollack et al., 2003).
(4) U02 shared environments with most items in booklet PS2. This provided the possibility to investigate the consistency of problem-solving strategies across items for each individual.

The present study used two datasets, the public-use background questionnaire (BQ) from PIAAC 2012 and the assessment's log file. The former dataset contains the original and derived variables from the BQ, cognitive response data, as well as sampling weights. The employment-related variables reflected different perspectives of the test taker's employment situation, such as employed or not, whether the test taker had a supervisor role, related work experience, computer use at work, and so on. The demographic variables included age, gender, the test taker's education level, the test taker's parents' education level, whether the assessment was given in the test taker's native language, and the number of books at home. Variables from the BQ, with a focus on those related to employment and work experience, were used to explore the relationship between patterns extracted from process data and respondents' employment situations. Variables measuring skills used at home, such as ICT and numeracy skills at home, were not considered since work-related background variables had stronger connections to employment situation

[1] PSTRE sample items can be found on the National Center for Education Statistics website: https://nces.ed.gov/surveys/piaac/sample_pstre.asp.

TABLE 1 | Summary of environments in each item.

Booklet	Order	Item	Email (MC)	Web (WB)	Word Processor (WP)	Spreadsheet (SS)
PS1	1	U01a	1			
PS1	2	U01b	1			
PS1	3	U03a		1		1
PS1	4	U06a		1		
PS1	5	U06b		1		
PS1	6	U21		1		
PS1	7	U04a	1			1
PS2	1	U19a	1			1
PS2	2	U19b			1	1
PS2	3	U07		1		
PS2	4	U02	1	1	1	
PS2	5	U16	1			
PS2	6	U11b	1			
PS2	7	U23	1	1		

(Organisation for Economic Co-operation and Development [OECD], 2016).

Additionally, scored responses, total response time, timing of first action, and number of actions were available for each item in the three domains. For each of the 3 domains, 10 plausible values were provided for each test taker (see more information in Organisation for Economic Co-operation and Development [OECD], 2013a, Chapter 17). The proposed analyses were conducted with and without sampling weights, and the differences were marginal. Therefore, we reported results with sampling weights only. Log files recorded the actions taken during the assessment, including actions taken during the assessment, such as sorting, clicking menu, opening a folder, using the help function, and so on.

The total sample size for the BQ was 5,010.[2] The descriptive statistics of age, gender, and education of all respondents in the BQ were reported in **Tables 2**, **3**. The distributions of age and gender are rather even. About 46% of the respondents obtained postsecondary education, 39% had upper secondary education, and 13% had lower secondary education or less.

Data Analyses

The present study was conducted in two phases: regression phase and feature identification phase (see **Figure 1** as an overview). In the first phase, we employed regression analyses to select background variables that could significantly predict respondents' PSTRE, literacy, and numeracy proficiency levels, response time, as well as response correctness in the example item. In the second phase, typical action sequences were identified by different subgroups using the chi-square feature selection model.

Regression Phase: Identifying Significant Employment-Related Background Variables

Regression analyses were conducted to examine which employment-related variables have significant associations with both person- and item-related outcome variables. The variables were carefully selected from the BQ, including 20 employment-related and 6 demographic variables. **Table 4** summarizes the description, number of non-missing categories, and the reference category for each variable (see **Appendix Table A1** for detailed descriptions for all levels of each variable). To avoid a dramatic decrease in sample size and incorporate information as much as possible in the regression analysis, we coded missing responses in the selected variables as an additional category and retained in the regression analyses. This method was popularized by Cohen and Cohen (1985) as a way to deal with missing responses in categorical variables. This method incorporates all the available information into the regression analyses, whereas other methods could heavily depend on data collection design and model specification (e.g., Howell, 2008). Compared to the deletion methods, the generalizability of the results to the United States population may also be retained using the proposed method. Moreover, it is the simplest approach to addressing missing data with some missing mechanisms being untestable (Carpenter and Goldstein, 2004; Horton and Kleinman, 2007).

With respect to the dependent variables, we used the respondents' scores in PSTRE, literacy, and numeracy as well as total response time and binary scores (correct as 1; incorrect as 0) in U02. To retain as much information as possible in the regression analysis, we included all respondents who had plausible values in each domain, resulting in a total sample size of 4,103, 4,898, and 4,898 for PSTRE, literacy, and numeracy scores, respectively. Further, we used total item response time (U02RT) and dichotomized scores (U02score) as item-related variables in the regression analysis for the Meeting Room Assignment item. Note that only 1,340 in the sample who had process data for this specific item were adopted in the current study, occupying one third of the whole sample size used in the regression analysis.

Of the five outcome variables, linear regression was conducted for the four continuous outcome variables—PSTRE, literacy, and numeracy scores, as well as the item response time on the Meeting Room Assignment item. For the dichotomized scores, a logistic regression was carried out. The regression analyses were conducted using the International Association for the Evaluation of Educational Achievement (IEA)'s International Database (IDB) Analyzer version 4.0.16.0 (International Association for the Evaluation of Educational Achievement [IEA], 2013) to interface with SAS 9.4 SAS Institute, 2015). In this study, each regression analysis was carried out using a full sample weight and 45 replicate weights, as well as 10 plausible values if the outcome

[2]For the regression analyses in study 1, all respondents with valid PSTRE/literacy/numeracy scores and background variables were used in the analyses. There are two reasons for this. First, it retains the generalizability to the whole United States population. Second, it enables the comparison among the significant variables for the PSTRE, literacy, and numeracy domains to explore the uniqueness of PSTRE skills. Only the sample that responded to PSTRE assessment was used when we further explored features from process data.

TABLE 2 | Descriptive statistics of age and gender for all respondents in BQ.

Level	Age	Count (%)	Gender	Count (%)
1	24 or less	837 (16.71%)	Male	2,323 (46.37%)
2	25–34	1,045 (20.86%)	Female	2,687 (53.63%)
3	35–44	978 (19.52%)		
4	45–54	1,084 (21.64%)		
5	55 plus	1,066 (21.28%)		

TABLE 3 | Descriptive statistics of education level for all respondents in BQ.

Level	Education	Count (%)
1	Lower secondary or less (ISCED 1,2, 3C short or less)	629 (12.55%)
2	Upper secondary (ISCED 3A-B, C long)	1,977 (39.46%)
3	Postsecondary, non-tertiary (ISCED 4A-B-C)	394 (7.86%)
4	Postsecondary, tertiary – professional degree (ISCED 5B)	414 (8.26%)
5	Postsecondary, tertiary – bachelor degree (ISCED 5A)	902 (18.00%)
6	Postsecondary, tertiary – master/research degree (ISCED 5A/6)	578 (11.54%)
7	Missing	116 (2.32%)

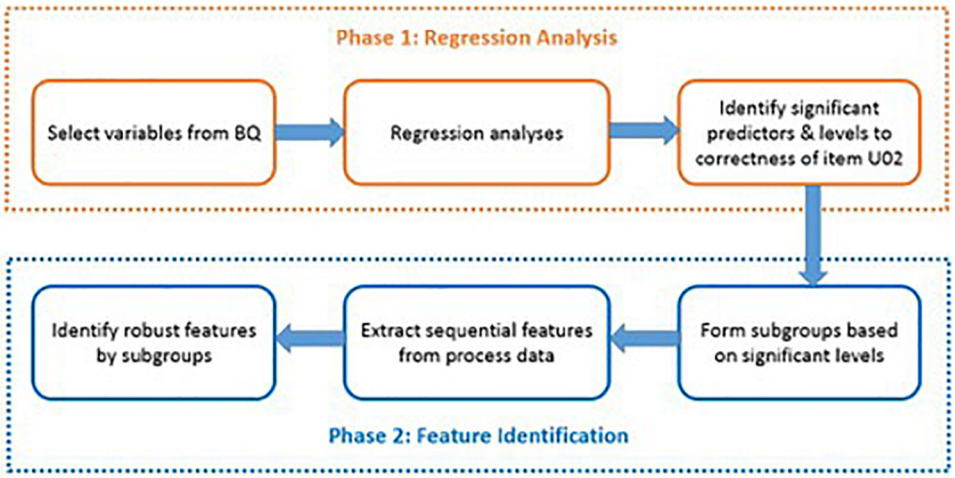


FIGURE 1 | An overview of the two-phase analysis.

TABLE 4 | Summary of BQ variables used in the present study.

No.	Variable	Description	Number of categories	Reference category
(1)	PAIDWORK[a]	Incidence of paid work experience	4	1. No work ever
(2)	C_D05	Employed/unemployed/out of labor force	3	1. Employed
(3)	D_Q04_T1[b]	Is an employee/supervisor/ self-employed/unpaid	4	1. Employee, not supervisor
(4)	D_Q08b	Managing how many employees	5	1. 1–5 people
(5)	D_Q12b	Education level sufficient/too high/too low to do job satisfactorily	3	1. Necessary
(6)	D_Q12c_RC[c]	Related work experience in years	4	1. No experience
(7)	F_Q05a	Incidence of solving simple problems	5	1. Never
(8)	F_Q05b	Incidence of solving complex problems	5	1. Never
(9)	F_Q07b	Need more training for skill use at work or not	2	1. Yes
(10)	G_Q06	Level of computer use	3	1. Straightforward
(11)	ISCOSKIL4	Skilled/semi-skilled/ elementary occupations	4	1. Skilled occupations
(12)	EARNMTHALLDCL	Monthly earning decile including all incomes	10	1. Lowest decile
(13)	LEARNATWORK_WLE_CA	Index of learning at work	6	0. All zero response
(14)	ICTWORK_WLE_CA	Index of use of ICT skills at work	6	0. All zero response
(15)	INFLUENCE_WLE_CA	Index of use of influencing skills at work	6	0. All zero response
(16)	NUMWORK_WLE_CA	Index of use of numeracy skills at work	6	0. All zero response
(17)	PLANNING_WLE_CA	Index of use of planning skills at work	6	0. All zero response
(18)	READWORK_WLE_CA	Index of use of reading skills at work	6	0. All zero response
(19)	TASKDISC_WLE_CA	Index of use of task discretion at work	6	0. All zero response
(20)	WRITWORK_WLE_CA	Index of use of writing skills at work	6	0. All zero response
(21)	AGEG10LFS	Age in 10-year bands	5	1. 24 or less
(22)	PARED	Highest of parents' level of education	3	1. Neither parent have attained upper secondary
(23)	NATIVELANG	Test language same as native language or not	2	0. Test language is not native language
(24)	EDCAT6	Highest level of formal education obtained	6	1. Lower secondary or less
(25)	GENDER_R	Gender	2	1. Male
(26)	J_Q08	Number of books	6	1. 10 books or less

[a]PAIDWORK is recoded from NOPAIDWORKEVER (never had paid work), PAIDWORK5 (have had paid work in the last 5 years) and PAIDWORK12 (have had paid work in the last 12 months). [b]D_Q04_T1 provides information about whether the respondent is an employee, supervisor, self-employed, or unpaid. This variable was chosen since research has shown that the relationship between some variables is different for wage earners and self-employed workers, such as earnings and conscientiousness (Organisation for Economic Co-operation and Development [OECD], 2016). [c]D_Q12c_RC is recoded from D_Q12c (related work experience in years) by collapsing category 2–4.

variable was the scores from one of the three domains. The final regression coefficient estimates were weighted averages of the coefficient estimates from each round. The standard errors of the coefficient estimates were pooled standard errors reflecting variability due to multiple imputation and/or sampling error. Then the significance of the coefficient estimates was determined by the relative magnitude of the final coefficient estimates and the pooled standard errors. Readers can refer to International Association for the Evaluation of Educational Achievement [IEA] (2013) for more information.

Feature Identification Phase: Identifying Typical Action Sequences by Subgroups

In the feature identification phase, process data were used to understand the inherent differences among respondents' action sequences in the test-taking process. Each individual's time-stamped action sequences in U02 were extracted from the log file and recoded into (mini-) sequences by *n*-grams.

An *n*-gram is defined as a contiguous sequence of *n* words in text mining; similarly, when analyzing action sequences from process data, an *n*-gram can be defined as a sequence of *n* adjacent actions (Manning and Schütze, 1999). For instance, a typical sequence for email review actions is recorded as "*MAIL_VIEWED_4, MAIL_VIEWED_2, MAIL_VIEWED_1*", the unigram is each of the three separate actions (e.g., "*MAIL_VIEWED_4*"), the bigram is the two adjacent actions as one unit, (e.g., "*MAIL_VIEWED_2, MAIL_VIEWED_1*"), and the trigram is the three adjacent actions as one unit (e.g., "*MAIL_VIEWED_4, MAIL_VIEWED_2, MAIL_VIEWED_1*"). In this study, we focused on unigrams, bigrams, and trigrams, which are adjacent action sequences of length 1, 2, and 3, respectively.

When retrieving information from the *n*-grams, a question regarding whether all terms could be considered equally important based on their raw frequencies needs to be addressed. In fact, certain terms have little or no discriminating power in determining relevance; it was recommended to give them less weight when classifying different subgroups (Manning and Schütze, 1999). We adopted term weights in this study to adjust for between- and within-individual differences in action frequencies. In terms of between-individual differences, a popular weighting method in text mining, inverse document frequency (IDF; Spärck Jones, 1972) that was renamed as inverse sequence frequency (ISF; He and von Davier, 2016) was adapted for estimating the weight of each *n*-gram. ISF is defined as $ISF_i = \log\left(N/sf_i\right) \geq 0$, where N denotes the total number of sequences in the sample, which is the same as the total number of respondents, and sf_{i} represents the number of sequences containing action, i.e., a large ISF reflects a rare action in the sample, whereas a small ISF represents a frequent one.

Within-individual differences occur when an individual takes some actions more often than others. Although more frequent sequences are usually more important than less frequent sequences, the raw frequencies of these action sequences often overestimate their importance (He and von Davier, 2015, 2016). To account for within-individual differences in the importance of action sequences, a weighting function was employed $f\left(tf_{\mathrm{ij}}\right) = 1 + \log\left(tf_{\mathrm{ij}}\right)$, where $tf_{\mathrm{ij}} > 0$ represents the frequency of action i in sequence j (Manning and Schütze, 1999). Combining the between- and within-individual weights, the final action weight can be defined as $weight\left(i, j\right) = \left[1 + \log\left(tf_{\mathrm{ij}}\right)\right]\log\left(N/sf_{\mathrm{i}}\right)$ for $tf_{\mathrm{ij}} \geq 1$. In contrast to raw frequency, this weighting mechanism was applied for attenuating the effect of actions or action vectors that occurred too often to be meaningful. (For more details of n-grams and term weights in process data analysis, refer to He and von Davier, 2015, 2016).

To answer the question regarding which actions or mini action sequences (i.e., *n*-grams) are the key factors that distinguish subgroups, we applied a commonly used tool in natural language processing—the chi-square feature selection model (Oakes et al., 2001)—to identify robust classifiers. The chi-square feature selection model is recommended for use in textual analysis due to its high effectiveness in finding robust keywords and for testing the similarity between different text corpora (e.g., Manning and Schütze, 1999; He et al., 2012, 2014, 2017). The definition of "robust" is different from what is defined in statistics; here, robust features are generally defined as the "best" features with high information gain in natural language processing (Joachims, 1998). Chi-square scores assigned to the features were ranked in a descending order, and those with the highest scores were defined as robust features. Specifically, frequencies and weights of certain actions for different employment statuses were used as input for the chi-square selection model.

Features extracted for different groups (e.g., income and employment type) were used to understand the inherent differences in typical sequences among subgroups. The package "tm" (Feinerer, 2017) in R version 3.3.3 (R Core Team, 2017) was utilized for applying chi-square selection model to identify robust features. We formed subgroups based on each significant employment-related predictor for the outcome variable U02score (i.e., binary variable correctness/incorrectness in the Meeting Room Assignment item). The significant level of the predictor was compared with the reference level of the predictor. For instance, if the fourth decile of EARNMTHALLDCL was significantly different from the reference group, two subgroups were formed by respondents in the lowest decile and in the fourth decile. Chi-square selection model was then applied to compare action sequences between these two subgroups and identify robust features to distinguish them.

RESULTS

Regression Phase

The distributions of the background variables were checked to ensure the representativeness of this sample. The difference between the percentages of each category of the background variables from the sample with valid scores on the U02 item (i.e., Meeting Room Assignment) and the total sample was usually around 1–2% (see **Appendix Table A2** for details). As such, we deemed that the differences were not substantially different.

The sample size and descriptive statistics of the five outcome variables—PSTRE, literacy, and numeracy scores, and the

TABLE 5 | Sample size and descriptive statistics of the outcome variables.

Variable name	Sample size	Minimum	Maximum	Mean	Standard deviation
PSTRE	4,103	113.56	425.01	277.98	43.11
Literacy	4,898	100.94	424.33	271.84	48.28
Numeracy	4,898	40.33	426.90	254.68	55.80
U02RT*	1,340	5.16	2704.38	215.67	208.63
U02score	1,340	0	1	0.30	0.21

**U02RT is reported in seconds.*

response time and dichotomized scores for item U02 —are reported in **Table 5**. Using scores from all three domains as dependent variables enabled us to explore the uniqueness of PSTRE skills. In other words, which employment-related variables were significant in predicting PSTRE scores but not literacy or numeracy scores. The significant predictors identified from regression analyses are summarized in **Table 6** with respect to each of the five outcome variables. **Table 7** presents the unstandardized coefficient estimates for the significant variables. The standardized coefficients for all variables were reported in **Appendix Table A3**, as a measure of variables' contributions to predicting the outcome that accounts for contributions of other independent variables (e.g., Menard, 1995, 2004; Zientek et al., 2008; Nathans et al., 2012). The rank ordering of the absolute values of these coefficients indicates the relative importance of the variables.

In general, all five outcome variables had one significant variable in common, EDCAT6, which means that the highest level of formal education is important for obtaining high scores in all three domains and on individual item responses, and it also contributes to longer item response time in this particular item. Among the three person-related dependent variables, more predictors were significant in predicting literacy and numeracy scores when compared with PSTRE scores. The significant variables for literacy and numeracy scores were more similar, though the three domains had 13

TABLE 6 | Summary of significant predictors.

No.	Variable	Description	PSTRE	Literacy	Numeracy	U02RT	U02score
(1)	PAIDWORK	Incidence of paid work experience		1*	1		
(2)	C_D05	Employed/unemployed/out of labor force					
(3)	D_Q04_T1	Is an employee/supervisor/self-employed/unpaid					
(4)	D_Q08b	Managing how many employees					
(5)	D_Q12b	Education level sufficient/too high/too low to do job satisfactorily	1	1	1		
(6)	D_Q12c_RC	Related work experience in years	1		1		
(7)	F_Q05a	Incidence of solving simple problems	1	1	1		
(8)	F_Q05b	Incidence of solving complex problems		1	1		
(9)	F_Q07b	Need more training for skill use at work or not		1	1		
(10)	G_Q06	Level of computer use	1	1	1	1	
(11)	ISCOSKIL4	Skilled/semi-skilled/elementary occupations	1	1	1		
(12)	EARNMTHALLDCL	Monthly earning decile including all incomes	1	1	1		1
(13)	LEARNATWORK_WLE_CA	Index of learning at work	1	1	1		
(14)	ICTWORK_WLE_CA	Index of use of ICT skills at work	1	1	1		
(15)	INFLUENCE_WLE_CA	Index of use of influencing skills at work		1	1		1
(16)	NUMWORK_WLE_CA	Index of use of numeracy skills at work	1	1	1	1	
(17)	PLANNING_WLE_CA	Index of use of planning skills at work		1	1	1	
(18)	READWORK_WLE_CA	Index of use of reading skills at work	1				1
(19)	TASKDISC_WLE_CA	Index of use of task discretion at work					1
(20)	WRITWORK_WLE_CA	Index of use of writing skills at work		1			1
(21)	AGEG10LFS	Age in 10-year bands	1	1	1		1
(22)	PARED	Highest of parents' level of education	1	1	1	1	
(23)	NATIVELANG	Test language same as native language or not	1	1	1		1
(24)	EDCAT6	Highest level of formal education obtained	1	1	1	1	1
(25)	GENDER_R	Gender	1		1		
(26)	J_Q08	Number of books	1	1	1		

**The number 1 indicates that at least one level of this variable was significantly different from 0 at the significance level of 0.05.*

TABLE 7 | Summary of unstandardized regression coefficients of significant variables.

Variable	Category	Description	PSTRE	Literacy	Numeracy	U02RT	U02score
Intercept		Intercept	214.93	162.42	126.44	380.70	
PAIDWORK	3	Have had paid work in 5 years but not 12 months		22.37	30.59		
	4	Have had paid work in 12 months			37.51		
D_Q12b	2	A lower education level would be sufficient	5.22	3.76			
	3	A higher education level would be needed	−8.17	−8.38	−9.35		
D_Q12c_RC	2	Less than 1 year of relevant work experience	−5.07		−5.49		
	4	More than 3 years of relevant work experience			6.76		
F_Q05a	2	Solve simple problems less than once a month		9.13			
	3	Solve simple problems less than once a week but at least once a month	13.53	10.91	17.92		
	4	Solve simple problems at least once a week but not everyday	12.58	13.80	17.11		
	5	Solve simple problems everyday	16.47	17.71	19.07		
F_Q05b	5	Solve complex problems every day		−11.01	−9.14		
F_Q07b	2	Do not need more training for skill use at work		6.20	7.28		
G_Q06	2	Moderate level of computer use	9.97	7.23	8.15	39.03	
	3	Complex level of computer use	15.44	10.35	15.96	74.98	
ISCOSKIL4	2	Semi-skilled white-collar occupations	−3.83	−4.47	−4.80		
	3	Semi-skilled blue-collar occupations	−7.29	−7.10	−6.24		
	4	Elementary occupations		−13.49	−14.11		
EARNMTHALLDCL	4	4th decile of monthly earning					2.00
	9	9th decile of monthly earning		7.88			
	10	10th decile of monthly earning	10.15	12.38	11.55		
LEARNATWORK_WLE_CA	4	More than 60–80% on index of learning at work	−7.60	−10.42	−9.74		
ICTWORK_WLE_CA	1	Lowest 20% on index of use of ICT skills at work		10.98	10.05		
	2	More than 20–40% on index of use of ICT skills at work	12.28	15.23	11.28		
	3	More than 40–60% on index of use of ICT skills at work	16.34	13.63	12.09		
	4	More than 60–80% on index of use of ICT skills at work	17.78	11.52	10.29		
	5	More than 80% on index of use of ICT skills at work	20.52	15.46	16.00		
INFLUENCE_WLE_CA	1	Lowest 20% on index of use of influencing skills at work					1.63
	5	More than 80% on index of use of influencing skills at work		−10.84	−10.24		
NUMWORK_WLE_CA	2	More than 20–40% on index of use of numeracy skills at work	9.12	5.70	8.76		
	3	More than 40–60% on index of use of numeracy skills at work			7.10		
	4	More than 60–80% on index of use of numeracy skills at work	8.51	8.33	10.47	50.43	
	5	More than 80% on index of use of numeracy skills at work	6.89		10.26		
PLANNING_WLE_CA	4	More than 60–80% on index of use of planning skills at work		9.33	12.72	55.12	
READWORK_WLE_CA	1	Lowest 20% on index of use of reading skills at work					2.23
	5	More than 80% on index of use of reading skills at work	−14.44				
TASKDISC_WLE_CA	4	More than 60–80% on index of use of task discretion at work					−0.43
WRITWORK_WLE_CA	3	More than 40–60% on index of use of writing skills at work		6.85			
	4	More than 60–80% on index of use of writing skills at work		7.93			1.68
AGEG10LFS	2	25–34	−17.20	−12.91	−12.10		
	3	35–44	−24.57	−17.21	−16.84		
	4	45–54	−31.89	−22.09	−20.77		0.63
	5	55 plus	−35.85	−23.64	−20.65		0.53

(Continued)

TABLE 7 | Continued

Variable	Category	Description	PSTRE	Literacy	Numeracy	U02RT	U02score
PARED	2	At least one parent has attained secondary and postsecondary, non-tertiary	10.31	10.15	13.06	46.16	
	3	At least one parent has attained tertiary	12.33	15.38	15.93	77.01	
NATIVELANG	1	Test language same as native language	13.92	17.41	9.99		1.29
EDCAT6	2	Upper secondary (ISCED 3A-B, C long)	10.09	16.72	21.31	52.29	7.70
	3	Postsecondary, non-tertiary (ISCED 4A-B-C)	14.18	19.91	27.92		7.70
	4	Tertiary – professional degree (ISCED 5B)	17.59	27.84	35.31		5.47
	5	Tertiary – bachelor degree (ISCED 5A)	24.00	35.74	45.59	66.62	10.16
	6	Tertiary – master/research degree (ISCED 5A/6)	28.60	44.53	55.83	79.51	15.18
GENDER_R	2	Female	−4.19		−12.89		
J_Q08	2	11–25 books		6.54			
	3	26–100 books	10.45	9.84	14.52		
	4	101–200 books	13.97	13.54	21.49		
	5	201–500 books	22.49	20.20	26.62		
	6	More than 500 books	14.13	19.74	23.33		

Coefficients reported in this table for U02score are odds ratios. U02RT is reported in seconds. All numbers shown in the table are significant regression coefficients. The missing cells or missing categories indicate insignificant coefficient values and therefore are not reported. Those cells in gray indicate coefficient estimates that are in consistent with our expectation.

significant variables in common. D_Q12c_RC, the related work experience in years, and GENDER_R (gender) were significant in predicting PSTRE and numeracy but not literacy, whereas WRITWORK_WLE_CA (index of use of writing skills at work) was only important for literacy scores. As the focus of this study, PSTRE scores had one unique significant variable—READWORK_WLE_CA (index of use of reading skills at work)—indicating that these skills are significantly related to PSTRE scores. This reflects that by item design, PSTRE items would require higher-level reading skill use at work to understand the item structure, follow the instructions, and browse the website.

Only five variables were significant in predicting the response time on the Meeting Room Assignment item. The regression coefficient estimates showed that respondents who were well-educated, had higher levels of computer use, used more numeracy and planning skills at work, and whose parents also obtained higher education degrees tended to spend more time on the item. Although some research has shown that people with higher ability need less time to finish an item (e.g., van der Linden, 2007; Klein Entink, 2009; Wang and Xu, 2015; Fox and Marianti, 2016), other studies demonstrated the opposite evidence, especially for non-speeded tests (e.g., Roberts and Stankov, 1999; Klein Entink et al., 2009). This observation is consistent with the fact that PIAAC was not a timed assessment; respondents were allowed to take as much time as needed.

Similarly, U02score did not have as many significant variables as the person-related outcomes either (i.e., PSTRE, literacy, and numeracy scores), where only eight variables were significant. It was also noted that not all variables were significant in predicting PSTRE scores. This might be because PSTRE scores are holistic measures of the PSTRE skills, which represent the common construct assessed by the 14 PSTRE items. As U02 only partially contributed to the PSTRE scores, it did not necessarily contain all aspects of the construct.

In terms of the coefficient estimates, most were consistent with our expectations. With respect to employment-related variables, respondents who had paid work, more related work experience, solved simple or complex problems more frequently, had higher level of computer use, had skilled occupations and higher monthly income, and/or had higher index variables tended to have higher scores in the three domains and higher odds of success in this specific item. For the demographic variables, younger male respondents who were well-educated and had many books at home would get higher scores when the test was given in their native languages.

However, some coefficient estimates were inconsistent with our expectations, which are highlighted in gray in **Table 7**. For example, we would expect respondents with more related work experience to perform better in general, but the estimates for the variable representing experience of less than a year were negative for PSTRE and numeracy scores. This indicates that, controlling for all other variables, having short work experience was not better than having no experience for these two outcomes. For F_Q05b (solve complex problems every day), coefficients for literacy and numeracy scores were also negative when comparing the highest category to the lowest, reflecting that a respondent who solved complex problems regularly might get a score lower than a respondent who never solved complex problems at work. These contradictory results may indicate some interactions among the predictors, which would be worthwhile for further exploration.

Feature Identification Phase

For the significant predictors for U02score, we further explored how the action sequences of the two groups were different from each other. We used two significant variables—monthly income and education—as concrete examples to show how the features from process data were identified. Given the limited space, we listed more detailed results in the **Appendix**.

Differences by Monthly Income Subgroup

The regression coefficient for the fourth decile of EARNMTHALLDCL (monthly earning decile including all incomes) was significant and positive, indicating that respondents with monthly income in that decile were more likely to get a score of 1 than those in the first (lowest) decile. As such, it is of interest to investigate how the respondents with monthly income in those deciles differed regarding their action sequences. In other words, what features did the two groups of respondents have in their test-taking behaviors that gave rise to higher or lower chances of answering the Meeting Room Assignment item correctly?

As demonstrated earlier, we conducted chi-square selection to identify the most distinguishable *n*-grams between the two groups. Specifically, the top five unigrams, bigrams, and trigrams with the highest chi-square scores were obtained for the focal group and the reference group, respectively. The description and frequency of 34 unigrams used in the present study were reported in **Table 8**. These robust features were used to understand the most distinctive action sequences between the two groups of respondents. The same procedure was carried out for all significant predictors for U02score. **Tables 9**, **10** demonstrate monthly income and education as two examples, respectively; the robust features for all the other predictors are reported in **Appendix Tables A4–A14** for more details. The interpretations of the actions were based on content experts who designed the item.

Table 9 presents the top five unigrams, bigrams, and trigrams for the respondents falling within the fourth and first (lowest) deciles of monthly earning. Among the unigrams, folder-related actions were found more often in the fourth decile group, such as fold, add, or delete a folder. There were a few folders in the email environment, though respondents were not required to perform any actions on them. The fourth decile group also applied more cancel-related actions, such as cancel sorting, cancel changing reservation, cancel switching to the

TABLE 8 | Description and frequency of unigrams.

No.	Features	Description	Frequency
(1)	*FOLDER_VIEWED*	View a folder	5762
(2)	*ENVIRONMENT_WB*	Go to web environment	4715
(3)	*ENVIRONMENT_MC*	Go to email environment	4317
(4)	*MAIL_VIEWED_1*	View 1st email	2725
(5)	*HISTORY_VIEWCALENDAR*	Go to calendar tab in web environment	2190
(6)	*MAIL_VIEWED_3*	View 3rd email	1968
(7)	*HISTORY_RESERVATION*	Go to reservation tab in web environment	1935
(8)	*COMBOBOX_ROOM*	Choose a room when filling out a room request	1891
(9)	*MAIL_VIEWED_4*	View 4th email	1698
(10)	*MAIL_VIEWED_2*	View 2nd email	1544
(11)	*MAIL_MOVE*	Move an email	1499
(12)	*NEXT_INQUIRY*	Go to next item	1371
(13)	*START*	Start item U02	1326
(14)	*COMBOBOX_START_TIME*	Choose start time when filling out a room request	1312
(15)	*COMBOBOX_END_TIME*	Choose end time when filling out a room request	1304
(16)	*COMBOBOX_DEPT*	Choose department when filling out a room request	1296
(17)	*HISTORY_MEETINGROOMS*	Go to meeting room details tab in web environment	1058
(18)	*ENVIRONMENT_WP*	Go to word processor environment	987
(19)	*SUBMIT_RESERVATION_FAILURE*	Submit a reservation request unsuccessfully	987
(20)	*SUBMIT_RESERVATION_SUCCESS*	Submit a reservation request successfully	971
(21)	*HISTORY_UNFILLED*	Go to unfilled tab in the web environment	551
(22)	*SUBMIT_UNFILLED*	Submit an unfilled request	414
(23)	*FOLDER*	Do folder-related actions (i.e., create/delete a folder)	332
(24)	*HISTORY_HOME*	Click on the home button in the web environment	244
(25)	*CHANGE_RESERVATION*	Change an existing reservation	227
(26)	*KEYPRESS*	Type in word processor environment	152
(27)	*REPLY*	Reply an email	118
(28)	*CANCEL*	Click on cancel button	111
(29)	*HELP*	Use help function	87
(30)	*COPY*	Use copy function	42
(31)	*SEARCH*	Use search function	38
(32)	*SORT*	Use sort function	21
(33)	*PASTE*	Use paste function	15
(34)	*BOOKMARK*	Do bookmark-related actions (i.e., add/delete a bookmark)	13

TABLE 9 | Top five features of action sequences selected for the 4th and 1st deciles of monthly earning groups.

Group	*N*-gram	Action sequences	Chi-square
4th decile of monthly earning	Unigram	*FOLDER*	39.08
		CANCEL	16.54
		BOOKMARK	7.44
		HISTORY_HOME	4.02
		HELP	1.84
	Bigram	*FOLDER_VIEWED, FOLDER*	24.06
		FOLDER, FOLDER_VIEWED	22.93
		FOLDER, FOLDER	22.67
		MAIL_VIEWED_1, MAIL_VIEWED_3	18.51
		NEXT_INQUIRY, CANCEL	17.64
	Trigram	*FOLDER, FOLDER_VIEWED, FOLDER_VIEWED*	21.87
		MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4	19.70
		FOLDER_VIEWED, FOLDER, FOLDER	18.67
		ENVIRONMENT_MC, MAIL_VIEWED_1, MAIL_VIEWED_3	17.25
		MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_3	16.68
1st decile of monthly earning	Unigram	*SEARCH*	11.72
		COPY	10.90
		KEYPRESS	10.51
		PASTE	5.89
		HISTORY_VIEWCALENDAR	2.81
	Bigram	*ENVIRONMENT_WB, COMBOBOX_START_TIME*	15.78
		COMBOBOX_END_TIME, HISTORY_VIEWCALENDAR	11.20
		HISTORY_RESERVATION, HISTORY_VIEWCALENDAR	10.50
		COPY, KEYPRESS	10.16
		HISTORY_UNFILLED, HISTORY_RESERVATION	9.79
	Trigram	*HISTORY_VIEWCALENDAR, HISTORY_RESERVATION, HISTORY_VIEWCALENDAR*	20.16
		MAIL_VIEWED_3, ENVIRONMENT_WB, HISTORY_UNFILLED	16.25
		ENVIRONMENT_MC, ENVIRONMENT_WB, ENVIRONMENT_MC	16.13
		ENVIRONMENT_MC, ENVIRONMENT_WB, COMBOBOX_START_TIME	15.87
		MAIL_VIEWED_4, ENVIRONMENT_WB, ENVIRONMENT_MC	14.88

next item, and so on. Though cancel actions are sometimes considered hesitative behaviors (He and von Davier, 2015), they could also indicate that the fourth decile group tried different options in the menu to figure out what could be done in the environment.

Other actions that the fourth decile group frequently used were actions associated with bookmarks, clicking the home button in the web environment, and help functions. The bookmarks were accessible via the dropdown menu or a button on the menu bar. Using the bookmark actions, respondents could easily access the pages that they considered important or useful. The home button was right next to the bookmark button on the menu bar, which is a convenient way to return to the main page of the web environment. The help functions were designed in both email and web environments. In the email environment, the help function provided information regarding actions taken for an email, for instance, write, reply, forward, or delete an email. In the web environment, the help function offered instructions on the menu bar items, such as home and bookmark. As expected, the fourth decile group appeared to take more exploratory actions to facilitate their problem-solving process compared to the first decile group.

The unigrams commonly adopted by the first decile group were entirely different. The most discriminating features included search, copy, keypress (pressing a key on the keyboard), paste, and click on the view calendar button. The search function was available in both email and web environments. However, the search function was not required to obtain a correct answer to U02, as the information in the two environments was displayed in short text or tables. The copy, keypress, and paste unigrams were used in the word processor environment solely, where respondents could take notes for the time and location of the meeting room requests and compare to the existing schedules. Similar to search, the three functions only existed to aid the synthesis of available information and conflict schedules. For the view calendar button, respondents used it to retrieve the schedules for each meeting room in a certain time period. Respondents were able to see not only the existing reservations, but also the reservations they made for the meeting room requests.

The lower odds of a correct answer to U02 for the first decile group indicated an association between these functions and lower performance in this group. One explanation for this phenomenon could be that the search function

and word processor environment were rather redundant for high-performing respondents since they could collect and synthesize information more efficiently. Applying such functions might be a sign that respondents were having difficulty in comprehending or solving U02. Additionally, the view calendar button seemed to suggest that respondents in this group were still in the process of figuring out the purpose of the item instead of working on solving the problem.

Compared to the unigrams, the robust bigrams and trigrams were often more closely related for a certain group. The bigrams for the fourth decile mainly involved folder-related actions, email-viewing actions, and cancel actions; the trigrams also contained similar information. The bigram "*FOLDER, FOLDER_VIEWED*" was found in the trigram "*FOLDER, FOLDER_VIEWED, FOLDER_VIEWED*"; the bigram "*MAIL_VIEWED_1, MAIL_VIEWED_3*" were also included in the robust trigram "*MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4*". This is because bigrams with high frequencies were also likely to appear more commonly when started with or followed by another action. Further, while the five robust unigrams tended to provide unique pieces of information, the five bigrams tended to have overlap due to the increase in sequence length, as did the trigrams. For instance, the top three robust bigrams for the fourth decile group were all folder-related actions, whereas three of the top five trigrams were email-viewing actions.

These mini-sequences of the fourth decile group, along with the unigrams, demonstrated evidence that respondents in this group were working on the item and trying to understand the meeting room requests. It is worth noticing that the emails viewed by the fourth decile group were the first, third, and fourth emails (i.e., *MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4*); the second email did not show up in any robust features. In fact, the second email was the only one irrelevant to meeting room requests among the four. Therefore, viewing only the three relevant emails was a strong indication that the respondents at least understood the goal of this item, and were able to filter out emails irrelevant to the goal.

For the first decile group, the respondents did a lot of switching among tabs in the web environment (e.g., *HISTORY_VIEWCALENDAR, HISTORY_RESERVATION, HISTORY_UNFILLED*), or switching among environments (e.g., *ENVIRONMENT_MC, ENVIRONMENT_WB*). Such switching actions indicated that the first decile did not devote much to solving the item. Instead, they seemed to be lost in the item or not interested in exploring more. Results based on unigrams, bigrams, trigrams all suggested that compared to the first decile, respondents in the fourth decile group were more engaged in solving the item. The fourth decile group also adopted more efficient problem-solving strategies, such as bookmark and help. This is consistent with the results from regression analysis that the fourth decile group was more likely to obtain a correct answer to the Meeting Room Assignment item (see **Table 7**).

Differences by Education Subgroups

Another example is the comparison between the robust features from the highest and lowest education groups, as presented in **Table 10**. Respondents in the highest education group obtained tertiary-master/research degrees, whereas the lowest education group obtained lower secondary education or less. The chi-square selection method also identified highly distinctive features for the two groups. The most discriminating unigrams for the highest education group were sorting, submitting filled reservation or unfilled request (i.e., *SUBMIT_RESERVATION_SUCCESS, UNFILLED_SUBMIT*), and filling out the room and the start time for the request (i.e., *COMBOBOX_ROOM, COMBOBOX_START_TIME*).

The sorting function was available in the email environment. Respondents could choose to sort by sender, subject, or receiver of the email. Although sorting was not a necessary step to the success of U02, well-educated respondents might consider sorting by subject as a more efficient approach to identifying the emails related to meeting room requests. The *COMBOBOX*-related actions showed evidence of filling out the details of meeting room requests using the dropdown menus. Then, if the requested room and time had no conflict with the existing schedule, one would receive a notice of submitting the reservation successfully. There was also one meeting room request that could not be filled given the current schedule, which needed to be recorded as well. *UNFILLED_SUBMIT* indicated that the test taker also submitted the unfilled request. Such actions were key to the correctness of the Meeting Room Assignment item, because one had to fill out the details of each room request and submit at least one reservation or unfilled request successfully to answer it correctly.

The lowest education group, however, mainly used redundant functions. Moving emails, viewing folder, pasting, copying, and searching were the most important unigrams, which coincidently were found as robust unigrams in the first decile monthly earning group as well. Both the lowest education group and first decile monthly earning group had lower performance on U02 compared with their peers. This finding suggested that lower-performing respondents might be prone to using these unnecessary functions (as defined by content experts), indicating they were unable to figure out a solution.

The robust bigrams and trigrams for the highest education group encompassed some action sequences that also related to filling and submitting the requests, as well as viewing emails, which were required procedures to obtain a correct answer. Some features indicating switching among tabs or environments also appeared. Though we interpreted similar actions for the first decile group as signs of low motivation, these actions could have different meanings for another group. When combined with other robust features for the highest education group, these actions served as connections among necessary steps to finish the item, such as filling in comboboxes and submitting requests. Therefore, the highest education group did not wander around aimlessly, but in fact attempted to synthesize information from multiple environments and make a successful reservation.

Email-moving and folder-viewing actions manifested themselves again in the robust bigrams and trigrams for the lowest education group. These action sequences identified by chi-square selection method demonstrated a clear distinction between the problem-solving processes of the two groups with

TABLE 10 | Top five features of action sequences selected for the highest and lowest education groups.

Group	*N*-gram	Action sequences	Chi-square
Tertiary – master/research degree	Unigram	*SORT*	14.04
		SUBMIT_RESERVATION_SUCCESS	7.24
		COMBOBOX_ROOM	6.88
		UNFILLED_SUBMIT	6.82
		COMBOBOX_START_TIME	6.22
	Bigram	*COMBOBOX_END_TIME, COMBOBOX_DEPT*	20.36
		ENVIRONMENT_WB, ENVIRONMENT_MC	17.06
		ENVIRONMENT_MC, MAIL_VIEWED_1	16.84
		SUBMIT_RESERVATION_SUCCESS, HISTORY_UNFILLED	16.73
		HISTORY_MEETINGROOMS, ENVIRONMENT_MC	16.57
	Trigram	*ENVIRONMENT_MC, MAIL_VIEWED_1, MAIL_VIEWED_2*	26.36
		HISTORY_RESERVATION, HISTORY_MEETINGROOMS, HISTORY_VIEWCALENDAR	23.26
		MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4	20.73
		COMBOBOX_END_TIME, COMBOBOX_DEPT, SUBMIT_RESERVATION_SUCCESS	19.96
		COMBOBOX_START_TIME, COMBOBOX_END_TIME, COMBOBOX_DEPT	19.35
Lower secondary or less	Unigram	*MAIL_MOVE*	197.12
		FOLDER_VIEWED	24.15
		PASTE	9.77
		COPY	7.73
		SEARCH	7.25
	Bigram	*FOLDER_VIEWED, MAIL_MOVE*	159.17
		MAIL_MOVE, FOLDER_VIEWED	156.81
		MAIL_VIEWED_3, MAIL_MOVE	104.55
		MAIL_VIEWED_4, MAIL_MOVE	90.67
		MAIL_MOVE, MAIL_VIEWED_4	90.10
	Trigram	*MAIL_MOVE, FOLDER_VIEWED, MAIL_MOVE*	148.51
		FOLDER_VIEWED, MAIL_MOVE, FOLDER_VIEWED	95.11
		MAIL_VIEWED_3, MAIL_MOVE, FOLDER_VIEWED	92.88
		FOLDER_VIEWED, MAIL_MOVE, MAIL_VIEWED_3	92.17
		MAIL_VIEWED_4, MAIL_MOVE, FOLDER_VIEWED	86.50

different education levels. While the highest education group was completing the item with clear subgoals, the lowest education group spent much time and effort moving the emails around and viewing the folder. As a result, these discriminating features identified from the action sequences were in fact strongly associated with the performance on the item.

Differences by Other Background Variables

Some general findings from other significant variables resembled the results from the two discussed examples. As presented above, higher income, higher level on the index variables (except for TASKDISC_WLE_CA, index of use of task discretion at work), and higher educational level were associated with higher probability of answering the Meeting Room Assignment item correctly. A younger respondent who took the test in the same language as his or her native language was also more likely to obtain a correct answer. Some background variables have more than one significant dummy variables, such as age and education. It is worth noticing that the features selected for the reference group did not need to be the same when the focal group changed since chi-square chose features that can best distinguish the reference and the focal groups.

Overall, groups with higher odds of a correct answer were likely to adopt the actions related to *SUBMIT* (submitting filled reservation or unfilled request), *COMBOBOX* (filling out the room and the start time for the request), help, and sort. Help and sort are two actions that might be indicative of more efficient problem-solving strategies. To complete the room requests in this item, respondents had to fill time slots for a specific room in the *COMBOBOX* and use one of the two submit buttons. These respondents demonstrated evidence that they went through the necessary steps to obtain correct answers to the Meeting Room Assignment item.

Groups with lower odds of a correct answer, however, used more actions such as *MAIL_MOVE* (moving email) and *SUBMIT_FAILURE* (failure to submit a room request). The occurrence of *MAIL_MOVE* and *SUBMIT_FAILURE* did not always mean that a respondent had trouble finishing an item. A respondent could have been categorizing emails, so he or she could discard those emails that were irrelevant to room requests. If *SUBMIT_FAILURE* was followed by some adjustments in *COMBOBOX* and *SUBMIT_RESERVATION_SUCCESS*, then the respondent made two attempts to submit a reservation and did self-correction. It is when the two actions appeared in the selected

features predominantly, and not accompanied by other useful actions, that they might not be able to solve the item.

For some significant dummy variables, *COMBOBOX*-related actions were in fact identified as robust features for the group with lower odds of a correct answer (e.g., INFLUENCE_WLE_CA and READWORK_WLE_CA, or lowest 20% on index of influencing skills at work, and lowest 20% on index of reading skills at work, respectively), while for others, the selected features were mainly associated with *MAIL_MOVE*. Adopting *COMBOBOX*-related actions could be a sign of understanding the purpose of the item and being able to figure out how to fill out the room requests. These respondents were considered closer to the borderline of a correct answer than the group with mostly *MAIL_MOVE* actions and might have had greater potential to get a score of 1 if proper interventions were given. On the contrary, if the majority of a respondent's actions were *MAIL_MOVE*, he or she might have needed more detailed guidance from the initial steps to submitting the requests.

An intriguing finding is that for the lowest age group (24 or less), the *MAIL_MOVE* action showed up in the top five robust features quite frequently, even though this group was more likely to answer correctly to the Meeting Room Assignment item compared to elder age groups. That is to say, given two respondents with the same occupation, work experience, work-related skills, and so on, the one who was 24 years old or younger would have had a higher probability of a correct response than the one who was 45 to 54, or 55 or older. However, the lowest age group often had different occupations and much less work experience than respondents who were 45 and above. The skills and experiences that the older age groups had accumulated might have enabled them to apply more efficient problem-solving strategies despite younger respondents having more advantage on information technologies. Another possible explanation is that using *MAIL_MOVE* was characteristic of the youngest age group as an action taken without realizing it. They could simply have been moving emails around as they went through the thinking process.

DISCUSSION

This study aimed at exploring the relationship between sequential problem-solving actions and employment-related variables, and identified the key features for respondents with different levels of employment-related variables. We focused on the data from BQ and log files for the United States population on one representative PSTRE item, the Meeting Room Assignment item, in the main study of 2012 PIAAC. The study was conducted in two phases: (a) use of regression analyses to identify background variables having significant associations with PIAAC performance, and (b) application of chi-square selection method to select robust features of the significantly different groups.

In general, most significant variables and their regression coefficients were consistent with our expectations. Respondents who were well-educated and young, and had more work experience and higher work-related skill use, tended to have higher scores in the three domains and higher odds of success in the Meeting Room Assignment item. Comparing scores in the three domains, the significant variables for literacy and numeracy scores were more similar. PSTRE scores had one unique significant variable—READWORK_WLE_CA (lowest 20% on index of use of reading skills at work)—indicating that PSTRE items might require higher-level reading skill use at work to understand the item structure, follow the instructions, and browse the website.

We further explored the process data to investigate what action sequences were associated with the variables that were significantly related to success in the Meeting Room Assignment item. Based on the final goal of submitting meeting room requests, there were five necessary steps in the problem-solving process for the studied item: (a) read emails; (b) choose the emails related to meeting room requests; (c) synthesize information from multiple environments; (d) determine the requests that could or could not be filled; (e) and submit filled reservations and unfilled requests. Similar to what He and von Davier (2015, 2016) found, respondents who had higher income, work-related skill use, and education level demonstrated clear subgoals in solving the item. For instance, respondents with higher income performed more *MAIL_VIEWED* actions; they were also able to focus on emails directly related to meeting room requests. *SUBMIT* and *COMBOBOX* actions were commonly applied by those with higher work-related skill use at work. Respondents with high education level and high writing skill use at work tended to use more sorting actions.

Some key actions were found more often in the groups with higher income and work-related skill use. Such group were generally prone to adopt *SUBMIT_RESERVATION_SUCCESS, UNFILLED_SUBMIT*, and actions related to *COMBOBOX*, *HELP*, and *SORT*. These actions demonstrated evidence that the respondents went through necessary steps to fulfill room requests in this item. Groups with lower income and work-related skill use, however, took more actions such as *MAIL_MOVE* and *SUBMIT_FAILURE*, which were either redundant or an indication of failing to complete a request.

The most important implication of the present study was that features identified from process data shed light on how much intervention a certain group of respondents might need. There was clear evidence from process data for the steps to read emails, filter the irrelevant email, and submit requests. For instance, respondents who adopted *COMBOBOX*-related actions but still failed to solve the item may have already mastered the majority of required PSTRE skills. Therefore simple instructions on the final steps might be sufficient for them to obtain a correct answer. In contrast, *MAIL_MOVE* and *FOLDER* could be a sign that the respondents needed more comprehensive guidance and training on PSTRE skill. However, evidence for synthesizing information and addressing conflicts were not as traceable. Given sufficient evidence for each required step, further analyses could potentially determine at which specific step an intervention was needed. It also provides the possibility of scoring complex items like PSTRE items base on process data in the future.

Overall, groups with different levels of background variables often demonstrated quite distinctive characteristics with respect to test-taking behaviors. Actions indicative of low PSTRE skill for

one group may not mean the same for another group. Therefore, it is important to establish a basic understanding of the common action sequences that a group would take before making decisions on the necessary training and interventions.

When interpreting the robust features identified from process data, it is recommended that one considers unigrams, bigrams, and trigrams simultaneously. This would provide a more holistic view of the respondents' problem-solving strategies. One example of this is the sequential action of switching among environments. This action could be indicative of aimless behaviors if it was predominant; it could also be the transition among required steps, such as reading emails and submitting requests, if a wide range of features appeared. Therefore, the diversity of the robust features was found informative regarding the interpretation of action sequences.

Despite innovations in this study, at least four limitations are worth mentioning. First, we restricted this study to United States respondents only. Findings related to test-taking behaviors and culture effects that might be learned from other countries were not taken into consideration. Nonetheless, the proposed research plan is applicable to data from other countries. Researchers may compare patterns and action sequences extracted from other countries to those from the United States sample to obtain further insights regarding cross-country differences.

Second, the study focused on process data from the PSTRE domain only. Considering the respondents who had scores in all three domains in the BQ dataset, the correlations between PSTRE scores and literacy/numeracy scores are about 0.81 and 0.76, respectively, for the United States in the 2012 PIAAC assessment (Organisation for Economic Co-operation and Development [OECD], 2013a, p. 7, Chapter 18). Given the strong correlations, the associations between respondents' sequential action patterns in PSTRE and other domains could be evaluated in future studies.

Third, we used the method suggested by Cohen and Cohen (1985) to deal with missing responses in the BQ, where missing responses were coded as another category for each variable. This method was employed in the present study to retain all available information when missingness occurred in the independent variables (Howell, 2008) and when the missing proportion was higher than 5% or 10% (Schafer, 1999; Bennett, 2001). However, the interpretability of the results becomes a problem (Howell, 2008). Some researchers have also found that this method may produce biased estimates for the regression coefficients under some circumstances, even though it produced reasonably accurate standard error estimates (Jones, 1996; Allison, 2001). Though comparing different approaches to dealing with missing data was not the focus of this study, more advanced methods might be considered in future studies, such as maximum likelihood and multiple imputation (e.g., Bennett, 2001; Howell, 2008).

Lastly, the present study investigated the sequential patterns for different subgroups on only one representative PSTRE item. As the action sequences in process data are highly context dependent (Rupp et al., 2010), the findings from this study need to be cross-validated using other items in a similar context. PSTRE items that share environments with U02 could be further explored to shed light on the consistency of problem-solving strategies across multiple items.

To summarize, this study provides critical evidence of relationships between employment-related background variables and sequential patterns in PSTRE using one example item based on the United States sample in PIAAC. It also provides information to education policy makers to find reasons for success and failure by different employment-related subgroups, thus helping to find an optimal solution to improve their PSTRE skills via a tailored approach. Such information would be key to improving adults' lifelong learning strategies. Further explorations have been done on multiple items, and similar patterns have been observed, but results were not included to avoid distracting from the main theme of the present study. We recommend to continue exploring the generalizability of results presented in this study across PSTRE items in future studies and to make comparisons across countries and language groups.

AUTHOR CONTRIBUTIONS

DL works on forming the research idea, conducting analyses, interpreting results, and write-up of the study. QH works on forming the research idea, interpreting results, and revising the manuscript. HJ works on revising the manuscript, providing suggestions and guidance regarding the overall technical quality of this study.

ACKNOWLEDGMENTS

The authors would like to thank the PIAAC group at American Institutes for Research (AIR) for the helpful discussions, Larry Hanover for his help in reviewing this article, and the joint funding support of the National Center for Education Statistics of the United States Department of Education, AIR PIAAC team, and National Science Foundation (NSF-1633353).

REFERENCES

Allison, P. D. (2001). *Missing Data: Sage University Papers Series on Quantitative Applications in the Social Sciences (07–136)*. Thousand Oaks, CA: Sage Publications, Inc.

Baker, R. S. (2007). "Modeling and understanding students' off-task behavior in intelligent tutoring systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, eds M. B. Rosson, D. J. Gilmore (San Jose, CA: The Interaction Design Foundation), 1059–1068. doi: 10.1145/1240624.1240785

Baker, R. S., Corbett, A. T., and Aleven, V. (2008). "Improving contextual models of guessing and slipping with a truncated training set," in *Proceedings of the 1st International Conference on Educational Data Mining* (Montreal, QC: EDM), 67–76.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* 25, 464–469. doi: 10.1111/j.1467-842X.2001.tb00294.x

Carpenter, J., and Goldstein, H. (2004). Multiple imputation in MLwiN. *Multilevel Modell. Newsl.* 16, 9–18.

Cohen, J., and Cohen, P. (1985). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences*, 2nd Edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.* 4, 253–278. doi: 10.1007/BF01099821

de Klerk, S., Veldkamp, B. P., and Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: a systematic review and a Bayesian network example. *Comput. Educ.* 85, 23–34. doi: 10.1016/j.compedu.2014.12.020

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., and Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Model. User-Adapt. Interact.* 18, 45–80. doi: 10.1007/s11257-007-9037-6

Feinerer, I. (2017). *Introduction to the TM Package Text Mining in R*. Available at: https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Fox, J.-P., and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *J. Educat. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Goodman, M., Sands, A., and Coley, R. (2015). *America's Skills Challenge: Millennials and the Future*. Princeton, NJ: Educational Testing Service.

He, Q., Glas, C. A., Kosinski, M., Stillwell, D. J., and Veldkamp, B. P. (2014). Predicting self-monitoring skills using textual posts on Facebook. *Comput. Hum. Behav.* 33, 69–78. doi: 10.1016/j.chb.2013.12.026

He, Q., Ling, H. K., Liu, J., and Ying, Z. (2018). Exploring relationship between sequence patterns in solving digital tasks and background variables: an empirical study using log data in PIAAC. *Paper presented at the 11th International Test Commission Conference*, Montreal, Canada.

He, Q., Veldkamp, B. P., and de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: a text mining approach. *Psychiatry Res.* 198, 441–447. doi: 10.1016/j.psychres.2012.01.032

He, Q., Veldkamp, B. P., Glas, C. A. W., and de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment* 24, 157–172. doi: 10.1177/1073191115602551

He, Q., and von Davier, M. (2015). "Identifying feature sequences from process data in problem-solving items with n-grams," in *Quantitative Psychology Research: Proceedings of the 79th Annual Meeting of the Psychometric Society*, eds A. van der Ark, D. Bolt, S. Chow, J. Douglas, and W. Wang (New York, NY: Springer), 173–190. doi: 10.1007/978-3-319-19977-1_13

He, Q., and von Davier, M. (2016). "Analyzing process data from problem-solving items with n-grams: insights from a computer-based large-scale assessment," in *Handbook of Research on Technology Tools for Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf (Hershey, PA: Information Science Reference), 749–776. doi: 10.4018/978-1-4666-9441-5.ch029

Horton, N. J., and Kleinman, K. P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am. Statist.* 61, 79–90. doi: 10.1198/000313007X172556

Howell, D. C. (2008). "The analysis of missing data," in *Handbook of Social Science Methodology*, eds W. Outhwaite and S. Turner (London: Sage).

Hurst, K., Casillas, A., and Stevens, R. H. (1997). *Exploring the Dynamics of complex Problem-Solving with Artificial Neural Network-Based Assessment Systems*. CSE Technical Report No. 444. Los Angeles, CA: University of California, Los Angeles.

International Association for the Evaluation of Educational Achievement [IEA]. (2013). *IDB Analyzer (Computer Software and Manual)*. Hamburg: Author.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *Mach. Learn. ECML-98 Lect. Notes Comput. Sci.* 1398, 137–142. doi: 10.1007/BFb0026683

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J. Am. Stat. Assoc.* 91, 222–230. doi: 10.1080/01621459.1996.10476680

Klein Entink, R. H. (2009). *Statistical Models for Responses and Response Times*. Doctoral dissertation, University of Twente, Enschede.

Klein Entink, R. H., Fox, J. P., and van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Manning, C. D., and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Vol. 999. Cambridge, MA: MIT Press.

Menard, S. (1995). *Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage.

Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *Am. Statist.* 58, 218–223. doi: 10.1198/000313004X946

Nathans, L. L., Oswald, F. L., and Nimon, K. (2012). Interpreting multiple linear regression: a guidebook of variable importance. *Pract. Assess. Res. Eval.* 17, 1–19.

Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, W. A. V., and Beaulieu, M. (2001). "A method based on chi-square test for document classification," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY: ACM), 440–441. doi: 10.1145/383952.384080

Organisation for Economic Co-operation and Development [OECD] (2013a). *Technical Report of the Survey of Adult Skills (PIAAC)*. Paris: OECD Publishing. Available at: http://www.oecd.org/skills/piaac/_technical%20report_17oct13.pdf

Organisation for Economic Co-operation and Development [OECD] (2013b). *Time for the U.S. to Reskill?* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD] (2012). *Literacy, Numeracy and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development [OECD] (2016). *Skills Matter: Further Results from the Survey of Adult Skills*. Paris: OECD Publishing.

Pavlik, P. I. Jr, Cen, H., and Koedinger, K. R. (2009). "Learning factors transfer analysis: using learning curve analysis to automatically generate domain models," in *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, 121–130.

Perry, A., Wiederhold, S., and Ackermann-Piek, D. (2016). How can skill mismatch be measured? New approaches with PIAAC. *Methods Data Anal.* 8, 137–174.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rampey, B. D., Finnegan, R., Goodman, M., Mohadjer, L., Krenzke, T., Hogan, J., et al. (2016). *Skills of U.S. Unemployed, Young, and Older Adults in Sharper Focus: Results from the Program for the International Assessment of Adult Competencies (PIAAC) 2012/2014: First look, NCES Report No. 2016–039. U.S. Department of Education*. Washington, DC: National Center for Education Statistics.

Roberts, R. D., and Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: toward a taxonomic model. *Learn. Ind. Diff.* 11, 1–12. doi: 10.1016/S1041-6080(00)80007-2

Rupp, A. A., Gushta, M., Mislevy, R. J., and Shaffer, D. W. (2010). Evidence-centered design of epistemic games: measurement principles for complex learning environments. *J. Technol. Learn. Assess.* 8, 1–47.

SAS Institute (2015). *Base SAS 9.4 Procedures Guide*. Cary, NC: SAS Institute.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statist. Methods Med. Res.* 8, 3–15. doi: 10.1177/096228029900800102

Scheuer, O., and McLaren, B. M. (2011). "Educational data mining," in *Encyclopedia of the Sciences of Learning*, ed. N. M. Seel (New York, NY: Springer).

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *J. Document.* 28, 11–21. doi: 10.1108/eb026526

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

Vanek, J. (2017). *Using the PIAAC Framework for Problem Solving in Technology-Rich Environments to Guide Instruction: An Introduction for Adult Educators*. Available at: https://piaac.squarespace.com/s/PSTRE_Guide_Vanek_2017.pdf

Vendlinski, T., and Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *J. Technol. Learn. Assess.* 1, 1–20.

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Mathemat. Statist. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wollack, J. A., Cohen, A. S., and Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *J. Educ. Measure.* 40, 307–330. doi: 10.1111/j.1745-3984.2003.tb01149.x

Zientek, L. R., Capraro, M. M., and Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: one look at the evidence cited in the AERA panel report. *Educ. Res.* 37, 208–216. doi: 10.3102/0013189X08319762

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Liao, He and Jiao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

TABLE A1 | Description of independent variables in the regression analyses.

No.	Variable	Category	Description
(1)	PAIDWORK	1	No work ever
		2	Have had paid work but not in 5 years
		3	Have had paid work in 5 years but not in 12 months
		4	Have had paid work in 12 months
(2)	C_D05	1	Employed
		2	Unemployed
		3	Out of the labor force
(3)	D_Q04_T1	1	Employee, not supervisor
		2	Employee, supervising fewer than 5 people
		3	Employee, supervising more than 5 people
		4	Self-employed or unpaid family worker
(4)	D_Q08b	1	1–5 people
		2	6–10 people
		3	11–24 people
		4	25–99 people
		5	100 or more people
(5)	D_Q12b	1	Current level is necessary
		2	A lower education level would be sufficient
		3	A higher education level would be needed
(6)	D_Q12c_RC	1	No experience
		2	Less than 1 year of relevant work experience
		3	1 or 2 years
		4	More than 3 years of relevant work experience
(7)	F_Q05a	1	Never solve simple problems
		2	Solve simple problems less than once a month
		3	Solve simple problems less than once a week but at least once a month
		4	Solve simple problems at least once a week but not everyday
		5	Solve simple problems everyday
(8)	F_Q05b	1	Never solve complex problems
		2	Solve complex problems less than once a month
		3	Solve complex problems less than once a week but at least once a month
		4	Solve complex problems at least once a week but not everyday
		5	Solve complex problems every day
(9)	F_Q07b	1	Need more training for skill use at work
		2	Do not need more training for skill use at work
(10)	G_Q06	1	Straightforward computer use
		2	Moderate computer use
		3	Complex computer use
(11)	ISCOSKIL4	1	Skilled occupations
		2	Semi-skilled white-collar occupations
		3	Semi-skilled blue-collar occupations
		4	Elementary occupations
(12)	EARNMTHALLDCL	1	1st decile of monthly earning
		2	2nd decile of monthly earning
		3	3rd decile of monthly earning
		4	4th decile of monthly earning
		5	5th decile of monthly earning
		6	6th decile of monthly earning
		7	7th decile of monthly earning
		8	8th decile of monthly earning

(Continued)

TABLE A1 | Continued

No.	Variable	Category	Description
		9	9th decile of monthly earning
		10	10th decile of monthly earning
(13)	LEARNATWORK_WLE_CA	0	All zero response
		1	Lowest 20% on index of learning at work
		2	More than 20–40% on index of learning at work
		3	More than 40–60% on index of learning at work
		4	More than 60–80% on index of learning at work
		5	More than 80% on index of learning at work
(14)	ICTWORK_WLE_CA	0	All zero response
		1	Lowest 20% on index of use of ICT skills at work
		2	More than 20–40% on index of use of ICT skills at work
		3	More than 40–60% on index of use of ICT skills at work
		4	More than 60–80% on index of use of ICT skills at work
		5	More than 80% on index of use of ICT skills at work
(15)	INFLUENCE_WLE_CA	0	All zero response
		1	Lowest 20% on index of use of influencing skills at work
		2	More than 20–40% on index of use of influencing skills at work
		3	More than 40–60% on index of use of influencing skills at work
		4	More than 60–80% on index of use of influencing skills at work
		5	More than 80% on index of use of influencing skills at work
(16)	NUMWORK_WLE_CA	0	All zero response
		1	Lowest 20% on index of use of numeracy skills at work
		2	More than 20–40% on index of use of numeracy skills at work
		3	More than 40–60% on index of use of numeracy skills at work
		4	More than 60–80% on index of use of numeracy skills at work
		5	More than 80% on index of use of numeracy skills at work
(17)	PLANNING_WLE_CA	0	All zero response
		1	Lowest 20% on index of use of planning skills at work
		2	More than 20–40% on index of use of planning skills at work
		3	More than 40–60% on index of use of planning skills at work
		4	More than 60–80% on index of use of planning skills at work
		5	More than 80% on index of use of planning skills at work
(18)	READWORK_WLE_CA	0	All zero response
		1	Lowest 20% on index of use of reading skills at work
		2	More than 20–40% on index of use of reading skills at work
		3	More than 40–60% on index of use of reading skills at work
		4	More than 60–80% on index of use of reading skills at work
		5	More than 80% on index of use of reading skills at work
(19)	TASKDISC_WLE_CA	0	All zero response
		1	Lowest 20% on index of use of task discretion at work
		2	More than 20–40% on index of use of task discretion at work
		3	More than 40–60% on index of use of task discretion at work
		4	More than 60–80% on index of use of task discretion at work
		5	More than 80% on index of use of task discretion at work
(20)	WRITWORK_WLE_CA	0	All zero response
		1	Lowest 20% on index of use of writing skills at work
		2	More than 20–40% on index of use of writing skills at work
		3	More than 40–60% on index of use of writing skills at work
		4	More than 60–80% on index of use of writing skills at work
		5	More than 80% on index of use of writing skills at work
(21)	AGEG10LFS	1	24 or less
		2	25–34
		3	35–44

(Continued)

TABLE A1 | Continued

No.	Variable	Category	Description
		4	45–54
		5	55 plus
(22)	PARED	1	Neither parent has attained upper secondary
		2	At least one parent has attained secondary and postsecondary, non-tertiary
		3	At least one parent has attained tertiary
(23)	NATIVELANG	0	Test language different from native language
		1	Test language same as native language
(24)	EDCAT6	1	Lower secondary or less (ISCED 1, 2, 3C short or less)
		2	Upper secondary (ISCED 3A-B, C long)
		3	Postsecondary, non-tertiary (ISCED 4A-B-C)
		4	Tertiary – professional degree (ISCED 5B)
		5	Tertiary – bachelor degree (ISCED 5A)
		6	Tertiary – master/research degree (ISCED 5A/6)
(25)	GENDER_R	1	Male
		2	Female
(26)	J_Q08	1	10 books or less
		2	11–25 books
		3	26–100 books
		4	101–200 books
		5	201–500 books
		6	More than 500 books

TABLE A2 | Difference in percentages of each category of background variables between the whole sample and the sample with U02 response.

	Levels											
Variables	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Missing**
PAIDWORK		1.35	1.42	−0.62	−4.40							2.25
C_D05		−4.86	−1.44	4.08	0.04							2.18
D_Q04_T1		0.13	−1.17	−1.49	−1.81							4.34
D_Q08b		−1.22	0.11	−0.86	−0.51	−0.33						2.81
D_Q12b		−3.28	−0.03	−0.15								3.47
D_Q12c		−0.76	−1.97	−0.46	0.13							
F_Q05a		1.97	0.45	0.38	−2.88	−4.83						4.92
F_Q05b		0.75	−1.02	−1.52	−1.70	−1.38						4.87
F_Q07b		−0.83	−4.05									4.88
G_Q06		−2.61	−4.92	−0.02								7.54
ISCOSKIL4		−5.49	−0.85	0.61	0.73							5.00
EARNMTHALLDCL		−0.44	−0.99	−0.13	−0.74	0.46	−0.32	0.63	−0.19	−1.32	−1.27	4.33
LEARNATWORK_WLE_CA	0.68	0.75	0.09	−2.12	−1.25	−1.54						3.39
ICTWORK_WLE_CA	0.71	−1.79	−0.70	−2.18	−0.95	−2.47						7.38
INFLUENCE_WLE_CA	0.58	1.33	0.21	−2.04	−1.86	−2.88						4.68
NUMWORK_WLE_CA	1.89	−0.91	−1.03	−1.14	−1.32	−2.17						4.68
PLANNING_WLE_CA	1.59	−0.40	−1.29	0.02	−1.40	−3.19						4.68
READWORK_WLE_CA	0.36	0.80	0.78	−1.63	−2.94	−2.05						4.68
TASKDISC_WLE_CA	0.75	0.10	−1.85	−3.27	0.03	−0.48						4.72
WRITWORK_WLE_CA	1.75	0.75	−1.38	−1.65	−1.15	−2.98						4.68
AGEG10LFS		−2.18	−1.34	−1.85	2.08	3.30						
PARED		4.18	−3.25	−4.49								3.57
NATIVELANG	3.05	−5.25										2.20
EDCAT6		3.22	−0.04	−0.79	−0.69	−2.24	−1.63					2.17
GENDER_R		−0.51	0.51									
J_Q08		3.48	−1.41	−1.00	−2.16	−0.97	−0.24					2.29

TABLE A3 | Standardized regression coefficients in the regression analyses.

No.	Variable	Category	PSTRE	Literacy	Numeracy	U02RT	U02score
(1)	PAIDWORK	2	0.02	0.02	0.04	0.05	0.46
		3	0.12	0.13*	0.15*	−0.14	0.41
		4	0.12	0.21	0.26*	−0.33	−0.17
(2)	C_D05	2	< 0.01	0.13	0.14	0.13	0.17
		3	0.06	0.23	0.26	0.11	0.15
(3)	D_Q04_T1	2	−0.05	0.02	0.02	−0.14	−0.45
		3	−0.11	−0.09	−0.09	< 0.01	1.91
		4	−0.09	−0.02	−0.05	< 0.01	−0.46
(4)	D_Q08b	2	0.02	0.07	0.07	−0.16	−0.54
		3	0.07	0.07	0.08	−0.12	−0.45
		4	0.04	0.04	0.07	−0.10	−0.38
		5	0.04	0.04	0.05	−0.07	−0.20
(5)	D_Q12b	2	0.05*	0.03*	0.03	0.01	0.03
		3	−0.05*	−0.04*	−0.04*	−0.03	−0.05
(6)	D_Q12c_RC	2	−0.04*	−0.03	−0.04*	−0.02	−0.14
		3	< 0.01	< 0.01	−0.01	< 0.01	< 0.01
		4	< 0.01	0.02	0.04*	0.04	< 0.01
(7)	F_Q05a	2	0.02	0.04*	0.04	0.03	0.41
		3	0.08*	0.06*	0.08*	0.04	0.37
		4	0.12*	0.11*	0.12*	0.04	0.76
		5	0.19*	0.18*	0.17*	0.12	0.90
(8)	F_Q05b	2	< 0.01	−0.02	< 0.01	−0.01	−0.63
		3	< 0.01	−0.03	−0.02	−0.01	−0.53
		4	−0.01	−0.02	< 0.01	< 0.01	−0.61
		5	−0.02	−0.07*	−0.05*	0.03	−0.44
(9)	F_Q07b	2	0.04	0.06*	0.06*	−0.02	0.04
(10)	G_Q06	2	0.11*	0.07*	0.07*	0.09*	0.64
		3	0.08*	0.05*	0.06*	0.08*	0.33
(11)	ISCOSKIL4	2	−0.04*	−0.04*	−0.04*	−0.01	0.09
		3	−0.06*	−0.05*	−0.04*	−0.02	0.01
		4	−0.03	−0.07*	−0.07*	−0.03	< 0.01
(12)	EARNMTHALLDCL	2	−0.02	< 0.01	< 0.01	−0.06	< 0.01
		3	−0.04	−0.04	−0.03	−0.06	−0.07
		4	−0.01	−0.01	< 0.01	< 0.01	0.09*
		5	−0.03	< 0.01	< 0.01	−0.06	−0.07
		6	−0.02	< 0.01	< 0.01	−0.08	−0.03
		7	< 0.01	0.03	0.03	−0.06	< 0.01
		8	< 0.01	0.02	0.02	−0.06	< 0.01
		9	0.02	0.04*	0.03	−0.08	0.02
		10	0.06*	0.06*	0.05*	−0.05	0.04
(13)	LEARNATWORK_WLE_CA	1	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
		2	< 0.01	< 0.01	< 0.01	−0.03	0.07
		3	−0.03	−0.01	< 0.01	−0.02	−0.02
		4	−0.04	−0.02	−0.02	−0.03	−0.03
		5	−0.07*	−0.08*	−0.07*	−0.05	−0.03
(14)	ICTWORK_WLE_CA	1	0.05	0.07*	0.06*	0.01	−0.28
		2	0.09*	0.09*	0.06*	< 0.01	−0.20
		3	0.12*	0.08*	0.06*	0.10	−0.16
		4	0.13*	0.07*	0.06*	0.07	−0.10
		5	0.17*	0.11*	0.10*	0.05	−0.15
(15)	INFLUENCE_WLE_CA	1	0.04	−0.01	< 0.01	0.02	0.08*
		2	0.06	−0.02	< 0.01	< 0.01	< 0.01
		3	0.03	−0.04	−0.02	< 0.01	< 0.01

(Continued)

TABLE A3 | Continued

No.	Variable	Category	PSTRE	Literacy	Numeracy	U02RT	U02score
		4	0.03	−0.04	−0.04	−0.06	−0.04
		5	0.03	−0.09*	−0.07*	−0.06	−0.05
(16)	NUMWORK_WLE_CA	1	0.04	0.01	0.01	0.04	< 0.01
		2	0.07*	0.04*	0.05*	0.04	< 0.01
		3	0.04	0.03	0.04*	0.06	< 0.01
		4	0.08*	0.06*	0.07*	0.09*	< 0.01
		5	0.06*	0.03	0.07*	0.06	< 0.01
(17)	PLANNING_WLE_CA	1	< 0.01	0.04	0.04	0.05	−0.03
		2	< 0.01	0.01	0.03	0.03	< 0.01
		3	< 0.01	0.02	0.03	0.05	< 0.01
		4	0.03	0.07*	0.08*	0.10*	0.04
		5	< 0.01	0.04	0.04	0.05	−0.03
(18)	READWORK_WLE_CA	1	0.02	0.03	0.02	0.06	0.14*
		2	< 0.01	0.02	0.02	0.02	0.04
		3	−0.05	0.01	0.01	< 0.01	−0.07
		4	−0.07	< 0.01	< 0.01	0.05	−0.02
		5	−0.14*	−0.02	−0.02	0.04	−0.09
(19)	TASKDISC_WLE_CA	1	−0.02	0.03	0.05	−0.05	< 0.01
		2	< 0.01	0.05	0.05	−0.09	−0.03
		3	< 0.01	0.06	0.05	−0.11	−0.09
		4	< 0.01	0.05	0.02	−0.07	−0.15*
		5	0.01	0.05	0.02	−0.11	−0.05
(20)	WRITWORK_WLE_CA	1	< 0.01	0.02	0.01	−0.05	−0.10
		2	0.01	0.04	0.02	< 0.01	< 0.01
		3	< 0.01	0.05*	0.03	< 0.01	< 0.01
		4	0.03	0.06*	0.04	< 0.01	0.10*
		5	< 0.01	< 0.01	−0.03	−0.04	0.03
(21)	AGEG10LFS	2	−0.16*	−0.11*	−0.09*	−0.07	0.05
		3	−0.23*	−0.14*	−0.12*	−0.05	0.01
		4	−0.30*	−0.19*	−0.15*	−0.04	−0.10*
		5	−0.31*	−0.19*	−0.14*	0.02	−0.13*
(22)	PARED	2	0.12*	0.10*	0.11*	0.11*	0.05
		3	0.14*	0.15*	0.13*	0.18*	0.09
(23)	NATIVELANG	1	0.10*	0.13*	0.06*	0.05	0.04*
(24)	EDCAT6	2	0.11*	0.17*	0.18*	0.12*	0.55*
		3	0.09*	0.11*	0.14*	0.06	0.33*
		4	0.12*	0.16*	0.18*	0.03	0.28*
		5	0.21*	0.27*	0.29*	0.12*	0.49*
		6	0.21*	0.28*	0.30*	0.12*	0.48*
(25)	GENDER_R	2	−0.05*	< 0.01	−0.11*	< 0.01	−0.03
(26)	J_Q08	2	0.04	0.05*	0.02	−0.03	0.22
		3	0.11*	0.09*	0.12*	< 0.01	0.29
		4	0.12*	0.1*	0.14*	−0.04	0.27
		5	0.16*	0.12*	0.14*	0.04	0.30
		6	0.08*	0.09*	0.09*	< 0.01	0.20

Significant variables are marked with asterisks. As the logit of the probability of U02score = 1 could range from negative to positive infinity, coefficients for U02score were standardized using $\pi/\sqrt{3}$ *as an approximate for the standard deviation of the dependent variable, as implemented in SAS (e.g., Menard, 1995, 2004).*

TABLE A4 | Top five features of action sequences selected for the groups with lowest 20% or all zero response on INFLUENCE_WLE_CA.

Group	*N*-gram	Action sequences	Chi-square
Lowest 20% on INFLUENCE_WLE_CA	Unigram	*COPY*	6.46
		REPLY	5.01
		SEARCH	4.56
		HELP	3.38
		PASTE	3.18
	Bigram	*SUBMIT_RESERVATION_SUCCESS, HISTORY_MEETINGROOMS*	11.12
		UNFILLED_SUBMIT, ENVIRONMENT_MC	8.10
		MAIL_VIEWED_3, FOLDER_VIEWED	7.31
		HISTORY_VIEWCALENDAR, COMBOBOX_ROOM	6.77
		ENVIRONMENT_WB, UNFILLED_SUBMIT	6.56
	Trigram	*COMBOBOX_DEPT, ENVIRONMENT_MC, ENVIRONMENT_WB*	12.09
		SUBMIT_RESERVATION_SUCCESS, HISTORY_MEETINGROOMS, HISTORY_VIEWCALENDAR	9.62
		HISTORY_UNFILLED, UNFILLED_SUBMIT, ENVIRONMENT_MC	9.58
		COMBOBOX_DEPT, SUBMIT_RESERVATION_SUCCESS, HISTORY_MEETINGROOMS	9.42
		ENVIRONMENT_MC, ENVIRONMENT_WB, ENVIRONMENT_MC	9.20
All zero response on INFLUENCE_WLE_CA	Unigram	*MAIL_MOVE*	17.83
		FOLDER_VIEWED	9.48
		CANCEL	7.97
		SUBMIT_RESERVATION_FAILURE	1.58
		HISTORY_HOME	0.54
	Bigram	*SUBMIT_RESERVATION_SUCCESS, SUBMIT_RESERVATION_FAILURE*	29.75
		MAIL_MOVE, FOLDER_VIEWED	23.45
		ENVIRONMENT_WB, HISTORY_HOME	21.24
		MAIL_VIEWED_3, MAIL_MOVE	20.17
		MAIL_VIEWED_2, MAIL_MOVE	18.77
	Trigram	*FOLDER_VIEWED, MAIL_VIEWED_1, MAIL_MOVE*	30.13
		MAIL_MOVE, FOLDER_VIEWED, FOLDER_VIEWED	23.99
		ENVIRONMENT_WB, COMBOBOX_ROOM, COMBOBOX_START_TIME	23.23
		COMBOBOX_DEPT, SUBMIT_RESERVATION_SUCCESS, SUBMIT_RESERVATION_FAILURE	22.58
		ENVIRONMENT_MC, NEXT_INQUIRY, CANCEL	22.00

TABLE A5 | Top five features of action sequences selected for the groups with lowest 20% or all zero response on READWORK_WLE_CA.

Group	*N*-gram	Action sequences	Chi-square
Lowest 20% on READWORK_WLE_CA	Unigram	*HELP*	4.29
		CANCEL	3.69
		COPY	3.59
		HISTORY_VIEWCALENDAR	3.28
		SEARCH	2.00
	Bigram	*ENVIRONMENT_MC, MAIL_VIEWED_3*	9.10
		HISTORY_VIEWCALENDAR, HISTORY_UNFILLED	5.45
		HISTORY_RESERVATION, COMBOBOX_ROOM	5.41
		ENVIRONMENT_WB, HISTORY_VIEWCALENDAR	5.15
		CHANGE_RESERVATION, CHANGE_RESERVATION	4.58
	Trigram	*ENVIRONMENT_MC, ENVIRONMENT_WB, HISTORY_VIEWCALENDAR*	6.70
		ENVIRONMENT_WB, HISTORY_RESERVATION, HISTORY_VIEWCALENDAR	6.59
		ENVIRONMENT_WB, HISTORY_RESERVATION, COMBOBOX_ROOM	6.27
		ENVIRONMENT_MC, MAIL_VIEWED_3, ENVIRONMENT_WB	6.06
		START, MAIL_VIEWED_1, MAIL_VIEWED_2	5.64

(Continued)

TABLE A5 | Continued

Group	*N*-gram	Action sequences	Chi-square
All zero response on READWORK_WLE_CA	Unigram	*MAIL_MOVE*	28.76
		FOLDER_VIEWED	4.66
		SUBMIT_RESERVATION_FAILURE	4.58
		HISTORY_HOME	4.45
		REPLY	1.07
	Bigram	*HISTORY_UNFILLED, HISTORY_UNFILLED*	35.28
		MAIL_VIEWED_3, ENVIRONMENT_MC	26.10
		FOLDER_VIEWED, MAIL_MOVE	22.78
		MAIL_MOVE, FOLDER_VIEWED	22.25
		HISTORY_HOME, HISTORY_MEETINGROOMS	20.55
	Trigram	*ENVIRONMENT_WB, COMBOBOX_END_TIME, COMBOBOX_DEPT*	47.51
		FOLDER_VIEWED, MAIL_MOVE, FOLDER_VIEWED	35.91
		MAIL_VIEWED_3, MAIL_MOVE, MAIL_VIEWED_2	34.97
		SUBMIT_RESERVATION_FAILURE, COMBOBOX_DEPT, SUBMIT_RESERVATION_FAILURE	33.24
		COMBOBOX_ROOM, SUBMIT_RESERVATION_FAILURE, COMBOBOX_DEPT	33.24

TABLE A6 | Top five features of action sequences selected for the groups with 60–80% or all zero response on TASKDISC_WLE_CA.

Group	*N*-gram	Action sequences	Chi-square
60–80% on TASKDISC_WLE_CA	Unigram	*FOLDER*	2.17
		HISTORY_HOME	1.87
		KEYPRESS	1.52
		HELP	0.88
		MAIL_MOVE	0.63
	Bigram	*HISTORY_RESERVATION, HISTORY_VIEWCALENDAR*	2.58
		COMBOBOX_END_TIME, ENVIRONMENT_MC	2.47
		MAIL_VIEWED_4, MAIL_VIEWED_3	2.13
		ENVIRONMENT_WB, ENVIRONMENT_WP	1.94
		ENVIRONMENT_WP, ENVIRONMENT_MC	1.91
	Trigram	*COMBOBOX_START_TIME, COMBOBOX_END_TIME, ENVIRONMENT_MC*	2.16
		ENVIRONMENT_MC, ENVIRONMENT_WB, COMBOBOX_DEPT	2.15
		COMBOBOX_END_TIME, ENVIRONMENT_MC, ENVIRONMENT_WB	1.74
		ENVIRONMENT_WB, COMBOBOX_DEPT, SUBMIT_RESERVATION_SUCCESS	1.60
		ENVIRONMENT_WB, HISTORY_RESERVATION, HISTORY_VIEWCALENDAR	1.44
All zero response on TASKDISC_WLE_CA	Unigram	*SUBMIT_RESERVATION_FAILURE*	3.94
		COMBOBOX_DEPT	0.69
		CHANGE_RESERVATION	0.41
		ENVIRONMENT_WP	0.19
		COMBOBOX_END_TIME	0.18
	Bigram	*CANCEL, HISTORY_MEETINGROOMS*	78.31
		ENVIRONMENT_WB, COMBOBOX_END_TIME	75.33
		COMBOBOX_DEPT, COMBOBOX_DEPT	34.89
		COMBOBOX_ROOM, ENVIRONMENT_MC	30.89
		START, NEXT_INQUIRY	30.31
	Trigram	*ENVIRONMENT_WB, COMBOBOX_END_TIME, COMBOBOX_DEPT*	138.78
		MAIL_VIEWED_4, MAIL_VIEWED_4, NEXT_INQUIRY	96.79
		ENVIRONMENT_WP, ENVIRONMENT_WB, COMBOBOX_END_TIME	95.64
		HISTORY_RESERVATION, COMBOBOX_START_TIME, ENVIRONMENT_MC	95.64
		MAIL_VIEWED_1, ENVIRONMENT_WP, ENVIRONMENT_WP	95.64

TABLE A7 | Top five features of action sequences selected for the groups with 60–80% or all zero response on WRITWORK_WLE_CA.

Group	*N*-gram	Action sequences	Chi-square
60–80% on WRITWORK_WLE_CA	Unigram	*CANCEL*	12.84
		SORT	12.34
		BOOKMARK	5.55
		UNFILLED_SUBMIT	5.41
		SEARCH	4.88
	Bigram	*SUBMIT_RESERVATION_SUCCESS, COMBOBOX_ROOM*	16.28
		COMBOBOX_END_TIME, COMBOBOX_DEPT	15.44
		ENVIRONMENT_WB, UNFILLED_SUBMIT	14.59
		COMBOBOX_END_TIME, SUBMIT_RESERVATION_SUCCESS	13.67
		CHANGE_RESERVATION, ENVIRONMENT_MC	12.38
	Trigram	*ENVIRONMENT_WB, HISTORY_VIEWCALENDAR, HISTORY_RESERVATION*	17.00
		HISTORY_VIEWCALENDAR, ENVIRONMENT_MC, NEXT_INQUIRY	16.99
		MAIL_VIEWED_4, ENVIRONMENT_WB, HISTORY_VIEWCALENDAR	16.20
		MAIL_VIEWED_1, MAIL_VIEWED_1, ENVIRONMENT_WB	15.83
		ENVIRONMENT_WB, ENVIRONMENT_MC, FOLDER_VIEWED	15.70
All zero response on WRITWORK_WLE_CA	Unigram	*MAIL_MOVE*	121.84
		FOLDER_VIEWED	12.97
		COPY	11.99
		PASTE	2.66
		MAIL_VIEWED_2	1.97
	Bigram	*MAIL_MOVE, FOLDER_VIEWED*	91.26
		FOLDER_VIEWED, MAIL_MOVE	89.43
		MAIL_MOVE, MAIL_VIEWED_1	74.40
		MAIL_VIEWED_4, MAIL_MOVE	72.36
		MAIL_MOVE, MAIL_VIEWED_3	69.92
	Trigram	*FOLDER_VIEWED, MAIL_MOVE, MAIL_VIEWED_1*	88.62
		MAIL_VIEWED_2, MAIL_VIEWED_2, MAIL_VIEWED_2	81.90
		MAIL_MOVE, FOLDER_VIEWED, MAIL_MOVE	76.13
		MAIL_MOVE, MAIL_VIEWED_1, MAIL_MOVE	68.14
		MAIL_MOVE, MAIL_VIEWED_4, MAIL_MOVE	59.57

TABLE A8 | Top five features of action sequences selected for the groups with age 45–54 and age 24 or less.

Group	*N*-gram	Action sequences	Chi-square
Age 45–54	Unigram	*HISTORY_HOME*	71.57
		HELP	51.53
		FOLDER	44.96
		REPLY	31.92
		COPY	11.36
	Bigram	*FOLDER_VIEWED, ENVIRONMENT_MC*	61.98
		HISTORY_HOME, HISTORY_HOME	56.88
		HELP, FOLDER_VIEWED	46.19
		FOLDER, FOLDER_VIEWED	42.16
		START, NEXT_INQUIRY	41.40
	Trigram	*FOLDER_VIEWED, FOLDER, FOLDER_VIEWED*	49.10
		HISTORY_HOME, HISTORY_HOME, ENVIRONMENT_MC	46.57
		MAIL_VIEWED_3, MAIL_VIEWED_4, ENVIRONMENT_WB	42.33
		FOLDER_VIEWED, FOLDER_VIEWED, ENVIRONMENT_MC	41.07
		MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4	40.06
Age 24 or less	Unigram	*MAIL_MOVE*	32.37
		HISTORY_UNFILLED	11.04
		CHANGE_RESERVATION	6.10
		SUBMIT_RESERVATION_FAILURE	3.41
		UNFILLED_SUBMIT	2.96
	Bigram	*MAIL_VIEWED_3, MAIL_MOVE*	56.40
		MAIL_VIEWED_1, MAIL_MOVE	41.25
		MAIL_MOVE, MAIL_VIEWED_2	40.61
		MAIL_MOVE, MAIL_VIEWED_4	38.06
		MAIL_VIEWED_4, MAIL_MOVE	37.37
	Trigram	*START, MAIL_VIEWED_1, MAIL_MOVE*	54.83
		MAIL_VIEWED_3, MAIL_MOVE, FOLDER_VIEWED	51.72
		MAIL_MOVE, MAIL_VIEWED_4, MAIL_MOVE	43.26
		MAIL_VIEWED_2, MAIL_VIEWED_2, MAIL_VIEWED_2	43.11
		MAIL_VIEWED_1, MAIL_MOVE, FOLDER_VIEWED	42.37

TABLE A9 | Top five features of action sequences selected for the groups with age 55 or more and age 24 or less.

Group	*N*-gram	Action sequences	Chi-square
Age 55 or more	Unigram	*HISTORY_HOME*	95.57
		HELP	69.64
		REPLY	56.68
		FOLDER	53.56
		SORT	22.21
	Bigram	*HISTORY_HOME, HISTORY_HOME*	103.26
		FOLDER_VIEWED, ENVIRONMENT_MC	84.40
		FOLDER_VIEWED, FOLDER	52.83
		HELP, FOLDER_VIEWED	48.34
		FOLDER_VIEWED, REPLY	39.27
	Trigram	*HISTORY_HOME, HISTORY_HOME, HISTORY_HOME*	75.62
		FOLDER_VIEWED, FOLDER_VIEWED, ENVIRONMENT_MC	60.11
		FOLDER_VIEWED, FOLDER_VIEWED, FOLDER	41.40
		FOLDER_VIEWED, ENVIRONMENT_MC, ENVIRONMENT_MC	39.70
		HELP, FOLDER_VIEWED, FOLDER_VIEWED	37.36
Age 24 or less	Unigram	*MAIL_MOVE*	40.37
		PASTE	10.44
		COPY	6.60
		ENVIRONMENT_WB	4.88
		HISTORY_MEETINGROOMS	3.82
	Bigram	*MAIL_MOVE, MAIL_VIEWED_2*	43.08
		FOLDER_VIEWED, MAIL_MOVE	34.00
		MAIL_VIEWED_1, MAIL_MOVE	32.85
		MAIL_MOVE, MAIL_VIEWED_3	31.82
		MAIL_MOVE, FOLDER_VIEWED	29.20
	Trigram	*MAIL_MOVE, MAIL_VIEWED_2, FOLDER_VIEWED*	34.83
		ENVIRONMENT_MC, ENVIRONMENT_WB, COMBOBOX_DEPT	33.78
		MAIL_VIEWED_2, FOLDER_VIEWED, MAIL_MOVE	33.19
		MAIL_MOVE, MAIL_VIEWED_1, MAIL_MOVE	32.86
		ENVIRONMENT_WB, COMBOBOX_DEPT, ENVIRONMENT_MC	28.72

TABLE A10 | Top five features of action sequences selected for the groups with test language same as native language or not same as native language.

Group	*N*-gram	Action sequences	Chi-square
Test language same as native language	Unigram	*SORT*	5.64
		PASTE	4.37
		HELP	3.00
		UNFILLED_SUBMIT	2.77
		FOLDER	2.29
	Bigram	*HISTORY_VIEWCALENDAR, ENVIRONMENT_WP*	17.51
		SUBMIT_RESERVATION_FAILURE, HISTORY_VIEWCALENDAR	10.59
		FOLDER, FOLDER_VIEWED	8.49
		MAIL_VIEWED_3, NEXT_INQUIRY	8.43
		COMBOBOX_DEPT, COMBOBOX_START_TIME	7.32
	Trigram	*ENVIRONMENT_WB, HISTORY_MEETINGROOMS, ENVIRONMENT_MC*	11.37
		FOLDER_VIEWED, FOLDER_VIEWED, ENVIRONMENT_WB	10.48
		ENVIRONMENT_WB, HISTORY_VIEWCALENDAR, HISTORY_UNFILLED	10.42
		HISTORY_UNFILLED, ENVIRONMENT_MC, ENVIRONMENT_WB	10.13
		HISTORY_VIEWCALENDAR, ENVIRONMENT_WP, ENVIRONMENT_WB	9.33
Test language not the same as native language	Unigram	*SEARCH*	16.28
		BOOKMARK	2.29
		MAIL_MOVE	1.16
		HISTORY_MEETINGROOMS	0.67
		COMBOBOX_DEPT	0.29
	Bigram	*NEXT_INQUIRY, KEYPRESS*	141.04
		COMBOBOX_ROOM, HISTORY_HOME	75.69
		BOOKMARK, HISTORY_VIEWCALENDAR	67.68
		HISTORY_RESERVATION, BOOKMARK	67.68
		SEARCH, KEYPRESS	49.98
	Trigram	*HISTORY_HOME, HISTORY_MEETINGROOMS, HISTORY_VIEWCALENDAR*	134.08
		SUBMIT_RESERVATION_FAILURE, ENVIRONMENT_MC, ENVIRONMENT_WP	132.41
		COMBOBOX_END_TIME, COMBOBOX_ROOM, ENVIRONMENT_MC	130.71
		FOLDER_VIEWED, SEARCH, SEARCH	129.22
		MAIL_VIEWED_2, ENVIRONMENT_WP, ENVIRONMENT_WB	128.67

TABLE A11 | Top five features of action sequences selected for upper secondary and lowest education groups.

Group	*N*-gram	Action sequences	Chi-square
Upper secondary	Unigram	*SUBMIT_RESERVATION_FAILURE*	6.88
		COMBOBOX_ROOM	5.08
		HISTORY_HOME	5.03
		COMBOBOX_START_TIME	3.49
		BOOKMARK	3.21
	Bigram	*MAIL_VIEWED_4, MAIL_VIEWED_2*	13.92
		FOLDER, FOLDER_VIEWED	13.77
		HISTORY_HOME, HISTORY_HOME	13.14
		FOLDER_VIEWED, FOLDER	12.18
		ENVIRONMENT_WB, ENVIRONMENT_MC	12.15
	Trigram	*ENVIRONMENT_WB, HISTORY_MEETINGROOMS, ENVIRONMENT_MC*	15.66
		ENVIRONMENT_MC, MAIL_VIEWED_1, MAIL_VIEWED_2	15.18
		MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4	14.87
		HISTORY_RESERVATION, HISTORY_MEETINGROOMS, HISTORY_VIEWCALENDAR	14.28
		FOLDER_VIEWED, FOLDER, FOLDER_VIEWED	13.07
Lower secondary or less	Unigram	*SEARCH*	32.49
		MAIL_MOVE	30.71
		CANCEL	14.37
		COPY	12.53
		PASTE	11.72
	Bigram	*SEARCH, FOLDER_VIEWED*	54.79
		HISTORY_UNFILLED, HISTORY_UNFILLED	54.12
		CANCEL, MAIL_MOVE	53.27
		CHANGE_RESERVATION, ENVIRONMENT_WP	52.96
		MAIL_MOVE, FOLDER	51.19
	Trigram	*FOLDER_VIEWED, MAIL_MOVE, MAIL_VIEWED_3*	74.80
		MAIL_VIEWED_4, FOLDER_VIEWED, MAIL_VIEWED_4	72.11
		MAIL_MOVE, MAIL_VIEWED_3, MAIL_MOVE	55.93
		HISTORY_UNFILLED, HISTORY_RESERVATION, HISTORY_UNFILLED	52.49
		MAIL_VIEWED_4, FOLDER_VIEWED, MAIL_VIEWED_1	52.28

TABLE A12 | Top five features of action sequences selected for postsecondary, non-tertiary, and lowest education groups.

Group	*N*-gram	Action sequences	Chi-square
Postsecondary, non-tertiary	Unigram	*HISTORY_HOME*	19.52
		SORT	8.20
		SUBMIT_RESERVATION_FAILURE	3.59
		BOOKMARK	3.29
		COMBOBOX_ROOM	1.85
	Bigram	*HISTORY_HOME, HISTORY_HOME*	26.85
		ENVIRONMENT_WB, ENVIRONMENT_MC	13.68
		SUBMIT_RESERVATION_SUCCESS, HISTORY_VIEWCALENDAR	11.67
		FOLDER_VIEWED, REPLY	11.06
		ENVIRONMENT_MC, MAIL_VIEWED_1	10.32
	Trigram	*ENVIRONMENT_MC, MAIL_VIEWED_1, MAIL_VIEWED_2*	18.60
		ENVIRONMENT_WB, ENVIRONMENT_WB, ENVIRONMENT_MC	18.46
		ENVIRONMENT_WB, ENVIRONMENT_MC, FOLDER_VIEWED	16.61
		ENVIRONMENT_WB, HISTORY_MEETINGROOMS, ENVIRONMENT_MC	16.19
		MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4	16.03
Lower secondary or less	Unigram	*SEARCH*	23.04
		MAIL_MOVE	19.31
		PASTE	6.78
		HISTORY_UNFILLED	2.58
		FOLDER_VIEWED	2.25
	Bigram	*MAIL_VIEWED_2, MAIL_MOVE*	32.76
		MAIL_MOVE, MAIL_VIEWED_1	25.84
		MAIL_VIEWED_3, MAIL_MOVE	25.74
		MAIL_MOVE, MAIL_VIEWED_2	25.42
		MAIL_VIEWED_1, MAIL_MOVE	21.10
	Trigram	*MAIL_MOVE, MAIL_VIEWED_2, MAIL_MOVE*	32.53
		FOLDER_VIEWED, MAIL_MOVE, MAIL_VIEWED_2	26.88
		MAIL_VIEWED_1, MAIL_MOVE, MAIL_VIEWED_1	23.64
		MAIL_VIEWED_2, MAIL_MOVE, FOLDER_VIEWED	23.23
		MAIL_VIEWED_3, MAIL_MOVE, FOLDER_VIEWED	21.26

TABLE A13 | Top five features of action sequences selected for tertiary-professional degree and lowest education groups.

Group	*N*-gram	Action sequences	Chi-square
Tertiary – professional degree	Unigram	*COMBOBOX_ROOM*	4.48
		SUBMIT_RESERVATION_SUCCESS	4.02
		COMBOBOX_DEPT	3.66
		COMBOBOX_START_TIME	3.65
		COMBOBOX_END_TIME	3.26
	Bigram	*ENVIRONMENT_WB, ENVIRONMENT_MC*	13.20
		MAIL_MOVE, ENVIRONMENT_WP	11.95
		ENVIRONMENT_WB, ENVIRONMENT_WB	10.87
		ENVIRONMENT_MC, MAIL_VIEWED_1	10.34
		SUBMIT_RESERVATION_FAILURE, HISTORY_MEETINGROOMS	9.98
	Trigram	*ENVIRONMENT_WB, ENVIRONMENT_WB, ENVIRONMENT_MC*	19.62
		MAIL_VIEWED_1, MAIL_VIEWED_4, ENVIRONMENT_WB	18.65
		ENVIRONMENT_MC, MAIL_VIEWED_1, MAIL_VIEWED_2	18.29
		ENVIRONMENT_WB, HISTORY_MEETINGROOMS, ENVIRONMENT_MC	17.63
		MAIL_VIEWED_3, MAIL_VIEWED_1, MAIL_VIEWED_4	15.67
Lower secondary or less	Unigram	*COPY*	30.70
		MAIL_MOVE	28.30
		FOLDER	22.22
		PASTE	20.33
		FOLDER_VIEWED	12.20
	Bigram	*MAIL_VIEWED_2, MAIL_MOVE*	31.13
		MAIL_MOVE, MAIL_VIEWED_3	28.74
		FOLDER_VIEWED, MAIL_MOVE	24.85
		MAIL_MOVE, MAIL_VIEWED_1	22.55
		MAIL_MOVE, MAIL_VIEWED_2	22.37
	Trigram	*MAIL_MOVE, MAIL_VIEWED_2, MAIL_MOVE*	38.57
		FOLDER_VIEWED, MAIL_MOVE, MAIL_VIEWED_3	38.13
		MAIL_MOVE, MAIL_VIEWED_1, MAIL_MOVE	30.98
		MAIL_VIEWED_2, MAIL_MOVE, FOLDER_VIEWED	29.04
		MAIL_VIEWED_1, MAIL_MOVE, MAIL_VIEWED_1	28.07

TABLE A14 | Top five features of action sequences selected for tertiary-bachelor degree and lowest education groups.

Group	*N*-gram	Action sequences	Chi-square
Tertiary – bachelor degree	Unigram	*REPLY*	7.10
		COMBOBOX_ROOM	6.50
		COMBOBOX_START_TIME	5.30
		SUBMIT_RESERVATION_SUCCESS	5.14
		COMBOBOX_DEPT	4.76
	Bigram	*MAIL_VIEWED_4, MAIL_VIEWED_1*	16.49
		MAIL_VIEWED_1, MAIL_VIEWED_3	16.24
		ENVIRONMENT_WB, ENVIRONMENT_MC	16.13
		HISTORY_MEETINGROOMS, ENVIRONMENT_MC	14.32
		MAIL_VIEWED_4, MAIL_VIEWED_2	13.87
	Trigram	*HISTORY_RESERVATION, HISTORY_MEETINGROOMS, HISTORY_VIEWCALENDAR*	20.75
		ENVIRONMENT_WB, HISTORY_MEETINGROOMS, ENVIRONMENT_MC	20.25
		ENVIRONMENT_MC, MAIL_VIEWED_1, MAIL_VIEWED_2	17.86
		MAIL_VIEWED_1, MAIL_VIEWED_3, MAIL_VIEWED_4	16.25
		MAIL_VIEWED_4, MAIL_VIEWED_1, ENVIRONMENT_WB	15.25
Lower secondary or less	Unigram	*MAIL_MOVE*	109.87
		PASTE	59.04
		FOLDER_VIEWED	14.04
		COPY	13.82
		SEARCH	8.05
	Bigram	*MAIL_VIEWED_1, MAIL_MOVE*	101.94
		FOLDER_VIEWED, MAIL_MOVE	98.18
		MAIL_MOVE, MAIL_VIEWED_2	90.67
		MAIL_MOVE, MAIL_VIEWED_3	90.67
		MAIL_MOVE, FOLDER_VIEWED	83.84
	Trigram	*FOLDER_VIEWED, MAIL_MOVE, MAIL_VIEWED_2*	103.42
		MAIL_MOVE, FOLDER_VIEWED, MAIL_MOVE	93.20
		MAIL_MOVE, MAIL_VIEWED_3, FOLDER_VIEWED	88.33
		MAIL_VIEWED_3, FOLDER_VIEWED, MAIL_MOVE	88.33
		MAIL_VIEWED_3, MAIL_MOVE, FOLDER_VIEWED	88.13

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 March 2021
doi: 10.3389/fpsyg.2021.469196

Variable Speed Across Dimensions of Ability in the Joint Model for Responses and Response Times

Peida Zhan[1], Hong Jiao[2], Kaiwen Man[3], Wen-Chung Wang[4] and Keren He[1]**

[1] Zhejiang Normal University, Jinhua, China, [2] University of Maryland, College Park, MD, United States, [3] University of Alabama, Tuscaloosa, AL, United States, [4] The Education University of Hong Kong, Tai Po, Hong Kong

Working speed as a latent variable reflects a respondent's efficiency to apply a specific skill, or a piece of knowledge to solve a problem. In this study, the common assumption of many response time models is relaxed in which respondents work with a constant speed across all test items. It is more likely that respondents work with different speed levels across items, in specific when these items measure different dimensions of ability in a multidimensional test. Multiple speed factors are used to model the speed process by allowing speed to vary across different domains of ability. A joint model for multidimensional abilities and multifactor speed is proposed. Real response time data are analyzed with an exploratory factor analysis as an example to uncover the complex structure of working speed. The feasibility of the proposed model is examined using simulation data. An empirical example with responses and response times is presented to illustrate the proposed model's applicability and rationality.

Keywords: response times, joint model, variable speed, multidimensional item response theory, hierarchical modeling framework

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Timothy R. Brick,
Pennsylvania State University, United States
Jean-Paul Fox,
University of Twente, Netherlands

****Correspondence:***
Keren He
kerenhe@zjnu.edu.cn
Peida Zhan
pdzhan@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 April 2019*
Accepted: *01 March 2021*
Published: *29 March 2021*

Citation:
Zhan P, Jiao H, Man K, Wang W-C and He K (2021) Variable Speed Across Dimensions of Ability in the Joint Model for Responses and Response Times. Front. Psychol. 12:469196. doi: 10.3389/fpsyg.2021.469196

INTRODUCTION

With the popularity of computer-based tests, the collection of item response times (RTs) has become a routine activity in large- and small-scale educational assessments. For example, the Programme for International Student Assessment (PISA) started using computer-based tests and recorded RTs data since 2012. RTs provide information about the working speed of respondents but also could be utilized to improve measurement accuracy because RTs are considered to convey a more synoptic depiction of the respondents' performance beyond what is obtainable based on correct responses alone (van der Linden et al., 2010; Bolsinova and Tijmstra, 2018).

Before making inferences by employing RTs, it is necessary to create an appropriate statistical model for RTs. Over the past few decades, various RT models have been presented based on cognitive/psychological theories and experimental research (for a review, see De Boeck and Jeon, 2019). Currently, the Bayesian hierarchical modeling framework (van der Linden, 2007) is one of the most flexible tools to explain the relationship between latent ability and working speed. This framework is gaining more recognition and is sufficiently generalized to integrate available measurement models for item response accuracy (RA) and RTs. Typically, in the hierarchical modeling of RTs and RA, the RT measurement model assumes that a respondent works at a constant speed throughout a test. Meanwhile, the RA measurement model assumes that a respondent puts

his or her best effort forward to solve a set of items correctly by using the required knowledge. Thus, the association between latent ability and working speed is assumed to be changeless for each respondent working on a test. In other words, each respondent is assumed to work at a constant pace given his or her invariant ability at that time (Fox and Marianti, 2016).

Currently, most joint models for RA and RTs only use unidimensional measurement models to capture the relationship between latent ability and working speed within a unidimensional test scenario (e.g., Klein Entink et al., 2009a,b; Wang et al., 2013; Fox et al., 2014; Molenaar et al., 2015, 2016; Wang and Xu, 2015; Fox and Marianti, 2016). In reality, however, multiple latent abilities are involved to correctly answer an item, especially in multidimensional tests (e.g., Tatsuoka, 1983; Reckase, 2009). Compared to unidimensional tests, one significant characteristic of multidimensional tests is that different test items may measure distinguish latent ability dimensions.

In educational and psychological measurements, working speed as a latent variable reflects a respondent's efficiency to apply a specific skill or a piece of knowledge to solve a problem. Therefore, latent speed should be discussed by considering the linkage to a particular dimension of latent ability. It is reasonable to assume that respondents could vary their working speeds across items that measure different dimensions of ability. In other words, the multidimensional structure for latent ability could be used to model the process of speed change, where the working speed is allowed to vary across dimensions of ability. For example, in a math test, the working speed on items that measure algebra problem-solving ability may differ from those measuring geometry problem-solving ability.

With the development of psychometrics, multidimensional measurement models for RA [e.g., multidimensional item response theory (MIRT) models and diagnostic classification models (DCMs)] have been well developed and widely used (see Reckase, 2009; Rupp et al., 2010). Recently, based on hierarchical modeling, a few studies have attempted to use MIRT models or DCMs for RA to capture the multidimensional structure of the latent trait when multidimensional tests are involved. But still, a unidimensional or single-factor RT (SRT) model is used to measure latent speed (Zhan et al., 2018; Man et al., 2019; Wang et al., 2019). Thus, in these studies, the relationships between multiple latent abilities and one single latent speed are assumed to be constant for each respondent working with a constant speed on different items. However, assuming identical working speeds across different dimensions of ability may be too restrictive to describe intricate data and thus may lead to ambiguous conclusions. It is desirable to release this limitation to allow each dimension of ability to be associated with a specific speed factor. As current joint models may be inappropriate for multidimensional tests, it is critical to develop a joint model that allows working speed to vary across dimensions of ability.

To model varying working speeds within different domains of ability, it is possible to use multiple-speed factors/dimensions to describe the speed process. Each speed factor corresponds to a specific dimension of latent ability. An individual speed process is assumed, describing the changes in speed across dimensions. Thus, respondents can work at different levels of speed on items within different dimensions of ability during multidimensional tests. Each individual speed process will be defined using a confirmatory multifactor structure, which in turn is defined by the dimensions of ability measured by items, according to the testing blueprint. Furthermore, it will be shown that the multifactor working speed model can be integrated with a MIRT model for latent ability. Under this new joint model, it is assumed that each respondent works at a unique speed corresponding to the dimension represented by an item.

We first extend the most popular single-factor lognormal RT (SLRT) model (van der Linden, 2006) to a multifactor working speed model that considers changing speed across dimensions. This is called the multifactor lognormal RT (MLRT) model. Second, a joint model of multidimensional latent ability and multifactor working speed will be proposed. Our paper starts with a brief review of the SLRT model, followed by presenting the proposed MLRT model. The proposed joint model is then presented. Next, a motivating example will be provided to demonstrate the multifactor structure of working speed and its compatibility with the multidimensional structure of latent ability. Moreover, two simulation studies will be conducted to evaluate the psychometric properties of the proposed joint model. An empirical example will also be analyzed to illustrate the application of the proposed joint model. Finally, we summarize our findings and discuss directions for future research.

MULTIFACTOR LOGNORMAL RESPONSE TIME MODEL

Let T_{ni} be the observed RT of person n (n = 1,..., N) to item i (i = 1,..., I). In the SLRT model, the logarithmic function is used to transform the positively skewed distribution of RT to a more symmetric shape and is assumed to be dominated by item i's time-intensity parameter ξ_i and person n's latent speed parameter τ_n as follows:

$$\log T_{ni} = \xi_i - \tau_n + \varepsilon_{ni}, \quad \varepsilon_{ni} \sim N(0, \quad \omega_i^{-2}), \tag{1}$$

or equivalently,

$$\log T_{ni} \sim N(\xi_i - \tau_n, \quad \omega_i^{-2}). \tag{2}$$

where ξ_i represents the time needed to complete item i, τ_n is the single-factor working speed of person n, and ε_{ni} is the normally distributed residual error term, with mean zero and varianceω_i^{-2}, where ω_i is the time-precision parameter.

In recent years, the SLRT model has been extended in some studies. For instance, Klein Entink et al. (2009a) included a time-discrimination parameter as a slope parameter for latent speed. Klein Entink et al. (2009b) proposed the Box-Cox transformation for RT modeling. Wang et al. (2013) proposed a linear transformation model for RTs. Furthermore, Fox and Marianti (2016) proposed a variable working speed model, which allows the respondents to adjust their working speed along the sequence of items throughout the test. Although Fox and Marianti's (2016) model relaxed the assumption of constant speed in the SLRT model, their variable speed was different from that

focused on in this study. One is to change speed as the item response progresses, and the other is to change speed as the dimension of ability examined by the item changes.

As mentioned previously, the kernel hypothesis of this study is that respondents can work with different levels of speed on items requiring different dimensions of ability during multidimensional tests. In other words, working speed has a multifactor structure, which is defined by the multidimensional structure of ability. In the multidimensional test, assuming there are K sub-dimensions of latent ability. In the current study, only the between-item multidimensionality (Adams et al., 1997) is considered, where each item measures a single dimension but different items measure different dimensions, so the multidimensionality occurs between items. To model variable speed across dimensions, we first relaxed the assumption of the SLRT model that each respondent works at a constant speed on all items throughout the test and allowed the instantaneous speed to be different on different items, that is, $\tau_n \rightarrow \tilde{\tau}_{ni}$. Then, a confirmatory multifactor structure was given to model the instantaneous speed at item i of person n, as

$$\tilde{\tau}_{ni} = \sum_{k=1}^{K} \tau_{nk} q_{ik}, \tag{3}$$

where $\tilde{\tau}_{ni}$ is the instantaneous speed at item i of person n, and τ_{nk} is the working speed factor of person n corresponding to kth-dimension (k = 1, 2,..., K) of ability. The Q-matrix (Tatsuoka, 1983) is an I-by-K confirmatory matrix with element q_{ik} indicating whether kth-dimension of ability is required to answer item i correctly: q_{ik} = 1 if the dimension is required, and q_{ik} = 0 otherwise. For between-item multidimensionality, only one dimension is measured by an item, namely, only one element in $\mathbf{q}_i$ equals to 1. In such cases, the MLRT model can be expressed as

$$\log T_{ni} = \xi_i - \tilde{\tau}_{ni} + \varepsilon_{ni} = \xi_i - \sum_{k=1}^{K} \tau_{nk} q_{ik} + \varepsilon_{ni}, \varepsilon_{ni} \sim N(0, \omega_i^{-2}) \tag{4}$$

or equivalently,

$$\log T_{ni} \sim N(\xi_i - \tilde{\tau}_{ni}, \ \omega_i^{-2}). \tag{5}$$

If only one dimension of ability is assumed to be measured by all items, the MLRT model reduces to the SLRT model.

JOINT MODEL FOR RESPONSE ACCURACY AND RESPONSE TIMES

Model Construction

Since both RA and RTs contain information about items and persons, it is advantageous to analyze them simultaneously. To this end, based on hierarchical modeling, we propose a new joint model called the multidimensional-multifactor joint (MMJ) model. For illustration purposes, in the MMJ model in this study, the MLRT model is used as the measurement model for RTs, and according to the 2012 PISA mathematics assessment framework (OECD, 2013), the multidimensional Rasch (MR) model (Adams et al., 1997) is employed as the measurement model for RA.

Besides observing RTs, let Y_{ni} be the observed RA for person n to item i. The MR model can be expressed as

$$\text{logit}(P(Y_{ni} = 1)) = \sum_{k=1}^{K} \theta_{nk} q_{ik} + d_i, \tag{6}$$

where logit(x) = log(x/(1–x)), $P(Y_{ni} = 1)$ is the probability of a correct response by person n to item i, θ_{nk} is the latent ability of person n on dimension k, d_i is the intercept or easiness of item i, and q_{ik} is the element of Q-matrix.

The multivariate normal distribution was used to describe the relationships among the multidimensional ability and multifactor speed:

$$\begin{pmatrix} \boldsymbol{\theta}_n \\ \boldsymbol{\tau}_n \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_\theta \\ \boldsymbol{\mu}_\tau \end{pmatrix}, \ \sum\nolimits_{\text{Person}} \right),$$

$$\sum\nolimits_{\text{Person}} = \begin{pmatrix} \sigma^2_{\theta_1} & & & & & \\ \vdots & \ddots & & & & \\ \sigma_{\theta_1\theta_K} & \cdots & \sigma^2_{\theta_K} & & & \\ \sigma_{\theta_1\tau_1} & \cdots & \sigma_{\theta_K\tau_1} & \sigma^2_{\tau_1} & & \\ \vdots & \cdots & \vdots & \vdots & \ddots & \\ \sigma_{\theta_1\tau_K} & \cdots & \sigma_{\theta_K\tau_K} & \sigma_{\tau_1\tau_K} & \cdots & \sigma^2_{\tau_K} \end{pmatrix}, \tag{7}$$

where $\boldsymbol{\theta}_n$ = (θ_{n1},..., θ_{nk},..., θ_{nK})' is the multidimensional latent ability vector; $\boldsymbol{\tau}_n$ = (τ_{n1},...,τ_{nk},...,τ_{nK})' is the multifactor working speed vector; $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\mu}_\tau$ are the population mean vector of multidimensional ability and the population mean vector of multifactor working speed, respectively; and Σ_{person} is a variance-covariance matrix of person parameters, where $\boldsymbol{\sigma}^2_{\theta_k}$ is the variance of θ_k, $\sigma^2_{\tau_k}$ is the variance of τ_k, $\sigma_{\theta_k\theta_{k'}}$ is the covariance of θ_k and $\theta_{k'}$, $\sigma_{\tau_k\tau_{k'}}$ is the covariance of τ_k and $\tau_{k'}$, and $\sigma_{\tau_k\tau_{k'}}$ is the covariance of θ_k and τ_k.

Furthermore, for the item parameters, a bivariate normal distribution was used to describe the relationship between item easiness and item time-intensity,

$$\begin{pmatrix} d_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_\xi \end{pmatrix}, \sum\nolimits_{\text{item}} \right), \sum\nolimits_{\text{item}} = \begin{pmatrix} \sigma^2_d & \\ \sigma_{d\xi} & \sigma^2_\xi \end{pmatrix}, \tag{8}$$

where μ_d and μ_ξ are the mean of item easiness and the mean of item time-intensity, respectively; and $\boldsymbol{\Sigma}_{item}$ is a variance-covariance matrix of item parameters, where σ^2_d and σ^2_ξ are the variance of item easiness and the variance of item time-intensity, respectively; $\sigma_{d\xi}$ is the covariance of item easiness and item time-intensity.. The residual error variance, ω_i^{-2}, is assumed to be independently distributed.

For the MMJ model, the latent scales of multidimensional ability and mutlifactor speed need to be identified. This can be accomplished by restricting the population mean of the ability and speed as $\boldsymbol{\mu}_\theta = \boldsymbol{\mu}_\tau = \mathbf{0}$.

Parameter Estimation

Parameters in the MMJ model can be estimated via the full Bayesian approach with the Markov Chain Monte Carlo (MCMC) method. In Bayesian estimation, prior distributions of model parameters and observed data likelihood produce a joint posterior distribution for the model parameters. In this study, the Just Another Gibbs Sampler (JAGS) software (Plummer, 2015) was used to estimate parameters. JAGS uses a default option of the Gibbs sampler (Gelfand and Smith, 1990), whose code for the proposed joint model is provided in the online **Supplementary Appendix**.

Under the assumption of local independence, Y_{ni} and $\log T_{ni}$ are independently distributed as

$$Y_{ni} \sim \text{Bernoulli}(P(Y_{ni}=1)) and \log T_{ni} \sim N(\xi_i - \tilde{\tau}_{ni}, \ \omega_i^{-2}).$$

Weakly but not non-informative priors are preferentially used in this study to increase the generalizability of our codes by imposing vague prior beliefs on estimating parameters. The setting of priors refers to that used by Zhan et al. (2018) and Man et al. (2019).

The priors of the person parameters are set as

$$\begin{pmatrix} \boldsymbol{\theta}_n \\ \boldsymbol{\tau}_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \ \sum\nolimits_{\text{person}} \right),$$

with a hyper prior

$$\sum_{\text{person}} \sim \text{InvWishart}(\boldsymbol{R}_{\text{person}}, \ K^{*}),$$

where $\mathbf{R}_{person}$ is a K^{*}-dimensional identity matrix, and K^{*} indicates the degree of freedom, which in this case is equal to the dimension of the $\mathbf{R}_{person}$.

In addition, the priors of item parameters are set as

$$\begin{pmatrix} d_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_d \\ \mu_\xi \end{pmatrix}, \sum_{\text{item}} \right), \ \omega_i^{-2} \sim \text{InvGamma}(1, \ 1)$$

.

Furthermore, the hyper priors are specified as

$$\mu_d \sim \text{Normal}(0, \ \ 2), \mu_\xi \sim \text{Normal}(4.3, \ \ 2),$$

$$\sum_{item} \sim \text{InvWishart}(R_{\text{item}}, \ 2),$$

where $\mathbf{R}_{item}$ is a two-dimensional identity matrix. Finally, the posterior mean is treated as the estimated value for model parameters.

A MOTIVATING EXAMPLE

To explore the multifactor structure of working speed, and to explore whether this structure matches the multidimensional structure of latent ability, a motivating example with the exploratory factor analysis (EFA) of RTs was presented first.

Data Description

The PISA 2012 computer-based mathematics RT data were analyzed. This data set was originally used by Zhan et al. (2018). In this study, there are N = 1,581 respondents and I = 9 items. The logarithm of RTs was computed before the analysis, and all zero RTs were treated as missing data. A Q-matrix (see **Table 1**) was specified based on the PISA 2012 mathematics assessment framework (OECD, 2013). Three dimensions that belong to the mathematical content knowledge were chosen, namely, change and relationships (θ_1), space and shape (θ_2), and uncertainty and data (θ_3). However, it should be noted that this Q-matrix was originally used to link items and latent abilities or to present the multidimensional structure of latent ability. In other words, this Q-matrix does not specify the latent structure of working speed unless the structure explored by the EFA of RTs matches it.

Exploratory Analysis and Results

The Mplus (version 8.1) (Muthén and Muthén, 2019) was used here. The EFA within a confirmatory factor analysis framework method was used by default in Mplus. In this study, the number of factors to retain was set as 1 to 5, which means 1- to 5-factor CFA models were all employed to fit RT data. Then, Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) were used as model-data fit indexes to help judge the number of factors/dimensions. Theoretically, correlations should exist among multiple dimensions; thus, oblique rotation was used. Other settings followed the default (e.g., the maximum likelihood was used as an extraction method).

Table 2 presents the model-data fit indexes of the EFA. According to previous studies, TLI $>$ 0.95, CFI $>$ 0.95, SRMR $\leq$ 0.08, and RMSEA $<$ 0.05 mean good model-data fit (Hu and Bentler, 1999; Steiger, 1990). The AIC preferred the 4-factor model, and the BIC preferred the 3-factor model after taking into account the penalty weighting of sample size. On the whole, the 3-factor model seems to fit the data better than the other models.

Table 3 presents the rotated factor loading matrix for the 3-factor model. Compared to the theoretically constructed Q-matrix for latent ability, there is only a difference in CM038Q03T. The rotated factor loading of CM038Q03T on Factor 3 is 0.300 (p $<$ 0.05), which also supports the

TABLE 1 | Q-Matrix for PISA 2012 released computer-based mathematics items.

Items	θ_1	θ_2	θ_3
CM015Q02D	1		
CM015Q03D	1		
CM020Q01		1	
CM020Q02		1	
CM020Q03		1	
CM020Q04		1	
CM038Q03T			1
CM038Q05			1
CM038Q06			1

Blank means zero.

TABLE 2 | Exploratory factor analysis model-data fit indexes for RT data.

Model	χ^2	*df*	TLI	CFI	AIC	BIC	SRMR	RMSEA (90% CI)
1-factor	462.79**	27	0.896	0.922	24592.15	24737.03	0.045	0.101 (0.093, 0.109)
2-factor	225.49**	19	0.930	0.963	24370.85	24558.65	0.032	0.083 (0.073, 0.093)
3-factor	32.66**	12	0.989	0.996	24192.02	24417.38	0.010	0.033 (0.020, 0.047)
4-factor	5.56	6	1.000	1.000	24176.92	24434.48	0.004	0.000 (0.000, 0.031)
5-factor	0.09	1	1.006	1.000	24181.44	24465.83	0.000	0.000 (0.000, 0.045)

$^{**}p < 0.01$; χ^2 = chi-square; df = degrees of freedom; TLI = Tucker-Lewis index; CFI = comparative fit index; AIC = Akaike information criterion; BIC = Bayesian information criterion; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CI = confidence interval.

theoretical structure to a certain extent. The results indicate that the latent structure of working speed might be a 3-factor structure, which is also consistent with the theoretical multidimensional structure of latent ability (i.e., the Q-matrix in **Table 1**).

Overall, the results of the EFA support the kernel hypothesis of this study. However, due to the limitations of the EFA, the estimation of parameters such as individual working speed cannot be realized. Therefore, further exploration and utilization of the proposed MMJ model are necessary.

SIMULATION STUDIES

Two simulation studies were conducted to evaluate the performance of the MMJ model under various conditions. The primary purpose of simulation study 1 was to examine whether the model parameters could be recovered accurately using the proposed Bayesian estimation algorithm, in which data were simulated from the MMJ model and analyzed with itself.

Man et al. (2019) has shown that, in multidimensional tests, the joint model that involves multidimensional ability and single-factor speed (denoted as MSJ model in this study) performs better than the joint model that involves unidimensional ability and single-factor speed (e.g., van der Linden, 2007). In this study, we focus on the comparison between the MMJ model and the MSJ model. Specifically, simulation study 2 was conducted to evaluate: (a) the consequences of ignoring the multifactor structure of working speed, in which the data were simulated from the MMJ model but analyzed with the MSJ model; and (b) the consequences of misspecifying a multifactor structure of working speed, in which the data were simulated from the MSJ model but analyzed with the MMJ model. Note that the results of simulation study 2 were omitted for brevity but can be found in the online **Supplementary Appendix** (see **Supplementary section S1**).

TABLE 3 | Rotated factor loading matrix for the 3-factor model for response times data.

Item	Factor 1	Factor 2	Factor 3
CM015Q02D	0.695*		
CM015Q03D	0.609*		
CM020Q01		0.565*	
CM020Q02		0.801*	
CM020Q03		0.642*	
CM020Q04		0.943*	
CM038Q03T		0.502*	
CM038Q05			0.985*
CM038Q06			0.621*

$^{*}p < 0.05$; absolute value of factor loading below 0.4 was omitted.

Design and Data Generation

In simulation study 1, four factors were manipulated including (a) sample size: N = 500 and 1,000, (b) test length: I = 15 and 30, (c) the correlation coefficient between latent ability and its corresponding working speed factor: $\rho_{\theta\tau}$ = −0.7 and −0.4, and (d) the number of dimensions of ability: K = 3 and 5. Q-matrices are presented in **Figure 1**. In addition, the true values of other parameters were generated according to the results of a data analysis using real data (Zhan et al., 2018). For item parameters, item easiness, d_i, and item time intensity, ξ_i, were generated from a bivariate normal distribution with mean vector (0, 4) and covariance matrix of [1, −0.2; −0.2, 0.25]. In such a setting, $\rho_{d\xi}$ = −0.4. The reciprocal of the standard deviation of the error term, ω, is set to 2 for all items. Person parameters were generated from$(\theta_n,\ \tau_n)' \sim N((\mathbf{0},\ \mathbf{0})', \sum_{Person})$, where

$$\sum_{Person} = \begin{pmatrix} \sigma^2_{\theta_1} & & & & & \\ \vdots & \ddots & & & & \\ \sigma_{\theta_1\theta_K} & \cdots & \sigma^2_{\theta_K} & & & \\ \sigma_{\theta_1\tau_1} & \cdots & \sigma_{\theta_K\tau_1} & \sigma^2_{\tau_1} & & \\ \vdots & \cdots & \vdots & \vdots & \ddots & \\ \sigma_{\theta_1\tau_K} & \cdots & \sigma_{\theta_K\tau_K} & \sigma_{\tau_1\tau_K} & \cdots & \sigma^2_{\tau_K} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & & & & & \\ \vdots & \ddots & & & & \\ 0.8 & \cdots & 1 & & & \\ 0.5\rho_{\theta\tau} & \cdots & 0.5\rho_{\theta\tau} & 0.25 & & \\ \vdots & \ddots & \vdots & \vdots & \ddots & \\ 0.5\rho_{\theta\tau} & \cdots & 0.5\rho_{\theta\tau} & 0.15 & \cdots & 0.25 \end{pmatrix}.$$

In such a case, the covariance of two latent abilities is $\sigma_{\theta\theta'} = 0.8$ (i.e., correlation coefficient $\rho_{\theta\theta'} = 0.8$) and the covariance of

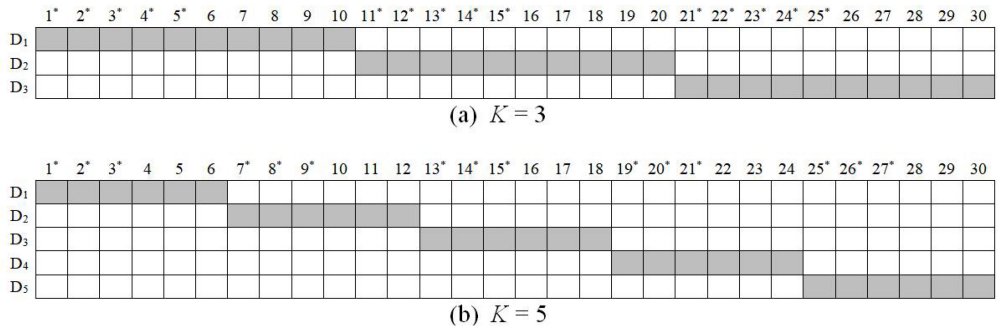


FIGURE 1 | *K-by-I Q' matrix in the simulation study 1. D* = dimension of latent ability; items with * are used for *I* = 15 conditions.

TABLE 4 | Recovery of item parameters in simulation study 1.

				Mean Bias			Mean RMSE			Cor		
I	*N*	$\rho_{\theta\tau}$	*K*	*d*	ξ	ω	*d*	ξ	ω	*d*	ξ	ω
15	500	−0.4	3	0.000	0.001	−0.013	0.106	0.021	0.075	0.995	0.999	NA
			5	0.011	−0.001	−0.024	0.110	0.023	0.077	0.995	0.999	NA
		−0.7	3	−0.006	0.000	−0.016	0.098	0.024	0.073	0.996	0.999	NA
			5	0.009	0.001	−0.017	0.114	0.022	0.085	0.994	0.999	NA
	1000	−0.4	3	−0.001	−0.001	−0.009	0.076	0.016	0.051	0.997	1.000	NA
			5	0.001	0.001	−0.012	0.074	0.015	0.056	0.998	1.000	NA
		−0.7	3	−0.002	0.000	−0.011	0.077	0.015	0.052	0.997	1.000	NA
			5	0.002	0.000	−0.014	0.077	0.016	0.053	0.997	1.000	NA
30	500	−0.4	3	−0.006	0.000	−0.015	0.110	0.022	0.070	0.994	0.999	NA
			5	0.003	0.000	−0.018	0.106	0.022	0.073	0.995	0.999	NA
		−0.7	3	−0.001	−0.001	−0.017	0.103	0.022	0.067	0.995	0.999	NA
			5	−0.003	0.000	−0.019	0.106	0.023	0.074	0.995	0.999	NA
	1000	−0.4	3	0.001	−0.001	−0.007	0.075	0.016	0.047	0.997	1.000	NA
			5	−0.003	0.000	−0.007	0.076	0.015	0.051	0.997	1.000	NA
		−0.7	3	0.000	0.000	−0.008	0.077	0.016	0.050	0.997	0.999	NA
			5	−0.002	0.000	−0.010	0.076	0.016	0.051	0.997	1.000	NA

I = test length; N = sample size; $\rho_{\theta\tau}$ *= correlation coefficient between ability and speed; K = number of dimensions of ability; d = item easiness;* ξ *= item time-intensity;* ω *= item time-discrimination; Mean Bias = mean bias across all items; Mean RMSE = mean root mean square error across all items; Cor = correlation between estimated and true values. Cor of* ω *is NA because of the variance of true* ω *is zero.*

two latent speeds is $\sigma_{\tau\ \tau'} = 0.15$ (i.e., correlation coefficient $\rho_{\tau\ \tau'} = 0.6$). Thirty data sets were generated.

Analysis

In simulation study 1, the MMJ model was fitted to each of the 30 replications. In each replication, two Markov chains with random starting points were used, and each chain ran 10,000 iterations with the first 5,000 iterations in each chain as burn-in. Finally, the remaining 10,000 iterations were used for the model parameter inferences. The potential scale reduction factor (PSRF; Brooks and Gelman, 1998) was computed to assess the convergence of each parameter. A PSRF with values smaller than 1.2 indicates convergence. Our studies indicated that the PSRF was smaller than 1.1 for all parameters, suggesting good convergence.

To evaluate parameter recovery, the bias and the root mean square error (RMSE) was computed as: $(\hat{\upsilon}) = \sum_{r=1}^{R} \frac{\hat{\upsilon}-\upsilon}{R}$ and $\text{RMSE}(\hat{\upsilon}) = \sqrt{\sum_{r=1}^{R} \frac{(\hat{\upsilon}-\upsilon)^2}{R}}$, where $\hat{\upsilon}_r$ is the estimated value of the model parameter in *r*th replication and υ is the true value of the corresponding model parameter, respectively; *R* is the total number of replications. The correlation between estimated and true values (Cor) was also computed.

Results

Table 4 presents the recovery of item parameters. All item parameters were well recovered. The recovery of time-intensity was the best, followed by time-discrimination, and then item easiness. An increasing sample size yielded a better recovery of item parameters. It seems that test length, the correlation coefficient between latent ability and latent speed, and the number of dimensions have a limited impact on the recovery of item parameters.

Tables 5, **6** present the recovery of ability and speed, respectively. First, the recovery of multiple speed factors was better than that of abilities. Increasing test length yielded a better recovery of person parameters; by contrast, increasing the number of dimensions yielded a worse recovery of person parameters. In addition, the higher the correlation coefficient between ability and speed, the better the recovery of latent abilities becomes; however, the correlation coefficient had little effect on the recovery of latent speeds.

Table 7 presents the recovery of the item mean vector and item variance-covariance. Increasing test length and sample size yielded a better recovery. However, the correlation coefficient between ability and speed and the number of dimensions had a limited effect on the recovery. Additionally, the recovery of covariances (omitted, due to space limitations) was better than that of variances of item parameters.

Tables 8, **9** present the recovery of variances of person parameters. Similar to the pattern of the recovery of ability and speed, the recovery of variances of multiple speed factors was better than that of abilities. Increasing test length, sample size, and the correlation coefficient between ability and speed yielded a better parameter recovery. By contrast, more dimensions led to a worse recovery of variances of person parameters. Additionally, the recovery of covariances (omitted, due to space limitations) was better than that of variances of person parameters.

In general, the recovery of time-related parameters (e.g., item intensity, the covariance of item easiness and time-intensity, speed factors, and covariance of ability and speed) was better than that of time-unrelated parameters (e.g., item easiness and latent abilities). Overall, simulation study 1 indicated that model parameters of the MMJ could be recovered very well via the proposed full Bayesian MCMC estimation algorithm.

AN EMPIRICAL EXAMPLE

Data Description and Analysis

In this section, the PISA 2012 computer-based mathematics RA and RT data were analyzed by using the MMJ model and the MSJ model to explore whether the former fits the data better than the latter when the test structure is multidimensional. Details about this data set were mentioned previously in the motivating example. The Q-matrix in **Table 1** was used. For each model, in each replication, the numbers of chains, burn-in iterations, and post-burn-in iterations were the same as those set in the simulation study. Convergence was well achieved according to the PSRF < 1.1.

Posterior predictive model checking (PPMC; Gelman et al., 2014) was used to evaluate model-data fit. A posterior predictive probability (*ppp*) value near 0.5 indicates that there are no systematic differences between the realized and predictive values, and thus an adequate fit of the model. In PPMC, the sum of the squared Pearson residuals for person n and item i (Yan et al., 2003) was used as a discrepancy measure to evaluate the overall fit of the RA model, which is written as

$$D(Y_{ni}; \theta_{nk}, d_i, q_{ik}) = \sum_{n=1}^{N}\sum_{i=1}^{I}\left(\frac{Y_{ni} - P(Y_{ni}=1)}{\sqrt{P(Y_{ni}=1)(1-P(Y_{ni}=1))}}\right)^2,$$

where $P(Y_{ni} = 1)$ has the same definition as that in Equation (6). The sum of the standardized error function of $\log T_{ni}$ for person n and item i was employed as a discrepancy measure of the RT model:

$$D(\log T_{ni}; \xi_i, \tilde{\tau}_{ni}, \omega_i) = \sum_{n=1}^{N}\sum_{i=1}^{I}\left(\omega_i(\log T_{ni} - (\xi_i - \tilde{\tau}_{ni}))\right)^2.$$

TABLE 5 | Recovery of multidimensional ability in simulation study 1.

				Mean Bias					Mean RMSE					Cor				
I	*N*	$\rho_{\theta\tau}$	*K*	θ_1	θ_2	θ_3	θ_4	θ_5	θ_1	θ_2	θ_3	θ_4	θ_5	θ_1	θ_2	θ_3	θ_4	θ_5
15	500	−0.4	3	0.000	0.000	0.000			0.599	0.599	0.598			0.798	0.800	0.800		
			5	0.000	0.000	0.000	0.000	0.000	0.623	0.627	0.624	0.625	0.623	0.780	0.779	0.782	0.781	0.781
		−0.7	3	0.000	0.000	0.000			0.520	0.518	0.519			0.854	0.854	0.854		
			5	0.001	0.000	0.000	0.000	0.000	0.522	0.529	0.526	0.523	0.524	0.853	0.849	0.851	0.853	0.850
	1000	−0.4	3	0.000	0.000	0.000			0.592	0.592	0.594			0.803	0.803	0.802		
			5	0.000	0.000	0.000	0.000	0.000	0.615	0.617	0.619	0.618	0.617	0.786	0.785	0.783	0.783	0.785
		−0.7	3	0.000	0.000	0.000			0.515	0.514	0.514			0.856	0.856	0.856		
			5	0.000	0.000	0.000	0.000	0.000	0.519	0.522	0.522	0.524	0.519	0.854	0.852	0.851	0.850	0.854
30	500	−0.4	3	0.000	0.000	0.000			0.497	0.495	0.497			0.866	0.867	0.866		
			5	0.000	0.000	0.000	0.000	0.000	0.540	0.536	0.536	0.534	0.526	0.840	0.842	0.843	0.844	0.849
		−0.7	3	0.000	0.000	0.000			0.448	0.450	0.449			0.893	0.892	0.892		
			5	0.000	0.000	0.000	0.000	0.000	0.474	0.474	0.470	0.473	0.470	0.879	0.880	0.881	0.880	0.881
	1000	−0.4	3	0.000	0.000	0.000			0.491	0.489	0.490			0.869	0.870	0.869		
			5	0.000	0.000	0.000	0.000	0.000	0.528	0.528	0.526	0.529	0.529	0.846	0.847	0.848	0.847	0.846
		−0.7	3	0.000	0.000	0.000			0.447	0.450	0.448			0.892	0.892	0.892		
			5	0.000	0.000	0.000	0.000	0.000	0.469	0.468	0.473	0.470	0.470	0.882	0.883	0.880	0.882	0.881

I = test length; N = sample size; $\rho_{\theta\tau}$ = correlation coefficient between ability and speed; K = number of dimensions of ability; θ = latent ability; Mean Bias = mean bias across all persons; Mean RMSE = mean root mean square error across all persons; Cor = correlation between estimated and true values.

TABLE 6 | Recovery of multifactor speed in simulation study 1.

				Mean Bias					Mean RMSE					Cor				
I	*N*	$\rho_{\theta\tau}$	*K*	τ_1	τ_2	τ_3	τ_4	τ_5	τ_1	τ_2	τ_3	τ_4	τ_5	τ_1	τ_2	τ_3	τ_4	τ_5
15	500	−0.4	3	0.000	0.000	0.000			0.191	0.194	0.194			0.922	0.920	0.920		
			5	0.000	0.000	0.000	0.000	0.000	0.225	0.225	0.227	0.227	0.226	0.891	0.888	0.890	0.889	0.890
		−0.7	3	0.000	0.000	0.000			0.191	0.190	0.188			0.922	0.924	0.925		
			5	0.000	0.000	0.000	0.000	0.000	0.221	0.221	0.221	0.224	0.224	0.895	0.895	0.895	0.892	0.893
	1000	−0.4	3	0.000	0.000	0.000			0.193	0.191	0.193			0.921	0.923	0.921		
			5	0.000	0.000	0.000	0.000	0.000	0.226	0.226	0.224	0.227	0.224	0.890	0.891	0.892	0.889	0.892
		−0.7	3	0.000	0.000	0.000			0.189	0.189	0.189			0.925	0.925	0.925		
			5	0.000	0.000	0.000	0.000	0.000	0.221	0.221	0.221	0.223	0.223	0.895	0.894	0.895	0.894	0.894
30	500	−0.4	3	0.000	0.000	0.000			0.144	0.145	0.146			0.957	0.956	0.956		
			5	0.000	0.000	0.000	0.000	0.001	0.177	0.177	0.177	0.176	0.177	0.934	0.935	0.934	0.935	0.934
		−0.7	3	0.000	0.000	0.000			0.143	0.144	0.144			0.957	0.957	0.957		
			5	0.000	0.000	0.000	0.000	0.000	0.175	0.175	0.176	0.173	0.175	0.935	0.936	0.935	0.937	0.936
	1000	−0.4	3	0.000	0.000	0.000			0.144	0.144	0.146			0.957	0.957	0.956		
			5	0.000	0.000	0.000	0.000	0.000	0.175	0.175	0.176	0.177	0.176	0.936	0.935	0.935	0.934	0.935
		−0.7	3	0.000	0.000	0.000			0.142	0.143	0.144			0.958	0.958	0.957		
			5	0.000	0.000	0.000	0.000	0.000	0.174	0.174	0.173	0.174	0.174	0.936	0.936	0.937	0.937	0.936

I = test length; N = sample size; $\rho_{\theta\tau}$ = correlation coefficient between ability and speed; K = number of dimensions of ability; τ = speed factor; Mean Bias = mean bias across all persons; Mean RMSE = mean root mean square error across all persons; Cor = correlation between estimated and true values.

TABLE 7 | Recovery of item mean vector and item variance-covariance in simulation study 1.

				Bias					RMSE				
I	*N*	$\rho_{\theta\tau}$	*K*	σ_d^2	$\sigma_{d\xi}$	σ_ξ^2	μ_d	μ_ξ	σ_d^2	$\sigma_{d\xi}$	σ_ξ^2	μ_d	μ_ξ
15	500	−0.4	3	0.155	0.021	0.095	0.020	0.006	0.169	0.027	0.096	0.027	0.007
			5	0.135	0.013	0.097	0.023	0.005	0.162	0.016	0.097	0.029	0.007
		−0.7	3	0.151	0.016	0.095	0.024	0.006	0.164	0.020	0.095	0.029	0.007
			5	0.142	0.021	0.094	0.026	0.005	0.163	0.026	0.094	0.032	0.007
	1000	−0.4	3	0.158	0.015	0.096	0.013	0.005	0.164	0.018	0.096	0.016	0.006
			5	0.124	0.015	0.095	0.015	0.004	0.136	0.020	0.095	0.018	0.006
		−0.7	3	0.150	0.018	0.096	0.016	0.005	0.159	0.022	0.096	0.019	0.006
			5	0.132	0.016	0.096	0.018	0.004	0.144	0.019	0.096	0.021	0.004
30	500	−0.4	3	0.070	0.012	0.046	0.024	0.004	0.083	0.016	0.046	0.027	0.005
			5	0.060	0.012	0.044	0.018	0.003	0.068	0.014	0.045	0.023	0.004
		−0.7	3	0.078	0.010	0.046	0.013	0.004	0.090	0.012	0.046	0.016	0.005
			5	0.056	0.012	0.043	0.019	0.003	0.066	0.014	0.043	0.023	0.004
	1000	−0.4	3	0.062	0.006	0.044	0.013	0.003	0.069	0.008	0.044	0.016	0.004
			5	0.053	0.008	0.045	0.013	0.002	0.060	0.010	0.045	0.017	0.003
		−0.7	3	0.068	0.010	0.046	0.009	0.003	0.079	0.012	0.046	0.012	0.004
			5	0.044	0.007	0.046	0.012	0.002	0.052	0.009	0.046	0.015	0.003

I = test length; N = sample size; $\rho_{\theta\tau}$ = correlation coefficient between ability and speed; K = number of dimensions of ability; σ_d^2 = variance of item easiness; σ_ξ^2 = variance of item time-intensity; $\sigma_{d\xi}$ = covariance of item easiness and item time-intensity; RMSE = mean root mean square error.

Additionally, two information criteria that suitable for Bayesian estimation, the deviance information criterion (DIC) and widely available information criterion (WAIC) (Gelman et al., 2014, Chapter 7), were computed for model selection. A smaller value of these two criteria indicates a better model-data fit.

Results

The DIC and WAIC both identified that the MMJ model fit the data better than the MSJ model, as shown in **Table 10**. In the MMJ model, the *ppp* values of the RA model and the RT model were 0.736 and 0.578, respectively, which indicates an adequate model-data fit. The results indicate that it is more appropriate to simultaneously consider the multidimensionality of latent ability and the multifactor structure of working speed for the multidimensional test.

Note that the parameter estimates of the MMJ model in the empirical example were omitted for brevity but can be found in the online **Supplementary Appendix** (see **Supplementary section S2**), mainly because this part of the content is not the main concern of the empirical study.

TABLE 8 | Recovery of the variance of ability in simulation study 1.

				Bias					RMSE				
I	*N*	$\rho_{\theta\tau}$	*K*	$\sigma_{\theta 1}^2$	$\sigma_{\theta 2}^2$	$\sigma_{\theta 3}^2$	$\sigma_{\theta 4}^2$	$\sigma_{\theta 5}^2$	$\sigma_{\theta 1}^2$	$\sigma_{\theta 2}^2$	$\sigma_{\theta 3}^2$	$\sigma_{\theta 4}^2$	$\sigma_{\theta 5}^2$
15	500	−0.4	3	0.002	−0.047	−0.026			0.152	0.154	0.142		
			5	−0.047	−0.088	−0.055	−0.056	−0.062	0.139	0.193	0.195	0.184	0.168
		−0.7	3	−0.007	0.000	0.010			0.140	0.121	0.142		
			5	−0.036	−0.066	−0.004	−0.016	−0.061	0.157	0.164	0.148	0.166	0.135
	1000	−0.4	3	−0.058	−0.033	−0.042			0.101	0.100	0.104		
			5	−0.072	−0.077	−0.023	−0.095	−0.092	0.123	0.147	0.116	0.139	0.140
		−0.7	3	−0.034	−0.018	−0.015			0.106	0.105	0.099		
			5	−0.071	−0.088	−0.067	−0.056	−0.045	0.148	0.139	0.117	0.131	0.118
30	500	−0.4	3	0.007	–0.035	0.010			0.090	0.099	0.078		
			5	−0.068	−0.085	−0.086	−0.054	−0.055	0.127	0.123	0.136	0.111	0.112
		−0.7	3	−0.014	−0.019	−0.017			0.082	0.097	0.080		
			5	−0.030	–0.075	−0.034	–0.070	−0.056	0.100	0.131	0.099	0.110	0.117
	1000	−0.4	3	−0.009	0.003	–0.040			0.060	0.057	0.063		
			5	−0.070	−0.033	−0.070	–0.084	–0.042	0.099	0.097	0.101	0.107	0.084
		−0.7	3	0.011	−0.032	−0.006			0.045	0.087	0.056		
			5	−0.050	−0.060	−0.069	−0.069	−0.072	0.100	0.091	0.110	0.113	0.101

I = test length; N = sample size; $\rho_{\theta\tau}$ *= correlation coefficient between ability and speed; K = number of dimensions of ability;* σ_θ^2 *= variance of ability; RMSE = mean root mean square error.*

TABLE 9 | Recovery of the variance of speed factor in simulation study 1.

				Bias					RMSE				
I	*N*	$\rho_{\theta\tau}$	*K*	$\sigma_{\tau 1}^2$	$\sigma_{\tau 2}^2$	$\sigma_{\tau 3}^2$	$\sigma_{\tau 4}^2$	$\sigma_{\tau 5}^2$	$\sigma_{\tau 1}^2$	$\sigma_{\tau 2}^2$	$\sigma_{\tau 3}^2$	$\sigma_{\tau 4}^2$	$\sigma_{\tau 5}^2$
15	500	−0.4	3	0.002	0.001	0.004			0.010	0.010	0.010		
			5	0.004	–0.003	0.003	0.001	−0.001	0.012	0.016	0.017	0.013	0.013
		−0.7	3	0.001	0.002	0.002			0.011	0.009	0.010		
			5	0.007	0.002	0.002	0.003	0.001	0.015	0.014	0.013	0.012	0.015
	1000	−0.4	3	0.000	0.003	0.000			0.006	0.008	0.008		
			5	–0.003	0.005	0.000	−0.001	−0.001	0.009	0.013	0.011	0.011	0.008
		−0.7	3	0.002	0.001	0.001			0.009	0.007	0.008		
			5	0.001	–0.002	0.002	−0.002	0.003	0.010	0.010	0.009	0.011	0.009
30	500	−0.4	3	0.004	0.002	0.003			0.008	0.008	0.008		
			5	0.002	0.003	0.001	0.002	−0.001	0.009	0.008	0.009	0.009	0.009
		−0.7	3	0.003	0.002	0.005			0.007	0.008	0.011		
			5	0.003	0.001	0.004	0.002	0.001	0.010	0.009	0.010	0.010	0.010
	1000	–0.4	3	0.003	0.002	0.000			0.006	0.005	0.005		
			5	0.000	0.002	0.003	0.001	0.000	0.006	0.008	0.009	0.008	0.006
		−0.7	3	0.001	0.002	0.002			0.005	0.007	0.006		
			5	0.000	0.000	0.002	0.001	−0.001	0.008	0.006	0.006	0.007	0.006

I = test length; N = sample size; $\rho_{\theta\tau}$ *= correlation coefficient between ability and speed; K = number of dimensions of ability;* σ_τ^2 *= variance of speed factor; RMSE = mean root mean square error.*

TABLE 10 | Model fit for the PISA 2012 computer-based mathematics data.

Analysis Model	DIC	WAIC	*ppp*_RA	*ppp*_RT
MMJ	**34853**	**34433**	0.736	0.578
MSJ	35910	35669	0.608	0.569

MMJ, multidimensional-multifactor joint model; MSJ, multidimensional-single-factor joint model; DIC, deviance information criterion; WAIC, widely available information criterion; ppp, posterior predictive probability value; RA, item response accuracy; RT, item response times. Relatively smaller values are indicated in bold.

DISCUSSION

The kernel hypothesis of this study is that respondents can work with different levels of speed on items that require different dimensions of ability for a multidimensional test. To model the varying speed across dimensions of ability, this study relaxed the assumption of many RT models in which it is assumed that respondents work with a constant rate throughout the test. As a result, a multifactor working speed model and a joint model for

multidimensional ability and multifactor speed were proposed.

First, a motivating example with the EFA of PISA 2012 computer-based mathematics RTs was presented. The results indicate that working speed has a multifactor structure, which is also consistent with the multidimensional structure of ability. Then, two simulation studies were used to evaluate the psychometric properties of the proposed joint model. The results indicate that (1) parameters of the proposed joint model could be well recovered using the proposed Bayesian MCMC approach, (2) misspecifying a multifactor structure of speed has limited effect on the recovery of model parameters, and (3) ignoring the multifactor structure of speed could lead to biased and imprecise estimation, especially for time-related parameters. The PISA 2012 computer-based mathematics RA and RT data were analyzed as well to illustrate the implications and applications of the proposed models. The results show that it is appropriate to consider the multidimensionality of latent ability and the multifactor structure of working speed, simultaneously, in multidimensional tests. Overall, considering the results of EFA, the simulation studies, and the empirical example, there are reasons to believe that the kernel hypothesis of this study is valid and the proposed model can reasonably jointly analyze RA and RTs in multidimensional tests.

The work presented in this article is only a first attempt to deal with the variable speed across dimensions of ability. Despite promising results, further exploration is encouraged. First, the proposed MLRT model is an extension of the classical lognormal RT model (van der Linden, 2006). Thus, there are some limitations of the current model. For instance, it assumes that RA and RTs are conditionally independent given all person parameters (Meng et al., 2015; Bolsinova and Maris, 2016); that after log-transformation, the log RTs follow a normal distribution (Klein Entink et al., 2009b); and that all respondents apply the same problem-solving strategy throughout the whole test (Wang and Xu, 2015).

Second, although the proposed model takes into account the differences in working speed across different dimensions of ability, it still assumes that the working speed of a respondent is constant on items within the same dimension. In future studies, this hypothesis can be further relaxed; that is, each respondent could be allowed to change his or her working speed in different dimensions, and could also be allowed to adjust his or her working speed within the same dimension according to the order of items.

Third, in the proposed joint model, a multivariate normal distribution was used to describe the relationships among multidimensional ability and multifactor speed. So, the number of total dimensions is twice as many as the number of dimensions that are measured by the test, which may increase the complexity of the model and the computational burden. If the ability and speed can each have a second-order (or bi-factor) structure, not only can the parameter estimation challenge be largely reduced, but the structures of ability and speed can be posited and tested.

Fourth, in this study, only the MR model and the MLRT model were used as measurement models for illustration. Given the "plug-and-play" nature of the hierarchical modeling, various MIRT models and multifactor working speed models can be adopted in the future.

Fifth, applications of the proposed model, such as detecting aberrant responses (e.g., rapid-guessing and cheating) in multidimensional tests, need further investigation.

Moreover, in Bayesian estimation, the prior distribution reflects the data analyst's beliefs and the known information about the data. In practice, we recommend that the data analyst select appropriate prior distributions based on the actual test scenario rather than copy those given in this study.

Last but not least, only the between-item multidimensional test was considered in this study. For the between-item multidimensional test, it is clear that working speed can vary across items when the items are related to different dimensions. However, the within-item multidimensional test is still possible in reality. For example, when respondents, especially non-native English speakers, take part in the GRE® Subject Test (e.g., Mathematics), at least two abilities are needed: one for understanding the questions (e.g., English reading ability), and one for solving the questions (e.g., the subject ability). Meanwhile, the corresponding two latent speed factors work; one reflects the working speed of reading, and the other one reflects the working speed of problem-solving. The introduction of within-item multidimensionality is bound to increase the complexity of the model and the difficulty of constructing the Q-matrix. Thus, the rationality and necessity of the within-item multifactor working speed model is still an open-ended question needed to be studied in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm and http://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm.

AUTHOR CONTRIBUTIONS

PZ contributed to the conception, design, and analysis of data as well as paper drafting and revising the manuscript. HJ contributed to the design and critically revising the manuscript. KM contributed to the critically revising the manuscript. W-CW contributed to conception, design, and revising the manuscript. KH contributed to the interpretation of data and critically revising the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 3190 0795).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.469196/full#supplementary-material

REFERENCES

Adams, R. J., Wilson, M., and Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Appl. Psychol. Meas.* 21, 1–23. doi: 10.1177/0146621697211001

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Bolsinova, M., and Maris, G. (2016). A test for conditional independence between response time and accuracy. *Br. J. Math. Stat. Psychol.* 69, 62–79. doi: 10.1111/bmsp.12059

Bolsinova, M., and Tijmstra, J. (2018). Improving precision of ability estimation: getting more from response times. *Br. J. Math. Stat. Psychol.* 71, 13–38. doi: 10.1111/bmsp.12104

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Graphical Stat.* 7, 434–455. doi: 10.2307/1390675

De Boeck, P., and Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Front. Psychol.* 10:102. doi: 10.3389/fpsyg.2019.00102

Fox, J.-P., Entink, R. K., and Timmers, C. (2014). The joint multivariate modeling of multiple mixed response sources: relating student performances with feedback behavior. *Multiv. Behav. Res.* 49, 54–66. doi: 10.1080/00273171.2013.843441

Fox, J.-P., and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multiv. Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85, 398–409. doi: 10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. New York, NY: Chapman & Hall.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Modeling* 6, 1–55. doi: 10.1080/10705519909540118

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/S11336-008-9075-Y

Klein Entink, R. H., van der Linden, W. J., and Fox, J.-P. (2009b). A Box-Cox normal model for response times. *Br. J. Math. Stat. Psychol.* 62, 621–640. doi: 10.1348/000711008X374126

Man, K., Harring, J. R., Jiao, H., and Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Appl. Psychol. Meas.* 43, 639–654. doi: 10.1177/0146621618824853

Meng, X.-B., Tao, J., and Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *J. Educ. Meas.* 52, 1–27. doi: 10.1111/jedm.12060

Molenaar, D., Oberski, D., Vermunt, J., and De Boeck, P. (2016). Hidden markov item response theory models for responses and response times. *Multiv. Behav. Res.* 51, 606–626. doi: 10.1080/00273171.2016.1192983

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *Br. J. Math. Stat. Psychol.* 68, 197–219. doi: 10.1111/bmsp.12042

Muthén, L. K., and Muthén, B. (2019). *Mplus. The Comprehensive Modeling Program for Applied Researchers: User's Guide*. Los Angeles, CA: Mutheìn & Mutheìn, 5.

OECD, (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing. doi: 10.1787/9789264190511-en

Plummer, M. (2015). *JAGS Version 4.0.0 User Manual*. Available online at: http://sourceforge.net/projects/mcmc-jags/ (accessed March 1, 2021).

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.

Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multiv. Behav. Res.* 25, 173–180. doi: 10.1207/s15327906mbr2502_4

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

van der Linden, W. J. (2006). A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J., Klein Entink, R., and Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Appl. Psychol. Meas.* 34, 327–347. doi: 10.1177/0146621609349800

Wang, C., Chang, H., and Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

Wang, C., Weiss, D. J., and Su, S. (2019). Modeling response time and responses in multidimensional health measurement. *Front. Psychol.* 10:51. doi: 10.3389/fpsyg.2019.00051

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Yan, D., Mislevy, R. J., and Almond, R. G. (2003). *Design and Analysis in a Cognitive Assessment* (ETS Research Report Series, RR-03-32). Princeton, NJ: ETS.

Zhan, P., Jiao, H., and Liao, D. (2018). Cognitive diagnosis modeling incorporating item response times. *Br. J. Math. Stat. Psychol.* 71, 262–286. doi: 10.1111/bmsp.12114

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Zhan, Jiao, Man, Wang and He. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 March 2021
doi: 10.3389/fpsyg.2021.579128

Improving the Precision of Ability Estimates Using Time-On-Task Variables: Insights From the PISA 2012 Computer-Based Assessment of Mathematics

Denise Reis Costa[1]*, *Maria Bolsinova*[2], *Jesper Tijmstra*[2] and *Björn Andersson*[1]

[1] Centre for Educational Measurement, Faculty of Educational Sciences, University of Oslo, Oslo, Norway, [2] Methodology and Statistics, Tilburg University, Tilburg, Netherlands

Log-file data from computer-based assessments can provide useful collateral information for estimating student abilities. In turn, this can improve traditional approaches that only consider response accuracy. Based on the amounts of time students spent on 10 mathematics items from the PISA 2012, this study evaluated the overall changes in and measurement precision of ability estimates and explored country-level heterogeneity when combining item responses and time-on-task measurements using a joint framework. Our findings suggest a notable increase in precision with the incorporation of response times and indicate differences between countries in how respondents approached items as well as in their response processes. Results also showed that additional information could be captured through differences in the modeling structure when response times were included. However, such information may not reflect the testing objective.

Keywords: log files, computer-based assessment, time on task, measurement precision, measurement invariance, PISA

OPEN ACCESS

Edited by:
Ian van der Linde,
Anglia Ruskin University,
United Kingdom

Reviewed by:
Qiwei He,
Educational Testing Service,
United States
Mark D. Reckase,
Michigan State University,
United States

***Correspondence:**
Denise Reis Costa
d.r.costa@cemo.uio.no

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *01 July 2020*
Accepted: *22 February 2021*
Published: *19 March 2021*

Citation:
Reis Costa D, Bolsinova M, Tijmstra J and Andersson B (2021) Improving the Precision of Ability Estimates Using Time-On-Task Variables: Insights From the PISA 2012 Computer-Based Assessment of Mathematics. Front. Psychol. 12:579128. doi: 10.3389/fpsyg.2021.579128

1. INTRODUCTION

Computers have become increasingly common implements used in classroom activities over the past few decades. As a reflection of this trend, large-scale educational assessments have moved from paper and pencil based tests to administrated computer assessments. In addition to being more efficient and reducing human error, computer-based assessments allow for a greater variety of tasks. Further, interactive computer environments can be used to generate log files, which provide easy access to information concerning the examinee response process. These log files contain time-stamped data that provide a complete overview of all communication between the user-interface and server (OECD, 2019). As such, it is possible to trace how respondents interact with the testing platform while gathering information about the amount of time spent on each task.

The first computer-based administration of the Programme for International Student Assessment (PISA) dates back to 2006 (OECD, 2010). However, more extensive studies involving log files were enabled through the release of the PISA 2009 digital reading assessment (OECD, 2011). In this context, time-on-task and navigating behaviors can be extracted from these log files as relevant variables. The information derived from variables of this type can help teachers further

understand the solution strategies used by students while also enabling a substantive interpretation of respondent-item interactions (Greiff et al., 2015; Goldhammer and Zehner, 2017). The variables taken from log files can also be included in sophisticated models designed to improve student proficiency estimations (van der Linden, 2007).

While log file data from computer-based assessments have been available for several years, few studies have investigated how they can be used to improve the measurement precision of resulting scores. Using released items from the 2012 PISA computer-based assessment of mathematics, this study thus explored the potential benefits of incorporating time-on-task variables when estimating student proficiency. We specifically compared three different models to advance the current understanding of what time-on-task adds to scores resulting from an international large-scale assessment program.

1.1. Time-On-Task and Item Responses

Several previous studies have investigated the relationship between time-on-task and item responses. For example, Goldhammer et al. (2015) studied the relationship between item responses and response times through a logical reasoning test, thus finding a non-linear relationship between reasoning skills and response times. Further, Goldhammer and Klein Entink (2011) investigated how time-on-task and item interactivity behaviors were related to item responses using complex problem-solving items. In addition, Naumann and Goldhammer (2017) found a non-linear relationship between time on task and performance on digital reading items from the PISA 2009 assessment. Finally, Goldhammer et al. (2014) studied the relationship between time-on-task, reading, and problem solving using PIAAC data. Results indicated that the association between time-on-task and performance varied from negative to positive depending on the subject matter and type of task.

In large-scale educational assessments, student proficiency is mainly estimated through the item response theory (IRT) framework (von Davier and Sinharay, 2013). Here, categorical item-response data are considered manifestations of an underlying latent variable that is interpreted as, for example, mathematics proficiency. While time-on-task can be incorporated in several different ways from an IRT perspective (van der Linden, 2007), the state-of-the-art view considers them as realizations of random variables, much like actual item responses (Kyllonen and Zu, 2016). A hierarchical model is most commonly used with time-on-task data. Specifically, a two-level structure is used to incorporate time-on-task, item responses, and latent variables into a single model (van der Linden, 2007). While the hierarchical modeling framework has the advantage of considering both response accuracy and response times as latent variables, it has practical limitations in that it requires specialized software for fitting the model. Molenaar et al. (2015) illustrated how the hierarchical model can be slightly simplified such that standard estimation techniques could be used. This type of formulation of the model allows the use of both generalized linear latent variable models (Skrondal and Rabe-Hesketh, 2004) and non-linear mixed models (Rijmen et al., 2003) with item-response and time-on-task data. Furthermore, the approach outlined by Molenaar et al. (2015) encompasses not only the standard hierarchical model (with the necessary simplification) but also its extensions which allow for more complex relationship between time-on-task and ability such as the model of Bolsinova and Tijmstra (2018). For these reasons, this study pursued the approach of Molenaar et al. (2015) for its analysis of PISA 2012 data.

1.2. The Present Study

This study investigated the utility of combining item responses with time-on-task data in the context of a large-scale computer-based assessment of mathematics. It also evaluated the properties of the employed model with respect to each participating country[1]. Specifically, the framework developed by Molenaar et al. (2015) was used to investigate how measurement precision was influenced by incorporating item responses and time-on-task data into a joint model. We also explored country-level heterogeneity in the time-on-task measurement model. As such, the model proposed for this analysis of computer-based large-scale educational assessments implied a different set of underlying assumptions than current procedures. Specifically, we viewed response-time data as comprising an extra information set that enabled us to gain additional insight regarding the latent construct of interest. This also implies that any inference regarding the underlying construct at the country level would potentially change through the proposed approach as opposed to current analysis methods, which this study also investigated. The three following research questions were thus proposed:

- RQ1: What changes occur in the overall ability estimates and their level of precision regarding PISA 2012 digital mathematics items when time-on-task data are included in the analysis?
- RQ2: How do time-on-task model parameters differ across items and countries?
- RQ3: What changes occur in country-level performance when time-on-task data are considered in the analysis?

Our findings should add to the current literature on the relationship between time-on-task and responses to performance items. Our results also have important implications for large-scale assessment programs in regard to evaluating the added measurement precision that is granted by incorporating additional data sources (e.g., time-on-task). Such investigations can inform large-scale assessment programs about whether and how time-on-task data should be included in models designed to generate operational results reports.

2. DATA AND METHODS

2.1. The 2012 PISA Computer-Based Assessment of Mathematics

PISA administered its first computer-based mathematics literacy assessment as part of its fifth program edition.

[1]We use "country" as a generic term referring to all countries and economies participating in the PISA study. A list of the countries codes is displayed in the **Supplementary Material.**

TABLE 1 | Sample size, mean score, and variation in student performance on all clusters, as well as sample size, percentage of female, average total time, and percentage of missing responses for the 10 released and valid log-file data from the PISA 2012 computer-based mathematics by country.

Country	All clusters (41 items)			Valid log-file data (10 items)			
	n	Mean	S.D.	*n*	% Female	Average total time (min)	% Missing response
SGP	2,873	566.02	98.34	453	49.89	16.13	3.22
QCN	2,409	562.26	93.64	393	49.87	16.13	0.64
KOR	2,675	552.57	90.15	433	44.34	13.92	1.20
HKG	2,714	549.64	86.71	421	45.37	15.37	2.45
MAC	3,147	542.90	82.85	522	50.00	17.92	3.41
JPN	6,351	539.01	87.80	982	46.44	15.65	3.21
TAP	3,063	537.26	88.80	513	51.27	15.13	2.51
CAN	10,817	522.85	91.92	1,527	51.34	14.28	4.16
EST	2,837	516.09	82.13	460	50.00	14.54	2.37
BEL	4,617	512.15	98.60	707	49.50	14.19	4.82
DEU	2,881	509.37	95.50	441	51.02	13.75	2.43
FRA	3,012	508.06	91.95	440	53.18	15.43	4.41
AUS	11,834	507.70	90.94	1,833	48.88	13.55	1.99
AUT	2,731	507.34	88.74	436	50.92	13.42	1.28
ITA	3,089	498.76	83.14	440	45.68	16.54	6.86
USA	2,572	498.03	88.75	402	46.77	14.57	1.89
NOR	2,924	497.56	87.25	413	48.67	13.48	2.20
SVK	3,145	497.34	86.07	505	44.75	16.24	5.88
DNK	4,149	496.19	86.41	629	51.83	13.51	1.43
IRL	2,613	493.08	80.50	389	51.41	14.85	3.26
SWE	2,671	489.93	86.06	423	52.48	13.84	3.62
RUS	3,186	489.15	79.83	531	50.28	16.36	4.24
POL	2,567	489.04	86.01	428	52.10	13.09	1.64
PRT	3,272	489.03	85.09	487	48.05	15.52	3.29
SVN	4,385	486.94	87.83	678	45.87	10.95	0.65
ESP	5,751	475.08	81.99	933	50.38	14.44	3.63
HUN	2,746	469.84	92.58	445	52.81	12.79	1.82
ISR	2,677	446.61	111.28	387	54.78	14.65	2.48
ARE	6,732	434.06	84.28	1057	51.09	14.03	4.07
BRA	3,172	420.74	83.85	480	50.00	16.40	9.92
COL	5,173	396.84	73.33	782	53.58	16.48	8.09
Overall mean	3,961			612	49.77	14.62	3.40

(1) Countries are displayed by the ISO three-letter code. Their correspondence names are available at the ***Supplementary Material****. (2) Countries are sorted by mean scores. (3) Mean scores and variation were retrieved by OECD (2014c) report, Annex B3. OECD does not provide the overall mean quantities. Remaining statistics were calculated by the authors using PISA 2012 micro and log-file data.*

A total of 32 countries participated in this effort. In this context, 40 min were allocated for the computer-based portion of the test, with math items arranged in 20 min clusters that were assembled with digital reading or problem-solving prompts (OECD, 2014a). A total of 41 math items were selected for this assessment. These items varied from standard multiple-choice to constructed response formats.

Table 1 presents the characteristics of the PISA sample by country (sample size, Math performance, and variation) for the whole computer-based of mathematics clusters (41 items) as well as to the subsample with available and valid log-file data (10 items).

We utilized data from a total of 18,970 students across 31 countries. We excluded data from Chile since log-file data for two of the analyzed items were unavailable (I20Q1 and I20Q3). Students with invalid information (e.g., those that did not receive final scores or had incomplete timing information) were also excluded from the analysis. On average, the sample size of each country is around 600 (S.D.= 333), the percentage of female is 50% across all the countries. The average total time on the 10 items varied from 10.95 to 17.92 min. Brazil was the country with the highest percentage of missing responses (9.92%) on the analyzed items.

The analyzed log-file data from 10 items were made publicly available on the OECD website. We thus extracted the time

TABLE 2 | Characteristics of the released PISA 2012 computer-based of mathematics items.

Item	Format	% correct	Thresholds - PISA scale		Average response time (min)	% Missing response
			1	2		
I15Q1	MC	59.02	498.51		1.36	0.81
I15Q2	CR	8.43	685.84	700.72	1.87	0.93
I15Q3	CR	29.02	577.18	658.58	1.98	3.25
I20Q1	CR	29.58	562.07	690.91	2.18	1.69
I20Q2	MC	47.42	549.29		0.96	1.93
I20Q3	CR	26.91	644.25		1.33	2.55
I20Q4	MC	44.12	565.73		0.84	3.30
I38Q3	MC	67.13	468.75		1.25	3.94
I38Q5	CR	27.75	641.05		1.82	6.60
I38Q6	CR	23.24	660.45		1.56	8.96

(1) Item are displayed by the position within the cluster. (2) MC = Multiple Choice and CR= Constructed Response item type. (3) The international percent of correct responses, and thresholds values were retrieved by OECD (2014b) report, Annex A1. Remaining statistics were calculated by the authors using PISA 2012 log-file data. (4) Threshold values in PISA 2012 were defined as the ability at which the probability of achieving that score or higher reaches 0.62 using a partial credit model.

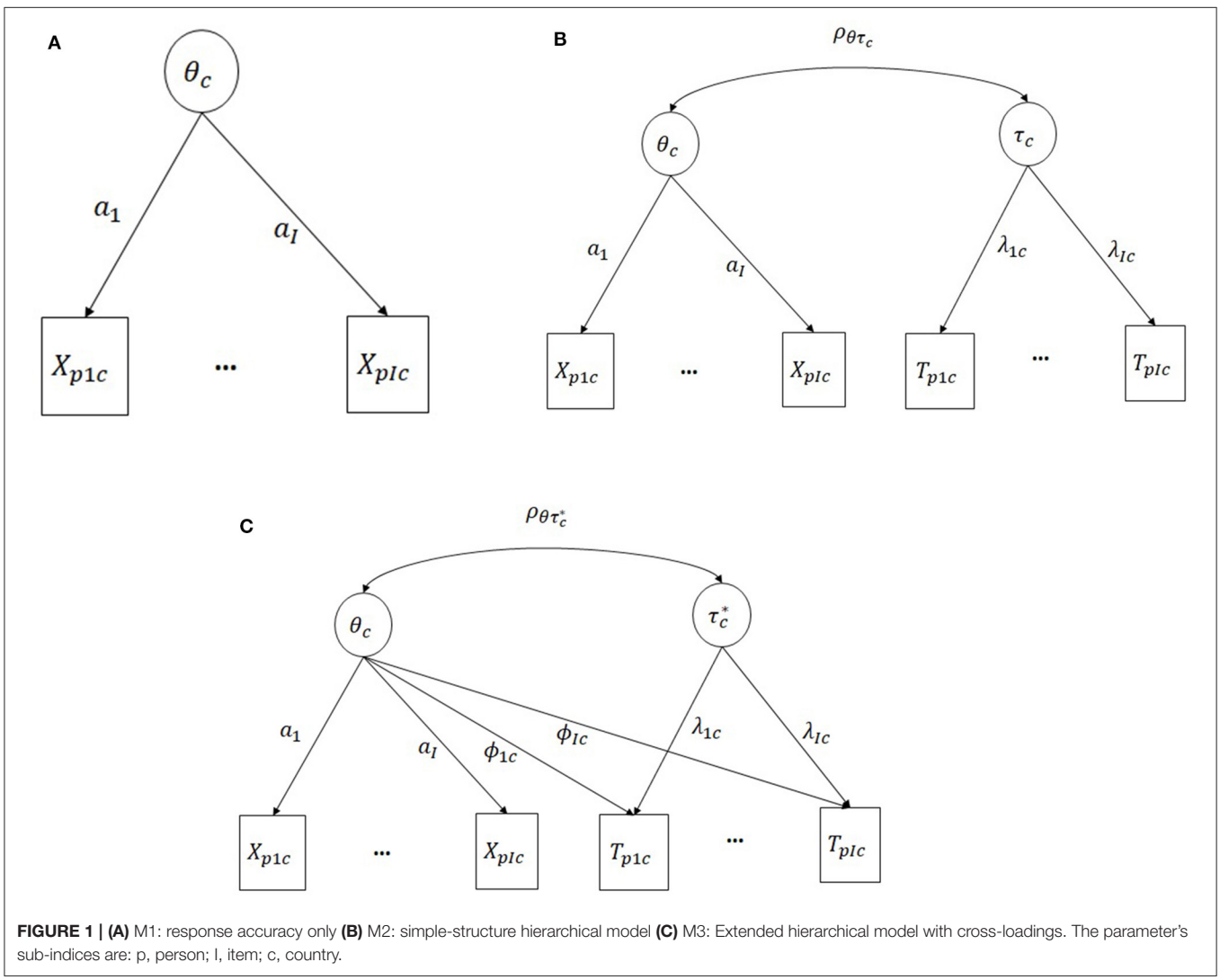


FIGURE 1 | (A) M1: response accuracy only **(B)** M2: simple-structure hierarchical model **(C)** M3: Extended hierarchical model with cross-loadings. The parameter's sub-indices are: p, person; I, item; c, country.

students spent on analyzed items and their final responses (i.e., response accuracy). All items were allocated in three units (CD production: items "I15Q1," "I15Q2," and "I15Q3"; Star points: items "I20Q1," "I20Q2," "I20Q3," and "I20Q4"; Body Mass Index: items "I38Q3," "I38Q5," and "I38Q6") and were administered in the same cluster.

Table 2 shows the reported item characteristics by OECD (international percent of correct responses, and thresholds used for scaling the items in PISA 2012) as well as the average response time and percentage of missing responses by item.

Although the effects of the item position were likely negligible due to the length of the computer-based assessment (OECD, 2014b), we were still able to determine that the percentages of missing data were larger for items located at the end of the cluster. We used the full information maximum likelihood approach (FIML) featured in Mplus version 7.3 (Muthén and Muthén, 2012) to incorporate all available data into our analyses. Doing this, the missing responses were treated as missing at random (MAR) and all the available data were incorporated in the modeling.

2.2. Statistical Analyses

This study compared three measurement models to estimate student proficiency based on the abovementioned PISA dataset. All these models can be seen as special cases of the framework of Molenaar et al. (2015). They are:

- Model 1 (M1): It provided a baseline and thus only included response accuracy in a unidimensional IRT framework. The model can be seen as a special case of the framework of Molenaar et al. (2015) in which it is assumed that there is no relationship between latent proficiency and response time data.
- Model 2 (M2): A multidimensional latent variable model for the response accuracy and response times, where the response accuracy are related to a latent proficiency and the response times are related to a latent speed. The latent factors are assumed to be correlated. This is a variant of the model described in Molenaar et al. (2015): Here the relationship between the latent proficiency and response times is specified through the relationship between the latent proficiency and latent speed.
- Model 3 (M3): A multidimensional latent variable model for the response accuracy and response times, where response accuracy is related to a latent proficiency and the response times are related to a latent speed and proficiency. This is also a special case of the approach of Molenaar et al. (2015)

TABLE 5 | Estimated means and variances of students' abilities, EAP reliability and average of the standard errors for the three measurement models.

Model	Mean ($\hat{\mu}_\theta$)	Variance ($\hat{\sigma}^2_\theta$)	EAP reliability	Average SE
M1	0.00	1.00	0.73	0.51
M2	−0.02	1.06	0.77	0.49
M3	−0.02	1.05	0.80	0.45

International results of the PISA 2012 digital math items.

TABLE 3 | Framework for the estimation of international parameters for each analyzed model.

Model	μ_θ	σ^2_θ	$\mu_{\tau c}$	$\sigma^2_{\tau c}$	a_is	b_is	ξ_is	λ_is	$\sigma^{2}{}_i$s	$\rho_{\theta\tau}$	ϕ_is
M1	0	1	-	-	Free	Free	-	-	-	-	-
M2	Free	Free	0	1	a_is M1	b_is M1	Free	Free	Free	Free	-
M3	Free	Free	0	1	a_is M1	b_is M1	Free	Free	Free	0	Free

For M3, the second latent variable (τ) was rotated to obtain the estimates of the transformed factor loadings (i.e., the factor loadings for speed correspond to the relationship between response time and the latent variable which has the same interpretation as τ in M2).*

TABLE 4 | Framework for the estimation of countries' parameters for each analyzed model.

Model	$\mu_{\theta c}$	$\sigma^2_{\theta c}$	$\mu_{\tau c}$	$\sigma^2_{\tau c}$	a_is	b_is	ξ_{ic}s	λ_{ic}s	σ^2_{ic}s	$\rho_{\theta\tau c}$	ϕ_{ic}s
M1_Full	Free	Free	-	-	a_is M1	b_is M1	-	-	-	-	-
M2_Full	Free	Free	Free	Free	a_is M1	b_is M1	ξ_is M2	λ_is M2	σ_i^2s M2	Free	-
M2_Strong	Free	Free	Free	Free	a_is M1	b_is M1	ξ_is M2	λ_is M2	Free	Free	-
M2_Weak	Free	Free	0	Free	a_is M1	b_is M1	Free	λ_is M2	Free	Free	-
M2_Struct	Free	Free	0	1	a_is M1	b_is M1	Free	Free	Free	Free	-
M3_Full	Free	Free	Free	Free	a_is M1	b_is M1	ξ_is M3	λ_is M3	σ_i^2s M3	Free	ϕ_is M3
M3_Strong	Free	Free	Free	Free	a_is M1	b_is M1	ξ_is M3	λ_is M3	Free	Free	ϕ_is M3
M3_Weak	Free	Free	0	Free	a_is M1	b_is M1	Free	λ_is M3	Free	Free	ϕ_is M3
M3_Struct	Free	$\sigma^2_{\theta c}$ M1_Full	0	1	a_is M1	b_is M1	Free	Free	Free	0	Free

(1) Since ϕ_{ic}s are not freely estimated in the models M3_Full, M3_Strong, and M3_Weak, the correlation between the latent variables is identified. For the M3_Struct model, however, we constrained the variance of the latent speed to be the same as the estimates from the M1_Full model to make sure that the correlations between X_is and θ will be the same as in model 1 and therefore θ will have similar interpretation as M1. (2) After obtaining the parameter estimates in the M3_Struct model, the second latent variable (τ) was rotated to match the latent speed variable in M2, and $\rho_{\theta\tau c}$ was calculated.*

TABLE 6 | Model fit statistics (BIC) by model and country.

Country	M2 models		M3 models	
	Range BIC	Preferred	Range BIC	Preferred
ARE	[29647.12–29774.20]	M2_Struct	[29537.07–29815.97]	M3_Weak
AUS	[54700.07–54766.34]	M2_Weak	[54328.41–54415.99]	M3_Full
AUT	[13147.99–13240.82]	M2_Full	[13049.54–13149.53]	M3_Full
BEL	[20621.68–20670.20]	M2_Full	[20487.24–20539.84]	M3_Weak
BRA	[12137.77–12214.90]	M2_Weak	[12091.05–12181.83]	M3_Weak
CAN	[43637.52–43803.38]	M2_Strong	[43440.33–43598.05]	M3_Weak
COL	[21367.18–21653.93]	M2_Weak	[21289.22–21653.15]	M3_Weak
DEU	[13002.93–13065.20]	M2_Full	[12919.52–13010.30]	M3_Full
DNK	[18226.95–18269.57]	M2_Full	[18090.74–18175.80]	M3_Full
ESP	[27485.45–27588.13]	M2_Full	[27287.14–27437.58]	M3_Full
EST	[12629.05–12761.87]	M2_Strong	[12482.02–12625.00]	M3_Strong
FRA	[12890.52–12913.41]	M2_Weak	[12772.27–12824.20]	M3_Weak
HKG	[13527.25–14034.09]	M2_Struct	[13479.78–13752.47]	M3_Struct
HUN	[12181.87–12197.26]	M2_Strong	[12110.29–12155.05]	M3_Strong
IRL	[10701.96–10754.46]	M2_Strong	[10602.46–10685.45]	M3_Strong
ISR	[12298.66–12436.20]	M2_Weak	[12229.09–12370.03]	M3_Weak
ITA	[12674.72–12748.87]	M2_Weak	[12559.46–12616.12]	M3_Full
JPN	[30542.91–31258.25]	M2_Struct	[30320.88–31046.07]	M3_Struct
KOR	[13060.99–13190.37]	M2_Struct	[12977.21–13138.38]	M3_Weak
MAC	[15390.05–15536.21]	M2_Weak	[15250.20–15396.36]	M3_Weak
NOR	[12867.52–12905.24]	M2_Strong	[12784.82–12845.24]	M3_Weak
POL	[12530.05–12563.72]	M2_Full	[12444.90–12511.19]	M3_Full
PRT	[14017.90–14079.03]	M2_Full	[13934.64–14045.98]	M3_Full
QCN	[11640.00–11840.43]	M2_Weak	[11579.94–11665.66]	M3_Struct
RUS	[15790.27–15843.04]	M2_Weak	[15683.03–15754.20]	M3_Weak
SGP	[13129.26–13240.15]	M2_Weak	[13016.65–13090.06]	M3_Strong
SVK	[13976.93–14004.75]	M2_Struct	[13877.89–13903.17]	M3_Full
SVN	[19994.64–20004.67]	M2_Weak	[19942.30–19986.59]	M3_Strong
SWE	[12449.46–12509.05]	M2_Full	[12325.18–12420.05]	M3_Full
TAP	[15235.19–15292.96]	M2_Struct	[15121.61–15158.37]	M3_Full
USA	[11209.98–11277.11]	M2_Weak	[11186.89–11239.37]	M3_Weak
Total BIC	[556031.15–557223.67]	M2_Weak	[552353.33–555491.33]	M3_Weak

(1) The range indicates the minimum and maximum values of the BIC statistic by country. (2) The suffix "_Full" indicates full measurement invariance (fixing all item parameters to be equal to the international estimates), "_Strong" indicates strong measurement invariance (country-specific residual variances are allowed to be estimated, while time intensity parameters and factor loadings are fixed to be equal to the international estimates), "_Weak" means weak measurement invariance model (country-specific residual variances and country-time intensity parameters are freely estimated, while factor loadings are fixed to be equal to the international estimates), and "_Struct," structural measurement invariance (wholly fitted time-related parameters, i.e., all time-related parameters are freely estimated in each country).

in which the relationship between latent proficiency and response times goes not only through the relationship between latent proficiency and latent speed, but also through the direct relationship between the ability and individual response times. For this model we employed a particular rotation approach described in Bolsinova and Tijmstra (2018).

Figure 1 shows the graphical representation of the models across PISA countries. For comparability purposes, the items' parameters for response accuracy were fixed from model 1 into models 2 and 3. This approach assures that the models are on the same scale since the relationship between response accuracy and latent proficiency will be the same across models.

This section discusses the mathematical formulations used in each model. The steps used to estimate model parameters for use with the PISA dataset and an analysis of measurement invariance across countries are discussed later.

2.2.1. Model Specification

Let $\mathbf{X} = (X_1, \ldots, X_I)$ be a random vector of responses on the I items and $\mathbf{T} = (T_{p1}, \ldots, T_{pI})$ be a random vector of response times on the same items with realizations $\mathbf{x}_{p\cdot} = (x_{p1}, \ldots, x_{pI})$ and $\mathbf{t}_{p\cdot} = (t_{p1}, \ldots, t_{pI})$, for each person p, respectively.

For response accuracy, we adopted the graded response model (GRM) used by Samejima (1969). This was done because some PISA items used a partial scoring method and, unlike

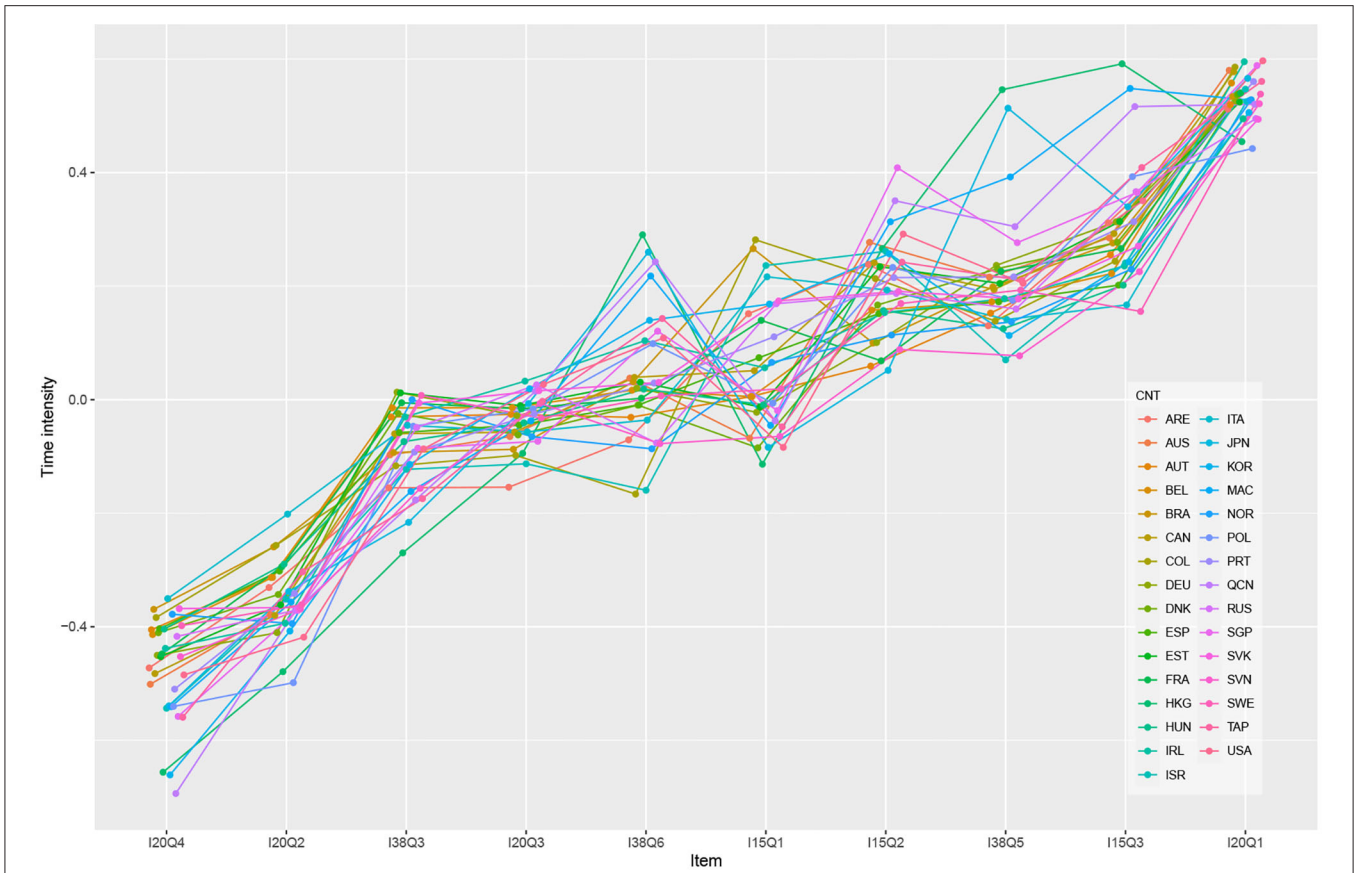


FIGURE 2 | Estimates of the countries' time intensity for model 2—Weak measurement invariance.

other IRT models used for polytomous data (e.g., the partial credit model), the GRM is equivalent to simple factor analytic models in application to discrete data and can therefore be fitted using standard factor analysis software and structural equation models. The differences between the various IRT models used for polytomous data are usually very small; in our case, only three items out of 10 allowed partial scoring. The GRM specifies the conditional probability to obtain each category $k \in [1:m]$, where m is the highest possible category for the item. The conditional probability of obtaining this score or higher, given the latent trait θ, is defined by

$$\Pr\left(X_i \geq k|\theta\right) = \frac{\exp[a_i\left(\theta - b_{ik}\right)]}{1 + \exp[a_i\left(\theta - b_{ik}\right)]}, \quad (1)$$

where a_i is the item factor loading/discrimination parameter, and b_{ik} is the item category threshold parameter[2]. The probability of obtaining a particular response category k is then

$$\Pr\left(X_i = k|\theta\right) = \Pr\left(X_i \geq k|\theta\right) - \Pr\left(X_i \geq k+1|\theta\right), \quad (2)$$

where $\Pr(X_i \geq 0) = 1$ and $\Pr(X_i \geq m+1) = 0$. When $m = 2$, the GRM reduces to the two-parameter logistic IRT model used by Birnbaum (1968), with only one difficulty parameter b_i per item instead of multiple threshold parameters. M1 defined exclusively by Equation (2).

There are also cases in which both responses and response times are used to estimate respondent proficiency. Here, instead of simply specifying the model for response accuracy, we must specify the full model for the joint distribution of response accuracy and response times. For Model 2, we thus adopted the hierarchical modeling approach used by van der Linden (2007), which requires not only the specification of the measurement model for response accuracy (in our case, the GRM) but also the specification of the measurement model for the response times, and the models for the relationship of the latent variables in the two measurement models. The model used for the relationship between item parameters in the two measurement models is often specified, as well. However, as shown by Molenaar et al. (2015), excluding this relationship does not substantially change the parameter estimates, especially when large sample sizes are involved. Furthermore, the use of standard estimation techniques is prevented when including a model for the item parameters. Given the very large sample sizes available in this analysis, we

[2]The logistic function was used here because it is more common than the cumulative normal function used in IRT applications, including for large-scale international assessments such as PISA.

thus specified a higher-order relationship on the person side (i.e., the model for latent variables), but did not do so on the item side.

The joint distribution of response accuracy and response times is conditional to both latent proficiency and speed (denoted by τ) in the hierarchical model. In this case, it is assumed to be a product of the marginal distribution of response accuracy, which only depends on latent proficiency, and the marginal distribution of response time, which only depends on latent speed. We refer to this as a simple-structure model because every observed variable therein is solely related to one latent variable. This differs from the extension of the hierarchical model used by Bolsinova and Tijmstra (2018), which includes direct relationships between response times and latent proficiency in addition to its relation to latent speed.

A lognormal model with item-specific loadings was used for the response times (Klein Entink et al., 2009). It is equivalent to the one-factor model used for log-transformed response times. The conditional distribution of response time on item i given the latent speed variable is defined by

$$T_i \sim \ln \mathcal{N}(\xi_i - \lambda_i \tau, \sigma_i^2), \tag{3}$$

which is the lognormal distribution in which the mean is dependent on the item time intensity ξ_i and the latent speed τ. The strength of the relationship between the response time and the latent speed depends on the factor loading λ_i. Meanwhile, σ_i^2 denotes the item-specific residual variance.

The dependence between the latent proficiency and the latent speed variables is modeled using a bivariate normal distribution with correlation parameter ρ. This correlation between the latent variables specifies the indirect relationship between response times and latent proficiency. In turn, this allows us to strengthen the measurement of proficiency (i.e., increase measurement precision) by using the information contained in the response times. The magnitude of the improvement in measurement precision is solely determined by the size of the correlation between the latent speed and latent proficiency (Ranger, 2013).

M3 employed the same model for response accuracy as that used in M1 and M2. However, a different model was used for response times. That is, the mean of the lognormal distribution of response time was dependent on two latent variables, as follows:

$$T_i \sim \ln \mathcal{N}(\xi_i - \lambda_i \tau^* + \phi_i \theta, \sigma_i^2), \tag{4}$$

where the cross-loading ϕ_i specifies the strength of the relationship between response time and proficiency. Here, an asterisk is used for the latent variable τ^* because it should be interpreted differently from the simple-structure model (M2). Since the cross-loadings between latent proficiency and response

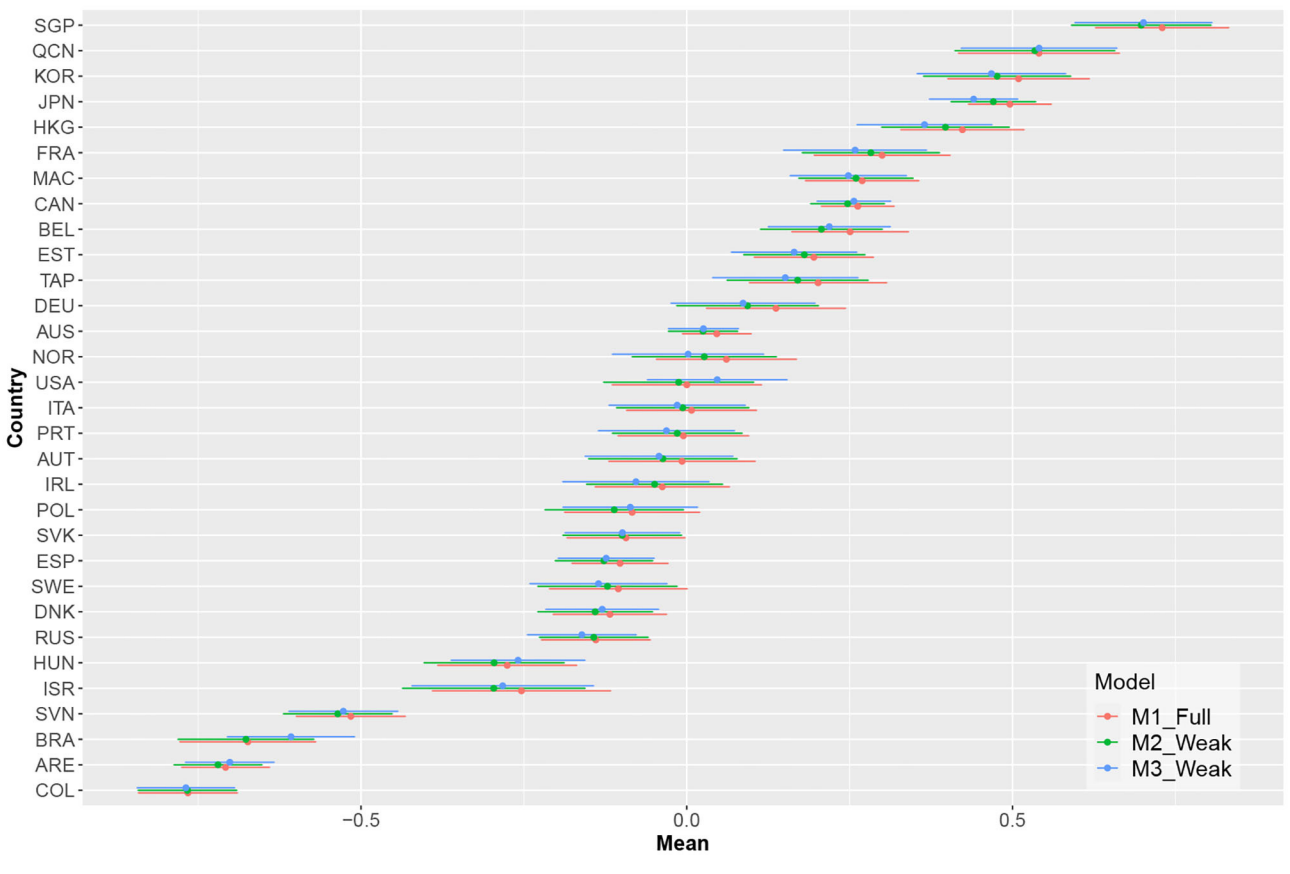


FIGURE 3 | Estimates of the countries' means and their respective confidence intervals for the different models.

time are freely estimated, the correlation between θ and τ^* is not identified and is instead fixed to zero so that τ^* can be interpreted as a latent variable, thus explaining the covariance of the response times that cannot be explained by latent proficiency. However, it is possible to rotate the latent variable τ^* to match the latent speed variable of the simple-structure model. Following Bolsinova and Tijmstra (2018), we will apply a rotation of the factors such that τ^* is the latent variable that explains most of the variance of response times. In that case, the correlation between latent proficiency and speed and the corresponding values for the transformed factor loading in the two dimensions can be calculated.

2.2.2. Analysis Strategies

We used the LOGAN R package version 1.0 (Reis Costa and Leoncio, 2019) to extract student response times and accuracy from the PISA 2012 log file containing data for 10 digital math items. We then conducted analyses according to two steps.

First, we fitted all three models by combining the sample consisting of 31 countries to estimate model parameters at an international level. Then, we analyzed the models across PISA countries by fixing specific parameters from previous analyses to allow cross-model comparisons. We also evaluated parameter invariance in the response time model. All model parameters were estimated using the restricted maximum likelihood method in Mplus version 7.3 (Muthén and Muthén, 2012).

Table 3 summarizes the analytical framework used in the first step. Item discrimination (a_is) and threshold parameters (b_is) were freely estimated for Model 1, with the proficiency mean (μ_θ) and variance (σ^2_θ) fixed to 0 and 1, respectively. To enable model comparisons, the item discrimination and threshold parameters were not estimated for M2 and M3 but were rather fixed to the parameter estimates from M1. For these models, the response time parameters (ξ_is, λ_is, σ^2_is, and ϕ_is) and the mean and variance of the proficiency were freely estimated.

All analyses were conducted assuming the same graded response model for the item-response modeling. We evaluated the fit of the GRM model for M1 (in which the item discrimination and threshold parameters were freely estimated) by calculating two approximate fit statistics [i.e., the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR)] using the complete dataset in the mirt R package (Chalmers, 2012). As a guideline, cutoff value close to 0.08 for SRMR and a cutoff value close to 0.06 for RMSE indicated acceptable fit (Hu and Bentler, 1999).

We conducted country-level analyses in the second step. **Table 4** shows the fixed and freely estimated parameters for each model. Here, models containing the suffix "_Full" indicate full

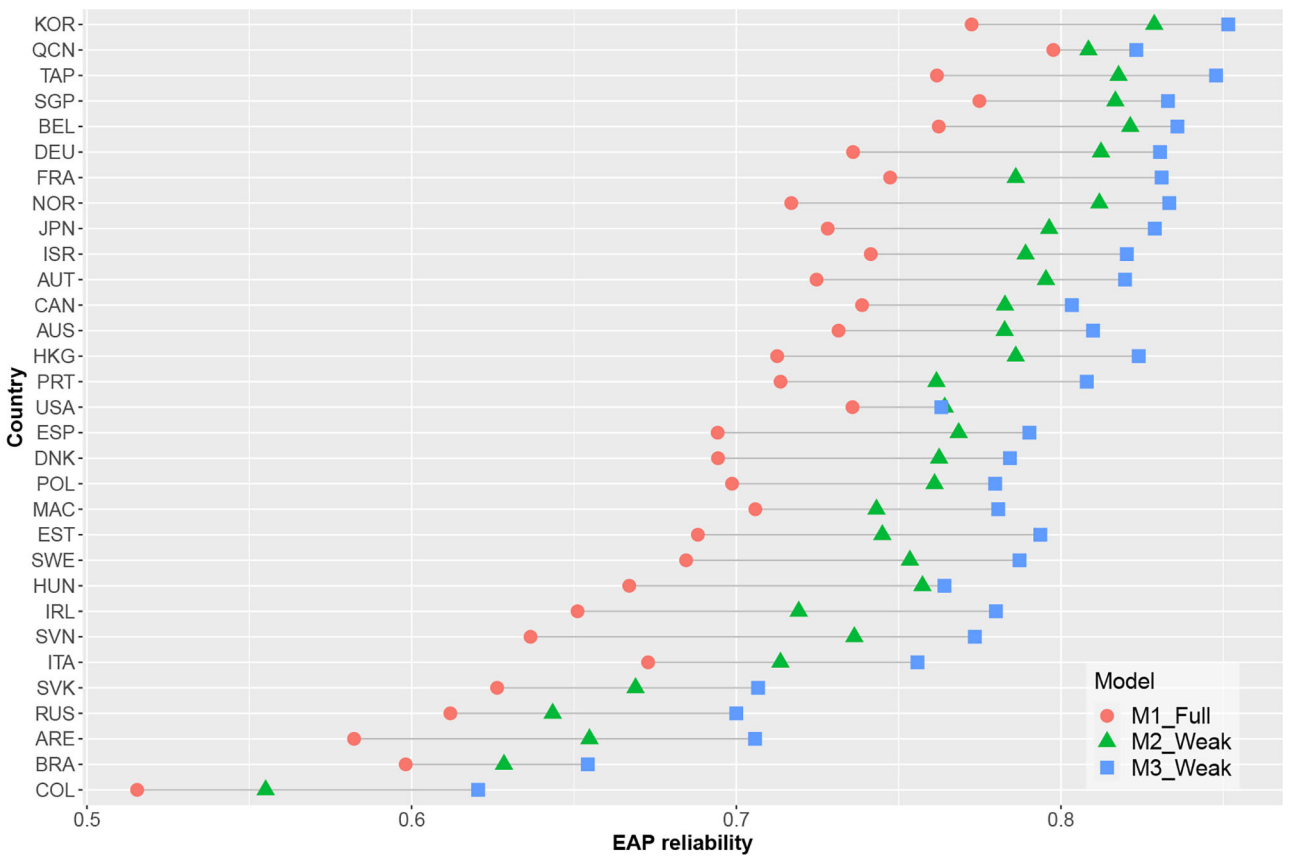


FIGURE 4 | EAP reliabilities estimates per country and model.

measurement invariance. That is, we estimated each country's mean and variance for the latent variables (θ_c, τ_c, or $\rho_{\theta\tau c}$), fixing all item parameters (a_is, b_is, ξ_{ic}s, λ_{ic}s, σ^2_{ic}s, or ϕ_{ic}s) with international estimates as derived in step one. Models containing the suffix "_Strong" indicate strong measurement invariance in which item-specific residual variances (σ^2_{ic}s) are allowed to be estimated, instead. Weak measurement invariance models contain the suffix "_Weak." Here, both the item-specific residual variance (σ^2_{ic}s) and item-time intensity parameters (ξ_{ic}s) were freely estimated. In this case, however, the mean of the latent speed variable was fixed to 0 for model identification. Lastly, structural measurement invariance (suffix "_Struct") indicates all time-related parameters are freely estimated (ξ_{ic}s, λ_{ic}s, σ^2_{ic}s or ϕ_{ic}s). For model identification, we fixed the mean and the variance of the latent speed variable to 0 and 1, respectively. We also incorporated a new constraint in model "M3_Struct" to allow the free estimation of the cross-loading parameter (ϕ_{ic}s). In this case, we constrained the variance of the latent speed to be the same as the estimates from the M1_Full model to make sure that the correlations between X_is and θ will be the same as in model 1 and therefore θ will have similar interpretation as in M1.

We estimated student abilities using the Expected a Posteriori (EAP) approach (Bock and Mislevy, 1982) and evaluated measurement precision using the EAP-reliability method (Adams, 2005) and the average of the standard errors of the ability estimates. Finally, we computed the Bayesian Information Criterion (BIC) for model selection (Schwarz, 1978).

3. RESULTS

We addressed our research questions by assessing the results according to the following three steps: (1) we estimated the overall ability estimates and their level of precision regarding PISA 2012 digital math items by the three measurement models, (2) presented our findings about the invariance of response-time model parameters across items and countries, and (3) showed changes in country-level performance when time-on-task was considered.

3.1. RQ1: Overall Performance

We first investigated the model fit for the graded response model. This model was assessed as having a good fit based on its SRMSR (0.036). It also exhibited acceptable fit according to its RMSEA (0.050). We thus concluded that our baseline model had sufficiently good overall fit for continued analyses, including those related to time-on-task variables.

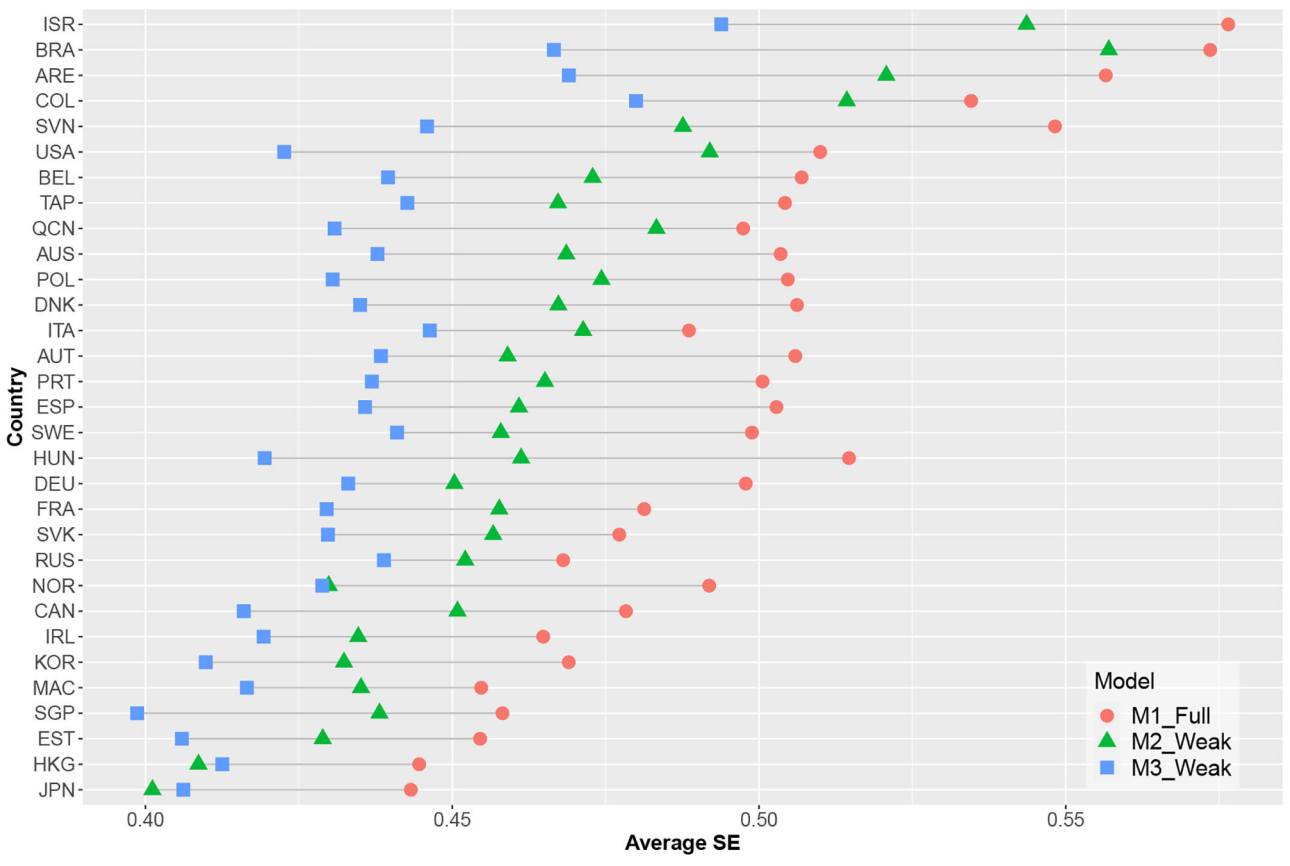


FIGURE 5 | Average standard errors of abilities' estimates per country and model.

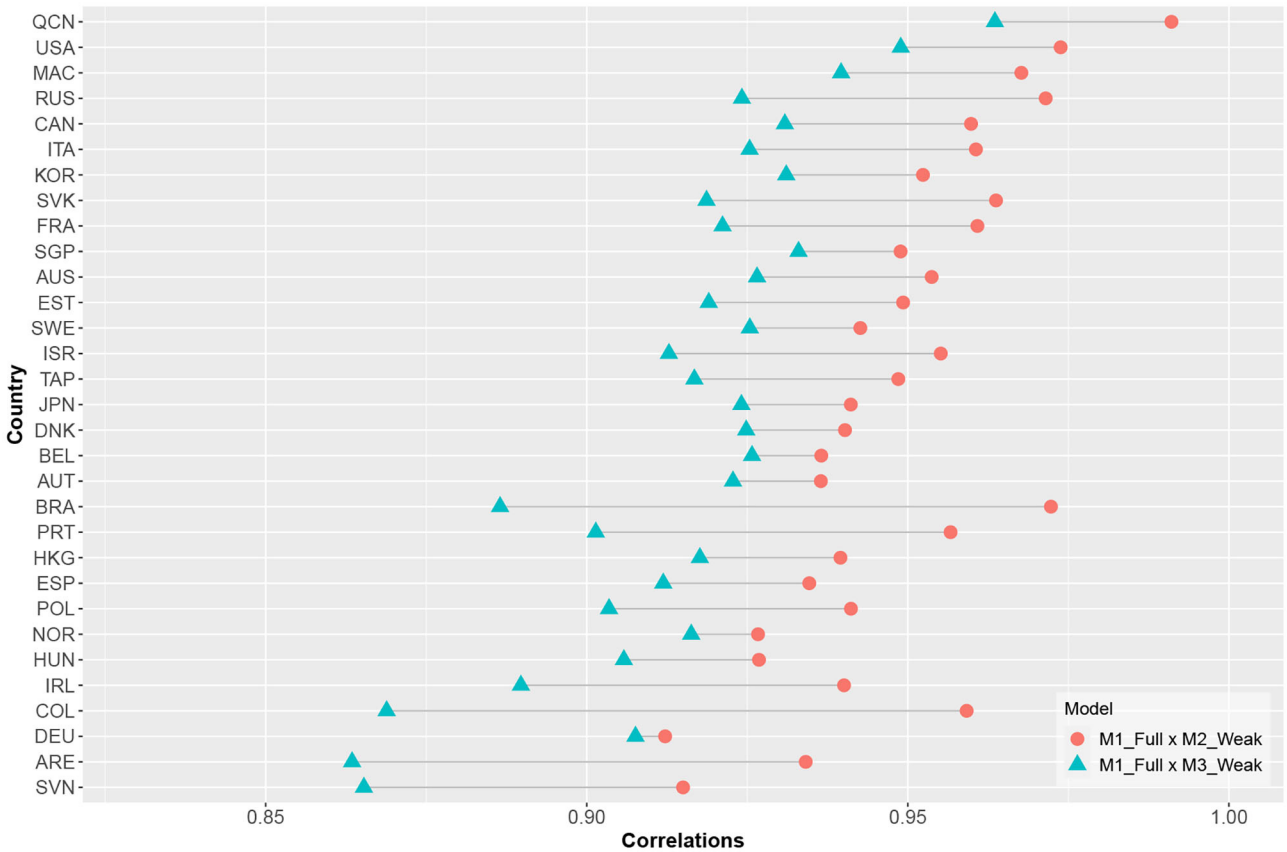


FIGURE 6 | Correlations between EAP estimates.

Table 5 shows the overall estimates for student abilities and the measurement precision of these estimates in relation to the PISA 2012 digital math items across the different models. Although there was no substantial difference, M2 and M3 (i.e., the simple-structure hierarchical model and the cross-loadings model, respectively) exhibited increased measurement precision (as captured by larger EAP reliability estimates and smaller average standard errors) when response times were included in the modeling framework.

3.2. RQ2: Measurement Invariance

We investigated measurement invariance of the time-on-task parameters for each country with both M2 and M3. We also calculated the BIC for each individual model and summarized these statistics to identify the level of invariance that best represented the data overall (**Table 6**). As such, the assumption of invariance of the model's parameters does not hold for most countries and models. Weak measurement invariance were preferred in most of the cases (i.e., there was country-specific heterogeneity in the time intensity (ξ_i) and residual variance (σ_i^2) parameters for the time-on-task measurement models).

To illustrate the differences in the time-on-task measurement model parameters, **Figure 2** presents the estimated time-intensity parameters for each item in each country as applied to the preferred model in the simple-structure framework (M2). The graph indicates that students in all analyzed countries placed the most effort into answering the first item, I20Q1, from Unit 20 (Star Point unit). However, the pattern of estimated time-intensity between different items varied according to country. For example, the estimated time intensity of item I38Q06 was larger than that of item I15Q01 for several countries, but the opposite was found for about just as many countries.

3.3. RQ3: Country-Level Performance

Figure 3 shows the estimated country means in computer-based mathematical literacy and the associated confidence intervals for the three measurement models. The estimated means did not show substantial discrepancies for the analyzed countries between the different models.

Figure 4 shows the estimated reliability of the EAP ability estimates for each country. Measurement precision increased for all countries when time-on-task variables were included; here, the model containing cross-loadings had the highest estimated EAP reliability. As illustrated in **Figure 5**, there was a decrease in average standard errors for ability estimates when time-on-task variables were included.

Figure 6 shows the correlations between EAP ability estimates from the baseline model and from those including time-on-task variables. Ability estimates from models that included cross-loadings generally had lower correlations with the baseline

model-based ability estimates as compared to models that did not include cross-loadings. This indicates that the ability estimates from model 3 captured an additional source of information. However, this may not have reflected the test objective (i.e., estimating student computer-based mathematical literacy).

4. SUMMARY AND DISCUSSION

This study examined the extent to which inferences about ability in large-scale educational assessments were affected by and improved by including time-on-task information in the statistical analyses. This issue was specifically explored using data from the PISA 2012 Computed-Based Assessment of Mathematics. In line with statistical theory, model-based measurement precision (as captured by the EAP reliability estimates) improved when using the standard hierarchical model as opposed to the response accuracy only model for each of the 31 considered countries that participated in the PISA program. This increase was notable for most countries, with many showing increases in estimated EAP reliability at or above 0.05. If such a version of the hierarchical model can adequately capture the data structure, then this suggests it can also provide a notable increase in precision over the default response-accuracy only models.

For practically all countries, model-based measurement precision was further increased when using the extended version of the hierarchical model, which allowed a direct link between response times and ability by including cross-loadings (i.e., rather than using the standard hierarchical model). This model successfully extended the hierarchical model by considering overall response speed as relevant to the estimated ability while also allowing individual item-response times to be linked to said ability if such patterns were present in the data. Thus, the model allowed time-on-task to provide more collateral information when estimating ability than was possible when using the standard hierarchical model. This increased precision was also notable for most countries (generally between 0.02 and 0.03). However, the increase was generally less sizable than those obtained by using the hierarchical model instead of a response-accuracy only model. Thus, the biggest gain in precision was already obtained by using a simple-structure hierarchical model; extending the model by incorporating cross-loadings generally only resulted in modest additional gains.

We investigated the extent to which time-on-task parameters could be considered invariant across countries for both the simple-structure hierarchical model and the extension that included cross-loadings. The results suggested that only weak measurement invariance existed. As such, full or strong measurement invariance did not hold. That is, our findings suggest that countries may differ both in item time-intensity (capturing how much time respondents generally spent on items) and the item-specific variability of the response times (i.e., the degree to which respondents differed in the amounts of time they spent on particular items). This suggests relevant differences between countries in regard to how respondents approached items as well as in their response processes.

Measurement precision improved for all countries when using the selected versions of M2 and M3 (i.e., over the precision levels obtained using M1). Since changing the model used to analyse the data may also affect model-based inferences, we also analyzed the extent to which such inferences would be affected by these changes. Here, no country showed a substantial change in estimated mean, thus suggesting that the overall assessment of proficiency levels for different countries was not heavily affected by a model change. However, the estimated correlations between the individual ability estimates obtained using M1, M2, and M3 showed small deviations from 1 for many countries, suggesting that the ability being estimated does not overlap perfectly across the three models. The differences between M1 and M3 were most notable in this regard. That is, they generally resulted in the lowest correlations between ability estimates. It is thus not surprising that these two models had the lowest correlation; they also had the largest differences in modeling structure. However, one should carefully consider which of the models best operationalizes the specific ability that will be estimated. Additional validation research is thus needed to determine whether the inclusion of time-on-task information results in overall improved measurement quality.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm.

AUTHOR CONTRIBUTIONS

DRC: conceptualization. DRC, MB, JT, and BA: methodology and paper writing. DRC: data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

The authors received financial support from the research group Frontier Research in Educational Measurement at University of Oslo (FREMO) for the publication of this article. JT was supported by the Gustafsson & Skrondal Visiting Scholarship at the Centre for Educational Measurement, University of Oslo.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.579128/full#supplementary-material

TABLE S1 | List of country codes.

REFERENCES

Adams, R. J. (2005). Reliability as a measurement design effect. *Stud. Educ. Eval.* 31, 162–172. doi: 10.1016/j.stueduc.2005.05.008

Birnbaum, A. L. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading: Addison-Wesley), 397–479.

Bock, R. D., and Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Meas.* 6, 431–444. doi: 10.1177/014662168200600405

Bolsinova, M., and Tijmstra, J. (2018). Improving precision of ability estimation: getting more from response times. *Brit. J. Math. Stat. Psychol.* 71, 13–38. doi: 10.1111/bmsp.12104

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* 48, 1–29. doi: 10.18637/jss.v048.i06

Goldhammer, F., and Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence* 39, 108–119. doi: 10.1016/j.intell.2011.02.001

Goldhammer, F., Naumann, J., and Greiff, S. (2015). More is not always better: the relation between item response and item response time in Raven's matrices. *J. Intell.* 3, 21–40. doi: 10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Goldhammer, F., and Zehner, F. (2017). What to make of and how to interpret process data. *Measurement* 15, 128–132. doi: 10.1080/15366367.2017.1411651

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equat. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74:21. doi: 10.1007/s11336-008-9075-y

Kyllonen, P., and Zu, J. (2016). Use of response time for measuring cognitive ability. *J. Intell.* 4:14. doi: 10.3390/jintelligence4040014

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *Brit. J. Math. Stat. Psychol.* 68, 197–219. doi: 10.1111/bmsp.12042

Muthén, B., and Muthén, L. (2012). *Mplus Computer Program Version 7.3*. Los Angeles, CA: Mplus.

Naumann, J., and Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learn. Individ. Diff.* 53, 1–16. doi: 10.1016/j.lindif.2016.10.002

OECD (2010). *PISA Computer-Based Assessment of Student Skills in Science*. Technical report, Paris. doi: 10.1787/9789264082038-en

OECD (2011). *PISA 2009 Results: Students On Line: Digital Technologies and Performance*. OECD Publishing, Paris. doi: 10.1787/9789264112995-en

OECD (2014a). *Pisa 2012 Results: Creative Problem Solving: Students' Skills in Tackling Real-Life Problems*. Vol. 5. Technical report, Paris.

OECD (2014b). *PISA 2012 Technical Report*. OECD Publishing, Paris.

OECD (2014c). *Results: What Students Know and Can Do Student Performance in Mathematics, Reading and Science*. OECD Publishing, Paris.

OECD (2019). *Beyond Proficiency: Using Lof Files to Understand Respondent Behaviour in the Survey of Adult Skills*. OECD Publishing, Paris. doi: 10.1787/0b1414ed-en

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der linden (2007). *Psychometrika* 78, 538–544. doi: 10.1007/s11336-013-9324-6

Reis Costa, D., and Leoncio, W. (2019). *LOGAN: Log File Analysis in International Large-Scale Assessments*. (R package version 1.0.0). Available online at: https://cran.r-project.org/web/packages/LOGAN/index.html

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychol. Methods* 8, 185–205. doi: 10.1037/1082-989X.8.2.185

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*. 34, 1–97. doi: 10.1007/BF03372160

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling:Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72:287. doi: 10.1007/s11336-006-1478-z

von Davier, M., and Sinharay, S. (2013). "Analytics in international large-scale assessments: item response theory and population models," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, eds L. Rutkowski, M. von Davier, and D. Rutkowski (Boca Raton, FL: CRC Press), 155–174.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Reis Costa, Bolsinova, Tijmstra and Andersson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 08 January 2020
doi: 10.3389/fpsyg.2019.02910

A Speed-Accuracy Tradeoff Hierarchical Model Based on Cognitive Experiment

Xiaojun Guo, Zhaosheng Luo and Xiaofeng Yu*

School of Psychology, Jiangxi Normal University, Nanchang, China

Most tests are administered within an allocated time. Due to the time limit, examinees might have different trade-offs on different items. In educational testing, the traditional hierarchical model cannot adequately account for the tradeoffs between response time and accuracy. Because of this, some joint models were developed as an extension of the traditional hierarchical model based on covariance. However, they cannot directly reflect the dynamic relationship between response time and accuracy. In contrast, response moderation models took the residual response time as the independent variable of the response model. Nevertheless, the models enlarge the time effect. Alternatively, the speed-accuracy tradeoff (SAT) model is superior to other experimental models in the SAT experiment. Therefore, this paper incorporates the SAT model with the traditional hierarchical model to establish a SAT hierarchical model. The results demonstrated that the Bayesian Markov chain Monte Carlo (MCMC) algorithm performed well in the SAT hierarchical model of parameters by using simulation. Finally, the deviance information criterion (DIC) more preferred the SAT hierarchical model than other models in empirical data. This means that it is indispensable to add the effect of response time on accuracy, but likewise should limit the effect on the empirical data.

Keywords: response time, accuracy, the speed-accuracy tradeoff, time limit, hierarchical model

OPEN ACCESS

Edited by:
Fernando Marmolejo-Ramos, University of South Australia, Australia

Reviewed by:
Jocelyn Holden Bolin, Ball State University, United States
Rubén Maneiro, Pontifical University of Salamanca, Spain

****Correspondence:***
Zhaosheng Luo
luozhaosheng@aliyun.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *19 July 2019*
Accepted: *09 December 2019*
Published: *08 January 2020*

Citation:
Guo X, Luo Z and Yu X (2020) A Speed-Accuracy Tradeoff Hierarchical Model Based on Cognitive Experiment. Front. Psychol. 10:2910. doi: 10.3389/fpsyg.2019.02910

INTRODUCTION

In any decision-making process, one of the most basic issues is the speed-accuracy tradeoff (SAT). In our various behaviors, the SAT is almost ubiquitous. From insects to primates, the changing trend of speed and accuracy in decision-making process is an inevitable problem. The SAT is defined as an individual's willingness to respond slowly and makes relatively fewer errors compared to their willingness to respond quickly and makes relatively more errors. This means that low speed corresponds to higher accuracy, or high speed corresponds to lower accuracy (Heitz, 2014).

In cognitive experiments, the SAT has been studied for a long time. The relationship between response time and accuracy can be obtained by different methods. In the traditional reaction time experiment, the SAT can be obtained by six basic methods: instructions, payoffs, deadlines, time bands, response signals, and partitioning of reaction times. However, it cannot obtain complete information processing dynamics and can only provide a single time point in different experimental condition. Unlike the traditional reaction time experiment, Reed (1973) and Wickelgren (1977) proposed a SAT experimental paradigm. Compared to the traditional reaction time experiment, the SAT experiment is a different experimental paradigm. In the SAT experiment, processing time is an independent variable or an experimental condition and each experimental condition

is applied to different processing times. Moreover, a speed-accuracy tradeoff model (SAT model) is used to fit the reaction time and accuracy in different experimental conditions. Therefore, SAT model can provide a complete dynamic relationship between reaction time and accuracy. After that, SAT experiment and the model were widely applied in cognitive experiments, such as conceptual processing (McElree et al., 2000), sentence comprehension (McElree, 2000; McElree et al., 2003), Memory (McElree, 1998) and Attention (McElree and Carrasco, 1999; Giordano et al., 2009).

In addition to the SAT model, the sequential sampling models are likewise used to analyze SAT experiments. In the sequential sampling models, the most popular model is the diffusion model. Furthermore, the diffusion model can interpret various SAT criterions by different parameters, such as boundary separation (Ratcliff and Rouder, 1998; Ratcliff et al., 2001; Ratcliff et al., 2003), drift rate (Starns et al., 2012; Rae et al., 2014). McElree and Dosher (1989) derived the expression for response time and accuracy from the diffusion model. In contrast with SAT model, the diffusion model was worse to fit the experimental data. In addition, the density function of the diffusion model is extremely complex (Cox and Miller, 1970), which makes more difficult to apply.

The relationship between response time and response accuracy represents an important area of study within educational testing. In educational testing, the most popular model is the hierarchical model of van der Linden (2007). Moreover, it is defined as the traditional hierarchical model in this paper. The traditional hierarchical model models the relations between speed and accuracy for a population of test takers separately from the impact of these parameters on the responses and times of the individual test takers. The same will be done for the relations between the time and response parameters of the items. Therefore, the relations between the response and time can be captured at a higher level of modeling. In other word, the traditional hierarchical model consists of two levels. The first level is two independent response models and response time model, and the second level is the joint distribution of the person parameters and the joint distribution of the item parameters. The hierarchical model links the correlation between ability and speed to account for the tradeoff between response time and accuracy. Additionally, the hierarchical model greatly promotes the analysis and application of response time and accuracy (Wang et al., 2013, 2018; Meng et al., 2014; Zhan et al., 2018). However, the traditional hierarchical model does not fully explain the relationship between the response time and accuracy. Because of this, Ranger and Ortner (2012) and Meng et al. (2015) further explained the relationship between response time and accuracy based on covariance. However, they cannot directly reflect the dynamic relationship between response time and accuracy. In contrast with covariance, a response moderation model took the residual response time as the independent variable of the response model (Bolsinova et al., 2016, 2017). Nevertheless, the response moderation model enlarged the time effect and ignored the influence of ability on accuracy.

In cognitive experiments, SAT model has obvious advantages, whereas the current hierarchical model has obvious shortcomings in the tradeoff between response time and accuracy. Therefore, a SAT hierarchical model integrates the SAT model with the traditional hierarchical model in this paper. The SAT hierarchical model not only reflects the dynamic relationship between response time and accuracy, but can also avoid the influence of expanding time on accuracy. The paper is organized as follows. Firstly, the SAT hierarchical model is described based on the SAT model. Secondly, a Bayesian estimation procedure is proposed and some simulation studies are used to evaluate parameter recovery. Thirdly, three hierarchical models are compared to an empirical data. Finally, the paper concludes with a discussion.

SAT HIERARCHICAL MODEL

In the paper, the SATHM is based on the hierarchical framework. In the SAT hierarchical model, the SAT response model is formulated by the previous response model and the SAT model. In addition, the other parts are the same with the traditional hierarchical model.

Response Time Model

For the response times, a lognormal model is linked by the latent speed variable (τ_i), the item time intensity (β_j) and the item residual variance (σ_j^2). Within Eq. 1, $\ln T_{ij}$ is the response time of examinee *i* on item *j* after a log transformation.

$$lnT_{ij} \sim N(\beta_j - \tau_i, \sigma_j^2) \tag{1}$$

SAT Response Model

In Eq. 2, SAT model is an exponential function (Reed, 1973, 1976).

$$d'(t) = \lambda \times (1 - \exp(-\varphi \times (t - \delta))) \quad t > \delta \text{ and } t \neq 0 \tag{2}$$

Where λ is the asymptotic level of accuracy, δ is the response time at which accuracy begins to grow above chance or non-decision time, φ represents the slope of the accuracy to asymptote. $d'(t)$ is the accuracy of different response time. In each experimental condition, the three parameters of the SAT model were fitted to each observer's response time and the average accuracy by the method of least squares. Moreover, the SAT model can determine the effect of experimental conditions by adding different parameters.

In the traditional hierarchical model, the basic assumption of the response model is that probability is not included time-limit effect. However, there is no doubt that time limits can detract from average examinee performance in that examinees correctly answered fewer items with the imposed time limits. Therefore, it is very necessary to model a response model that takes into account the impact of response time and ability. In the SAT model, λ is the asymptotic level of accuracy with no time limit. It is consistent with the assumptions of the response model in the traditional hierarchical model. Because of this, the lambda (λ) of the SAT model is defined as two-parameter logistic model (2PLM):

$$\lambda_{ij} = P(\eta_{ij} = 1 | a_j, b_j; \theta_i) = \frac{\exp(a_j \times \theta_i - b_j)}{1 + \exp(a_j \times \theta_i - b_j)} \tag{3}$$

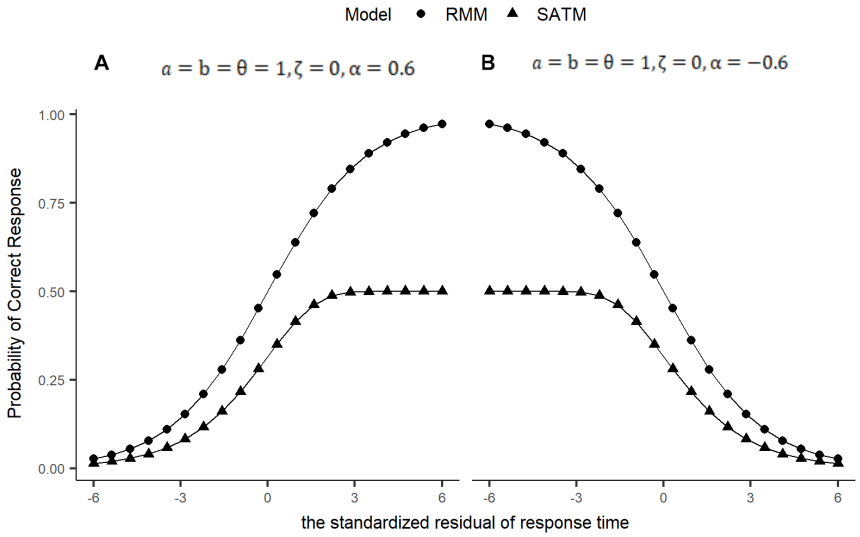


FIGURE 1 | The relationship between residual time and accuracy. **(A)** $\alpha > 0$, **(B)** $\alpha < 0$

Where η_{ij} is the latent response of examinee i for item j. θ_i denotes the ability parameter, a_j and b_j are the discrimination and difficulty for item j.

In the educational test, the tradeoffs of different test takes on the item may be different. Therefore, the parameters of the SAT model should be reconstructed. For the time term $\varphi \times (t - \delta)$, it can be replaced with the term $\alpha_j Z_{ij} + \zeta$. Z_{ij} is the standardized residual log-response time of examinee i for item j, which reflects the difference between the observation time and the expected time (Eq. 4). α_j is the slope of residual time for item j, and ζ is the intercept of the effect of residual time on the test. Due to the condition of $t - \delta > 0$, the exponential transformation is added in the term $(\alpha_j Z_{ij} + \zeta)$. Finally, the SAT response model is established (SATM, Eq. 5). Furthermore, when the time is sufficient, the SAT response model is transformed to 2PLM. Due to response time as a random variable, the response time may be different if an examinee on the item can be answered more than once. The SATM can describe the theoretical relationship between the different response time and accuracy.

$$Z_{ij} = \frac{lnT_{ij} - (\beta_j - \tau_i)}{\sigma_j} \tag{4}$$

$$P_{ij}(U_{ij} = 1|Z_{ij}) = \lambda_{ij}(1 - \exp(-\exp(\alpha_j Z_{ij} + \zeta)) \tag{5}$$

In order to compare the SAT response model with other models, response moderation model was Eq. 6 (RMM, Bolsinova et al., 2017). **Figure 1** showed the relationship between residual time and accuracy of SATM and RMM. In figure A and B, the parameters of the two models were the same. However, there were significant differences between the two models on the asymptotic level of accuracy. The probability of RMM can always close to 1 by the increase in response time. Therefore, it means that response time has a crucial impact on accuracy. Although examinees' ability are extremely low, they can also get a high score in the difficult item by increasing the time. In SATM, the accuracy is affected not only by response time, but also by ability. Even if the time is enough, the accuracy of SATM is also low for low-ability examinee.

$$P(U_{ij} = 1|Z_{ij}) = \frac{\exp(a_j \times \theta_i - b_j + \alpha_j Z_{ij})}{1 + \exp(a_j \times \theta_i - b_j + \alpha_j Z_{ij})} \tag{6}$$

Hierarchical Model Framework

The SAT hierarchical model also consists of two levels. At the first level, SATM and the response time model are two independent models. At the second level, the person parameters and item parameters are assumed to draw from a multivariate normal distribution with mean vector and covariance matrix, respectively (Eq. 7).

$$\mu_I = [\mu_\theta \mu_\tau] \text{ and } \Sigma_P = \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{bmatrix} \tag{7}$$

$$\mu_J = [\mu_b \mu_\beta] \text{ and } \Sigma_I = \begin{bmatrix} \sigma_b^2 & \sigma_{b\beta} \\ \sigma_{b\beta} & \sigma_\tau^2 \end{bmatrix} \tag{8}$$

ESTIMATION AND MODEL SELECTION

Identifying Restrictions

To identify the SAT hierarchical model, the parameters should be fixed to $\mu_\theta = \mu_\tau = 0$ and $\sigma_\theta^2 = \sigma_\tau^2 = 1$ (van der Linden, 2007).

Prior Distributions

The SAT hierarchical model is estimated by a fully Bayesian Markov chain Monte Carlo (MCMC) method. The prior

for the item parameters a_j,$1/\sigma_j$, and α_j all follow the left-truncated normal distribution $N(0, 1)\,I(0,)$. The prior for ζ is follows the standard normal distribution $N(0, 1)$. Moreover, the item parameters b_j and β_j are assumed to follow the normal distribution $N(0.001, 0.001)$. The covariance matrix Σ_I selects an inverse-Wishart distribution InvWishart(R_2, 2), where R_2 is a binary unit matrix. Due to identifying restrictions, the correlation $\rho_{\theta\tau}$ is equal to the covariance $\sigma_{\theta\tau}$, and $\rho_{\theta\tau} \in [-1, 1]$. A doubly truncated normal distribution is selected as the prior distribution of the covariance $\sigma_{\theta\tau} \sim N(0, 1)\,I(-1, 1)$ (Meng et al., 2015).

Model Fit for the Hierarchical Models

On the model selection criteria, the deviance information criterion (DIC, Spiegelhalter et al., 2002) is selected. Based on the posterior distribution of the log-likelihood or the deviance, DIC is calculated from the samples generated by the MCMC simulation. $DIC = \bar{D} + pD$, where pD is the effective number of parameters, $\bar{D}$ is the posterior mean of deviance (i.e., $-2 \times$ Log-likelihood). The smaller the DIC, the better the model is fitted in the empirical data.

SIMULATION STUDY

Design of the Simulation Study

To verify the parameter recovery with the proposed estimation method, a simulation study was carried out based on the test length (m = 30, 60) and the sample size (N = 500, 1000). There were 30 replications for each condition. For different item parameters, they were separately drawn from different distributions: $a_j \sim N(0, 1)I(0)$, $\frac{1}{\sigma_j} \sim N(0, 1)I(0)$, $\alpha_j \sim N(0, 1)I(0)$, $\zeta \sim N(0, 1)$, and $[b_j, \beta_j] \sim MVN\left([0, 3], \begin{bmatrix} 1 & .25 \\ .25 & .25 \end{bmatrix}\right)$. The person parameters θ and τ were sampled from a bivariate normal distribution with $\sigma_{\theta\tau} = 0.5$. The chosen parameters, test length and sample size are the most commonly used settings (Wang et al., 2013; Meng et al., 2015; Bolsinova et al., 2017).

Results of the Simulation Study

The item and person parameters were measured by the Mean squared error (MSE) and average bias (Bias).

$$MSE(\hat{\xi}) = \frac{\sum_{r=1}^{R}\sum_{j=1}^{m}(\hat{\xi} - \xi)^2}{R \times m} \quad (9)$$

$$Bias(\hat{\xi}) = \frac{\sum_{r=1}^{R}\sum_{j=1}^{m}(\hat{\xi} - \xi)}{R \times m} \quad (10)$$

Where $\hat{\xi}$ and ξ are the estimated and true values of model parameters, respectively. R is the number of replications and m is the test length.

The estimated results of the item parameters are displayed in **Table 1**. The MSE for the item parameters decreased when the sample size N increased. For the condition with N = 1000, m = 60, the MSE of b decrease from 0.0592 to 0.034, and the other parameters were less than 0.032. The absolute Bias of the item parameters were close to 0.07. Therefore, the results of item parameters were acceptable for all conditions.

Alternatively, **Table 1** shows the result of the person parameters. The MSE of the speed parameter was below 0.03 within each condition. However, the result of the ability decreased from 0.17 to less than 0.10 with the increase of the test length. On the other hand, the Bias of the person parameters fluctuated around zero. Consequently, the person parameters were likewise acceptable.

EMPIRICAL EXAMPLE

Data and Method

We analyzed data from the Raven's Standard Progressive Matrices (SPM). The SPM includes five sets (A to E) and 12 items in each set. The valid sample size was 320 and the difficulty of the items was disorderly. In the process of responding, examinees could only answer questions in the order of the presented, and were not allowed to be returned. The time limit of this test was 40 min.

Three models were fitted to the empirical data using Gibbs samplers (30000 iterations, 10000 burn-in, 2 chains and 2 thinning). The multivariate potential scale reduction factor (Brooks and Gelman, 1998) was used to monitor the convergence diagnostic and required less than 1.1.

RESULTS

The SPM data was fitted by the traditional hierarchical model (van der Linden, 2007, M0), the RMM hierarchical model (RMHM) and the speed-accuracy tradeoff hierarchical model (SATHM), respectively. According to the DIC, SATHM was the smallest (DIC = 47306.22), RMHM was followed (DIC = 47677.85) and the largest was M0 (DIC = 48069.24). Therefore, it means that considering the effect of response time on accuracy can improve model fit. Furthermore, SATHM fitting is superior to RMHM, so it is necessary to limit the effect of response time on accuracy. The remainder of this section will focus on the results of SATHM.

The results of the hyperparameters and the intercept parameter (ζ) are presented in **Table 2**. With the 95% credible interval for the correlation $\sigma_{\theta\tau}$, speed was negatively correlated with ability. The mean of the intercept parameter (ζ) was 2.6431 and the mean of b was -2.646. Meanwhile, the correlation of item parameters b and β was highly positive.

Finally, the relationship between b and alpha is presented in **Figure 2**. The dotted line of the horizontal axis is the mean of b. From **Figure 2**, when α was less than 0, b was greater than or approaching the mean of b for all items. Therefore, the effect of residual response time is more likely to be negative for medium-difficulty items. The result is slightly different from that of Bolsinova et al. (2017; Figure 3). It may be related to the difficulty of the test, because the test is relatively simple.

TABLE 1 | MSE and Bias for the item parameters.

		N = 500, *m* = 30		*N* = 1000, *m* = 30		*N* = 500, *m* = 60		*N* = 1000, *m* = 60	
Model parameters		**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**
Item parameters	*a*	0.0592	−0.0279	0.0315	−0.0338	0.0460	−0.0334	0.0314	−0.0350
	b	0.0592	0.0423	0.0295	0.0115	0.0564	−0.0003	0.0340	−0.0046
	ζ	0.0032	0.0186	0.0092	0.0611	0.0110	−0.0100	0.0004	0.0047
	α	0.0551	0.0102	0.0663	0.0019	0.0590	0.0085	0.0256	0.0148
	σ	0.0018	−0.0007	0.0008	0.0018	0.0016	−0.0007	0.0009	0
	β	0.0048	−0.0068	0.0019	0.0146	0.0068	−0.0116	0.0025	−0.0087
Person parameters	θ	0.1710	0.0174	0.1770	−0.0019	0.0918	0.0031	0.0948	−0.0031
	τ	0.0298	−0.0002	0.0258	0.0050	0.0151	0.0053	0.0147	0.0142

DISCUSSION

The accuracy of completing a task has always been the main evaluation index in the educational assessment. During a variety of task situations, all the indexes indicating the quality of examinees are extremely important, including the correctness of the result as well as the timeliness of the decision-making process. Moreover, most tests are administered within an allocated time. Due to the time limit, examinees might have different tradeoffs on different items. However, current models cannot effectively analyze the effect of the SAT. In cognitive experiments, SAT model is more superior to describing the dynamic relationship between reaction time and accuracy than other models. Therefore, this paper incorporates the SAT model with the traditional hierarchical model to establish the SATHM. In addition, the parameters of SATHM can be performed well using the MCMC algorithm and the DIC more preferred the SATHM than other models in empirical data.

Some other issues should be further researched. Firstly, the SATHM merely explains the item-specific tradeoff. However, it is simple to extend to the tradeoff of between-person differences with reference to Bolsinova et al. (2016, 2017). Secondly, the lognormal response model was selected to model the response time in SATHM, but it not always satisfies the normality assumption. Therefore, some other models should be investigated, such as Shifted Wald distribution (Anders et al., 2016) and the semi-parameter model (Wang et al., 2013). Finally, Chen et al. (2018) have explored the relationships between response time and accuracy and found that there may be a curvilinear dependency. Accordingly, a curvilinear SATHM can be obtained with some extensions.

TABLE 2 | Posterior means and 95% credible intervals of the hyperparameters and the intercept parameter (ζ) under the SAT hierarchical model.

	Mean	**s.d**	**95% credible interval**
$\sigma_{\theta\tau}$	−0.7361	0.0544	[−0.8239 −0.6109]
σ_b^2	2.2172	0.4404	[1.5157 3.2314]
$\sigma_{b\beta}$	0.8467	0.1818	[0.5612 1.2672]
σ_β^2	0.4849	0.0916	[0.3367 0.6966]
ζ	2.6431	0.1385	[2.3960 2.9316]
μ_b	−2.6460	0.2019	[−3.0521 −2.2519]
μ_β	2.4868	0.0983	[2.2950 2.6783]

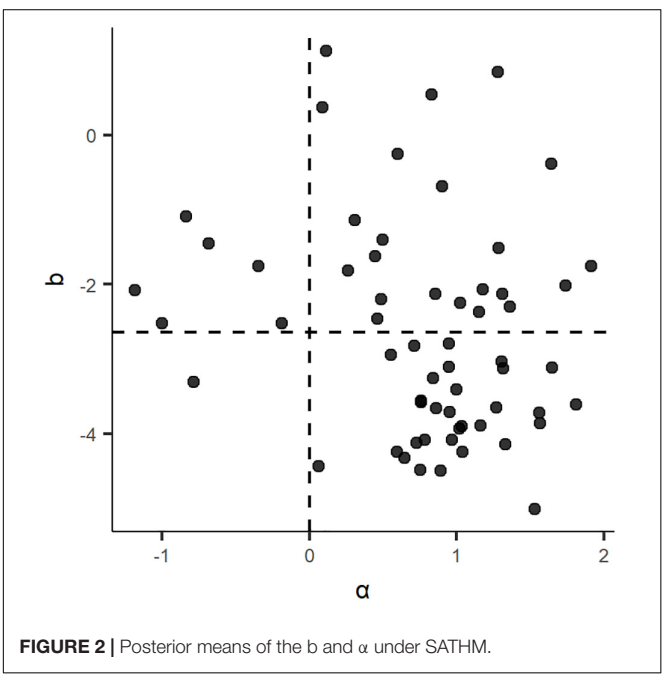


FIGURE 2 | Posterior means of the b and α under SATHM.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

XG: design the study, data analysis, manuscript writing, and revision. ZL: preliminary idea construction, manuscript revision, and proofreading. XY: manuscript revision, and proofreading.

FUNDING

This study was partially supported by the Humanities and Social Sciences Research Foundation of Ministry of Education of China (Grant 17YJC190029), the Social Science Planning Program

of Jiangxi (16JY11), the Humanities and Social Sciences Program (Grant XL1509), the Education Science and Technology Research Program (Grant GJJ160309) of Jiangxi Provincial Department of Education, and the National Natural Science Foundation of China (Grant 31660279).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02910/full#supplementary-material

REFERENCES

Anders, R., Alario, F.-X., and Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychol. Methods* 21, 309–327. doi: 10.1037/met0000066

Bolsinova, M., De Boeck, P., and Tijmstra, J. (2016). Modelling conditional dependence between response time and accuracy. *Psychometrika* 82, 1126–1148. doi: 10.1007/s11336-016-9537-6

Bolsinova, M., Tijmstra, J., and Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *Br. J. Math. Stat. Psychol.* 70, 257–279. doi: 10.1111/bmsp.12076

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Grap. Stat.* 7, 434–455. doi: 10.2307/1390675

Chen, H., De Boeck, P., Gradya, M., Yanga, C.-L., and Waldschmidta, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence* 69, 16–23. doi: 10.1016/j.intell.2018.04.001

Cox, D. R., and Miller, H. D. (1970). *The Theory of Stochastic Processes*. London: Methuen.

Giordano, A. M., McElree, B., and Carrasco, M. (2009). On the automaticity and flexibility of covert attention: a speed-accuracy trade-off analysis. *J. Vis.* 9, 1–10. doi: 10.1167/9.3.30

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front. Neurosci.* 8:150. doi: 10.3389/fnins.2014.00150

McElree, B. (1998). Attended and non–attended states in working memory, Accessing categorized structrures. *J. Mem. Lang.* 38, 225–252. doi: 10.1006/jmla.1997.2545

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *J. Psycholinguist. Res.* 29, 111–123. doi: 10.1023/A:1005184709695

McElree, B., and Carrasco, M. (1999). The temporal dynamics of visual search, evidence for parallel processing in feature and conjunction searches. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 1517–1539. doi: 10.1037/0096-1523.25.6.1517

McElree, B., and Dosher, B. A. (1989). Serial position and set size in short-term memory: time course of recognition. *J. Exp. Psychol. Gene.* 18, 346–373. doi: 10.1037//0096-3445.118.4.346

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91. doi: 10.1016/S0749-596X(02)00515-6

McElree, B., Jia, G. X., and Litvak, A. (2000). The time–course of conceptual processing in three bilingual populations. *J. Mem. Lang.* 42, 229–254. doi: 10.1006/jmla.1999.2677

Meng, X. B., Tao, J., and Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *J. Educ. Measur.* 52, 1–27. doi: 10.1111/jedm.12060

Meng, X. B., Tao, J., and Shi, N.-Z. (2014). An item response model for Likert-type data that incorporates response time in personality measurements. *J. Stat. Comp. Simu.* 84, 1–21. doi: 10.1080/00949655.2012.692368

Rae, B., Heathcote, A., and Donkin, C. (2014). The hare and the tortoise: emphasizing speed can change the evidence used to make decisions. *J. Exp. Psychol.* 40, 1226–1243. doi: 10.1037/a0036801

Ranger, J., and Ortner, T. (2012). The case of dependency of responses and response times: a modeling approach based on standard latent trait models. *Test Assess. Model.* 54, 128–148.

Ratcliff, R., and Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychol. Sci.* 9, 347–356. doi: 10.1111/1467-9280.00067

Ratcliff, R., Thapar, A., and McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychol. Aging* 16, 323–341. doi: 10.1037/0882-7974.16.2.323

Ratcliff, R., Thapar, A., and McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Percept. Psychophys.* 65, 523–535. doi: 10.3758/BF03194580

Reed, A. V. (1973). Speed–accuracy trade-off in recognition memory. *Science* 181, 574–576. doi: 10.1126/science.181.4099.574

Reed, A. V. (1976). List length and the time course of recognition in immediate memory. *Mem. Cogn.* 4, 16–30. doi: 10.3758/BF03213250

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–616. doi: 10.1111/1467-9868.00353

Starns, J. J., Ratcliff, R., and McKoon, G. (2012). Evaluating the unequal-variability and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cogn. Psychol.* 64, 1–34. doi: 10.1016/j.cogpsych.2011.10.002

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

Wang, C., Chang, H. H., and Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

Wang, S.-Y., Zhang, S.-S., Douglas, J., and Culpepper, S. (2018). Using response times to assess learning progress: a joint model for responses and response times. *Meas. Interdiscip. Res. Perspect.* 16, 45–58. doi: 10.1080/15366367.2018.1435105

Wickelgren, W. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychol.* 41, 67–85. doi: 10.1016/0001-6918(77)90012-9

Zhan, P., Jiao, H., and Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *Br. J. Math. Stat. Psychol.* 71, 262–286. doi: 10.1111/bmsp.12114

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Guo, Luo and Yu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

METHODS
published: 05 August 2019
doi: 10.3389/fpsyg.2019.01675

Bayesian Covariance Structure Modeling of Responses and Process Data

Konrad Klotzke and Jean-Paul Fox*

Faculty of BMS, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, Enschede, Netherlands

A novel Bayesian modeling framework for response accuracy (RA), response times (RTs) and other process data is proposed. In a Bayesian covariance structure modeling approach, nested and crossed dependences within test-taker data (e.g., within a testlet, between RAs and RTs for an item) are explicitly modeled. The local dependences are modeled directly through covariance parameters in an additive covariance matrix. The inclusion of random effects (on person or group level) is not necessary, which allows constructing parsimonious models for responses and multiple types of process data. Bayesian Covariance Structure Models (BCSMs) are presented for various well-known dependence structures. Through truncated shifted inverse-gamma priors, closed-form expressions for the conditional posteriors of the covariance parameters are derived. The priors avoid boundary effects at zero, and ensure the positive definiteness of the additive covariance structure at any layer. Dependences of categorical outcome data are modeled through latent continuous variables. In a simulation study, a BCSM for RAs and RTs is compared to van der Linden's hierarchical model (LHM; van der Linden, 2007). Under the BCSM, the dependence structure is extended to allow variations in test-takers' working speed and ability and is estimated with a satisfying performance. Under the LHM, the assumption of local independence is violated, which results in a biased estimate of the variance of the ability distribution. Moreover, the BCSM provides insight in changes in the speed-accuracy trade-off. With an empirical example, the flexibility and relevance of the BCSM for complex dependence structures in a real-world setting are discussed.

Keywords: process data, educational measurement, Bayesian modeling, covariance structure, marginal modeling, cross-classification, response times, latent variable modeling

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Paul De Boeck,
The Ohio State University, United States
Maria Bolsinova,
University of Amsterdam, Netherlands

****Correspondence:***
Konrad Klotzke
k.klotzke@utwente.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *15 September 2018*
Accepted: *02 July 2019*
Published: *05 August 2019*

Citation:
Klotzke K and Fox J-P (2019) Bayesian Covariance Structure Modeling of Responses and Process Data. Front. Psychol. 10:1675. doi: 10.3389/fpsyg.2019.01675

1. INTRODUCTION

Computer-based assessments (CBAs) provide the opportunity to gather responses times (RTs) and other process data in addition to the test-takers' responses. Empirical research has shown that in combination with response patterns, RTs can lend valuable insight into interesting test-taker, item and test characteristics, such as pre-knowledge of items, motivation, time-pressure or differential speededness (Bridgeman and Cline, 2004; Wise and Kong, 2005; Meijer and Sotaridona, 2006; van der Linden et al., 2007; van der Linden and Guo, 2008; Marianti et al., 2014; Qian et al., 2016). New types of process data have been explored lately that carry the potential to lend additional insight into (latent) response processes and to improve inferences about constructs of interest (e.g., Azevedo, 2015; He et al., 2016; Goldhammer and Zehner, 2017; Maddox, 2017). To make valid inferences

from process data, innovative joint models are needed that are capable of utilizing test-taker data beyond RAs and RTs, while accounting for complex relationships in multiple data types.

An important concept is the speed-accuracy trade-off, which states that, on average, a test-taker's ability suffers from an increased working speed (van der Linden, 2009). Test scores depend on test-takers' speed during the test and ignoring this within-subject relationship threatens the validity of inferences about their ability level. In experimental cognitive psychology, the speed-accuracy trade-off can be modeled for individual persons as the relationship between the proportion of correct tasks and the average time spent on the tasks (Luce, 1986). In educational measurement, learning effects can be expected when presenting the same item to a test-taker multiple times (Butler, 2010). Hence, in practical applications, often only a single measurement of RA and RT is obtained for each combination of test-taker and item. Therefore, it is common to assume a certain homogeneity in the speed-accuracy trade-off within a group of test-takers and how they are affected by the condition of interest (Thissen, 1983; Klein Entink et al., 2008; Glas and van der Linden, 2010; Ranger and Kuhn, 2013; Goldhammer and Kroehne, 2014; Goldhammer et al., 2014; Loeys et al., 2014; Molenaar et al., 2015; van der Linden and Fox, 2016). Alternatively, in certain experimental settings, the researcher can control the test-takers' working speed (by imposing time limits) and thereby exclude the person-level working speed variable from the regression equation (Goldhammer and Kroehne, 2014).

More general and flexible approaches to model and test the within-subject dependence structure have been achieved through the generalized linear mixed model (GLMM) (McCulloch, 2003) and mixture models. The within-subject mixture models allow subject-specific changes in the speed-accuracy trade-off across different states. However, in practice the number of states is very limited (Wang and Xu, 2015; Molenaar et al., 2016) to obtain identifiable and stable estimation results. In GLMMs, the measurement model for the RAs or the RTs is extended by including either the person level variable (ability or working speed) or the dependent variable of the respective other measurement model as a covariate in the regression equation. Item-specific person-level and person-specific item-level variables allow the speed-accuracy trade-off to vary between items and allow item parameters to vary across persons, respectively (e.g., Goldhammer et al., 2014, 2015). Furthermore, a non-linear relation between RAs and RTs can be specified (e.g., Molenaar et al., 2015; Bolsinova and Molenaar, 2018).

However, the complexity of a GLMM is drastically increased when including other process data and extending the GLMMs with additional person-level variables. It is therefore questionable whether the GLMM approach can manage the challenges of utilizing new types of process data in complex CBAs. Currently, GLMMs are limited in the amount of process data information that can be utilized to make inferences due to restrictions on the model complexity and the sample size. Furthermore, GLMMs are also limited in how the information is utilized. For instance, correlations between RAs and different types of process data may vary depending on item characteristics or test design. In that case, interaction effects are needed to model item and/or testlet-specific dependences, but this will significantly increase the complexity of the GLMM. To prevent over-parameterization and weak numerical stability, techniques such as principle component analysis, latent class analysis, or various model selection algorithms (e.g., backward elimination, forward selection or all subsets regression) (Thomas, 2002; Efron et al., 2004; Wetzel et al., 2015) have been proposed to reduce the number of covariates in the regression equation. However, this complicates a straightforward modeling approach and can lead to arbitrary assumptions and *ad hoc* decisions. It is well-known that ignoring correlations in test-taker data may cause violations of local independence assumptions and can result in biased inferences about parameters, the reliability of the test, and hinder test equating (e.g., Yen, 1984; Ackerman, 1987; Chen and Thissen, 1997; Bradlow et al., 1999; Baker and Kim, 2004; Jiao et al., 2005, 2012; Wang and Wilson, 2005; Wainer et al., 2007). Therefore, when including new types of process data, care must be taken in modeling the dependence structure to avoid making biased inferences.

The proposed Bayesian Covariance Structure Model (BCSM) can handle different types of nested and cross-classified dependence structures for multiple types of test-taker data. The BCSM extends the marginal model for hierarchically structured item RT data of Klotzke and Fox (2018). In the model of Klotzke and Fox (2018), dependences that follow from nested classifications (e.g., item clusters in a testlet design) are directly modeled as covariances without including random effects. The methodology is extended to classifications across multiple data types. Thus, in addition to modeling nested classifications (within a data type), relationships in data across different types (e.g., RTs and dichotomous responses) are modeled through cross-classifications in the dependence structure. In the same manner as the nested classifications, crossed classifications are modeled explicitly as covariance parameters. Without the inclusion of random effects, the parsimony of the BCSM is preserved, where dependences between each cluster of observations can be modeled with a single covariance parameter. The BCSM assumes a multivariate normal distribution for the data, either directly or through a threshold specification (i.e., for categorical or count data), and allows distinct modeling of the mean and covariance structure. The BCSM parameters can be estimated with an efficient Gibbs-sampling algorithm, even for a reasonably small sample size. Modeling local dependences via covariance parameters instead of modeling dependences through random effects (i.e., the random effect variance defines the covariance between clustered observations) has two advantages: first, covariances can be negative or positive, which allows more flexibility in specifying complex dependence structures than random effect variances. The latter can only model positive dependences. Second, tests for local independence under the BCSM framework do not require testing at the boundary of the parameter space (i.e., the null hypothesis states that the covariance parameter is equal to zero). This stands in contrast to a random effect variance, which is *a-priori* restricted to be positive. In the BCSM, this means that the prior distributions for the covariance parameters are less informative, i.e., they don't assume beforehand that the covariance parameters are

greater than zero. Therefore, more objective inferences about the dependence structure can be made. Finally, contrary to common marginal modeling approaches such as generalized estimating equations (GEE) (Liang and Zeger, 1986; Diggle et al., 2013), the dependence structure is fully modeled in an additive covariance structure. This allows testing for interaction effects (e.g., local dependence within testlets) (Lee and Neider, 2004), and to estimate random person/group effects *post-hoc* from the residuals of the model. The latter is of utility if the random effects structure cannot be estimated in the traditional way (fitting a random effects model) due to for instance sample size limitations. For example, test-taker ability estimates can be obtained under a complex within-subject dependence structure, while accounting for various types of process data information.

The paper is organized as follows: first, the BCSM is introduced. Next, an additive covariance structure is defined that can be utilized to explicitly model dependences in data from different types (RAs, RTs, and other process data). Five well-known dependence structures are presented under the BCSM. An approach to model the interdependence of categorical data through truncated conditional univariate normal distributions of latent variables is specified. Closed-form expressions for the conditional posterior distributions of the covariance parameters are derived through truncated shifted inverse-gamma priors, where the truncation point ensures the positive definiteness of the additive covariance matrix. Samples from the full joint posterior are obtained using a Gibbs sampler. In a simulation study, a BCSM for RAs and RTs is compared to the hierarchical model of van der Linden (2007) (LHM) given a situation in which the test-taker's working speed and ability are allowed to vary over the course of the test, thus violating the assumption of local independence in models that presume a fixed working speed and ability. In an empirical example, data from the Programme for the International Assessment of Adult Competencies (PIAAC) study (OECD, 2013) is analyzed with a BCSM for RAs and two types of process data. Finally, the results, limitations and future prospects of the proposed framework for educational measurement applications are discussed.

2. MODELING COVARIANCE STRUCTURES

Data can be dependent on different levels. For example, in a testlet structure data of items within a testlet may be correlated stronger than data across items. Hence, data may be locally dependent on a testlet level. Furthermore, more than one data point may be available for an item (e.g., dichotomous responses, RTs and additional process data). The relationship between two data points (of the same test-taker) for an item can be positive or negative. The corresponding local dependence can either increase (when positive) or decrease (when negative) the total correlation of data for an item. This multilevel dependence structure is specified through a cross-classification matrix. In BCSM, an additive covariance structure forms the link between the covariance parameters and the cross-classification matrix.

Observations within a group can be more alike than observations across groups. In the BCSM framework, this local dependence is modeled in an additive covariance structure. Each source of local dependence, i.e., the effect of each grouping on the association of components, is represented by a covariance parameter and a layer in the additive structure. Group membership is specified by a $N_t \times N_c$ classification matrix $\boldsymbol{u}$, where N_t is the number of layers in the additive covariance structure and N_c is the number of components. Each row in $\boldsymbol{u}$ thus corresponds to a layer in the covariance structure and the columns define the local dependence of components in that layer. Components that are grouped together within a layer are marked by a 1, ungrouped components are marked by a 0.

The components are assumed to be multivariate normally distributed with a N_c-dimensional mean vector $\boldsymbol{\mu}$ and a $N_c \times N_c$-dimensional covariance matrix $\boldsymbol{\Sigma}$. The inclusion of person level random effects is not necessary as the covariance structure implied by the usual person level variables (such as ability and working speed) is directly modeled. As a consequence, the mean structure consists of intercepts on the group and item level. Intercepts on the group level are for example the average working speed or ability in a group of test-takers. Intercepts on the item level are commonly denoted as item time intensity and item difficulty parameters. Furthermore, test-taker background variables can be included as covariates.

The covariance matrix consists of a base layer $\boldsymbol{\Sigma}_0$ and N_t additive layers. In the base layer, the measurement error variance is modeled, whereby $\boldsymbol{\Sigma}_0 = diag\left(\sigma_1^2, \ldots, \sigma_{N_c}^2\right)$. Each additive layer t is constructed out of a covariance parameter θ_t and the t-th row of the classification vector, i.e., $\boldsymbol{u}_t$:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \sum_{t=1}^{N_t} \theta_t \boldsymbol{u}_t \boldsymbol{u}_t^T. \quad (1)$$

On a mathematical level, no qualitative difference is made between the covariance parameters $\boldsymbol{\theta} = \{\delta, \tau, \omega, \phi, \boldsymbol{\nu}, \boldsymbol{\Delta}\}$. However, for the sake of clarity, in this text δ, τ, and ω represent the covariance between, respectively, the RTs, the RAs and additional process data of a test-taker. The local dependence that follows from grouping observations from different data types on a person level is represented by the covariance parameter ϕ. The vector $\boldsymbol{\nu}$ contains the cross-covariances between components of different types (e.g., RAs and RTs). Furthermore, $\boldsymbol{\Delta}$ are the covariances that follow from blocks within components of one type (e.g., testlets within RT data).

Five examples of models for responses and process data are described that can be constructed within the BCSM framework. Each example is illustrated for a test size of $p = 6$ items. The utilized data types are RTs ($\boldsymbol{RT}$), RAs that are manifested as discrete variables through a threshold specification ($\boldsymbol{RA}$) and additional process data ($\boldsymbol{W}$). The observed categorical data are denoted as $\boldsymbol{D}$. Finally, the scalability of BCSMs given a growing number of items and extensions to the classification structure is discussed.

2.1. The BCSM for Speed and Ability

The BCSM for speed and ability follows the classification as implied by the LHM with binary factor loadings. In this model,

a test-taker's RTs are grouped by the latent factor working speed, and the RAs are grouped by the latent factor ability. Furthermore, observations are grouped across the two data types on a person level, which represents a correlation between a test-taker's ability and working speed. **Table 1** shows the classification matrix and covariance parameters of the BCSM for speed and ability.

2.2. Variable Speed-Accuracy Trade-Off

For the variable speed-accuracy trade-off model, the BCSM for speed and ability is extended with an item-specific cross-covariance between a test-taker's RTs and RAs. This allows to investigate how the speed-accuracy trade-off within a group of test-takers varies between items. Thereby, a certain homogeneity in the relevant response processes is assumed, which leads to test-takers within a group sharing a common speed-accuracy trade-off. The classification diagram for the variable speed-accuracy trade-off model is shown in **Figure 1**. **Table 2** extends **Table 1** with the additional classification rules and covariance parameters implied by a variable speed-accuracy trade-off.

2.3. Blocked Structures of Cross-Covariances

Just as the variable speed-accuracy trade-off model, the blocked structures of cross-covariances model extends the BCSM for speed and ability with a varying cross-covariance between a test-taker's RTs and RAs. However, the cross-covariance is defined to change per blocks of (here: two) items. A possible application for this model is test-taking under varying time-pressure conditions. In such a scenario, it is reasonable to assume local dependence for components (i.e., RTs and RAs) within a block of items that belong to the same time-pressure condition. In the variable speed-accuracy trade-off model on the other hand, the local dependence is defined per individual item. **Table 3** extends **Table 1** with the additional classification rules and covariance parameters of the blocked structures of cross-covariances model.

2.4. Differential Blocked Structures of Cross-Covariances Across Factors

The within-subject dependence structure can also be specified for components within a single data type. In the differential blocked structures of cross-covariances across factors model, the variable speed-accuracy trade-off model is extended with a separate testlet structure for each the RTs and the RAs. The testlet structures are defined independently of each other. **Table 4** extends **Tables 1**, **2**

TABLE 1 | The additive covariance structure of the BCSM for speed and ability is implied by the random effects structure of the LHM with binary factor loadings.

	Classification matrix u											
Covariance	**Response times**						**Response accuracies**					
δ	1	1	1	1	1	1	0	0	0	0	0	0
τ	0	0	0	0	0	0	1	1	1	1	1	1
ϕ	1	1	1	1	1	1	1	1	1	1	1	1

TABLE 2 | The additive covariance structure of the variable speed-accuracy trade-off model is an extension of the BCSM for speed and ability with item-specific cross-covariances between RTs and RAs.

	Classification matrix u											
Covariance	**Response times**						**Response accuracies**					
ν_1	1	0	0	0	0	0	1	0	0	0	0	0
ν_2	0	1	0	0	0	0	0	1	0	0	0	0
ν_3	0	0	1	0	0	0	0	0	1	0	0	0
ν_4	0	0	0	1	0	0	0	0	0	1	0	0
ν_5	0	0	0	0	1	0	0	0	0	0	1	0
ν_6	0	0	0	0	0	1	0	0	0	0	0	1

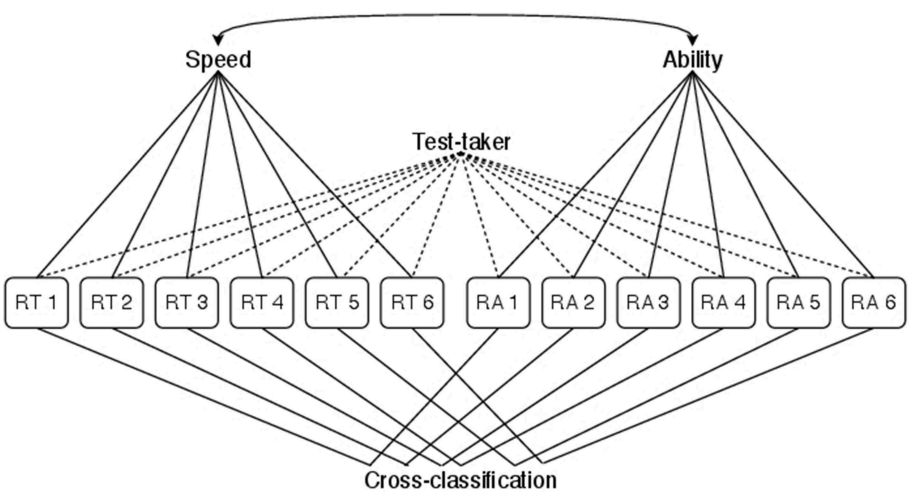


FIGURE 1 | Classification diagram for the variable speed-accuracy trade-off model. The classification implied by the LHM is extended by grouping components item-wise. This allows the group level speed-accuracy trade-off to vary between items.

TABLE 3 | The additive covariance structure of the blocked structures of cross-covariances model is an extension of the BCSM for speed and ability with block-wise cross-covariances between RTs and RAs.

Covariance	Classification matrix u											
	Response times						Response accuracies					
ν_1	1	1	0	0	0	0	1	1	0	0	0	0
ν_2	0	0	1	1	0	0	0	0	1	1	0	0
ν_3	0	0	0	0	1	1	0	0	0	0	1	1

TABLE 4 | The additive covariance structure of the differential blocked structures of cross-covariances across factors model is an extension of the variable speed-accuracy trade-off model with independent testlet structures for separate data types.

Covariance	Classification matrix u											
	Response times						Response accuracies					
Δ_1	1	1	0	0	0	0	0	0	0	0	0	0
Δ_2	0	0	1	1	0	0	0	0	0	0	0	0
Δ_3	0	0	0	0	1	1	0	0	0	0	0	0
Δ_4	0	0	0	0	0	0	1	1	1	0	0	0
Δ_5	0	0	0	0	0	0	0	0	0	1	1	1

with the additional classification rules and covariance parameters of the differential blocked structures of cross-covariances across factors model.

2.5. More Than Two Data Types

A BCSM is not limited to RTs and responses. Additional process data can carry information relevant to the research. In this example, additional process data is available for each combination of test-taker and item. Therefore, $p = 6$ components are added to the model. In the illustrated model, an item-specific cross-covariance between components of all types is assumed. That means for example that RTs and RAs to an item may correlate in a different way than RAs and process data, to the same item. Furthermore, ϕ_1, ϕ_2, and ϕ_3 represent the 3-by-3 covariance of the three latent factors (e.g., ability, working speed, and speed first action) that are related to the three types of data. **Table 5** shows the classification matrix and covariance parameters of the more than two data types model.

2.6. Model Scalability

The models constructed in the BCSM framework are scalable with respect to the length of the test, the number of data types and the specified dependence structure. The number of columns of $\boldsymbol{u}$ corresponds to the number of data components (N_c). If a single observation is available for each combination of test-taker, item and data type, the number of data components is the product of the number of items (p) and the number of data types (N_d), i.e., $N_c = p * N_d$. Consequently, extending the test length with one item increases the number of columns of $\boldsymbol{u}$ by N_d. Similarly, introducing an additional data type increases the number of columns by p.

TABLE 5 | The additive covariance structure for a BCSM that incorporates additional process data, next to RTs and RAs.

Covariance	Classification matrix u																	
	Response times						Response accuracies						Process data					
δ	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
τ	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0
ω	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
ϕ_1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
ϕ_2	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1
ϕ_3	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
ν_1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ν_2	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ν_3	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ν_4	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ν_5	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
ν_6	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
ν_7	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ν_8	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ν_9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ν_{10}	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ν_{11}	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
ν_{12}	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
ν_{13}	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
ν_{14}	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
ν_{15}	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
ν_{16}	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
ν_{17}	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
ν_{18}	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

The number, if any, of additional rows of $\boldsymbol{u}$ depends on the specified classification structure. For example, under the structure specified in **Table 1**, a change in the number of data components does not affect the number of rows of $\boldsymbol{u}$. Instead, the existing groupings are extended to include the new data components.

In other situations, the number of groupings depends on the number of data components. For example, given the item-specific cross-classifications as defined in **Table 2**, each additional item leads to one additional classification rule (the RA and RT of a test-taker to one item are grouped together) and therefore inserts one row into $\boldsymbol{u}$. Thus, if the variable speed-accuracy trade-off joint-model is applied to a test with $p_2 = 100$ instead of $p_1 = 10$ items, the number of columns increases by $(p_2 - p_1) * N_d = (100 - 10) * 2 = 180$ and the number of rows increases by $p_2 - p_1 = 90$.

3. CATEGORICAL OUTCOME DATA

When recording the test-takers' responses during a test, discrete realizations of latent response variables are observed. The multivariate normally distributed RA data (latent responses) are linked through a threshold specification to their discrete

realizations. However, truncating a multivariate normal distribution in high dimensions is non-trivial (Botev, 2017) and simply truncating independently for each dimension does not lead to the intended multivariate joint-distribution (Horrace, 2005).

The proposed solution is to derive the univariate normal distribution of each latent response component, conditional on all other components. The univariate normal distribution is derived by partitioning the additive covariance structure $\boldsymbol{\Sigma}$, as defined in Equation 1, into four parts. The upper left part, $\boldsymbol{B}_{11}$, gives the variance of the k-th component and the diagonal parts, $\boldsymbol{B}_{12}$ and $\boldsymbol{B}_{21}$, contain the covariance of the k-th component with the remaining components. Finally, $\boldsymbol{B}_{22}$ describes the covariance structure of all components but the k-th:

$$\boldsymbol{\Sigma} = \begin{matrix} \\ Y_{ik} \\ \tilde{\boldsymbol{Y}}_i \end{matrix} \overset{\begin{matrix} Y_{ik} & \tilde{\boldsymbol{Y}}_i \end{matrix}}{\begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}}, \tag{2}$$

where $\boldsymbol{Y}$ is a $N \times N_c$-dimensional matrix, containing data from all N_c components and N test-takers. A tilde, i.e., a $\sim$, above a vector or matrix indicates that the k-th component is excluded from the data structure. Based on the partitioned covariance matrix, the means and variance of the conditionally univariate normal distribution of the k-th component are derived for each test-taker:

$$\boldsymbol{\mu}_{Y_k|\tilde{Y}} = \boldsymbol{\mu}_{Y_k} + \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}(\tilde{\boldsymbol{Y}} - \boldsymbol{\mu}_{\tilde{Y}}), \tag{3}$$

$$\sigma^2_{Y_k|\tilde{Y}} = \boldsymbol{B}_{11} - \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21}. \tag{4}$$

A closed-form expression for $\boldsymbol{B}_{22}^{-1}$ is derived through the Sherman-Morrison formula (e.g., Lange, 2010, p. 261):

$$\boldsymbol{A}_{t+1}^{-1} = (\boldsymbol{A}_t + \lambda\boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{A}_t^{-1} - \frac{\boldsymbol{A}_t^{-1}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{A}_t^{-1}}{1/\lambda + \boldsymbol{v}^T\boldsymbol{A}_t^{-1}\boldsymbol{v}}, \tag{5}$$

where $\boldsymbol{A}_t^{-1} = \tilde{\boldsymbol{\Sigma}}_t^{-1}$ is the inverse of the additive covariance structure for all but the k-th component at the t-th layer, $\lambda = \theta_{t+1}$ is the covariance parameter for the added layer and $\boldsymbol{v} = \tilde{\boldsymbol{u}}_{t+1}$ contains the classification structure for the new layer. Given that the inverse of $\boldsymbol{A}_0^{-1} = \tilde{\boldsymbol{\Sigma}}_0^{-1}$, i.e., the inverse of the diagonal matrix consisting of the measurement error variance parameters for all but the k-th component, is known, the inverse for any additional layer can be derived recursively.

4. BAYESIAN INFERENCE

In line with the approach suggested by Fox et al. (2017) and Klotzke and Fox (2018), closed-form expressions for the conditional posterior distributions of the variance and covariance parameters are derived through truncated shifted inverse-gamma priors. For each of the N_t layers of the additive covariance matrix, a truncation point tr_t is derived by applying the Sherman-Morrison formula (Lange, 2010, p. 260–261). Enforcing the truncation through the indicator function $\mathbb{1}_{tr}$ ensures that the covariance matrix is positive definite at any layer t. This leads to a lower bound for each covariance parameter ($\theta_t > tr_t$) conditional on the classification structure and the inverse of the covariance matrix at the underlying layer ($t-1$):

$$tr_t = -1/\boldsymbol{u}_t^T\boldsymbol{\Sigma}_{t-1}^{-1}\boldsymbol{u}_t. \tag{6}$$

For the measurement error variance parameters (the diagonal terms of $\boldsymbol{\Sigma}_0$) a truncation sets the probability of negative values a-priori to zero.

The reasoning behind the shift parameters is based upon two premises: (1) a draw of θ_t is obtained through sampling $\theta_t + \psi_t$ and subtracting the shift parameter ψ_t iteratively within the Markov chain Monte Carlo (MCMC) (Gilks et al., 1995) algorithm, (2) the probability distribution of $\theta_t + \psi_t$ must incorporate all information that is available in the data about θ_t. It is shown in Equations (7) and (8) that the probability distribution of the person level means across that are grouped together in $\boldsymbol{u}_t$ contains all available information about the covariance parameter θ_t. Note that the person level means are constructed as the mean of (correlated) random normal variables and are therefore univariate normally distributed.

Conditional on the classification structure and the additive covariance matrix at its highest layer ($\boldsymbol{\Sigma}_{N_t}$), the variance of the person level means is derived through the property that the variance of the sum of correlated random variables is the sum of their covariances:

$$\begin{aligned} Var\left(\bar{Y}_{i(k\in\boldsymbol{u}_t)}|\boldsymbol{\Sigma}_{N_t},\boldsymbol{u}\right) &= Var\left(\sum_{k\in\boldsymbol{u}_t} Y_{ik}/\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)\right) \\ &= \left[\left(\sum_{k=1}^{N_c}\sigma_k^2 u_{tk} + \theta_t\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)^2 \right.\right. \\ &\quad \left.\left. + \sum_{j\neq t}\theta_j\left(\mathbf{1}_{N_c}^T(\boldsymbol{u}_j\odot\boldsymbol{u}_t)\right)^2\right)/\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)^2\right] \\ &= \theta_t + \left[\left(\sum_{k=1}^{N_c}\sigma_k^2 u_{tk} + \sum_{j\neq t}\theta_j\left(\mathbf{1}_{N_c}^T(\boldsymbol{u}_j\odot\boldsymbol{u}_t)\right)^2\right)/\left(\mathbf{1}_{N_c}^T\boldsymbol{u}_t\right)^2\right] \\ &= \theta_t + \psi_t, \end{aligned} \tag{7}$$

where $\odot$ denotes the Hadamard product and $\mathbf{1}_{N_c}$ is a N_c-dimensional vector of ones. A sufficient statistic for $Var\left(\bar{Y}_{i(k\in\boldsymbol{u}_t)}|\boldsymbol{\Sigma}_{N_t},\boldsymbol{u}\right) = \theta_t + \psi_t$ is therefore the sum of squares of the deviations of the conditional person level means from the conditional grand mean,

$$SSB_t = \sum_{i=1}^{N}\left(\bar{Y}_{i(k\in\boldsymbol{u}_t)} - \bar{Y}_{.(k\in\boldsymbol{u}_t)}\right)^2. \tag{8}$$

Similarly, the within-component sum of squares is a sufficient statistic for $Var(Y_{ik}) = \sigma_k^2 + \sum_{t=1}^{N_t}\theta_t u_{tk}$, namely

$$SSW_k = \sum_{i=1}^{N}\left(Y_{ik} - \bar{Y}_{.k}\right)^2. \tag{9}$$

From Equations (8) and (9) follow $N_t + N_c$ sufficient statistics for the N_t covariance and N_c variance parameters, out of which the additive covariance structure, as specified in Equation (1), is composed. The model is therefore identified under the condition that the rows of the classification matrix $\boldsymbol{u}$ are mutually distinct.

The truncated shifted inverse-gamma prior extends the default inverse-gamma prior for variance components with a shift and a truncation parameter; the former allowing a covariance parameter to take on negative values, the latter ensuring the positive definiteness of the additive covariance matrix at any layer:

$$IG(x, \alpha_0, \beta_0, \psi_t, tr_t) = \left[\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(x+\psi_t)^{-\alpha_0-1} exp\left(-\frac{\beta_0}{x+\psi_t}\right)\right] \cdot \mathbb{1}_{tr}(x > tr_t), \tag{10}$$

where the truncation point (tr_t) and shift parameter (ψ_t) are computed according to Equations (6) and (7).

Note that conjugacy between the extended inverse-gamma prior and the likelihood function of a normal distribution is preserved, thus leading to truncated shifted inverse-gamma posteriors for the covariance and measurement error variance parameters:

$$\theta_t \sim IG(x, \alpha_0 + N/2, \beta_0 + SSB_t/2, \psi_t, tr_t), \tag{11}$$

$$\sigma_k^2 \sim IG(x, \alpha_0 + N/2, \beta_0 + SSW_k/2, \sum_{t=1}^{N_t} \theta_t u_{tk}, 0). \tag{12}$$

The a-priori restriction of $\sigma_k^2 > 0$ is thus enforced by fixing the truncation point for the measurement error variance parameters to zero.

See Appendix A in **Supplementary Material** for an outline of the MCMC algorithm and the corresponding sampling steps.

5. SIMULATION STUDY

In a simulation study, the within-subject dependence structure under a model for RTs and dichotomous responses is estimated. A comparison is made between a BCSM and the LHM. In the BCSM framework, the dependence structure is directly modeled in an additive covariance matrix. In the LHM framework, the dependence structure is implied by the random effect structure and in particular the random effect variances. Therefore, the focus of this simulation study is the precision and bias of the (co)variance parameter estimates.

In the simulated experiment, across two conditions, $N = 200$ and $N = 1,000$ randomly selected persons are taking a test that consists of $p = 12$ items. Furthermore, the time-pressure on the test-takers systematically changes after every two items. This is assumed to affect the response processes within the group of test-takers over the course of the test. For example, under a perceived high time-pressure, guessing may become more likely. The change in response processes is reflected by the within-subject dependence structure, i.e., the speed-accuracy trade-off may vary between blocks of two items and is common across test-takers.

The length of the test is fixed across the 100 replications for both conditions of the simulation. Within each condition, all test-takers are part of the same group. Within each replication, test-taker data are generated and the BCSM as well as the LHM are fitted with 5000 MCMC iterations and a burn-in phase of 10%. The LHM is fitted using the R-package LNIRT (Fox et al., 2018).

5.1. LHM for Fixed Speed and Ability

On the first level of the hierarchical framework, separate measurement models for the RTs and RAs are specified. The item discrimination parameters are fixed to 1, which gives the following first level models for the RTs (RT) and RAs (RA) of test-taker i and item k:

$$RT_{ik} = \beta_k - \zeta_i + e_{RT_{ik}}, \tag{13}$$

$$RA_{ik} = \theta_i - b_k + e_{RA_{ik}}, \tag{14}$$

where $\zeta_i \sim \mathcal{N}(\mu_\zeta, \delta)$ and $\theta_i \sim \mathcal{N}(\mu_\theta, \tau)$ are random variables on a person level, representing the variation in working speed and ability between test-takers. The time intensity and item difficulty parameters β_k and b_k are item level intercepts and are not given further attention in this simulation study. Finally, $e_{RT_{ik}} \sim \mathcal{N}(0, \sigma_k^2)$ and $e_{RA_{ik}} \sim \mathcal{N}(0, 1)$ are the measurement errors. On the second level, a model for the joint-distribution of the person parameters (working speed and ability) is defined:

$$\boldsymbol{\Sigma}_p = \begin{pmatrix} \delta + \phi & \phi \\ \phi & \tau + \phi \end{pmatrix}. \tag{15}$$

Note that the LHM assumes a constant working speed and ability across the test for a test-taker. From this follows a test-wide cross-covariance between a test-taker's RTs and RAs ϕ.

5.2. BCSM for Variable Speed and Ability

In the BCSM, the within-subject dependence structure is modeled directly in an additive covariance structure with 9 layers. The covariance structure is defined in Equation (1), where $\boldsymbol{\theta} = \{\delta, \tau, \phi, \nu_1, \ldots, \nu_6\}$ are the (cross-)covariance parameters and the classification matrix is specified in **Table 6**. A truncated shifted inverse-gamma prior with $shape = 10^{-3}$ and $scale = 10^3$ is defined for the variance and covariance parameters.

5.3. Data Generation

Data are generated under a generalization of the models specified in Equations (13)–(15) that allows the test-takers' working speed and ability to vary over the course of the test:

$$RT_{ik} = \beta_k - \zeta_{it(k)} + e_{RT_{ik}}, \tag{16}$$

$$RA_{ik} = \theta_{it(k)} - b_k + e_{RA_{ik}}. \tag{17}$$

$$\boldsymbol{\Sigma}_{pk} = \begin{pmatrix} \delta + \phi + \nu_{t(k)} & \phi + \nu_{t(k)} \\ \phi + \nu_{t(k)} & \tau + \phi + \nu_{t(k)} \end{pmatrix}, \tag{18}$$

where $t(k)$ denotes item k in classification group t. The population values of the (co)variance parameters are $\delta = 0.5$, $\tau = 0.5$, $\phi = 0.5$ and $\boldsymbol{\nu} = \{0, -0.05, -0.1, 0.4, 0.2, 0.3\}$. The item level intercepts ($\boldsymbol{\beta}$ and $\boldsymbol{b}$) are set to zero. Finally, the population values of the measurement error variances are generated from a uniform distribution with lower bound 0.5 and upper bound 1.5.

TABLE 6 | The additive covariance structure of the BCSM allows a varying speed-accuracy trade-off between blocks of two items.

Covariance	Classification matrix u																							
	Response times												Response accuracies											
δ	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
τ	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
ϕ	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ν_1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
ν_2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
ν_3	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ν_4	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
ν_5	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0
ν_6	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	1

TABLE 7 | Means and standard deviations of posterior mean estimates across 100 simulated replications of data for 200 and 1,000 test-takers and 12 items.

	Mean (SD) of posterior mean estimates			
	N = 200		N = 1,000	
Cov	BCSM	LHM	BCSM	LHM
$\delta = 0.5$	0.49 (0.03)	0.51 (0.03)	0.50 (0.01)	0.52 (0.01)
$\tau = 0.5$	0.51 (0.07)	0.45 (0.06)	0.48 (0.03)	0.42 (0.03)
$\phi = 0.5$	0.51 (0.03)	0.50 (0.03)	0.50 (0.02)	0.49 (0.01)
$\nu_1 = 0$	0.00 (0.04)		−0.01 (0.02)	
$\nu_2 = -0.05$	−0.05 (0.04)		−0.06 (0.03)	
$\nu_3 = -0.1$	−0.07 (0.03)		−0.11 (0.03)	
$\nu_4 = 0.4$	0.39 (0.06)		0.39 (0.03)	
$\nu_5 = 0.2$	0.18 (0.05)		0.19 (0.02)	
$\nu_6 = 0.3$	0.28 (0.06)		0.30 (0.02)	

A comparison is made between a BCSM and the LHM. In the BCSM framework, the full within-subject dependence structure is modeled.

5.4. Results

Under the LHM, the test-wide cross-covariance and the variance of the test-taker working speed distribution are successfully estimated. The variance of the ability distribution (τ) is underestimated for both sample size conditions under the LHM, which can be attributed to ignoring the block-wise deviations from the test-wide cross-covariance. Under the BCSM, the full within-subject dependence structure is successfully estimated. Cross-covariances near zero (ν_1, ν_2, and ν_3) are estimated without bias regardless of sample size, which can be attributed to the non-informative truncated shifted inverse-gamma priors. The standard deviations of the posterior mean estimates are comparable for both models. Increasing the sample size leads to smaller standard deviations of the posterior mean estimates for both models. Under the BCSM, an average correlation of 0.99 (SD: 0.01) is observed under both conditions between the simulated measurement error variance parameters and their posterior mean estimates. The results of the simulation study are summarized in **Table 7**.

6. EMPIRICAL EXAMPLE: PIAAC 2012

The Programme for the International Assessment of Adult Competencies (PIAAC) study deploys a computer-based large scale assessment to gain insight into adult competencies across the domains of numeracy, literacy and problem solving (OECD, 2013). The computer-based nature of the assessment allows recording behavioral process data, in addition to the scored responses. It is assumed that the process data correlate with the scored responses and therefore contain information about the latent competencies of interest. Describing these correlations requires paying attention to local dependences within the data. Local dependences follow from shared item characteristics (e.g., response mode), the test design (e.g., testlets), the manner the process data is obtained (e.g., a single measurement per type, test-taker and item, multiple measurements or aggregated data) and the latent factor structure (e.g., data components load on test-takers' ability and working speed). Furthermore, test-taker characteristics such as computer experience or gender may affect the associations of data components (e.g., the correlation of RTs and RAs of an item may differ between test-takers with and without computer experience). It will be shown that a BCSM can be constructed that (a) takes the complex dependence structure within test-taker data into account, (b) allows correcting for between-subject differences in the dependence structure by including test-taker background variables, and (c) can be estimated given a reasonable sample size.

6.1. Data Set

The data set consists of responses and process data for $N = 745$ Canadian test-takers and $p = 15$ items. For each combination of item and test-taker, three data points are available: the scored dichotomous response, the total (log) RT it took the test-taker to complete the item and the (log) time it took the test-taker until they took their first action on that item. Nine of the items measure numeracy competencies, the remaining six items measure literacy competencies. Furthermore, the items differ in their response mode. See **Table 8** for an overview of the included items and their characteristics. Moreover, the test-takers' gender (0: male, 1: female), computer experience (0: no, 1: yes), whether or not they are a native speaker (0: no, 1: yes)

TABLE 8 | Id, name, domain, and response mode of the 15 PIAAC items included in the data analysis of the empirical example.

Item id.	Name	Domain	Response mode
1	Wine 1	Numeracy	Number match
2	Wine 2	Numeracy	Stimulus clicking
3	Gas gauge	Numeracy	Number match
4	Photo 1	Numeracy	Number match
5	Photo 2	Numeracy	Stimulus clicking
6	Photo 3	Numeracy	Exact match
7	Urban population	Numeracy	Number match
8	Tiles	Numeracy	Exact match
9	Package	Numeracy	Stimulus clicking
10	Baltic stock market 1	Literacy	Stimulus clicking
11	Baltic stock market 2	Literacy	Stimulus highlighting
12	Baltic stock market 3	Literacy	Stimulus clicking
13	Baltic stock market 4	Literacy	Stimulus clicking
14	TMN antitheft 1	Literacy	Stimulus highlighting
15	TMN antitheft 2	Literacy	Stimulus highlighting

and their educational level (1: low, 2: medium, 3: high) were recorded. Further information on test-taker demographics and item characteristics can be found in Statistics Canada (2013).

6.2. Dependence Structure

Data that are naturally grouped may be stronger correlated than (conditionally) unrelated data. In the data set at hand, items are grouped through their domain (numeracy or literacy) and their response mode (number match, exact match, stimulus clicking or stimulus highlighting). For each grouping, three layers are defined: one for each pair of data types. This allows to explore how the dependences between, respectively, RAs and RTs, RAs and times to first action (TAs), and RTs and TAs vary across item domains and response modes, while controlling for the rest of the dependence structure. Furthermore, data components that load on a common latent factor may be correlated. Latent factors are the test-taker's ability, working speed and speed first action. The correlation between the latent factors is modeled in separate layers. **Figure 2** illustrates the classifications that follow from the groupings. Data within each classification group may be locally dependent. The corresponding classification matrix for the $N_c = 45$ data components and $N_t = 24$ classification groups is shown in Appendix B (**Supplementary Material**).

6.3. Statistical Model

Under the BCSM framework, a model for response and process data is constructed. In the mean structure of the joint-model, test-taker background data are modeled as predictor variables. The dependence structure is modeled through an additive covariance matrix that defines the relationship of the multivariate normally distributed error terms:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{B} + \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}_{N_c}, \boldsymbol{\Sigma}), \tag{19}$$

where $\boldsymbol{Y} = \{\boldsymbol{RA}, \boldsymbol{RT}, \boldsymbol{TA}\}$ is a $N \times N_c$-dimensional matrix containing the RAs that underlie the scored dichotomous responses (RA), the total RTs per item for each test-taker (RT), and the time passed until the test-taker's first action per item (TA). The $N \times 5$-dimensional matrix $\boldsymbol{X}$ contains the grand-mean centered test-taker background variables (gender, computer experience, native speaker and education level) and a vector of ones as first column. $\boldsymbol{B}$ is a $5 \times N_c$ matrix containing the regression weights for each of the four covariates on the N_c data components, and the intercepts. The first column of $\boldsymbol{B}$ contains the item-specific intercepts, which can be interpreted as item difficulty, time intensity and average time to first action parameters. The weights and intercepts are thus modeled for each data component and are equal across test-takers, therefore representing fixed effects. Note that no random variance components are associated with fixed effects, whereby they don't enter the modeled dependence structure. The $N_c \times N_c$-dimensional additive covariance matrix $\boldsymbol{\Sigma}$ consists of $N_t = 24$ layers that correspond to the specified dependence structure:

$$\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}) + \sum_{t=1}^{N_t} \theta_t \boldsymbol{u}_t \boldsymbol{u}_t^T. \tag{20}$$

For the RA components, the measurement error variance parameters are fixed to one. Furthermore the scale of the IRT model is set by fixing the mean of the item-specific intercepts (i.e., the mean of the item difficulty parameters) to zero. The classification matrix $\boldsymbol{u}$ is shown in Appendix B (**Supplementary Material**). A truncated shifted inverse-gamma prior with $shape = 10^{-3}$ and $scale = 10^3$ is defined for the variance and covariance parameters. No a-priori information about the regression weights is used: the prior guesses for the scale matrix and the mean matrix of $\boldsymbol{B}$ equal the identity matrix and a matrix of zeros, respectively.

6.4. Results

The model parameters are estimated with a single MCMC chain of 55,000 iterations from which the first 15,000 iterations are discarded as burn-in period. Visual inspection of traceplots and applying the Heidelberger and Welch' criterion (Heidelberger and Welch, 1983) using the R-package coda (Plummer et al., 2016) indicate a satisfying exploration of the parameter space and do not provide evidence against convergence of the MCMC algorithm. The posterior means and standard deviations of the twenty-four covariance parameters in the additive covariance structure are summarized in **Table 9**. **Figure 3** shows the corresponding 95%-Highest Posterior Density (HPD) intervals.

Given the observed data, it can be concluded that the probability of local dependence in the ability and speed first action latent factor classification groups is at least 95%. Furthermore, a positive interdependence in the higher order relationship between RTs and TAs is found. This implies that on average, test-takers who work overall faster also lose less time before making the first move in the item solving process. The results indicate that it is necessary to model the implied covariance structure of the correlated person effects on each type of test-taker data (RAs, RTs, and TAs). The variation in the data explained on a person level that is captured by the latent factors (ability, working speed, and speed first action) and

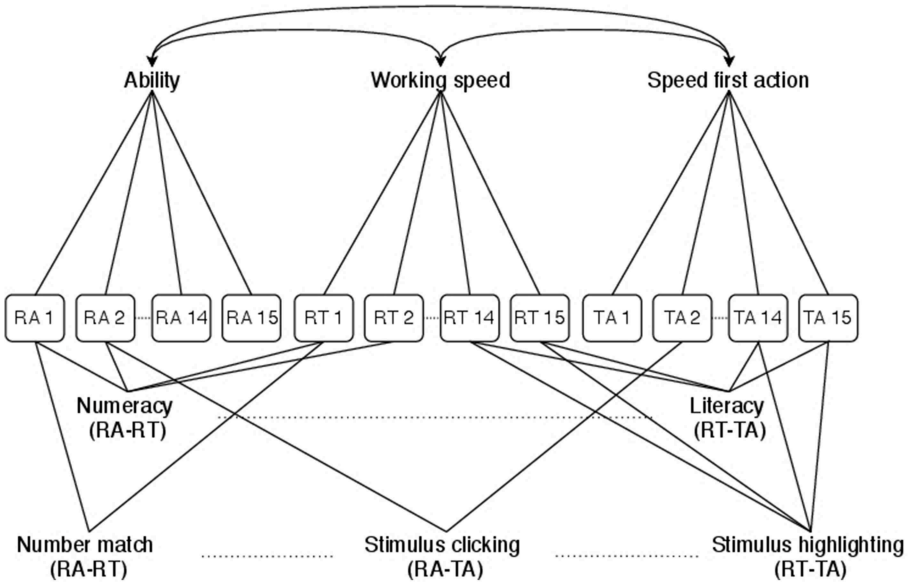


FIGURE 2 | Classification diagram for the PIAAC 2012 BCSM. The classification structure specifies dependences between scored responses and behavioral process data for varying item characteristics (domain and response mode) and a correlated latent factors structure. RA, RAs that underlie the scored dichotomous responses; RT, total RTs per item; TA, times to first action per item.

their correlation is estimated through the corresponding layers in the additive covariance structure: modeling the person effects themselves is not required.

Neither variations in the item domains, nor in the response modes caused local dependence in the data. For each domain and response mode, three sources of local dependence are independently evaluated: the relationships between, respectively, (1) RAs and RTs, (2) RAs and TAs, and (3) RTs and TAs. Modeling the 3-by-3 covariances for each specified subset of items shows that the interdependences across data types and the investigated item characteristics are sufficiently captured by the covariance layers through which the dependences of the latent factors structure are specified. It can therefore be concluded that the items' domain and response mode do not explain a noticeable amount of variance in the test-taker data when controlling for the rest of the dependence structure.

The occurrence of a vast number of covariance parameter estimates close to, or approximately equal to, zero highlights the importance of the truncated shifted inverse-gamma prior specification that avoids boundary effects by moving the edge of the parameter space away from zero. For instance, a default inverse-gamma prior would presume that $\theta_{20} > 0$ and would therefore be informative with regard to the probability of local dependence caused by the cross-relationship of RAs and TAs that belong to items with the stimulus clicking response mode: it decreases the estimated probability of local independence, i.e., the (estimated) probability that the true value of θ_{20} is zero, and can thereby provoke false conclusions about the underlying response processes. Finally, measurement error variance parameters are estimated for the fifteen RT components (mean: 0.62, SD: 0.47) and the fifteen TA components (mean: 0.31, SD: 0.19).

7. DISCUSSION

A novel Bayesian framework to model local dependences in test-taker data is proposed. The BCSM allows specifying dependences across different types of data (RAs, RTs and other process data) and multiple levels (e.g., within a testlet, clustered data per item and test-taker). The local dependences are specified through a cross-classification structure and are explicitly modeled as covariance parameters. In an additive covariance structure, nested and/or cross-classified data structures are modeled through covariance parameters.

Recording test-taker data during CBAs is not limited to scored responses and RTs. For researchers and assessors these additional process data are of utility: they can increase the precision of test-taker ability estimates and lend new insights into underlying response processes. However, using process data to draw inferences is problematic in the GLMM framework: each additional type of data requires the inclusion of new person-level variables. If interaction effects occur, the model's complexity further increases drastically. A highly complex model is prone to over-parameterization and weak numerical stability, which may strongly limit its utility in practical applications.

The BCSM framework allows the construction of parsimonious models without requiring random effects (on a person or group level) to model data dependences. Contrary to common marginal modeling approaches such as GEE, the dependence structure is however fully modeled in an additive covariance structure. This allow testing for interaction effects and to estimate the random effects *post-hoc* from the residuals of the model. By estimating random effects *post-hoc*, inferences

TABLE 9 | Posterior means and standard deviations of the $N_t = 24$ covariance parameters in the additive covariance structure.

Layer	Classification	Level	Posterior distribution	
			Mean	SD
1	Ability	Latent factor	0.47	0.16
2	Working speed	Latent factor	0.01	0.03
3	Speed first action	Latent factor	0.05	0.02
4	Ability-Working speed	Latent factor	0.01	0.01
5	Ability-Speed first action	Latent factor	−0.03	0.02
6	Working speed-Speed first action	Latent factor	0.12	0.02
7	Numeracy: RA-RT	Item domain	0.01	0.03
8	Numeracy: RA-TA	Item domain	0.05	0.03
9	Numeracy: RT-TA	Item domain	0.04	0.02
10	Literacy: RA-RT	Item domain	0.00	0.03
11	Literacy: RA-TA	Item domain	0.01	0.03
12	Literacy: RT-TA	Item domain	0.03	0.02
13	Exact match: RA-RT	Response mode	−0.01	0.17
14	Exact match: RA-TA	Response mode	0.11	0.11
15	Exact match: RT-TA	Response mode	0.07	0.10
16	Number match: RA-RT	Response mode	−0.02	0.05
17	Number match: RA-TA	Response mode	0.07	0.07
18	Number match: RT-TA	Response mode	0.04	0.04
19	Stimulus clicking: RA-RT	Response mode	0.01	0.04
20	Stimulus clicking: RA-TA	Response mode	0.00	0.03
21	Stimulus clicking: RT-TA	Response mode	0.00	0.02
22	Stimulus highlighting: RA-RT	Response mode	−0.02	0.04
23	Stimulus highlighting: RA-TA	Response mode	0.01	0.03
24	Stimulus highlighting: RT-TA	Response mode	0.02	0.02

Each layer of the covariance structure corresponds to one classification. Classifications are made across three data types (RA, response accuracies that underlie the scored dichotomous responses; RTs, response times; TAs, times to first action taken) based on (correlated) latent factors, item domains, and item response modes.

about test-taker characteristics can be made conditional on a complex within-subject dependence structure that follows from combining various auxiliary process data types into a coherent model. There is no theoretical limitation to the number of data types to combine, or in the number of components within each type (e.g., test length).

Modeling local dependences through covariance parameters instead of random effect variance parameters results in an extended parameter space. This allows more flexibility in specifying complex dependence structures (covariances can be negative, zero or positive). Compared to default inverse-gamma priors for variance parameters, truncated shifted inverse-gamma priors for the covariance parameters are less informative and allow more objective inferences about the dependence structure. The truncation is furthermore used to ensure the positive definiteness of the additive covariance structure, and can be utilized for inequality hypothesis testing (e.g., $\theta_1 < \theta_2 < \theta_3$). Through conjugacy of the proposed priors, BCSMs can be fit with an efficient Gibbs-sampling algorithm.

In a simulation study, a complex within-subject dependence structure was successfully estimated under a BCSM for responses and RTs. The model used for data generation allowed the test-takers' working speed and ability to vary over the course of a test. The LHM was not capable to capture this variation and showed bias in the variance estimate of the ability distribution. Under the BCSM, variation in test-takers' working speed and ability did not violate the condition of local independence: the dependence structure was extended to account for the variation. Furthermore, by estimating the extended dependence structure, insight into the development of the speed-accuracy trade-off on group level across the test was obtained.

The empirical example based on the PIAAC study showed a complex real-world dependence structure in response and process data. Covariance, measurement error variance and item parameters were estimated conditional on a dependence structure that took into account the classifications across three data types (scored dichotomous responses, RTs, TAs), item characteristics (domain, response mode), and the latent factor structure (data components load on the correlated factors ability, working speed and speed first action). Furthermore, test-taker background variables were included as covariates to correct for between-subject differences in the dependence structure. Through additive layers in a single covariance matrix, 3-by-3 covariance structures were modeled for each specified subset of items. This allowed to evaluate the cross-dependence between all pairs of data types individually for each of the item domains and response modes. The results indicated, that the interdependences across data types and the investigated item characteristics were sufficiently captured by the covariance layers through which the dependences of the latent factors structure were specified. The empirical example illustrates how, in the BCSM framework, the modeled dependence structure can be flexibly adapted to the design and the underlying theoretical constructs of an assessment. Furthermore, the vague nature of the truncated shifted inverse-gamma prior specification promotes unbiased inferences about the dependence structure. In the empirical example, this was in particular important due to the vast number of covariance parameter estimates close to, or approximately equal to zero. In this situation, a prior specification that does not take boundary effects into account artificially increases the estimated probability of local independence and hence provokes false conclusions about the dependence structure and the underlying response processes.

In addition to integrating multiple types of test-taker data, dependences can follow from the test design, item properties, the (sub-)population of test-takers, test-taking modes, test-taking conditions, and from an interaction of these characteristics. Examples are testlet structures, in which data within a testlet is often more alike then data across testlets (e.g., Wainer and Kiely, 1987; Yen, 1993; Wainer et al., 2007), or the interaction of culturally loaded concepts in items and diverse (sub-)populations of test-takers (e.g., with and without migration background) (e.g., Steele and Aronson, 1995; Paniagua, 2000; Good et al., 2003; Robinson, 2010). The resulting dependences in test-taker data form a threat for the flawless psychometric equivalence of an assessment, if not accounted for Helms (1992).

In educational measurement, factor loadings, or slope parameters, are utilized to assess differential item functioning (DIF) across groups, test-taking modes and over time (Millsap, 2010), allow multidimensional item response theory (MIRT)

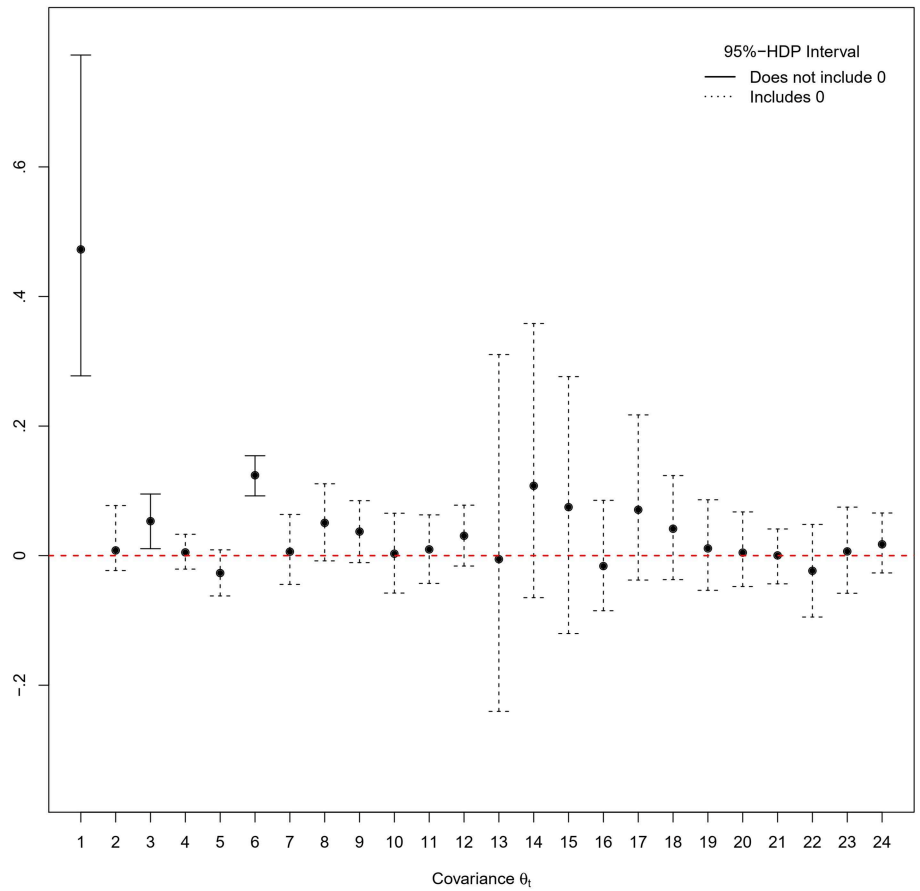


FIGURE 3 | 95%-Highest Posterior Density (HPD) intervals for the $N_t = 24$ covariance parameters in the additive covariance structure. Black dots correspond to posterior mean estimates.

(Reckase, 2009), and are used to represent the quality of an item to discriminate between distributions of test-takers with a different level of ability or speed (van der Linden, 2007; Klein Entink et al., 2008). As discussed by Klotzke and Fox (2018), factor loadings integrate seamlessly into the proposed modeling framework. In fact, the inclusion of factor loadings solely removes the restriction of values being either zero or one in the classification matrix, hence keeping the modeling structure and the therein derived equations intact. However, while this allows to include pre-calibrated factor loadings into the model, no estimation procedure has been described so far. In a conditional-BCSM hybrid model, the factor loadings can also be modeled in the mean structure instead of in the covariance structure. For example, a 2PL-IRT model with item-discrimination parameters can be specified in the mean structure and the dependences implied by a testlet structure can be explicitly modeled in the multivariate distribution of the error terms. This approach is straightforward and suited for practical applications. A downside is, that a trade-off is been made between the parsimony of the model and the number of person level variables included in the mean structure. In the empirical PIAAC data example showcased, the factor loadings were predefined given the test design and item characteristics. Freeing the factor loadings will further increase the flexibility in the modeled dependence structure and thereby the utility of BCSM for practical applications in educational measurement.

It has been shown that modeling a non-linear relationship between RAs and RTs can be beneficial (e.g., Molenaar et al., 2015; Bolsinova and Molenaar, 2018). Through the additive covariance structure in BCSM, the conditional dependence between RTs and RAs is not limited to vary solely based on item membership (i.e., data points that belong to the same item are conditionally more alike), but is allowed to change based on item characteristics (e.g., domain and response mode), test form (e.g., computer based vs. paper-and-pencil) and test design (e.g., a testlet structure). Individual test-taker characteristics that may cause between-subject differences in the dependences of RTs and RAs are controlled for through modeling test-taker background variables as covariates in the mean structure (e.g., the relationship between RTs and RAs may vary based on the test-takers' age or a pre-test speed categorization). This differs from methods that model a non-linear relationship between RTs and RAs through a predefined function that involves person-specific random components

and/or item parameters (Molenaar et al., 2015; Bolsinova et al., 2017; Bolsinova and Molenaar, 2018): in BCSM, test-taker characteristics that may affect the relationship between data types are controlled for in the mean structure, and the person-specific random effects are not modeled. Item characteristics are modeled in the mean structure (e.g., item difficulty parameters) and through additive layers in the covariance structure (e.g., item response mode). It is an interesting future prospect to see in how far the BCSM framework can be extended for covariance structures that follow from curvilinear functions for the relationship between data types. Furthermore, the BCSM approach must be distinguished from methods that model a person-specific covariance matrix (e.g., Meng et al., 2015): by their nature, models that explicitly specify a covariance matrix for each test-taker heavily increase in complexity with growing sample size and thus must impose strong restrictions on the modeled dependence structure to achieve model identification. In contrast, BCSM aims at designing parsimonious models that are easily identified when complex dependence structures are modeled.

The BCSM framework is not limited to RTs and dichotomous responses. Dependences between dichotomous responses and RTs were modeled through latent continuous variables. Expressions for the mean and variance of the conditional normal distribution of a latent variable were obtained by partitioning the additive covariance matrix and analytically deriving its inverse. Information from the observed responses (whether or not a test-taker responded correctly to an item) was utilized by truncating the respective distribution. Modeling dependences through latent continuous variables can be extended to data with more than two ordered or unordered response categories (e.g., Castro et al., 2012). This extends the range of process data that can be integrated into a BCSM. For example, sequential action patterns can be operationalized as count variables through N-grams (He et al., 2016). It is interesting to see under which conditions a BCSM allows to draw inferences about the interdependence between responses, RTs and action patterns and which new insights into latent response processes can be obtained. Further future prospects of BCSMs are the application to additional real-world empirical settings, extensions to unbalanced data and nested classifications on a person level (e.g., a test-taker is part of a school and classroom), and evaluating the utility of estimating test-taker effects *post-hoc*. Finally, it is of interest to compare the plausibility of different dependence structures in a Bayesian model selection framework (e.g., Kass and Raftery, 1995).

AUTHOR CONTRIBUTIONS

KK wrote the manuscript, developed the software and performed the analysis. J-PF and KK developed the modeling framework and the MCMC algorithm. J-PF contributed to the structure of the manuscript, and gave suggestions for the analysis and writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01675/full#supplementary-material

REFERENCES

Ackerman, T. A. (1987). *The Robustness of LOGIST and BILOG IRT Estimation Programs to Violations of Local Independence (Research Report No. 87-14)*. Iowa City: The American College Testing Program.

Azevedo, R. (2015). Defining and measuring engagement and learning in science: conceptual, theoretical, methodological, and analytical issues. *Educ. Psychol.* 50, 84–94. doi: 10.1080/00461520.2015.1004069

Baker, F. B., and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques, 2nd Edn*. New York, NY: CRC Press.

Bolsinova, M., and Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Front. Psychol.* 9:1525. doi: 10.3389/fpsyg.2018.01525

Bolsinova, M., Tijmstra, J., and Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *Brit. J. Math. Stat. Psychol.* 70, 257–279. doi: 10.1111/bmsp.12076

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *J. R. Stat. Soc. Ser. B* 79, 125–148. doi: 10.1111/rssb.12162

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika* 64, 153–168. doi: 10.1007/BF02294533

Bridgeman, B., and Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *J. Educ. Meas.* 41, 137–148. doi: 10.1111/j.1745-3984.2004.tb01111.x

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *J. Exp. Psychol. Learn. Mem. Cogn.* 36, 1118–1133. doi: 10.1037/a0019902

Castro, M., Paleti, R., and Bhat, C. R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: application to pre- dicting crash frequency at intersections. *Transport. Res. B Methodol.* 46, 253–272. doi: 10.1016/j.trb.2011.09.007

Chen, W.-H., and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *J. Educ. Behav. Stat.* 22, 265–289. doi: 10.3102/10769986022003265

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2013). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* 32, 407–499. doi: 10.1214/009053604000000067

Fox, J.-P., Klotzke, K., and Klein Entink, R. H. (2018). *LNIRT: LogNormal Response Time Item Response Theory Models*. Available online at: https://CRAN.R-project.org/package=LNIRT

Fox, J.-P., Mulder, J., and Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika* 82, 979–1006. doi: 10.1007/s11336-017-9577-6

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: CRC Press.

Glas, C. A., and van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *Brit. J. Math. Stat. Psychol.* 63(Pt 3), 603–626. doi: 10.1348/000711009X481360

Goldhammer, F., and Kroehne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: experimental time limits, posterior time limits, and response time modeling. *Appl. Psychol. Meas.* 38, 255–267. doi: 10.1177/0146621613517164

Goldhammer, F., Naumann, J., and Greiff, S. (2015). More is not always better: the relation between item response and item response time in Raven's matrices. *J. Intell.* 3, 21–40. doi: 10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated

by task diffi- culty and skill: insights from a computer-based large-scale assessment. *J. Educ. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Goldhammer, F., and Zehner, F. (2017). What to make Of and how to interpret process data. *Measurement* 15, 128–132. doi: 10.1080/15366367.2017.1411651

Good, C., Aronson, J., and Inzlicht, M. (2003). Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. *J. Appl. Dev. Psychol.* 24, 645–662. doi: 10.1016/j.appdev.2003.09.002

Gupta, A. K., and Nagar, D. K. (1999). *Matrix Variate Distributions*. New York, NY: CRC Press.

He, Q., and Von Davier, M. (2016). "Analyzing process data from problem-solving items with n-grams: insights from a computer-based large-scale assessment," in *Handbook of Research on Technology Tools for Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf (New York, NY: Chapman and Hall/CRC Press), 749–776. doi: 10.4018/978-1-4666-9441-5.ch029

Heidelberger, P., and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operat. Res.* 31, 1109–1144. doi: 10.1287/opre.31.6.1109

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *Am. Psychol.* 47, 1083–1101. doi: 10.1037//0003-066X.47.9.1083

Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *J. Multivar. Anal.* 94, 209–221. doi: 10.1016/j.jmva.2004.10.007

Jiao, H., Kamata, A., Wang, S., and Jin, Y. (2012). A multilevel testlet model for dual local dependence. *J. Educ. Meas.* 49, 82–100. doi: 10.1111/j.1745-3984.2011.00161.x

Jiao, H., Wang, S., and Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *J. Appl. Meas.* 6, 311–321.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2008). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74:21. doi: 10.1007/s11336-008-9075-y

Klotzke, K., and Fox, J.-P. (2018). *Response Times in a Bayesian Marginal Modeling Framework*. LSAC Research Report Series. Newport, PA: Law School Admission Council (USA).

Lange, K. (2010). *Numerical Analysis for Statisticians, 2nd Edn*. New York, NY: Springer Publishing Company, Incorporated.

Lee, Y., and Neider, J. A. (2004). Conditional and marginal models: another view. *Stat. Sci.* 19, 219–228. doi: 10.1214/088342304000000305

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22. doi: 10.1093/biomet/73.1.13

Loeys, T., Legrand, C., Schettino, A., and Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *Brit. J. Math. Stat. Psychol.* 67, 304–327. doi: 10.1111/bmsp.12020

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

Maddox, B. (2017). Talk and gesture as process data. *Measurement* 15, 113–127. doi: 10.1080/15366367.2017.1392821

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., and Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *J. Educ. Behav. Stat.* 39, 426–451. doi: 10.3102/1076998614559412

McCulloch, C. E. (2003). "Generalized linear mixed models," in *NSF-CBMS Regional Conference Series in Probability and Statistics* (Beachwood, OH), Vol. 7, i-84.

Meijer, R. R., and Sotaridona, L. S. (2006). *Detection of Advance Item Knowledge Using Response Times in Computer Adaptive Testing (info:eu-repo/semantics/report No. CT 03-03)*. Newton, PA: Law School Admission Council.

Meng, X.-B., Tao, J., and Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *J. Educ. Meas.* 52, 1–27. doi: 10.1111/jedm.12060

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: an introduction. *Child Dev. Perspect.* 4, 5–9. doi: 10.1111/j.1750-8606.2009.00109.x

Molenaar, D., Oberski, D., Vermunt, J., and De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivar. Behav. Res.* 51, 606–626. doi: 10.1080/00273171.2016.1192983

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivar. Behav. Res.* 50, 56–74. doi: 10.1080/00273171.2014.962684

OECD (2013). *Technical Report of the Survey of Adult Skills (PIAAC) (Tech. Rep.)*. Paris: OECD Publishing.

Paniagua, F. A. (2000). *Handbook of Multicultural Mental Health: Assessment and Treatment of Diverse Populations*. San Diego, CA: Academic Press.

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., et al. (2016). *CODA: Output Analysis and Diagnostics for MCMC*. Available online at: https://CRAN.R-project.org/package=coda

Qian, H., Staniewska, D., Reckase, M., and Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educ. Meas.* 35, 38–47. doi: 10.1111/emip.12102

Ranger, J., and Kuhn, J.-T. (2013). Analyzing response times in tests with rank correlation approaches. *J. Educ. Behav. Stat.* 38, 61–80. doi: 10.3102/1076998611431086

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer-Verlag.

Robinson, J. P. (2010). The effects of test translation on young english learners' mathematics performance. *Educ. Res.* 39, 582–590. doi: 10.3102/0013189X10389811

Statistics Canada (2013). *Skills in Canada: First Results From the Programme for the International Assessment of Adult Competencies (PIAAC) (Monograph)*. Ottawa, ON: Statistics Canada.

Steele, C. M., and Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *J. Pers. Soc. Psychol.* 69, 797–811. doi: 10.1037/0022-3514.69.5.797

Thissen, D. (1983). "9 - Timed testing: an approach using item response theory," in *New Horizons in Testing*, ed D. J. Weiss (San Diego, CA: Academic Press), 179-203.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika* 67, 33–48. doi: 10.1007/BF02294708

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72:287. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *J. Educ. Meas.* 46, 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., Breithaupt, K., Chuah, S. C., and Zhang, Y. (2007). Detecting differential speededness in multistage testing. *J. Educ. Meas.* 44, 117–130. doi: 10.1111/j.1745-3984.2007.00030.x

van der Linden, W. J., and Fox, J.-P. (2016). "Joint hierarchical modeling of responses and response times," in *Handbook of Item Response Theory, Volume One, Models*, ed W. J. van der Linden (New York, NY: Chapman and Hall/CRC Press), 481–500.

van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. New York, NY: Cambridge University Press.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *J. Educ. Meas.* 24, 185–201. doi: 10.1111/j.1745-3984.1987.tb00274.x

Wang, C., and Wilson, M. (2005). The rasch testlet model. *Appl. Psychol. Meas.* 29, 126–149. doi: 10.1177/0146621604271053

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Brit. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wetzel, E., Xu, X., and von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educ. Psychol. Meas.* 75, 739–763. doi: 10.1177/00131644 14558843

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Meas.* 8, 125–145. doi: 10.1177/014662168 400800201

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *J. Educ. Meas.* 30, 187–213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Klotzke and Fox. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 18 February 2019
doi: 10.3389/fams.2019.00002

Invariance of the Response Processes Between Gender and Modes in an Assessment of Reading

Ulf Kroehne[1]*, *Carolin Hahnel*[1,2] and *Frank Goldhammer*[1,2]

[1] DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany, [2] Centre for International Student Assessment (ZIB), Frankfurt am Main, Germany

In this paper, we developed a method to extract item-level response times from log data that are available in computer-based assessments (CBA) and paper-based assessments (PBA) with digital pens. Based on response times that were extracted using only time differences between responses, we used the bivariate generalized linear IRT model framework (B-GLIRT, [1]) to investigate response times as indicators for response processes. A parameterization that includes an interaction between the latent speed factor and the latent ability factor in the cross-relation function was found to fit the data best in CBA and PBA. Data were collected with a within-subject design in a national add-on study to PISA 2012 administering two clusters of PISA 2009 reading units. After investigating the invariance of the measurement models for ability and speed between boys and girls, we found the expected gender effect in reading ability to coincide with a gender effect in speed in CBA. Taking this result as indication for the validity of the time measures extracted from time differences between responses, we analyzed the PBA data and found the same gender effects for ability and speed. Analyzing PBA and CBA data together we identified the ability mode effect as the latent difference between reading measured in CBA and PBA. Similar to the gender effect the mode effect in ability was observed together with a difference in the latent speed between modes. However, while the relationship between speed and ability is identical for boys and girls we found hints for mode differences in the estimated parameters of the cross-relation function used in the B-GLIRT model.

Keywords: reading ability, computer-based assessment, mode-effects, response times, log data, paper-based assessment

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Minjeong Jeon,
University of California, Los Angeles,
United States
Carol Eckerly,
Educational Testing Service,
United States

***Correspondence:**
Ulf Kroehne
kroehne@dipf.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *31 August 2018*
Accepted: *14 January 2019*
Published: *18 February 2019*

Citation:
Kroehne U, Hahnel C and Goldhammer F (2019) Invariance of the Response Processes Between Gender and Modes in an Assessment of Reading. Front. Appl. Math. Stat. 5:2. doi: 10.3389/fams.2019.00002

INVARIANCE OF THE RESPONSE PROCESSES BETWEEN GENDER AND MODES IN AN ASSESSMENT OF READING

Technology-based assessments offer the possibility to collect additional log data, including response times as the amount of time test-takers spend responding to particular questions or tasks. Primarily, response times provide information about test-takers' speed, which—if not considered—might confound comparisons of test-takers' ability [2] due to inter-individual differences in the speed-ability compromise [3]. Response times can allow inferences about response processes [4, 5]. Although at the individual level response times are influenced by multiple factors [6], the comparison of characteristics at the group level, such as the between-person relationship of speed

and ability (e.g., [7]), can provide information about the comparability of underlying processes. Yet, response times and the relationship between speed and ability might also be considered as criteria for a comparison of response processes between test administrations with different properties [8]. Such mode effect studies often focus on the degree of measurement invariance (e.g., [9]), typically excluding effects of between-person characteristics, such as gender, either as differential item functioning (DIF, e.g., [10]) or ability difference at the population level (e.g., [11, 12]).

Research on mode effects and gender differences both consider mainly comparisons concerning the correctness of responses (response accuracy). Time-related criteria, such as the number of not reached items or response times at the item level, are rarely considered (see [13], for an exception) for the investigation of equivalence between modes or gender. Instead, as summarized by Kong et al. [14], reading speed is often analyzed for the comparison between modes.

Focusing on differences in the response accuracy requires that the underlying response processes are identical. However, the differences in reading comprehension found between computer-based assessment (CBA) and paper-based assessment (PBA, e.g., [15]), and boys and girls [e.g., [16]] might also be caused by construct-related differences such as reading strategies (e.g., [17]) between mode or gender. These differences might be associated with different test-taking processes in the assessments, resulting in a certain response behavior which allows investigating the relationship between speed and ability as characteristic of the response process.

Since PISA introduced CBA in 2015 [18], some authors have questioned the comparability of CBA and PBA in PISA, in particular, concerning trend estimates (e.g., [19, 20]). Although a detailed review of the current literature regarding mode effects is beyond the scope of this paper (see, e.g., [8]), we aim at providing a method for investigating response time differences between modes, which is currently missing. The potential benefit of a method that allows investigating speediness as an additional criterion of equivalence for mode effects is 2-fold: Allowing the removal of potential confounding of inter-individual differences in the speed-ability compromise concerning the comparability across modes and developing models for the explanation of mode effects beyond their simple quantification.

Comparing response processes regarding response times between modes requires the accessibility of response time measures at the item-level. While the availability of response times is often mentioned as one of the benefits of CBA (e.g., [21]), response times are metered only at an aggregate level in PBA (e.g., [22]). This lack of detailed time information from PBA instruments restrains the investigation of mode effects concerning time-related equivalence criteria and results in limited knowledge about the comparability of response processes between modes. The current study aims at filling this gap by using time measures collected with so-called digital pens in comparison to time measures extracted from log files of CBA test administration. Hence, gender differences which are well-known to exist concerning reading ability are investigated at first demonstrating the usefulness of the derived time measures for the comparison of response times of boys and girls in CBA and PBA separately. Afterward, the time measures are used to investigate mode effects assuming that the operationalization of the time measures is comparable between CBA and PBA. To summarize, we use B-GLIRT models, which are introduced in the next section, to investigate speed and ability using data from a technology-based assessment conducted as add-on study to PISA 2012 in Germany.

Modeling Response Times With the B-GLIRT Framework

Recent psychometric models incorporate response times in item response theory (IRT) models, either with a constant speed assumption (hierarchical modeling, e.g., [23]; bivariate generalized linear IRT, B-GLIRT, [1]) or without (e.g., [24]). In this paper, the B-GLIRT approach is focused, because this framework provides promising features: (i) it can be adapted to various relationships between responses and response times, (ii) it can be estimated with standard SEM software, and (iii) it can be applied to investigate measurement invariance [1]. The B-GLIRT model as shown in **Figure 1** [1] models responses to dichotomous items using a linear model and a link function, known to result in the normal ogive version of the 2-parameter model

$$E\left(Z_{pi}\right) = \Phi^{-1}\left[E\left(X_{pi}\right)\right] = \alpha_i\theta_p + \beta_i \tag{1}$$

with a slope parameter α_i and an intercept parameter β_i for each item i. In addition to the responses X_{pi}, the response times T_{pi} are included in the B-GLIRT model, log-transformed, as factor model

$$E\left(\ln\ T_{pi}\right) = \lambda_i + \varphi_i\tau_p + f\left(\theta_p; \boldsymbol{\rho}\right) \tag{2}$$

with $\sigma^2_{\omega i}$ as the variance of the residual variables ω_i. The intercepter parameter λ_i is the time intensity parameter for item i and a slope parameter φ_i as time discrimination relates the latent speed factor τ_p to the (log-) response time.

The flexible nature of the B-GLIRT model comes into play when the cross-link function $f\left(\theta_p; \rho\right)$ is specified, allowing the estimation of models that are equivalent or similar to different psychometric models for the simultaneous estimation of response accuracy and response times. A B-GLIRT model for response times that corresponds to the model proposed by Thissen [25] with an additional time discrimination parameter is

$$E\left(\ln\ T_{pi}\right) = \lambda_i + \varphi_i\tau_p - \rho_1\alpha_i\theta_p \tag{3}$$

with ρ_1 as the slope parameter for the regression of (log-) response time on the latent ability variable θ_p [see [1], for the derivation of the cross-link function]. We will refer to this model as the B-GLIRT regression model.

A second model can be specified using the following form of the cross-link function:

$$E\left(\ln\ T_{pi}\right) = \lambda_i + \varphi_i\tau_p - \varphi_i\rho_1\theta_p \tag{4}$$

The model in Equation (4) is equivalent to the model suggested by van der Linden [23] as the hierarchical model, extended by the

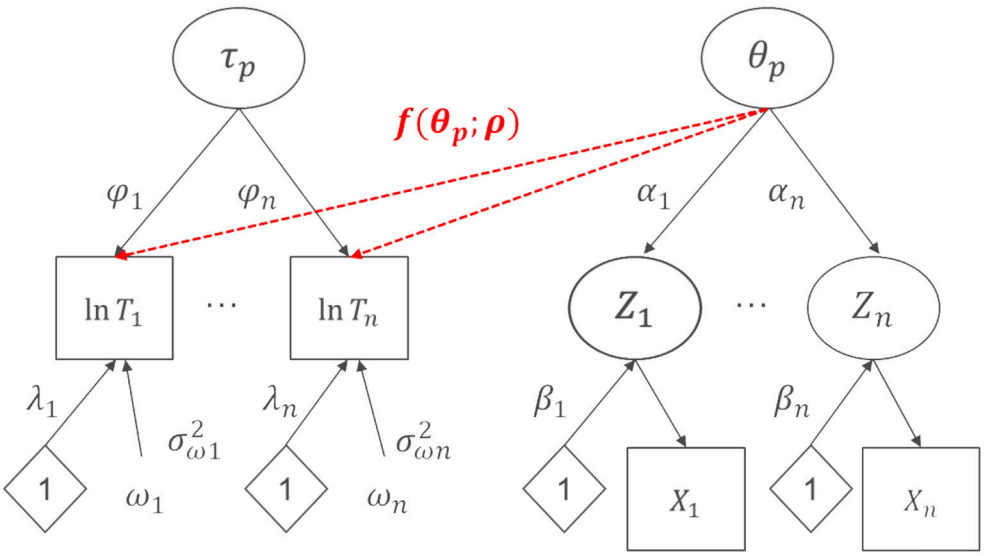


FIGURE 1 | Schematic display of the B-GLIRT Model by Molenaar et al. [1] used to investigate measurement invariance.

slope parameter φ_i for the time discrimination. The parameter ρ_1 can be interpreted as correlation, depending on the constraint used to identify the factor model. We will refer to this model as the B-GLIRT hierarchical model.

A third model, which is to some extent equivalent to a between-subject version of the model suggested by Partchev and De Boeck [26], can be specified by adding an interaction term in the cross-relation function as follows:

$$E\left(\ln\ T_{pi}\right) = \lambda_i + \varphi_i\tau_p + \rho_1\theta_p + \rho_2\tau_p\theta_p. \tag{5}$$

In this model, the parameter ρ_1 corresponds to the linear regression coefficient, while ρ_2 is the regression coefficient for the latent interaction term $\tau_p\theta_p$. We will refer to this model as the B-GLIRT interaction model.

Item-Level Response Times

As we will describe in this section, time differences between subsequent responses were used to derive comparable item-level response times for PBA and CBA. The main idea that allows extracting time measures from PBA and CBA is to focus on time between responses while taking into account the order of responses. For instance, a proxy for the item-level response time for a question "Q2" (the second question of a hypothetical test) is the time difference between the responses to question "Q2" and "Q1" if both questions are answered consecutively. When all questions in a booklet are answered in a strictly ascending order, time measures for all questions could be derived from the time difference between responses (besides the very first question "Q1").

Reconstructing the Response Sequence Using States

Omitted responses and answers in non-ascending order have to be dealt with to obtain quantities that allow an interpretation as item-level response times measures in real data applications. The strategy described in this paper requires partitioning the whole testing time into segments, which start and end with the selection of answers in consecutive questions. The necessary theoretical justification for the treatment of the self-selected order of responses can be provided by a general framework that uses log data to distinguish meaningful states of the test-taking process [27]. In this framework, log events are processed algorithmically by reconstructing the sequence of states using, for instance, a finite state machine. Using this formal method allows implementing a procedure that can be applied to extract response times from the gathered raw log events that takes the response sequences and omitted responses as different "states" into account. The considered states correspond to sections of the test-taking processes that can be interpreted with respect to reading the text or answering a particular question.

Creating States Using Answer-Change Events Only

In the following, the theoretical framework will be applied to create meaningful sequences, by considering only so-called answer-change events (i.e., events that occur when the response to a task is changed). **Figure 2** presents an example for three different state sequences for the first three questions of a test.

For the first sequence with the response pattern "R-R-R-…" (i.e., a response R is given for the first three items, and no response is missing), the first answer-change event is used to separate the state "Reading Unit Text & Answering Q1" from the state "Answering Q2." Since it is impossible to extract from the time differences between responses, at which point a test-taker exactly started a unit, no item-level time measure can be extracted for the very first question in this approach. However, when the second question is answered, the time between the first and the second response can be counted as time component for question Q2.

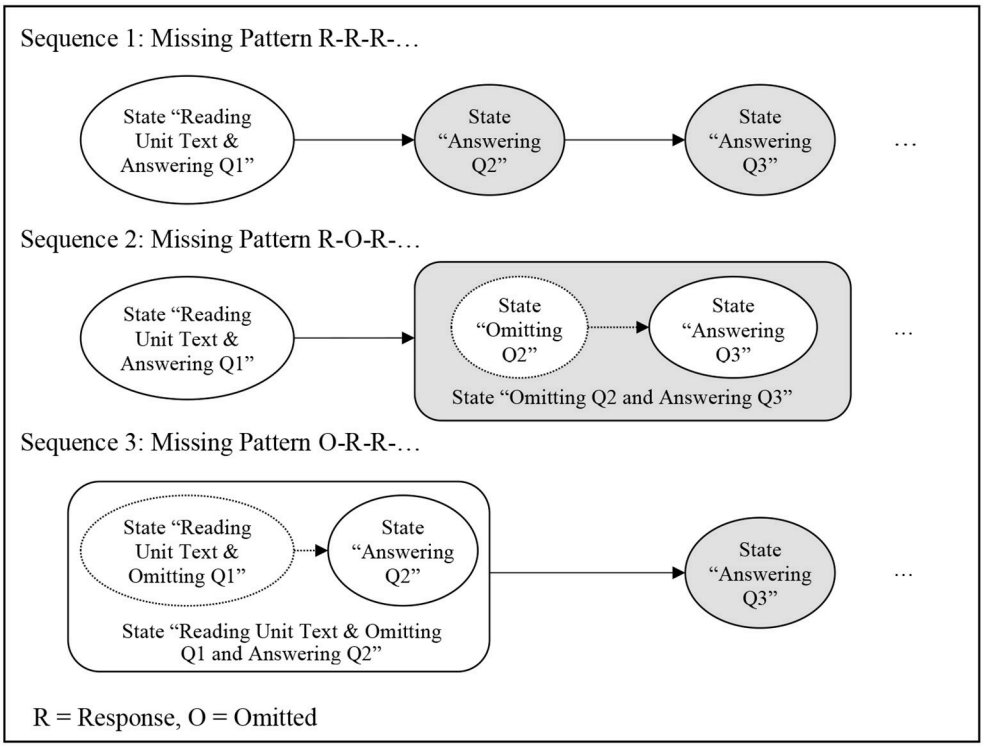


FIGURE 2 | State sequences for the responses to the first 3 questions for 3 different missing response pattern.

The second pattern illustrates the effect of an omitted item. In this sequence with the pattern "R-O-R-…" (i.e., the second item was omitted, O), the first answer-change event separates the states "Reading Unit Text & Answering Q1" from the state "Omitting Q2 and Answering Q3." According to the nature of missing responses, no time measure for the two states "Omitting Q2" and "Answering Q3" can be extracted, because the two underlying states collapse into the state "Omitting Q2 and Answering Q3" due to the missing answer-change event for question Q2. Hence, it depends on additional reasoning whether and how the time measure for this collapsed state "Omitting Q2 and Answering Q3" should be used. For sequences with more than one omitted response (e.g., R-O-O-R, not included in **Figure 1**), the structure of the problem remains identical: the time differences between two answer-change events do not allow to identify states with a clear meaning if states collapse due to omitted responses. As we will show later, sequences with collapsed states resulting from the process of omitting one or multiple (further) questions can either be ignored (option "strict") or counted as time components for the first questions after the omission (option "liberal").

The third sequence in **Figure 1** illustrates the missing value pattern "O-R-R-…" (i.e., the first question is omitted). In this situation, the observed answer-change event can be used to identify the transition from the state "Reading Unit Text & Omitting Q1 and Answering Q2" and the state "Answering Q3." Even though no item-level response time for the collapsed state "Reading Unit Text & Omitting Q1 and Answering Q2" is available, the time between the responses to question Q2 and question Q3 allows extracting an item-level response time for question Q3 with a clear interpretation. Note that all sequences start in the state "Reading Unit Text & Answering Q1" because the reading text was shown on the first page(s) of the CBA instrument and the reading text was printed on the first page(s) of the PBA booklet, respectively. The use of the finite state machine approach for analyzing log data assumes that the test-taking process can be described as a progression of states that corresponds to one question at a time, starting with the reading text and the first question for each unit. By starting the finite state machine used to reconstruct the sequence of states for each test-taker in the state "Reading Unit Text & Answering Q1," knowledge about the CBA and PBA instrument is included in the analysis using the finite state machine approach introduced in [27].

Defining the Interpretation of Time Measures Using Bigrams

To generalize from the sequences shown in **Figure 2** to all possible sequences, we consider subsequences of length two (i.e., pairs of subsequent answers). Subsequences can be described using the terminology of n-grams of all potential sub-sequences. Pairs of subsequent answers create n-grams of size 2, labeled as bigrams. The bigrams Q1-Q2 and Q2-Q3, for instance, correspond to the sequence one as shown in **Figure 2**. Item-level response times can be extracted using bigrams with the finite state machine approach, in which the time between responses is counted as time component for the second part of each bigram.

That is, the time difference between the observed responses in the bigram Q1-Q2 is used as time component counting for the item-level response time for question Q2.

The representation of subsequences as bigrams created from the reconstructed sequence of states points to the different options for the handling of omitted responses and non-ascending response orders. If no additional navigation-related log data are incorporated (i.e., if only time differences between responses are used that create answer-change events)[1], an assumption is required for a "liberal" treatment, namely the assumption that the time between two responses can always be counted for the question which was answered last. Time measures for collapsed states can also be derived under this assumption (e.g., the item-level response time for omitting question Q2 and answering question Q3 after question Q1, see the second sequence in **Figure 2**). However, depending on the frequency of omitted responses, time measures for collapsed states (i.e., bigrams resulting from omitted responses or jumps to previous questions), that occur only rarely in an empirical application might not be possible to include in a psychometric model due to data sparseness. Therefore, we consider the following two options to deal with bigrams that represent either omitted responses (such as Q1-Q3) or non-ascending self-selected response orders (such as Q3-Q1):

- Option 1 (Strict): Only time measures of adjacent tasks are extracted, using missing values for response times when questions are omitted, or answers are given in a self-selected, non-ascending order for a particular test-taker. Item-level response times extracted in this strict way have the clearest interpretation but are only available for a subset of responses. They cannot be computed for responses that are given next to an omitted response and they are missing for questions answered in non-ascending order.
- Option 2 (Liberal): Time measures for all bigrams are extracted and interpreted regardless of the previous response. For each bigram of two responses, the time difference between the two responses is counted as time component for the second response of the bigram, regardless whether the two responses represented in the bigram are direct neighbors or any pair of questions, answered one after the other. This option results in item-level response times that always contain the true response times, but also additional time for omitting one or multiple responses. Accordingly, more noise is absorbed into the response time measures using option 2.

Accordingly, the following interpretations of item-level response times are possible for the exemplary sequences with omissions in **Figure 2**. For the second sequence, only the bigram Q1-Q3 is observed because Q2 is omitted. The time between the response to the first question Q1 and the response to the third question Q3 could be interpreted as time measure for the third response Q3 (option "liberal"). Alternatively, the time measure could be modeled separately or excluded from the analysis, as such a time measure is only available from persons showing this particular pattern (omitting Q2). For the third sequence, only the bigram Q2-Q3 can be observed in the data, because Q1 was omitted. Accordingly, no time measure for Q1 can be extracted for persons showing such a pattern. In summary, the derivation of item-level response times can be described as follows: The test-taking process is segmented into states based on the answer-change events. The treatment of the time measure for a particular state is derived from the bigram of two subsequent states. A liberal option to extract item-level response times ignores the first state in the bigram, a strict option creates time measures for identical bigrams only, resulting in more missing values.

Cumulating Response Times for Multiple Answer-Changes

In the empirical application, the within-unit navigation was neither restricted in CBA nor PBA[2]. Accordingly, any bigram, for instance, Q2-Q3, could occur multiple times if answers to question Q2 and Q3 were selected and changed again. Hence, the item-level response times had to be aggregated within test-takers before using them in the psychometric model. Aggregating response times over multiple exact identical bigrams (option 1) is only expected for a small number of cases. In contrast, for option 2, item-level response times from all bigrams that end with a particular question are aggregated to derive total item-level response times for a particular question. In both options, the response times are cumulated.

Hypotheses

A linear cross-relation function in the B-GLIRT model seems plausible for ability tests if higher underlying abilities are related to faster responses (i.e., if working more fluently and faster reflects higher ability level). Spending more time on a particular task may also indicate a more careful work on the task, predicting a linear relationship in the opposite direction. From the "*time on task*" literature (e.g., [28]) it was derived that the linear relationship between speed and ability should be negative for reading items. However, it is known from previous research that the time on task effect increases with increasing ability level [28]. Hence, it is expected that for a linear cross-relation function the correlation is negative (H1a) and that a speed-ability interaction, as modeled in the interaction term $\tau_p\theta_p$ of the cross-relation function f(.) in Equation (5), can be found (H1b). Regarding the two different options to operationalize response times at the item level, we have no specific hypothesis, i.e., we investigate the robustness of the findings regarding both options for hypothesis H1a and H1b.

We expected to replicate gender differences for reading ability that were repeatedly found in previous research [e.g., [12]] in CBA (H2a) and PBA (H2b). However, we did not expect different response processes between boys and girls, i.e., equal parameters

[1]In this paper we focus on possibilities that are applicable to log data gathered in PBA and CBA, ignoring approaches that use, for instance, navigation-related paradata or other information that indicate, how long a page with an omitted task was visible on a computer screen.

[2]Note that the between-unit navigation was not restricted in PBA (i.e., students were able to go back to a previous unit within the booklet), but between-unit navigation was not possible in the CBA implementation. As discussed, for instance, by Kroehne and Martens [8] the different possibilities to navigation between units could contribute to the differences between modes.

of the cross-relation function are expected for boys and girls, again for CBA (H3a) as well as for PBA (H3b), when the model takes ability and speed differences into account.

Analyzing PBA and CBA together, we also expected equal parameters of the cross-relation function for both modes (H4), after taking inter-individual differences in the accuracy mode effect into account which were found in previous research [Kroehne et al. (submitted)].

To provide empirical evidence regarding the validity of the extracted item-level response times (a) and the response times collected using digital pens (b), we disentangled the analyses of gender and mode effects. After selecting the parameterization fitting best for the cross-relation function we first analyzed data for CBA concerning gender effects. Subsequently, we repeated the analyses of gender effects using PBA data aiming at similar patterns and, in particular, equal parameters of the cross-relation function. Based on these analyses we proceeded by analyzing CBA and PBA simultaneously concerning mode effects.

METHOD

Instruments

Items measuring reading comprehension investigated in this study were taken from the PISA 2009 reading assessment [18]. Two intact clusters with non-overlapping items had been selected and computerized with the CBA-ItemBuilder [29]. The two clusters comprised five polytomously scored items with multiple score categories and 32 dichotomously scored items (eight units in total). Polytomously scored items were dichotomized in this study by merging full credits and partial credits. Kroehne et al. (submitted) presents a detailed description of various properties of the test administration that constituted the assessment in both modes. Specifically, PBA was conducted with digital pens [for technical details see [13]] allowing to record time stamps and digital traces of strokes that provide the basis for the comparison of response processes as indicated by response times.

Sample

In this study, 856 students (aged from 15.33 to 16.33, $M = 15.82$, $SD = 0.29$) were assessed (48.67% female). The subset of students was sampled randomly from the sample of German PISA 2012 main study schools and none of the sampled schools were excluded due to technical problems. The sample contained 33.9% students from the academic track, 15.89% were immigrants in the first or second generation, and 8.86% reported that German is not their language at home.

Design

An experimental design with random assignment of test-takers to modes was implemented. A between-subject design was supplemented by an additional within-subject component to investigate construct-related changes and cross-mode correlations. For that purpose, a subset of 440 test-takers answered reading items in CBA and PBA (i.e., one cluster in each mode). Those students had to change the administration mode (i.e., switching between modes in the middle of the testing session was implemented). The sequence of modes (CBA-PBA and PBA-CBA) was balanced between the 440 test-takers to avoid confounding of mode and position effects. None of the test-takers answered both clusters in the identical mode. Consequently, the 416 test-takers in the between-subject part of the design which were administered only one cluster in one mode had missing values by design for all units of the other cluster. In both modes the reading assessment was administered self-paced with a time limit at the cluster level and only the mode was randomly assigned, while the speed for reading texts and answering tasks was self-selected.

Data Analysis

As described above, item-level response times were extracted from the CBA log data and the digital traces of strokes gathered with digital pens using time differences between responses only and further prepared using the strict and the liberal option as described above. Specifically, we operationalized the time of an answer-change event comparable between both modes as the point in time when the last response-related action to a particular question was observed (mouse click for complex and simple multiple-choice items or typing for text response in the CBA mode; last XY-coordinate event of a stroke in the PBA mode). Focusing on the last response-related action results in time measures that absorb the answering time itself (i.e., the time that was required to write, type, or select an answer).

The log-transformed response times were used in latent variable models for the speed factor, after trimming item-level time measures by recoding response times that were larger than 300 s[3] as missing values. Time measures for the response to the first question in each cluster were not included in the mode due to data sparseness (i.e., the models were built using up to 30 responses and up to 28 response times for each test-taker).

B-GLIRT models were estimated in Mplus 7.4 [30] using dichotomous indicators of the item responses for the ability factor. Multi-group structural equation models with latent interaction term were estimated using the MLR estimator and the Knownclass-option of Mplus. Example inputs for the different models are provided in the digital supplement.

Absolute model fit measures were not available for this estimation of the B-GLIRT models. Therefore, the analyses were based on the assumption that the ability part fit the particular IRT model in each mode (see Kroehne et al. (submitted), for a detailed investigation of mode effects in the responses, ignoring response times). For the B-GLIRT models, which also contain a speed part, the information criteria AIC and BIC were used for relative model comparisons. As discussed by Vrieze [31], the choice between AIC and BIC depends on the researchers' notion of the true model and the assumption whether the true model is one of the candidate models. Hence, the selection of AIC or BIC also reflects the result of weighing efficiency (i.e., finding the model that minimizes the mean squared error of prediction; AIC) against consistency (i.e., finding asymptotically the true model, if it is one of the candidate models; BIC).

In the first step of the analysis we aim at finding the best fitting parametrization of the cross-relation function.

[3]The value of 300 s was selected as a threshold to remove outliers without any further theoretical justification.

For this comparison we do not necessarily expect that the true model is part of the candidate models, so we prefer the AIC. The set of investigated parameterizations of the cross-relation function is chosen as the models that fit into the framework of the B-GLIRT models and that were used in previous research to investigate the relationship of speed and ability. However, we formulate a particular hypothesis specifically for one of the possible parameterizations (H1a).

In the second step, we investigate measurement invariance of the response and speed parts of the B-GLIRT model with the best fitting parameterization of the cross-link function selected in the first step, regarding gender groups. In order to find the degree of measurement invariance, we compare model fit criteria between different restrictive models starting with an unconstraint multi-group model as a baseline model. Technically we follow Millsap and Yun-Tein [32] using a stepwise procedure to find the best fitting model by constraining (1) slope parameters, (2) intercept parameters, and (3) residual variances across groups. Doing so we include constraints on measurement parameters and relaxed constraints about the equality of latent variables variances and means. For this comparison we prefer AIC for the same reasons as in the first step, but we also report BIC to allow an independent evaluation of the model comparisons.

Mplus code to fit the described models can be found in the digital **Appendix**.

RESULTS

Descriptive summaries of the variability of response time and correctness of responses across items are provided in **Figure 3** and **Figure 4**. The (log-) response time (upper part) for the 28 responses and the proportion correct (lower part) for the 30 items are plotted in **Figure 3** for males and females and in **Figure 4** for PBA and CBA. The descriptive plots in **Figure 3** suggest systematic gender differences, especially for response times. For a multitude of items, systematic differences in the response times can also be supposed for the mode comparison presented in **Figure 4**.

Table 1 contains the information criteria for selecting the best-fitting parameterization for the cross-relation function of the B-GLIRT models. Consistent for both modes (PBA vs. CBA) and both options of aggregating item-level response times (strict vs. liberal), the cross-link function of the B-GLIRT interaction model that corresponds to Partchev et al. [26], including an interaction between ability and speed parameter, was the best fitting model according to the AIC (and BIC) criterion.

Table 1 also presents the estimated parameters (with standard errors in parenthesis) for all considered parameterizations of the cross-relation function. For a linear relationship between speed and ability (i.e., the parameter of the B-GLIRT hierarchical model), we found the expected negative relationship (confirming hypothesis H1a). Note that the negative relationship is found to be slightly stronger for CBA (e.g., −0.614 for the strict option for PBA compared to −0.766 for PBA). However, as the B-GLIRT interaction model provides the better model fit for PBA and CBA regarding AIC (and BIC, confirming H1b), all further analyses will be based on the parameterization of the cross-relation function for the B-GLIRT interaction model as shown in Equation (5).

The observed patterns of regression coefficients are identical for the two investigated options to aggregate response time measures to item-level time measures (strict vs. liberal). Accordingly, all subsequent models are based on the more liberal option 2 as slightly smaller standard errors are observed for this option with fewer missing time measures. It should be acknowledged that the estimated regression coefficients are small overall with only minor differences between modes (PBA vs. CBA): 0.074 (0.023) vs. 0.190 (0.021) for ρ_1 and −0.133 (0.020) vs. −0.192 (0.020) for ρ_2.

The results of the second analysis step, the investigation of measurement invariance of parameters estimated in the B-GLIRT interaction model for gender and mode, are reported in the next two subsections.

Gender

Table 2 presents the information criteria for the investigation of measurement invariance between gender groups for data gathered in the CBA mode using the B-GLIRT interaction model. Model comparisons with respect to the information criteria are conducted relative to a baseline model with 293 estimated parameters which were allowed to vary freely between groups (Model C1). The model comparisons are grouped into three sets of models: ability (successively constraining discrimination and difficulty parameters; Model C2-C7), speed (successively constraining time intensity and time discrimination parameters; Model C8-C16) and ability and speed (successively constraining all parameters; Model C17-C25). The AIC (and BIC) can be compared across the different model specifications and the model with the lowest information criterion will be selected and interpreted in this step of the analysis.

Constraining loadings for the response model α_i improved the model fit (Model C2). Relaxing the equality constraint of the latent variances for the ability factor θ_p (Model C3) decreased the fit slightly, indicating that there are no gender differences with respect to the ability variance. This pattern was also observed for the restriction of the intercept parameters β_i in Model C4, which fits slightly better than the model with relaxed equality constraint on the latent variance (Model C5). The model fitting best for the response part in terms of AIC (and BIC) is Model C6, in which, in addition to Model C3, group differences in the latent mean of the ability factor θ_p are estimated (but different from Model C7 equal ability variances are specified).

Table 2 also informs about possible differences that can be estimated using a particular set of constraints. Using Model C6, the best fitting model in the set of models constraining parameters of the ability part (Model C2-C7), we estimated the expected ability differences to be 0.426 (0.092) between boys and girls (confirming H2a).

Model C8 to Model C16 add parameter constraints of the speed part of the B-GLIRT interaction model, starting with

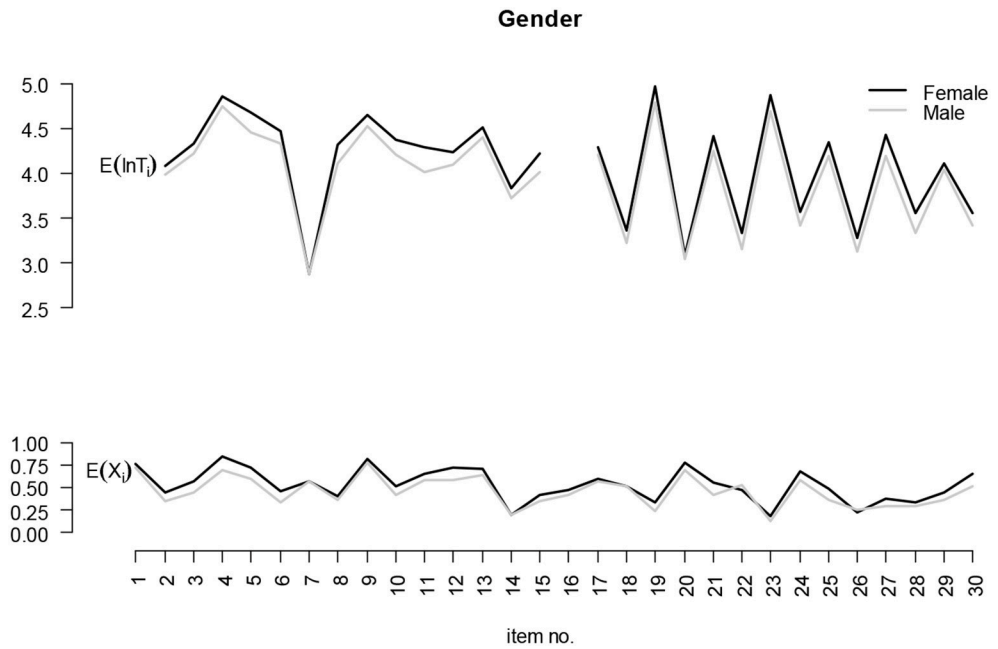


FIGURE 3 | Gender differences with respect to the log-response time and the proportion correct responses.

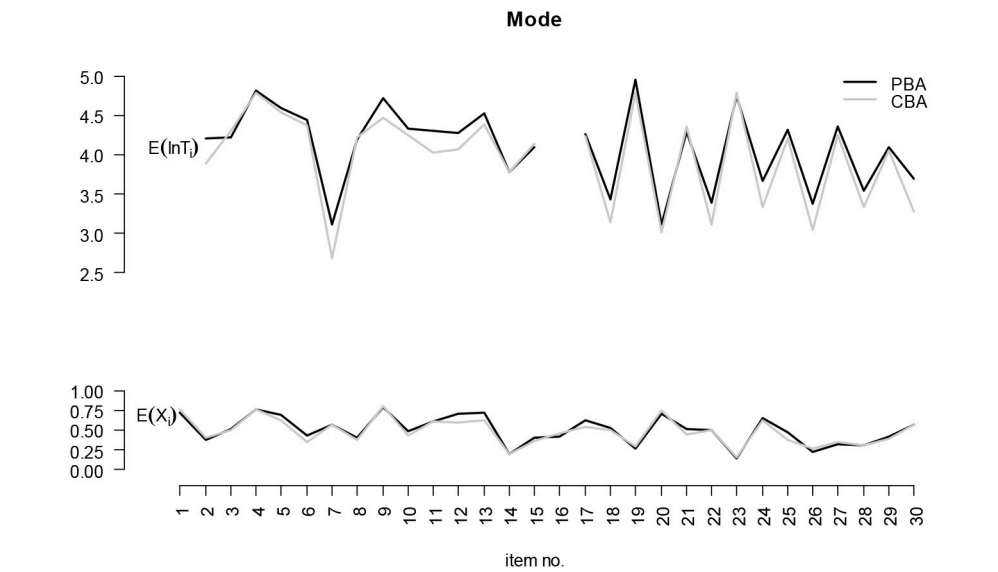


FIGURE 4 | Mode differences with respect to the log-response time and the proportion correct responses.

constraining the slope parameters φ_i, which resulted in a better model fit compared to the baseline Model C1. Similar to the ability part, we observed that allowing for group differences with respect to the variance of the latent speed factor τ_p consistently decreases model fit. AIC and BIC disagree with respect to the invariance of the intercept parameter λ_i for CBA. Constraining λ_i across groups resulted in an increased AIC (Model C8 vs. C10 and Model C9 vs. C11), while the BIC decreased. However, the best fitting model according to the AIC criterion in this set of models with constrained parameters of the speed part is model C12 (vs. Model C15 using BIC).

Corresponding to the observed ability differences, speed differences in CBA between boys and girls of 0.471/0.454 (0.134/0.130) were estimated using Model C12 and Model C15, respectively. However, interpreting these speed differences requires accepting invariance of time intensity parameters

TABLE 1 | Information criteria for the model comparison (different parameterizations of the cross-relation).

Mode	Time option	B-GLIRT cross-link function	Npar	AIC	BIC	Cross parameter(s)	
						ρ_1	ρ_2
PBA	Strict	Regression	145	20982.961	21632.344	−0.122 (0.016)	
PBA	Strict	Hierarchical	145	20938.633	21588.017	−0.614 (0.048)	
PBA	Strict	Interaction	146	**20911.934**	**21565.796**	0.092 (0.026)	−0.131 (0.030)
PBA	Liberal	Regression	145	24144.109	24793.493	−0.105 (0.017)	
PBA	Liberal	Hierarchical	145	24102.078	24751.462	−0.534 (0.058)	
PBA	Liberal	Interaction	146	**24042.836**	**24696.698**	0.074 (0.023)	−0.133 (0.020)
CBA	Strict	Regression	145	25485.367	26134.082	−0.197 (0.025)	
CBA	Strict	Hierarchical	145	25396.492	26045.207	−0.766 (0.030)	
CBA	Strict	Interaction	146	**25319.898**	**25973.086**	0.206 (0.022)	−0.204 (0.021)
CBA	Liberal	Regression	145	26513.802	27162.516	−0.188 (0.025)	
CBA	Liberal	Hierarchical	145	26447.617	27096.332	−0.741 (0.033)	
CBA	Liberal	Interaction	146	**26378.315**	**27031.503**	0.190 (0.021)	−0.192 (0.020)

For the AIC and BIC, best values are in bold. Identical patterns were observed for both time options (strict vs. liberal) and for all time thresholds (300, 500, 1500 s), not printed. Standard errors for cross parameters are in parenthesis. The column Npar contains the number of estimated parameters.

TABLE 2 | Measurement invariance of the B-GLIRT model with linear interaction term in the cross-relation function with respect to gender for data from computer-based assessment.

Model	Part	Constrained	Estimated differences	Npar	AIC	BIC	$\chi^2_{\Delta\rho}$
C1	Baseline	-	-	293	26244.851	27543.231	7.252
C2	Ability	α_i		263	26207.033	27372.473	7.879
C3		α_i	σ^2_θ	264	26208.507	27378.379	7.332
C4		α_i, β_i		233	26197.810	27230.310	8.344
C5		α_i, β_i	σ^2_θ	234	26199.670	27236.602	8.113
C6		α_i, β_i	μ_θ	234	*26178.163*	*27215.094*	8.130
C7		α_i, β_i	σ^2_θ, μ_θ	235	26179.916	27221.279	7.791
C8	Speed	φ_i		265	26231.190	27405.493	18.347
C9		φ_i	σ^2_τ	266	26232.850	27411.585	16.365
C10		φ_i, λ_i		237	26246.339	27296.565	3.357
C11		φ_i, λ_i	σ^2_τ	238	26248.039	27302.696	2.297
C12		φ_i, λ_i	μ_τ	238	*26228.993*	27283.650	2.002
C13		φ_i, λ_i	σ^2_τ, μ_τ	239	26230.999	27290.088	1.143
C14		φ_i, λ_i, $\sigma^2_{\omega i}$		209	26258.705	27184.853	3.900
C15		φ_i, λ_i, $\sigma^2_{\omega i}$	μ_τ	210	26242.394	*27172.973*	2.439
C16		φ_i, λ_i, $\sigma^2_{\omega i}$	σ^2_τ, μ_τ	211	26244.365	27179.376	1.154
C17	Ability & Speed	α_i, φ_i		235	26192.054	27233.416	19.606
C18		α_i, φ_i	σ^2_θ, σ^2_τ	237	26195.487	27245.712	17.006
C19		α_i, φ_i, β_i	μ_θ	206	26155.924	27068.778	15.533
C20		α_i, φ_i, β_i, λ_i		177	26179.820	26964.166	3.449
C21		α_i, φ_i, β_i, λ_i	μ_θ, μ_τ	179	**26144.712**	26937.921	1.188
C22		α_i, φ_i, β_i, λ_i	σ^2_τ, μ_τ	179	26182.876	26976.084	2.738
C23		α_i, φ_i, β_i, λ_i	σ^2_θ, σ^2_τ, μ_θ, μ_τ	181	26147.798	26949.869	0.972
C24		α_i, φ_i, β_i, λ_i, $\sigma^2_{\omega i}$	μ_θ, μ_τ	151	26158.674	**26827.805**	1.384
C25		α_i, φ_i, β_i, λ_i, $\sigma^2_{\omega i}$	σ^2_θ, σ^2_τ, μ_θ, μ_τ	153	26162.001	26839.994	0.998

Column $\chi^2_{\Delta\rho}$ contains the value of the test statistic for the hypothesis that the estimated parameters of the cross-relation function are equal between groups (df = 2). For the AIC and BIC best values are in bold, and italic values indicate best values within parts of the model. The column Npar contains the number of estimated parameters.

between groups, which is not supported based on the AIC criterion for the speed part of the B-GLIRT interaction model. Taking BIC as the criterion for simultaneously investigating speed and ability provides weak reasoning for assuming measurement invariance concerning the time intensity parameters λ_i (Model C20 or Models C15/C24). Moreover,

the BIC also shows improved model fit for constraining the residual variances $\sigma^2_{\omega i}$ across groups (Model C24).

The last column of **Table 2** shows the value of the χ^2 statistic provided as Wald-test by Mplus, specified to test the hypothesis that the parameters ρ_1 and ρ_2 of the cross-relation function of the B-GLIRT interaction model are equal between groups ($\chi^2_{\Delta\rho}$, with $df = 2$). The estimated parameters ρ_1 and ρ_2 involved in this Wald-test (see **Table 3**, upper part) from Model C12 and Model C15 did not differ statistically significant between boys and girls, as soon as time intensity parameters were assumed to be invariant. Note that the impact of the time discrimination parameter λ_i is a *post-hoc* explanation motivated by the observation that the smallest $\chi^2_{\Delta\rho}$ value of models with unconstrained time discrimination parameters (C2-C9; C17-C19) is 7.879, while the largest value of all models with constrained discrimination parameters (C10-C16; C20-C25) is 3.900. However, invariance is not supported by the AIC criterion, which was preferred by theoretical considerations. Accordingly, we only partially confirmed hypothesis H3a.

We now describe the results for investigating measurement invariance of the B-GLIRT interaction model between gender groups for data gathered in the PBA mode.

Table 4 replicates the measurement invariance investigation for the data gathered in the paper-based test administration. Concerning the ability model, AIC and BIC favor the models with constrained slope and intercept parameters. The latent variance of the ability factor in the female group is estimated 0.758 (0.126) in Model P7 (the model with the lowest AIC value), but BIC slightly favors Model P6 which assumes equal latent ability variances between groups. However, we found full measurement invariance for the ability model in the PBA administration according to the AIC when the variance and the mean of the latent ability variable are freely estimated in the female group.

Concerning the speed part of the B-GLIRT interaction model for the PBA data we observed a small improvement in model fit when constraining the slope parameters φ_i across groups (see Model P8 vs. Model P1 in **Table 4**) and a small additional improvement for constraining the time discrimination parameters λ_i (see Model P10 vs. Model P8). The best fitting model of the speed component for PBA in terms of AIC was achieved by acknowledging latent mean differences in the speed factor between gender groups, estimated as 0.426 (0.133) for PBA using Model P12. Constraining the residual variances $\sigma^2_{\omega i}$ across groups increased the AIC (i.e., measurement invariance with respect to the residual variances of the speed factor could not be achieved using the AIC as the criterion). Similar to the CBA data, taking BIC as the criterion would allow establishing measurement invariance with respect to the residuals (see Model P22 in **Table 4**)[4].

Combining ability and speed for the PBA data resulted in the best fitting model regarding AIC when α_i, φ_i, β_i, and λ_i are constrained across groups (Model P21/Model P23 in **Table 4**). However, measurement invariance with respect to the residual variances $\sigma^2_{\omega i}$ was neither achieved using AIC nor BIC. Speed and ability differences estimated using Model P21 were 0.316 (0.099) and 0.405 (0.161) demonstrating that girls have a higher ability (confirming H2b) while taking more time to respond in PBA. However, as the last column in **Table 4** reveals, the estimated parameters of the cross-relation function ρ_1 and ρ_2 did not differ statistically significantly between boys and girls (see **Table 3**, lower part, for the estimated values of ρ_1 and ρ_2), confirming H3b. Moreover, **Table 3** contains the descriptive result that the coefficients ρ_1 for the linear main effect of θ_p and ρ_2 for the interaction $\tau_p\theta_p$ are smaller for PBA compared to CBA, estimated in separate B-GLIRT interaction models for both modes.

TABLE 3 | Estimated parameters of the cross-relation function for the B-GLIRT interaction model.

Mode	Model	Group	ρ_1		ρ_2	
CBA	C12	Male	0.234	(0.028)	−0.189	(0.024)
		Female	0.208	(0.033)	−0.183	(0.034)
	C15	Male	0.242	(0.028)	−0.200	(0.027)
		Female	0.190	(0.030)	−0.158	(0.025)
	C21	Male	0.226	(0.026)	−0.187	(0.023)
		Female	0.200	(0.029)	−0.180	(0.028)
PBA	P12	Male	0.118	(0.030)	−0.157	(0.034)
		Female	0.063	(0.031)	−0.078	(0.044)
	P15	Male	0.117	(0.031)	−0.158	(0.034)
		Female	0.070	(0.035)	−0.083	(0.042)
	P21	Male	0.108	(0.030)	−0.142	(0.034)
		Female	0.073	(0.028)	−0.102	(0.051)

Mode

In the following, we report the results for investigating measurement invariance regarding mode. For these analyses we examined the data for boys and girls together, without grouping by gender. In order to test the equality of the cross-relation parameters of the B-GLIRT interaction model between modes, we considered a combined model for CBA and PBA simultaneously (see **Figure 5** for a schematic path diagram). Due to the within-subject component of the experimental design, a single group structural equation model with indicators for PBA and CBA items can be specified.

Measurement invariance as investigated for the comparison of groups regarding the person-level variable "gender" is not strictly necessary when modeling mode effects since the mode effect can be identified making use of the randomized assignment of test-taker to CBA vs. PBA. To model the mode effect as a latent difference variable, we re-parameterized the model as a *method effect model* (e.g., [33]) within the B-GLIRT framework (see **Figure 6**)[5]. Doing so, we assumed measurement invariance concerning the ability part of the model by constraining α_i and β_i across modes. This allows identifying a latent difference variable θ_{Mode} as the difference between $\theta_{p;PBA}$ and $\theta_{p;CBA}$ (included with equal loadings for each item administered in CBA mode, see

[4]Measurement invariance with respect to the residual variances $\sigma^2_{\omega i}$ of the speed model is not required for a valid interpretation of gender differences.

[5]The symmetrical reformulation of the speed model was avoided taking into account that, although first evidence was provided regarding the comparability of time measures, the time indicators $T_{i;PBA}$ and $T_{i;CBA}$ were created using completely different log data.

TABLE 4 | Measurement invariance of the B-GLIRT model with linear interaction term in the cross-relation function with respect to gender for data from paper-based assessment.

Model	Part	Constrained	Estimated Differences	Npar	AIC	BIC	$\chi^2_{\Delta\rho}$
P1	Baseline	-	-	293	24040.658	25340.450	2.896
P2	Ability	α_i		263	24026.589	25193.297	3.036
P3		α_i	σ^2_θ	264	24027.350	25198.494	2.948
P4		α_i, β_i		233	24024.846	25058.469	3.013
P5		α_i, β_i	σ^2_θ	234	24023.531	25061.590	2.853
P6		α_i, β_i	μ_θ	234	24014.357	*25052.416*	2.957
P7		α_i, β_i	σ^2_θ, μ_θ	235	*24013.158*	25055.653	2.870
P8	Speed	φ_i		265	24038.974	25214.554	7.613
P9		φ_i	σ^2_τ	266	24039.335	25219.351	6.160
P10		φ_i, λ_i		237	24021.565	25072.933	1.989
P11		φ_i, λ_i	σ^2_τ	238	24022.290	25078.094	1.483
P12		φ_i, λ_i	μ_τ	238	*24008.566*	25064.369	2.181
P13		φ_i, λ_i	σ^2_τ, μ_τ	239	24008.621	25068.861	1.162
P14		φ_i, λ_i, $\sigma^2_{\omega i}$		209	24032.007	24959.162	2.112
P15		φ_i, λ_i, $\sigma^2_{\omega i}$	μ_τ	210	24019.856	*24951.448*	1.998
P16		φ_i, λ_i, $\sigma^2_{\omega i}$	σ^2_τ, μ_τ	211	24016.573	24952.601	0.764
P17	Ability & Speed	α_i, φ_i		235	24023.956	25066.452	7.915
P18		α_i, φ_i	σ^2_θ, σ^2_τ	237	24024.405	25075.773	5.554
P19		α_i, φ_i, β_i	μ_θ	206	24005.692	24919.539	6.913
P20		α_i, φ_i, β_i, λ_i		177	23995.572	24780.771	2.228
P21		α_i, φ_i, β_i, λ_i	μ_θ, μ_τ	179	**23974.340**	24768.411	1.336
P22		α_i, φ_i, β_i, λ_i	σ^2_τ, μ_τ	179	23996.141	**23996.141**	1.169
P23		α_i, φ_i, β_i, λ_i	σ^2_θ, σ^2_τ, μ_θ, μ_τ	181	23974.766	24777.709	0.681
P24		α_i, φ_i, β_i, λ_i, $\sigma^2_{\omega i}$	μ_θ, μ_τ	151	23985.216	24655.075	0.990
P25		α_i, φ_i, β_i, λ_i, $\sigma^2_{\omega i}$	σ^2_θ, σ^2_τ, μ_θ, μ_τ	153	23983.940	24662.671	0.346

Column $\chi^2_{\Delta\rho}$ contains the value of the test statistic for the hypothesis that the estimated parameters of the cross-relation function are equal between groups (df = 2). For the AIC and BIC, best values are in bold, and italic values indicate best values within parts of the model. The column Npar contains the number of estimated parameters.

Figure 6) and estimating the latent correlation between θ_p and θ_{Mode}. The ability difference between modes is obtained as the mean of this latent difference variable θ_{Mode}.

Starting with an unconstrained baseline model (Model M1) for the combined estimation of B-GLIRT interaction models for CBA and PBA data, **Table 5** contains information criteria for the sequence of constrained models that allow to evaluate measurement invariance. As the comparison AIC reveals (see the ability part in **Table 5**) a model with equal item difficulty parameters and equal item discrimination parameters (Model M3) fits the data almost equally well-compared to mode-specific item difficulties (Model M2) or mode-specific item parameters (Model M1).

The lower part of **Table 5** contains results from the combined estimation of B-GLIRT interaction models with constrained parameters of the measurement models for the latent speed factors ($\tau_{p;PBA}$ and $\tau_{p;CBA}$) and the latent ability factors ($\theta_{p;PBA}$ and $\theta_{p;CBA}$), which are re-parameterized as the latent ability factor (θ_p) and the latent difference variable (θ_{mode}). To ensure the necessary comparability concerning the measurement model of the speed factor for comparing ρ_1 and ρ_2 between CBA and PBA, we estimated a sequence of models with different equality constraints for φ_i and λ_i between modes (Model M4 to Model M8 in **Table 5**). While constraining φ_i (Model M4) led to an improved model fit in terms of AIC (and BIC), the fit declined, when time intensity parameters λ_i were constrained (Model M6 vs. Model M4), as indicated by an increased value of the AIC (and the BIC).

As **Figure 7** shows, time intensity parameters estimated using Model M4 under the specification of zero and equal latent means for the speed factors $\tau_{p,CBA}$ and $\tau_{p,PBA}$ are slightly higher for PBA for most items (only 6 out of 28 time intensity parameters are smaller for PBA). Similar to the model specification used for the mode effect in the ability part of the B-GLIRT interaction model we simplified the mode effect for the time factor by constraining the differences in the time intensity parameter to a single parameter. Using an explicit identification based on an equality constraint of λ_i between modes, the parameter $\mu_\tau^{(PBA)}$ (i.e., the mean of the latent speed factor $\tau_{p,PBA}$) was estimated in Model M6 as 0.561 (0.097), $p < 0.05$. A similar effect of 0.508 (0.089), $p < 0.05$, was estimated under the best fitting model in terms of BIC (Model M7). The mean $\mu_\tau^{(PBA)}$ of the speed factor $\tau_{p,PBA}$ corresponds to the mode effect with respect to speed, because the mean of the latent speed factor was constrained to zero for the CBA speed factor $\tau_{p,CBA}$. Allowing the mean of the latent speed factor to vary between modes worsened the model

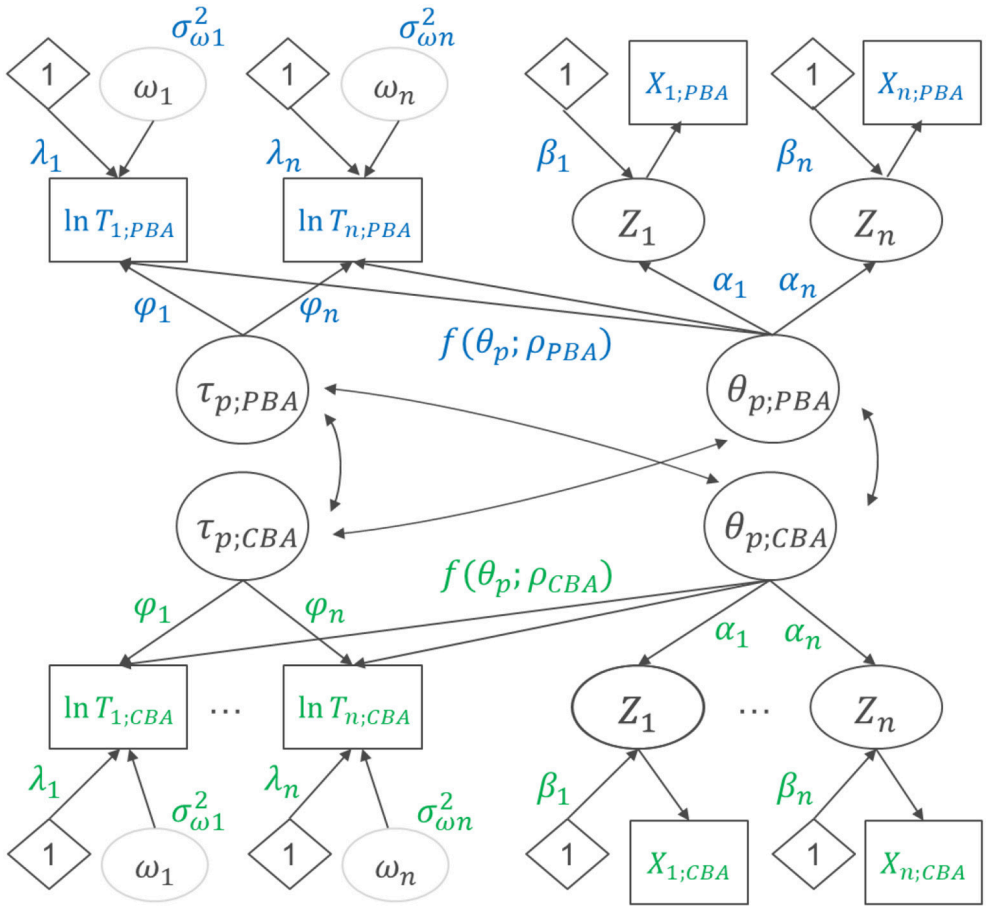


FIGURE 5 | B-GLIRT model for responses and response times gathered in a within-subject design with planned missing values in paper-based and computer-based assessment.

fit in terms of AIC (and BIC), but strongly affected the estimated parameters in the cross-relation function. This is illustrated in the last column of **Table 5**. The χ^2 statistics, provided as a Wald-test by Mplus for tests of the hypothesis that the parameters ρ_1 and ρ_2 of the cross-relation function differ between modes ($\chi^2_{\Delta\rho}$), are highly sensitive to this part of the model (mis-) specification.

Finally, **Table 6** shows the estimated parameters of the cross-relation function and the estimated mode effect parameters concerning speed and ability. In both models M6 and M8 the combined Wald-test for the constraint that the parameters ρ_1 and ρ_2 differ between CBA and PBA resulted in a χ^2-value that indicates with $df = 2$ statistical significance. Inspecting the estimated values revealed that the difference is mainly due to the different interaction effect ρ_2 between modes, rejecting H4. Moreover, we found that the latent speed factor $\tau_{p,PBA}$ has a variance different from 1 and a mean different from 0 (i.e., we found speed differences between modes). Test-takers tend to work slower in PBA and the speed factor has a smaller variance in PBA (i.e., test-takers are more similar with respect to speed in PBA compared to CBA). This completes the picture that is described with an ability mode effect: the latent mode effect variable θ_{mode} with an estimated mean $\mu_{\theta,mode}$ different from 0 indicates an overall shift in the estimated ability (lower ability estimated from CBA). Inter-individual differences in the mode effect regarding the ability are modeled as latent difference variable. This latent difference variable is negatively correlated with reading ability, indicating a higher ability mode effect for test-takers with low reading abilities. Finally, we consistently found a moderate correlation between the latent speed factors for PBA and CBA across the different models. This moderate correlation indicates that responding in a particular speed to PBA administered reading units is only moderately related to the speed chosen to respond to CBA administered units in the self-paced assessment.

DISCUSSION

In this study, we applied psychometric modeling of response times gathered from CBA and PBA with digital pens for two selected clusters of PISA 2009 print reading assessment. For that purpose, we developed a method to derive comparable response times at item-level from log data that can be obtained from CBA and PBA (using digital pens). Subsequently, we used the time measures as indicators for the response speed in latent variable

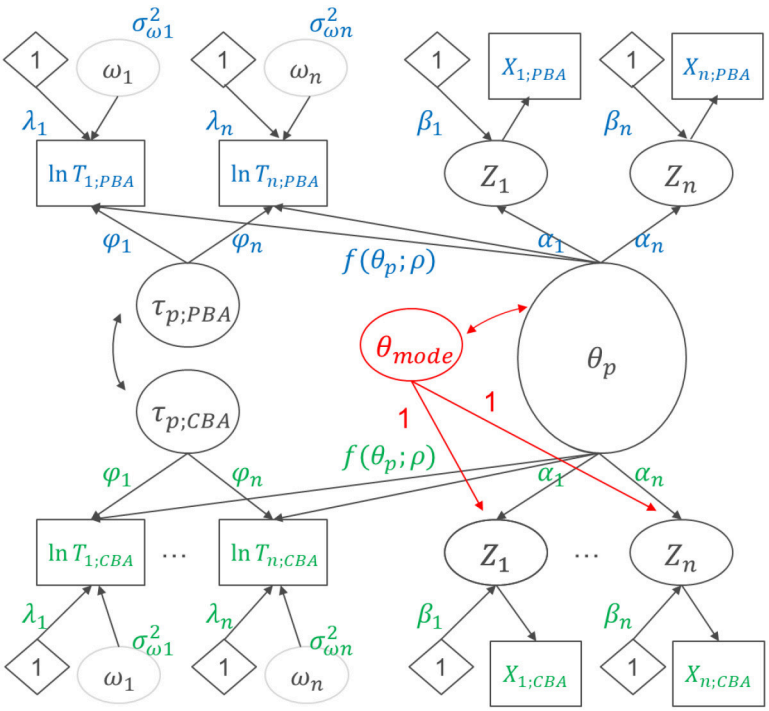


FIGURE 6 | Re-parameterized B-GLIRT model with a latent difference variable for inter-individual differences in the mode-effect and correlated, mode-specific speed factor.

TABLE 5 | Measurement invariance of the B-GLIRT model with linear interaction term in the cross-relation function between computer-based and paper-based assessment.

Model	Part	Constrained	Estimated differences	Npar	AIC	BIC	$\chi^2_{\Delta\rho}$
M1	Baseline	-	-	295	50207.795	51619.876	22.798
M2	Ability	α_i	$\sigma^2_{\theta mode}$	265	50197.487	51465.967	26.496
M3		α_i, β_i	$\mu_{\theta mode}$, $\sigma^2_{\theta mode}$	236	50198.058	51327.723	25.940
M4	Ability & Speed	α_i, β_i, φ_i	$\mu_{\theta mode}$, $\sigma^2_{\theta mode}$	208	**50196.661**	51192.299	43.474
M5		α_i, β_i, φ_i, λ_i	$\mu_{\theta mode}$, $\sigma^2_{\theta mode}$	180	50385.100	51246.709	12.886
M6		α_i, β_i, φ_i, λ_i	$\mu_{\theta mode}$, $\sigma^2_{\theta mode}$, $\mu_\tau^{(PBA)}$, $\sigma_\tau^{2\ (PBA)}$	182	50320.474	51191.656	7.966
M7		α_i, β_i, φ_i, λ_i, $\sigma^2_{\omega i}$	$\mu_{\theta mode}$, $\sigma^2_{\theta mode}$	153	50500.148	51232.516	12.171
M8		α_i, β_i, φ_i, λ_i, $\sigma^2_{\omega i}$	$\mu_{\theta mode}$, $\sigma^2_{\theta mode}$, $\mu_\tau^{(PBA)}$, $\sigma_\tau^{2\ (PBA)}$	155	50432.544	**51174.485**	7.294

Column $\chi^2_{\Delta\rho}$ contains the value of the test statistic for the hypothesis that the estimated parameters of the cross-relation function are equal between modes (df = 2). For the AIC and BIC, best values are in bold. The column Npar contains the number of estimated parameters.

models and investigated differences in relationship of speed and ability across gender and mode. The analyses were grounded in the underlying idea that differences in the relationship between speed and ability could provide hints for actual differences in the true response process.

We found an overall negative correlation between speed and ability. The best relative fit was observed for a B-GLIRT interaction model, meaning a B-GLIRT model with a cross-relation function that included an interaction term between speed and ability in the regression of the (log-) response time on speed and ability. The regression coefficient for this interaction was estimated negatively for CBA and PBA. Using this specification of the cross-relation function in a multi-group analysis we found measurement invariance of the B-GLIRT interaction model between boys and girls concerning the slope and the intercept parameters of the speed and the ability part of the measurement model, but not concerning residual variances of the speed part of the measurement model. In line with previous research, we found ability differences between boys and girls. We also found that the gender differences in the reading ability coincide with speed differences between boys and girls for both modes (girls responded slower while tending to obtain higher scores than boys). Due to the nature of the study (only the mode was randomly assigned while the speed was self-selected) we can only assess that gender differences exist concerning speed and ability for PBA as well as for CBA. Using the available data, the specified

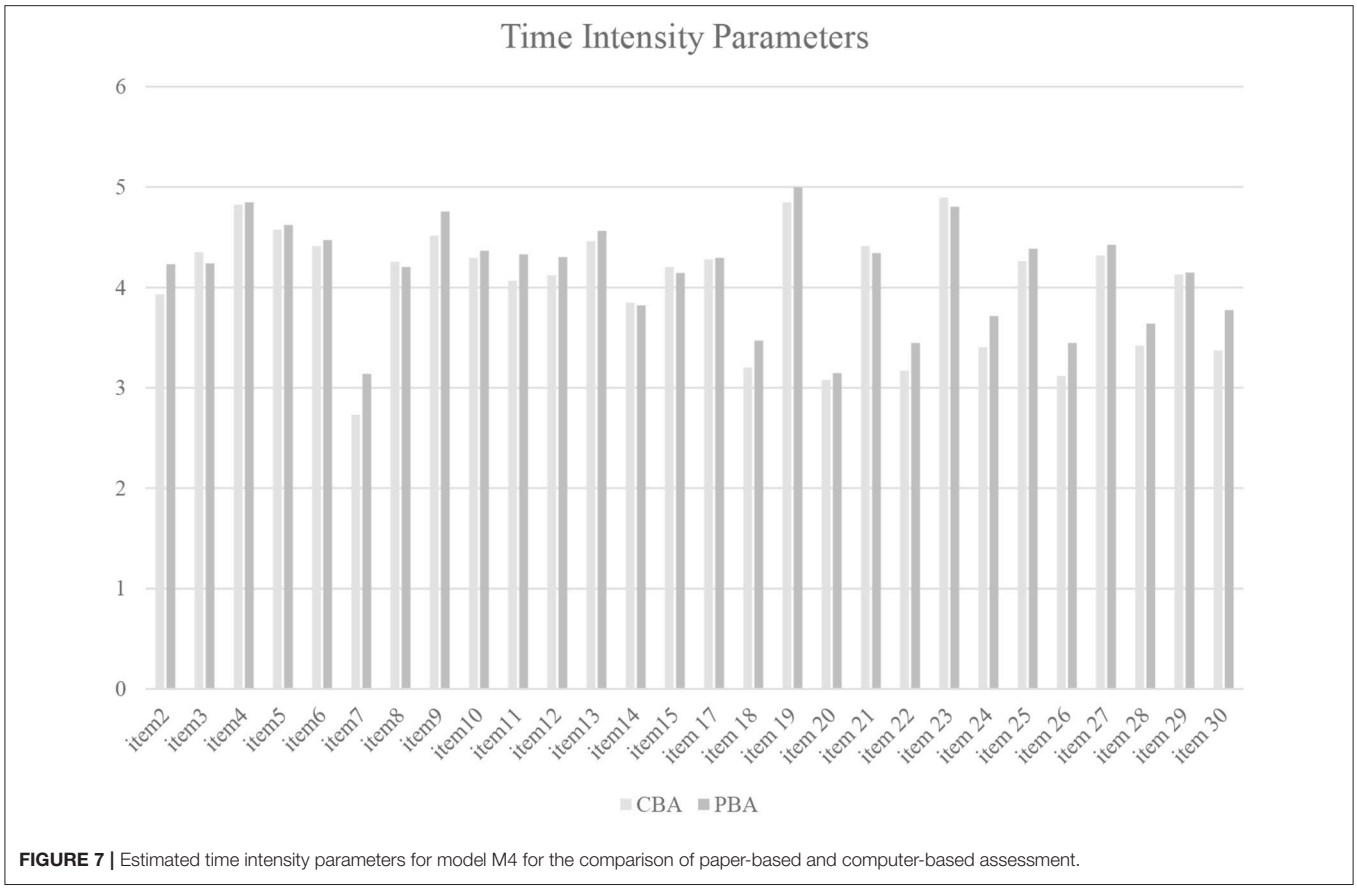


FIGURE 7 | Estimated time intensity parameters for model M4 for the comparison of paper-based and computer-based assessment.

TABLE 6 | Estimates of selected model parameters of the combined B-GLIRT interaction model for CBA and PBA data (see **Figure 6**).

Model	CBA		PBA		Mode effect (Speed)		Mode effect (Ability)		Correlations	
	ρ_1	ρ_2	ρ_1	ρ_2	$\mu_{\tau;\text{PBA}}$	$\sigma^2_{\tau;\text{PBA}}$	$\mu_{\theta;\text{mode}}$	$\sigma^2_{\theta;\text{mode}}$	$\text{Cor}(\tau_{\text{CBA}}, \tau_{\text{PBA}})$	$\text{Cor}(\theta_{\text{mode}}, \theta)$
M4	0.152(0.017)	−0.167(0.016)	0.036(0.017)	−0.146(0.027)	–	–	−0.105(0.047)	0.143(0.051)	0.422(0.066)	−0.253(0.139) n.s.
M6	0.166(0.026)	−0.172 (0.022)	0.172 (0.032)	−0.237 (0.025)	0.503(0.089)	0.586 (0.080)	−0.106 (0.047)	0.238 (0.069)	0.407(0.071)	−0.358 (0.120)
M8	0.157 (0.019)	−0.163 (0.016)	0.157 (0.038)	−0.215 (0.041)	0.484 (0.080)	0.622 (0.115)	−0.113 (0.047)	0.165 (0.055)	0.399(0.070)	−0.262 (0.117)

Estimated mean of the mode effect variable $\mu_{\theta;mode}$ is unstandardized. Mean ($\mu_{\tau;PBA}$) and variance ($\sigma^2_{\tau;CBA}$) of the mode effect (speed) are parameterized relative to the mean and variance of the latent speed variable, that was fixed to mean zero ($\mu_{\tau;CBA} = 0$) and variance ($\sigma^2_{\tau;CBA} = 1$). All estimated parameters not marked with "n.s." are statistically significant different from zero ($p < 0.05$).

Wald-tests for the comparison of the parameters in the cross-relation function between boys and girls showed no statistically significant difference in these parameters. Hence, our results give no reason to assume different response processes (as captured by the cross-relation function of the B-GLIRT model) for boys and girls so far.

Analyzing data for CBA and PBA together allowed investigating mode effects with respect to speed (in addition to the typical investigated comparability of ability estimates or item parameters). Resting on the invariance assumptions of time discrimination parameters estimated from PBA and CBA data, which was only supported by BIC, we found mean differences in speed in addition to mean differences in ability. Test-takers tend to answer items in PBA slower and more correctly in comparison to CBA. Moreover, we found hints that the relationship between speed and ability, modeled using the cross-relation function of the B-GLIRT model with an interaction term, differs between modes. However, the estimated regression coefficients for the interaction term are of small magnitude.

Our results reason further investigations of mode and gender differences in the response process and, in particular, a focus on the role of the self-selected speed component when changing the administration mode and when investigating gender differences in reading ability. However, as already investigated with this dataset Kroehne et al. (submitted) the mode effects found in the ability model give no reason to reject the hypothesis that CBA and PBA are construct-equivalent. Specifically, the latent correlation of PBA and CBA was not found to differ statistically significantly from the expected correlation of two test parts measuring the identical construct within one mode (for instance, PBA).

Limitations and Further Research

Up to the authors' knowledge, the current study is the first investigation of mode effects for self-paced assessments conducted under large-scale assessment conditions that includes a comparison of PISA reading assessments concerning response times. As the method to derive comparable item-level time measures for PBA and CBA was not previously available, the current study has some major limitations. For instance, the time measure for the first question of each cluster could not be derived in PBA, because no timestamp was available that indicated the start of the assessment. Future use of digital pens as an assessment device to collect log data from PBA might modify the assessment instrument by requesting the test-taker to draw a cross right before starting to read the reading text for the first unit in a booklet.

The conducted comparison of different parameterizations of the cross-link function within the framework of B-GLIRT models and the investigation of measurement invariance in parameters of the best fitting model is limited as both are based on the assumptions that the items fit the ability part and that the unconstrained combined model shows acceptable absolute model fit. An additional limitation regarding the applied modeling is that we restricted the complexity by investigating gender effects and mode effects separately in this study. Current literature (e.g., [34]) give reasons to assume an interplay between the mode and gender effect. Further research might focus on a multi-group model by gender with both modes. Due to the small sample size, estimation problems and the required estimation time, this has not been investigated yet. Moreover, as the latent interaction model is demanding on a computational level and the sample size is rather small, a replication of the finding that a cross-relation function including the interaction between speed and ability using different data would be beneficial.

Regarding the interpretation of the study results, a technical limitation is that we have not computed any measures of effect size and practical significance. Moreover, the relationship of the speed differences with other covariates such as an ICT-related measure of basic computer skills is subject to additional investigations. Also, since we did not control for the answering time as the time to give the response (writing vs. typing), the observed response time differences between modes might be caused by the mechanical process of "answering."

Conceptually, the role of speed as a potential mediating variable of mode effects is open for further research, for instance, by analyzing mode and gender differences within subgroups of test-takers with comparable (self-selected) speed. In the current form, the study is limited to the descriptive finding that ability differences coexist together with speed differences.

The validity of the reported results regarding the response time models rest on the operationalization of item-level response times applied in this paper. The chosen approach was mainly driven by the desire to create time measures that give insights into the comparison of the response process between modes. However, only limited validity evidence exists regarding the interpretation of the time measures derived from paper-based assessments using digital pens. More detailed analyses are needed, for instance, comparing item-level response time measures derived using different operationalization using all available log data for CBA.

ETHICS STATEMENT

The data collection was carried out with permission of the Centre for International Student Assessment (ZIB) as national add-on study to PISA 2012. As per OECD guidelines and German national regulations (KMK) no new ethics approval was required.

AUTHOR CONTRIBUTIONS

The paper uses data gathered in a national extension study of the Centre for International Student Assessment (ZIB) to PISA 2012. CH computerized the reading units using the CBA ItemBuilder and coordinated the data collection and the preparation of raw data. The authors jointly developed the study design. UK was responsible for the integration of the CBA instrument and coordinated data collection using digital pens for the paper-based assessment. UK developed the strategy for analyzing log data from digital pens and conducted the statistical analyses. UK wrote the first draft of the manuscript and FG and CH reviewed the draft and contributed to the final manuscript.

FUNDING

This work was funded by the Open Access Fund of the Leibniz Association.

ACKNOWLEDGMENTS

This study was previously presented at the annual meeting of the National Council on Measurement in Education (NCME) 2018 in New York. The authors would like to thank Katie McClarty, the discussant of the session Diving into Data with Response Process Research as well as Cordula Artelt from the Leibniz Institute for Educational Progress (LIfBi) for their helpful feedback.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2019.00002/full#supplementary-material

REFERENCES

1. Molenaar D, Tuerlinckx F, van der Maas HLJ. A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivar. Behav. Res.* (2015) **50**:56–74. doi: 10.1080/00273171.2014.962684
2. Goldhammer F, Kroehne U. Controlling individuals' time spent on task in speeded performance measures: experimental time limits, posterior time limits, and response time Modeling. *Appl. Psychol. Measure.* (2014a) **38**:255–67. doi: 10.1177/0146621613517164
3. Goldhammer F. Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental

control. *Measure. Interdiscipl. Res. Perspect.* (2015) **13**:133–64. doi: 10.1080/15366367.2015.1100020
4. Molenaar D. The value of response times in item response modeling. *Measure. Interdiscipl. Res. Perspect.* (2015) **13**:177–81. doi: 10.1080/15366367.2015.1105073
5. Li Z, Banerjee J, Zumbo BD. Response time data as validity evidence: has it lived up to its promise and, if not, what would it take to do so. In: Zumbo BD, Hubley AM, editor. *Understanding and Investigating Response Processes in Validation Research*. Cham: Springer International Publishing (2017). p. 159–177.
6. Kyllonen P, Zu J. Use of response time for measuring cognitive ability. *J. Intellig.* (2016) **4**:14. doi: 10.3390/jintelligence4040014
7. Goldhammer F, Klein Entink RH. Speed of reasoning and its relation to reasoning ability. *Intelligence* (2011) **39**:108–19. doi: 10.1016/j.intell.2011.02.001<underline
8. Kroehne U, Martens T. Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft* (2011) **14**:169–86. doi: 10.1007/s11618-011-0185-4
9. Schroeders U, Wilhelm O. Equivalence of reading and listening comprehension across test media. *Edu. Psychol. Measure.* (2011) **71**:849–69. doi: 10.1177/0013164410391468
10. Grover RK, Ercikan K. For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Appl. Measure. Edu.* (2017) **30**:178–95. doi: 10.1080/08957347.2017.1316276
11. Marks GN. Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxford Rev. Edu.* (2008) **34**:89–109. doi: 10.1080/03054980701565279
12. Lynn R, Mikk J. Sex differences in reading achievement. *Trames J. Humanit. Soc. Sci.* (2009) **13**:3. doi: 10.3176/tr.2009.1.01
13. Dirk J, Kratzsch GK, Prindle JP, Kroehne U, Goldhammer F, Schmiedek F. Paper-based assessment of the effects of aging on response time: a diffusion model analysis. *J. Intellig.* (2017) **5**:12. doi: 10.3390/jintelligence5020012
14. Kong Y, Seo YS, Zhai L. Comparison of reading performance on screen and on paper: a meta-analysis. *Comp. Edu.* (2018) **123**:138–49. doi: 10.1016/j.compedu.2018.05.005
15. Mangen A, Walgermo BR, Brønnick K. Reading linear texts on paper versus computer screen: Effects on reading comprehension. *Int. J. Edu. Res.* (2013) **58**:61–8. doi: 10.1016/j.ijer.2012.12.002
16. Schwabe F, McElvany N, Trendtel M. The school age gender gap in reading achievement: examining the influences of item format and intrinsic reading motivation. *Read. Res. Q.* (2015) **50**:219–32. doi: 10.1002/rrq.92
17. Bouchamma Y, Poulin V, Ruel C. Impact of reading strategy use on girls' and boys' achievement. *Read. Psychol.* (2014) **35**:312–31. doi: 10.1080/02702711.2012.724043
18. OECD. *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing (2016). Available online at: http://dx.doi.org/10.1787/9789264266490-en
19. Jerrim J, Micklewright J, Heine JH, Salzer C, McKeown C. *PISA 2015: How Big Is the 'Mode Effect' and What Has Been Done About It?* Oxford Review of Education (2018). doi: 10.1080/03054985.2018.1430025
20. Robitzsch A, Luedtke O, Koeller O, Kroehne U, Goldhammer F, Heine, JH. Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien [Challenges in estimations of trends in large-scale assessments: a calibration of the German PISA data]. *Diagnostica* (2016) **63**:148–65. doi: 10.1026/0012-1924/a000177
21. Drasgow F. (ed.). *Technology and Testing: Improving Educational and Psychological Measurement*. New York, NY: Routledge (2016).
22. Hecht M, Siegle T, Weirich S. A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments. *J. Edu. Res. Online* (2017) **9**:32.
23. van der Linden WJ. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* (2007) **72**:287–308. doi: 10.1007/s11336-006-1478-z
24. Fox J-P, Marianti S. Joint modeling of ability and differential speed using responses and response times. *Multivar. Behav. Res.* (2016) **51**:1–14. doi: 10.1080/00273171.2016.1171128
25. Thissen D. Timed testing: an approach using item response testing. In: Weiss DJ, editor. *New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing*. New York, NY: Academic Press (1983). p. 179–203.
26. Partchev I, De Boeck P. Can fast and slow intelligence be differentiated? *Intelligence* (2012) **40**:23–32. doi: 10.1016/j.intell.2011.11.002
27. Kroehne U, Goldhammer F. How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika* (2018) **45**: 527–63. doi: 10.1007/s41237-018-0063-y
28. Goldhammer F, Naumann J, Stelter A, Tóth K, Rölke H, Klieme E. The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *J. Edu. Psychol.* (2014b) **106**:608–26. doi: 10.1037/a0034716
29. Roelke H. The ItemBuilder: a graphical authoring system for complex item development. *Presented at the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (2012). p. 344–53.
30. Muthén LK, Muthén BO. *Mplus Version 7.4 [Computer software]*. Los Angeles, CA: Muthén & Muthén (2015).
31. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol. Methods* (2012) **17**:228–43. doi: 10.1037/a0027127
32. Millsap RE, Yun-Tein J. Assessing factorial invariance in ordered-categorical measures. *Multivar. Behav. Res.* (2004) **39**:479–515. doi: 10.1207/S15327906MBR3903_4
33. Pohl S, Steyer R. Modeling common traits and method effects in multitrait-multimethod analysis. *Multivar. Behav. Res.* (2010) **45**:45–72. doi: 10.1080/00273170903504729
34. Zehner F, Goldhammer F, Lubaway E, Sälzer C. Unattended consequences: how text responses alter alongside the mode change from PISA 2012 to 2015. *Edu. Inquiry* (2018) 1–22. doi: 10.1080/20004508.2018.1518080

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Kroehne, Hahnel and Goldhammer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

SYSTEMATIC REVIEW
published: 06 February 2019
doi: 10.3389/fpsyg.2019.00102

An Overview of Models for Response Times and Processes in Cognitive Tests

Paul De Boeck[1,2]* *and Minjeong Jeon*[3]

[1] Department of Psychology, Ohio State University, Columbus, OH, United States, [2] KU Leuven, Leuven, Belgium, [3] Graduate School of Education and Information Studies, University of California, Los Angeles, Los Angeles, CA, United States

Response times (RTs) are a natural kind of data to investigate cognitive processes underlying cognitive test performance. We give an overview of modeling approaches and of findings obtained with these approaches. Four types of models are discussed: response time models (RT as the sole dependent variable), joint models (RT together with other variables as dependent variable), local dependency models (with remaining dependencies between RT and accuracy), and response time as covariate models (RT as independent variable). The evidence from these approaches is often not very informative about the specific kind of processes (other than problem solving, information accumulation, and rapid guessing), but the findings do suggest dual processing: automated processing (e.g., knowledge retrieval) vs. controlled processing (e.g., sequential reasoning steps), and alternative explanations for the same results exist. While it seems well-possible to differentiate rapid guessing from normal problem solving (which can be based on automated or controlled processing), further decompositions of response times are rarely made, although possible based on some of model approaches.

Keywords: response time, response accuracy, cognitive tests, cognitive processes, psychometric models, local dependencies, automated and controlled processes

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Mark D. Reckase,
Michigan State University, United States
Peida Zhan,
Zhejiang Normal University, China

***Correspondence:**
Paul De Boeck
deboeck.2@osu.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *02 September 2018*
Accepted: *14 January 2019*
Published: *06 February 2019*

Citation:
De Boeck P and Jeon M (2019) An Overview of Models for Response Times and Processes in Cognitive Tests. Front. Psychol. 10:102. doi: 10.3389/fpsyg.2019.00102

INTRODUCTION

Cognitive tests are meant to measure abilities. Abilities refer to levels of performance, whereas processes are the activities involved in reaching a performance outcome. Typically, cognitive tests do not yield process measures. It is perfectly possible to measure an ability without knowledge of the processes that are involved, but then the resulting measure only describes the level of performance, which is not always satisfying because it leaves why questions unanswered. Explanation requires a narrative of how something comes about. Processes provide such a narrative. Processes do not only help for understanding, they also help for more informative feedback and knowing the processes may help for interventions and remediation. Process information is also relevant to make validity inferences in the positive sense if the inferred processes support the interpretation of the intended ability, and in the negative sense, for example, because unintended processes can invalidate a measurement result. An important example of an invalidating process is guessing. Like it is possible to measure without investigating processes, it is also possible to investigate processes without measuring the related abilities, and a combination of the two is also possible.

Processes have the intrinsic feature that they take time. Therefore, response times are natural and evident kinds of data to investigate processes. Other kinds of data can also be informative regarding processes involved in reaching or not reaching a certain performance level.

In fact, the responses themselves may be informative. For example, based on a cognitive theory stipulating the processes involved in finding the correct response to a set of test items, a model can be developed for the probability of a correct response based on the mastery of the process skills required to successfully respond to the items. This is the basic principle behind cognitive diagnostic modeling (Rupp et al., 2010). Mediation research can also contribute to process research because the mediation variable functions as a process in the narrative of how the level of a dependent variable comes about (Hayes, 2017). It may explain why mediation analysis has become so popular. As far as types of data are concerned, eye movement data are an interesting source of information regarding processes (Cho et al., 2018), because it may be assumed that the mind follows the eyes, or the eyes fixate the stimuli the viewer is processing. Furthermore, brain activation and EEG data can be useful, as well as actions such as clicking and moving on the computer screen to find an answer to a question.

Here we will focus on response times, the time a respondent takes to respond to individual items in a cognitive test. Making use of response times in modeling test data can lead to the identification and measurement of processes, but, as will be discussed, the use of response time information does not necessarily imply it leads to inferences regarding the processes which are involved. The scope of this article comprises modeling approaches in which response times are used and cognitive process inferences can be made. For more general reviews of the use and importance of response time and of time available to make a test, see reviews by Lee and Chen (2011); Kyllonen and Zu (2016) and Schnipke and Scrams (2002).

Response time modeling approaches can be classified into four very broad possibly overlapping and not necessarily homogeneous categories. The categories are partly inspired by an overview made by van der Linden (2009). Before listing the categories, we introduce a symbolic notation for the models:

T_{pi} for the response time of person p and item i;
A_{pi} for the response accuracy of person p and item i;
$\leftarrow$ to indicate which variable is the dependent or independent variable; for example, $T_{pi} \leftarrow$ means that response time is the dependent variable.

(a) *Response time models*: response times as the sole end variable ($T_{pi} \leftarrow$);
(b) *Joint models*: response times as one of the end variables, jointly with another kind of variable (e.g., accuracy) ($[T_{pi}, A_{pi}] \leftarrow$);
(c) *Dependency models*: joint models in which response times and other data (e.g., response accuracy) are jointly modeled with the possibility of dependencies beyond dependencies captured by latent variables and item parameters ($[T_{pi} \leftrightarrow A_{pi}] \leftarrow$);
(d) *Response times as covariate models*: response times as an origin variable and another kind of variable (e.g., accuracy) as the end variable ($A_{pi} \leftarrow T_{pi}$).

An *end variable* is an outcome variable, also called dependent variable, the last variable in a dependency network. For example, in a simple measurement model for speed, the observed response times are modeled as a function of a latent speed variable and item time parameters. More than one variable can have the status of an end variable. For example, response time and response accuracy (correct vs. incorrect) can be joint end variables. An *origin variable* is a covariate, also called independent variable, a variable in the dependency network that is not explained by any other variable. More than one variable can have the status of origin variable.

RESPONSE TIME MODELS

Three subtypes of modeling will be discussed for the $T_{pi} \leftarrow$ case, and thus with response time as the sole end variable: (1) distribution models, (2) explanatory models, and (3) models with response accuracy as a covariate.

Distribution Models for Response Times

Not only the mean but also the distribution of response times is informative (e.g., Van Zandt, 2002). In most studies response times turn out to be distributed with a variance that increases with the mean. Many types of distributions have this feature or can accommodate this feature: gamma, inverse Gaussian, ex-Gaussian, and ex-Wald, lognormal, Weibull, and Gumble, while in fact also the normal distribution has been used even though it does not have the feature. Distributions are in the first place used as a tool to make a model work, which for some of these distributions means deciding on a link function or a transformation (Lo and Andrews, 2015). However, the distributions have also been interpreted in terms of generating processes and these processes may have cognitive interpretations.

- *Gamma distribution*: is generated when the response process consists of a set of sequential processes with an exponential time distribution, suggesting that the underlying processes are sequential. For example, Maris (1993) has used gamma distribution models to model response times for mental rotation items.
- *Inverse Gaussian distribution*: is generated from an information accumulation process with a single stopping criterion. For illustrations of this and other distributions, see Lo and Andrews (2015).
- *Weibull* and *Gumbel distributions*: are generated from parallel processes with a stopping rule based on the first process that reaches the information accumulation criterion (a decision threshold). The Weibull distribution has been used by Loeys et al. (2011) for a joint model of response time and accuracy.
- *Ex-Gaussian distribution*: is generated by the sum of a normally distributed random variable and an exponentially distributed random variable. It has three parameters: μ and σ for the normal distribution, and τ for the exponential distribution. The exponential distribution explains the skew. The Gaussian component has been interpreted as reflecting automatic processes and the exponential component as reflecting more controlled processes. There also seems to be a relationship of τ with cognitive efficiency (based on the drift rate parameter of the drift diffusion model, see Ratcliff,

1978; Ratcliff and McKoon, 2008) and working memory (Schmiedek et al., 2007). Based on simulation studies by Matzke and Wagenmakers (2009) it seems that all three ex-Gaussian parameters are sensitive to the decision threshold (the boundary separation from the diffusion model) but that primarily τ is sensitive to differences in cognitive efficiency (the drift rate parameter of the diffusion model).

- *Shifted Wald distribution*: is generated by an accumulation process with a certain rate and threshold, and with a shift parameter. The shift parameter can also be added to other distributions to account for the fact that the lower response time boundary is not zero but slightly higher (a zero response time is impossible). The shifted Wald distribution has been used by Anders et al. (2016).

It was Luce's (1986) purpose to derive underlying processes from response time distributions, but he came to the conclusion that the relationship between processes and distribution is not as clear as one would like (p. 173–174), and additionally, differentiating between the distributions is not always easy. The relationship between distributions and processes is also discussed by Van Zandt and Ratcliff (1995).

For the practical purpose of measurement and because it often fits the data very well, the *lognormal distribution* has become popular for cognitive test response times (van der Linden, 2006, 2007) without process interpretation claims. In some other applications, practical considerations have led to an approach based on the proportional hazard principle (e.g., Ranger and Kuhn, 2012, 2014; Ranger and Ortner, 2012; Wang and Xu, 2015; Kang, 2017). Burbeck and Luce (1982) explain that the normal, Gumbel, and ex-Gaussian distributions have a monotone non-decreasing hazard function, while the exponential distribution (a special case of the Weibull) has a constant hazard function, and the Weibull distribution can accommodate a decreasing, constant, and increasing function. Finally, a peaked hazard function applies to the lognormal and the inverse Gaussian. The hazard function approach may be more than just practical for fitting the data. The actual shape of the function (increasing, decreasing, constant, curvilinear) may imply suggestions for the kind of process. As an alternative for the proportional hazards model, the response times can also be categorized so that a generalized linear mixed model approach can be used (Molenaar et al., 2018), and a Box-Cox transformation is another option (Klein Entink et al., 2009a).

Explanatory Response Time Models

There is a tradition in cognitive psychology to decompose response times based on hypothesized sequential processes (Donders, 1869; Sternberg, 1969). The most extensive work is conducted by Sternberg (1977b, 1985). He started his work with analogy items (Sternberg, 1977a,b) and later extended it to other cognitive problems, such as deductive reasoning problems (Sternberg, 1980, 1986).

His theory, models, and analyses are briefly described here. Suppose an analogy problem "Son is to aunt as daughter is to ?.." (A:B :: C:? ..), with D as the correct response. The hypothesized processes are: encoding, inference, mapping, and application. First, there are three terms to be encoded ("son," "aunt," and "daughter"). Second, an inference needs to be made, based on a comparison of A and B ("son" and "aunt") which implies two differences (sex and generation). Third, mapping consists of comparing A and C ("son" and "daughter"), which implies one difference (sex). Finally, application consists of applying the A:B relationship to C to find D, which implies two differences (sex and generation). A basic assumption in the model is that a difference between terms takes time. To differentiate the number of feature differences to be processed for inference and application and to vary the number of terms to be encoded, one can present the respondents with A and B before the response time is recorded, so that the task requires only the encoding of one term (C), and the feature differences relevant for mapping and application (assuming A and B have already been encoded and an inference is made). The example item with a full item format leads to the following equation:

$$RT = intercept + aX_a + bX_b + cX_c + dX_d + \varepsilon, \qquad (1)$$

where RT is the response time, $X_a = 3$ (encoding of A, B, C), $X_b = 2$ (differences between A and B), $X_c = 1$ (differences between A and C), $X_d = 2$ (differences between C and D), and a, b, c, and d are parameters referring to the time spent per process, while ε is a residual term. For the reduced item format, with A and B presented before the response time is registered, the equation would be:

$$RT = intercept + aX_a + cX_c + dX_d + \varepsilon, \qquad (2)$$

where $X_a = 1, X_c = 1, X_d = 2$.

When a person is presented with a large set of problems with different values for the different X-variables, regression analyses can be conducted, one per respondent, which is what Sternberg (1977a) did at a time when mixed models were not yet common practice. Based on this approach, he was able to estimate the time each hypothesized process takes per person.

Around the same time as Robert Sternberg did his research, Susan Embretson (Whitely, 1976, 1977) was doing very similar work but with binary accuracy as the dependent variable, using item response (IRT) models. In fact, Fischer (1973) had formulated an IRT model with the potential to do just that. His Q-matrix contains the X-variables from the above equations. Within IRT this has further led to the test design idea (Embretson, 1985), cognitive diagnosis modeling (CDM) (Rupp et al., 2010) and explanatory item response models (De Boeck and Wilson, 2004). An important difference between CDM and the other approaches is that process inferences are discrete (often binary) and refer to mastery of skills that may be related to hypothesized processes; but see Zhan et al. (2018c) for mastery in probabilistic terms. However, because response times are not involved in these approaches, we will not follow up on these developments here.

Explanatory response time models have also been embedded in models discussed elsewhere in this article. For example, Maris (1993) has used item covariates in his gamma model, Klein Entink et al. (2009b) have used item covariates in the

hierarchical model of van der Linden (2007) to be discussed in Section Distribution Models for Response Times, and van Breukelen (2005) did the same in a related model. However, such applications with the possibility for process inferences are rather rare, whereas they have clear potential for the study of response times, just as they have for response accuracy. Possibly, the extension of CDM with response time data (Zhan et al., 2017) can lead to a further interest in this approach.

Response Time as a Function of Response Accuracy

Usually response time is considered as the independent variable for response accuracy and not the other way around. However, there is some literature on how the type of incorrect response is an indication for response time and for the underlying processes. For example, Novikov et al. (2017) hypothesize based on the literature that errors either stem from lack of cognitive control (deemed to be premature responses) and would lead to short response times (error speeding) or from attentional lapses and uncertainty. The study by Novikov et al. (2017) concerns an auditory discrimination task and the use of EEG to locate oscillations in different regions of interest in the brain. On average the response times were shorter for correct responses than for incorrect responses, a common finding for complex attentional tasks (Wilding, 1971; Luce, 1986) and slow errors are found to be an indication of attentional lapses and uncertainty. The empirical results turned out to be roughly in line with the hypothesis about fast and slow errors based on EEG oscillations in regions of interest in the brain known to be informative about the hypothesized processes.

JOINT MODELS

It has become common practice to register response times for all item responses, so that *parallel data* are available: response accuracy and response time per pair of respondent and item. This allows then for ($[T_{pi}, A_{pi}] \leftarrow$) models, where time and accuracy are joint end variables. The parallel data concept is broader than response time and response accuracy. Although the applications are rare or even non-existing, parallel data can also include eye-movement data, brain activation data (BOLD signals) and EEG data for one or more regions of interest (ROI).

Molenaar et al. (2015) have discussed a broad framework for joint models, called the bivariate *generalized linear item response theory modeling* (B-GLIRT) framework. As shown by Molenaar et al. (2015), these models are basically IRT versions of two-dimensional confirmatory factor analysis (CFA) models: one factor for ability and another (correlated) factor for speed. Guessing and random item parameters are thus far not used in factor models, but they can be and have been included in the IRT versions. The prototypical model in the category is the hierarchical model (van der Linden, 2007), which has inspired related models with a different response time distribution (e.g., Loeys et al., 2011; Wang et al., 2013; Kang, 2017), with a multidimensional extension of the measurement model (Zhan et al., 2018a), and with item response time varying in a systematic way during the test (Fox and Marianti, 2016). An interesting feature of the B-GLIRT framework is that Thissen's (1983) joint model can also be accommodated into B-GLIRT although it may not look like a typical CFA model. Another feature is that polytomous responses can also be dealt with.

The B-GLIRT models are measurement models but not process models. The primary function of response times is to strengthen ability measurement. However, two other types of joint models exist with the ambition to model cognitive processes based on parallel data regarding response time and response accuracy: diffusion models (Ratcliff, 1978) and race models (Townsend and Ashby, 1978). Tuerlinckx and De Boeck (2005) have shown that both these cognitive models can be approximately re-parameterized as item response models and thus as measurement models for test data. Since then, van der Maas et al. (2011) have developed a version of the diffusion model for cognitive test data (see Ranger and Kuhn, 2018, for estimation methods), and Rouder et al. (2015) and Ranger et al. (2014), have developed race models for joint response accuracy and response time data from cognitive tests. The diffusion model and the race model as process models are discussed after the hierarchical model is presented. Finally, there is a beginning research line of using parallel data for cognitive diagnostic modeling (Zhan et al., 2017, 2018b) with the possibility of accommodating local dependencies (Zhan et al., 2018b). These models offer the possibility of extending the hierarchical model and dependency models to another popular type of psychometric models.

The Hierarchical Model

The most popular method to analyze parallel data is van der Linden's (2007) hierarchical model and it is a member of the B-GLIRT family. Roughly speaking it is a two-dimensional model, with one dimension for accuracy (correct vs. incorrect) interpreted as ability and another dimension for response time (log of response time) interpreted as speed. The model is more complex, because the ability dimension is based on the three-parameter logistic (3PL) model with random items parameters for accuracy as well as for response time. The model is a hierarchical model because of the multivariate distribution for ability and speed and for the item parameters of response accuracy and response time. Furthermore, van der Linden (2009) notes that the ability would be higher and the speed lower if the respondent would make the same test with more focus on accuracy. Therefore, the ability and speed as measured are "effective" ability and speed for an unknown speed-accuracy tradeoff from the part of the respondent. Although the model is very useful as a measurement model, it is not a process model. It is a measurement model with the advantage that the measurement of ability can benefit from the response time information. If the two dimensions are related, the measurement of each of them gains strength from the data for the other.

The assumption of van der Linden (2007) model is that response times follow a lognormal distribution. Loeys et al. (2011) have used the lognormal distribution and the shifted Weibull, while for example Wang et al. (2013) and Kang (2017) have used a semi-parametric proportional hazards model which gives the opportunity to accommodate most types of

distributions and deviations from these. As far as the distribution can be interpreted in process terms, the proportional hazard approach can function as an explorative approach for cognitive processes.

Diffusion Model

The drift diffusion model has been presented in an explicit way as an alternative for the hierarchical model by van der Maas et al. (2011). The model is a modification of the original drift diffusion model (Ratcliff, 1978; Ratcliff and McKoon, 2008; Ratcliff et al., 2016) so that it can be used for multiple-choice data from cognitive tests. The primary process is information accumulation in response to a stimulus (an item) that comes with a binary choice question (e.g., "is the number of asterisks you see smaller or larger than 50?"). The restriction to binary choices is removed in the van der Maas et al. (2011) version. The information accumulation process is not a straight-line process, instead it is a random walk process between two boundaries (one for each response option) with a trend in the direction of one of both but with the possibility to end up at the boundary opposite to the trend because of the random character of the process. When a decision boundary is reached, the corresponding response follows. The trend parameter is called the drift parameter. The other parameters are boundary separation, bias, and non-decision time. The boundary separation represents the speed-accuracy balance (how certain one wants to be before responding), bias depends on where the process starts (in the middle or closer toward and thus in favor of one of the boundaries), and the non-decision time is the time not taken by the information accumulation.

Although the diffusion model is a process model, it is basically a one-process model, with the one process being information accumulation, governed by three parameters (drift, boundary separation, and starting point). The non-decision time is a rest category for processes involved in the perception of the stimulus and the act of responding.

For rather simple binary choice tasks with on average extremely fast responses—much faster than cognitive test responses—it makes sense that only one process is involved, while this is less likely for more complex cognitive tasks as presented in cognitive tests. Information accumulation may be a basic elementary component, but if it is, it would need to be repeated in each of the processes involved in more complex tasks, for example, in each of the processes Sternberg (1977a) has found to play a role in analogy tasks. Such an extension is a serious complication and cannot yet be dealt with in model formulation and estimation.

Still, van der Maas et al. (2011) have shown that latent variable modeling (including item parameters) is possible for the diffusion model assuming just one diffusion process. The major two latent variables in the model are cognitive efficiency (drift rate of the process) which is always positive in the van der Maas et al. model, and cautiousness (boundary separation for the process). Cognitive efficiency makes one respond faster and with a higher probability of a correct response, whereas cautiousness makes one respond slower and with a higher probability of a correct response. Therefore, and roughly speaking one can expect that these two dimensions are a rotation of the ability and speed dimensions of the hierarchical model, with cognitive efficiency in between ability and speed and with cautiousness in between ability and the opposite of speed.

In sum, although the diffusion model has several advantages (a process model, more fine-grained, taking the speed-accuracy balance into account), it is based on a one-process assumption, and as far as the latent variables are concerned, it is roughly speaking a rotation of the hierarchical model. Conceptually speaking, the cognitive efficiency as measured in the diffusion model, shows clear similarities to Spearman's (1927) view on intelligence and how the speed-accuracy balance plays a role in the response process (p. 250).

Race Models

Race models are based on the notion of a competitive race between accumulators, one for each response option. The Rouder et al. (2015) model has a shift parameter for response time but it has only one latent variable: the ratio of the rate of information gain and response boundary, and for the application Rouder et al. (2015) describe, this one latent variable is highly correlated with effective ability from the hierarchical model. The Ranger et al. (2014) model has two latent variables (but not a shift parameter): one for information accumulation in support of the correct response, and one for misinformation accumulation (supporting the incorrect response). The amount of processing capacity is the sum of these two and accounts for response time, whereas the discrepancy between the two accounts for response accuracy. The authors show that the speed-accuracy trade-off is a complicated function of these two. Because the two latent variables can be approximately re-parameterized as effective speed and effective ability, this race model is equivalent to the recognition of speed and ability as basic latent variables. We have empirical evidence for this conceptual analysis. From our own analysis of data, it was found that for the Ranger et al. latent variables the multiple correlations with effective ability are 0.886 and 0.833 (two different sets of items were used) and with effective speed they are 0.979 and 0.962. In other words, although the models have very different functional forms, the latent variables that are being extracted belong roughly to the same two-dimensional space.

The race models share with the diffusion model that they are process models, that they are more fine-grained, and that they have a solution for the speed-accuracy issue, but as far as latent variables are concerned, they seem to work with roughly the same two-dimensional space as the hierarchical model. In other words, the difference with the hierarchical model is primarily an interpretation difference. The diffusion model and race models both assume one primary process: either information accumulation between boundaries, or a race among different accumulators.

LOCAL DEPENDENCY MODELS

Local dependency models are models in which response time and response accuracy are jointly modeled but in which they are also related to each other beyond the relationship of their corresponding latent variables and item parameters so that they imply or can explain an extra dependency (of the type $[T_{pi} \leftrightarrow$

$A_{pi}] \leftarrow$). While T_{pi} and A_{pi} are end point variables, they also are covariates to explain the local dependency.

Types of Models

There is clear evidence for local dependencies between response time and accuracy (Bolsinova and Maris, 2016). The inclusion of dependencies in a joint model can be realized through the introduction of local dependency parameters or through models with different classes of responses (based on different response mechanisms). The former models are *latent variable models with remaining dependencies*. Either the item response time has a direct effect on the corresponding item accuracy (Bolsinova et al., 2017a; De Boeck et al., 2017) or vice versa (van der Linden and Glas, 2010), or the relationship is modeled as a symmetrical residual dependency. The alternative type of models are class models with two *classes of responses* corresponding to two response modes: a fast mode and a slow mode. The classes are classes of item responses (not of items and neither of persons), each with a different model and thus with different processes to arrive at a response. Examples of such models are described by Partchev and De Boeck (2012) (for manifest classes) and by Molenaar and De Boeck (2018), Wang and Xu (2015), Molenaar et al. (2016) for latent classes.

In the models presented in the former two articles with class models, either the observed item response time determines which model applies for accuracy (Partchev and De Boeck, 2012) (it is a manifest class model) or the item response time is a covariate for the probability of the model that applies for accuracy (Molenaar and De Boeck, 2018) (it is a latent class model). In both these models there is only one sub-model (one class) for response times, but there are two for accuracy. Which of the two applies depends on the response time, in a deterministic way in the former model and in a stochastic way in the latter.

In the other two models the response classes are associated with different models for response accuracy and response time. In the Wang and Xu (2015) model, one class represents the regular problem solving process and the other class is a rapid guessing class, while in the Molenaar et al. (2016) model, the two classes represent fast and slow problem solving processes (with a Markov transition between the two), respectively, but none of the two corresponds to guessing.

Two other models may seem similar to the latter two, but they are in fact person class models and not response class models. First, Meyer (2010) has also published a model for response time and response accuracy with two classes, a regular problem solving class and a rapid guessing class, for problem solvers and rapid guessers. Second, Jeon and De Boeck (2018) also work with person classes, each with its own accuracy model and with item response times as covariates of the class probabilities. The resulting classes are interpreted by the authors as a regular problem solving class and one or two automatic knowledge retrieval classes.

Findings

Based on the latent variable models with remaining dependencies, the main finding is a negative dependency between response time and response accuracy. Fast responses (short response times) have a higher accuracy (Bolsinova et al., 2017a,b; De Boeck et al., 2017). The dependency cannot be explained by the fact that easy items require less response time because the relationship across items (and persons) is taken care of through the item parameters (and the latent variables). The results are supported by the response class models with a fast and slow class. Items are easier in the fast response class than in the slow item response class (Partchev and De Boeck, 2012; DiTrapani et al., 2016; Molenaar et al., 2016; Molenaar and De Boeck, 2018). The rapid guessing mixture model cannot explain these results because it implies a positive dependency (slower responses are more correct). It is possible that the two types of response class models inform us about different underlying phenomena in the same data. Rapid guessing is considered an important phenomenon in educational measurement. It has been linked to lack of motivation, and in line with this hypothesis a response time effort (RTE) index has been developed (Wise and Kong, 2005; Wise and Gao, 2017) to identify motivation issues.

The negative dependency does not show in all studies, for example, in one of the two datasets in Bolsinova et al. (2017b), the dependency is positive. The exceptions can be explained by another rather robust finding that the dependency is positively correlated with the difficulty of the items (Meng et al., 2015; Bolsinova et al., 2017a,b; De Boeck et al., 2017; Molenaar and De Boeck, 2018). The easier (more difficult) the items are the stronger (weaker) the negative dependency is, and for more difficult items the dependency can be positive.

The negative dependency can be interpreted as the consequence of attention variation during the test. This would imply a variation of cognitive efficiency and thus a higher (lower) accuracy paralleled by shorter (longer) response time. The link with item difficulty can be explained if one assumes, in line with the diffusion model, that dominant responses are faster. The easier an item is, the more dominant the correct response is, and thus faster. For the difficult items, there may be one or more dominant incorrect responses raising the chances of an incorrect response being faster. Therefore, a variation of cognitive efficiency may lead to an association of fast with correct or with incorrect, depending on the difficulty of an item.

There are some alternative explanations for the same findings. First, on average easy items come with faster responses, but if easiness also depends on the respondent this would lead to a negative dependency between response time and response accuracy. At the same time, difficult items come with slower responses, but it is likely that respondents guess more on difficult items, which would lead to fast responses with a small probability of being correct. Second, it is also possible that, again on average, for easy items one relies more on automated processes, such as knowledge retrieval, which can be very fast, whereas difficult items require more controlled processing, which takes time. The latter explanation can be found in Goldhammer et al. (2014) for results that will be discussed in the next section on studies with response time as a covariate. For a further discussion of possible explanations, see Bolsinova et al. (2017c).

Based on the studies cited here, the residual dependencies are a robust finding, in low-stakes and high-stakes tests, for open-ended as well as multiple-choice items, for children and

adults, for educational tests as well as for intelligence tests. They are an intriguing phenomenon in the investigation of cognitive processes because they are derived from a more fine-grained analysis than the common models with latent variables and item parameters. Latent variables inform us about rather general individual differences in speed and ability and their association seems to vary depending on the test (Schnipke and Scrams, 2002; Klein Entink et al., 2009c; van der Linden, 2009). They can stem from differences in the speed-accuracy balance and other confounding variables. With respect to correlations across items, overall item differences in time intensity and difficulty and the fact that more difficult items take more time are rather self-evident findings. However, the dependencies are a new category of findings obtained after controlling for general differences and associations across persons and items; they refer to the more specific relationship between response time and accuracy (Bolsinova et al., 2017c).

One further and even more specific finding, although not based on joint modeling of response times and response accuracy, but on double-centering of response times instead (an explorative technique) is that the residual relationships between response time and difficulty may be curvilinear (Chen et al., 2018). The curvilinear relationship including its precise shape is confirmed with a fine-grained modeling approach by Bolsinova and Molenaar (2018). Naumann and Goldhammer (2017) also obtained curvilinear relationships with a method described in Section Local Dependency Models, and van Breukelen (2005) found indications of curvilinearity for some types of items with a related model.

Another and very recent joint latent variable model with dependencies is the generalized speed-accuracy response model for dichotomous items (van Rijn and Ali, 2017, 2018). It is a model with only one latent variable (a capacity variable) for when a scoring rule is used described by Maris and van der Maas (2012). Starting from the scoring rule, a corresponding model is formulated, by way of reversed engineering. The scoring rule implies that correct (incorrect) responses are rewarded (penalized) more the shorter the response time is. Responses, whether correct or incorrect do contribute less to the score the slower they are. When all the available time to respond is used (response time equal to the time limit) the response has no effect on the score. The model is at the same time a model with local dependence between response time and response accuracy, which is not surprising given that it is a model for a scoring rule that combines correctness and response time. Interestingly this model is applied by the authors to data from respondents who were not aware of the scoring rule. Therefore, the implicit assumption is that the rule they were using reflects their actual speed-accuracy balance. The speed-accuracy balance is of a different kind than the one defined by the boundary separation in the diffusion model. The latter implies that the larger the boundary separation is, the larger the value discrepancy is between a success and a failure. Instead, following the Maris and van der Maas scoring rule, the value of success and failure depends on the response time. The model does not allow for individual differences and item differences with respect to the speed-accuracy balance, but such an extension could lead to an estimation of the balance. A further interesting implication of the model is that the relationship between response probability and response time is curvilinear.

The findings from the class models are partly overlapping with those from latent variable models with residual dependencies in that the negative dependency and the link with item difficulty are supported as explained earlier. On the other hand, the class models seem to provide evidence for a dual-processing view. This is easy to understand for rapid guessing as a processing mode (Meyer, 2010; Wang and Xu, 2015), even though it might be necessary to distinguish between rapid guessing and cheating (Wang et al., 2018) because cheating can also be fast. Class models may be more difficult to understand for other distinctions between processes (if not prior suspects such as rapid guessing or cheating are available). A first obstacle is that the latent variable for accuracy is the same or highly correlated in the two classes in class models for slow and fast responses (Partchev and De Boeck, 2012; Coomans et al., 2016; DiTrapani et al., 2016; De Boeck et al., 2017; Molenaar and De Boeck, 2018). It means that, although the processes seem different, as one may infer from a difference in item parameters, the underlying abilities cannot be differentiated. When a respondent switches from one mode to another, which is modeled through a Markov model in Molenaar et al. (2016), an empirically not distinguishable ability is being used. This may seem odd, but it is possible indeed that, for example, the abilities for automated processing and controlled processing are empirically extremely highly correlated and nearly identical, even though the actual processes are different. A second obstacle is that the differences between the two classes have not much been explored in terms of item features or kinds of error. Based on the only effort we know of (Coomans et al., 2016), there is evidence for a qualitative difference between the response errors in the fast and slow response classes. For the two example items (multiplication items) given in Table 5 of the article, fast errors seem to be typos or negligent responses based on the correct or a related arithmetic operation, whereas slow errors can be reconstructed based on an unrelated kind of operation. For example, for $100 \times 3000 =?$, 3,0000 is a popular fast error, and 400,000 and 1,300,000 are more typically slow errors. Similarly, for $2 \times 80?$, 40 is more popular as a fast than as a slow error and the reverse is true for 600. Whereas, fast errors seem to be slips, slow errors seem based on complicated incorrect operations or slow guesses.

RESPONSE TIMES AS COVARIATE MODELS

Finally, there are studies in which response times are used as a covariate, in all cases with response accuracy as the dependent variable (models of the type $A_{pi} \leftarrow T_{pi}$). Response time is the origin variable and accuracy is the end variable. We will first discuss models inspired by the speed-accuracy tradeoff (SAT) and next the generalized linear mixed model (GLMM) approach of Goldhammer and colleagues will be covered. A combination of both can be found in van Breukelen (2005) and his analysis of mental rotation data.

SAT-Based Models

Perhaps the most well-known phenomenon that relates response time to accuracy is the speed-accuracy trade-off (Heitz, 2014). The SAT implies that the success rate shows an exponential growth to a limit as a function of time. The curve has been described by Wickelgren (1977) and is very similar to the curve that can be derived from the diffusion model (Wagenmakers et al., 2004). Lohman (1989) has used the curve for test data and has estimated the corresponding person parameters, such as the growth rate and the upper asymptote. It does make sense that with increasing time available, the accuracy rate goes up. A quite different question is whether the success rate goes up with the time a respondent takes to respond.

Roskam (1987) and Verhelst et al. (1997) make the assumption that a similar growth curve as the SAT curve applies to the time a respondent takes to respond (Roskam, 1987, 1997) and to minus the actual speed of a respondent (Verhelst et al., 1997). Wang and Hanson (2005) make the same assumption as Roskam although for a more complex model. A very nice feature of the Wang and Hanson (2005) model and of Lohman's (1989) approach is that the growth rate can be interpreted as speed (accuracy gain per unit of time, analogous to miles per hour) and the upper asymptote can be interpreted as power in the sense of the maximum accuracy one can reach. While it is undoubtedly true that the probability of success increases as a function of releasing time pressure or extending the available response time (e.g., Semmes et al., 2011; Davison et al., 2012; Goldhammer and Kroehne, 2014; Goldhammer et al., 2017; Chen et al., 2018), it also seems empirically the case that the accuracy curve does often not increase with the observed response time, as will be discussed in the following.

GLMM Based Covariate Models

In a series of studies, Goldhammer and colleagues (Goldhammer et al., 2014, 2015, 2017; Naumann and Goldhammer, 2017) have investigated the relationship of time on task with response accuracy, inspired by a dual-processing theory. The basic findings obtained with GLMM are that the association between response time and response accuracy controlling for the latent accuracy variable and for accuracy item parameters depends on the kind of task. However, it was always the case that the association is less negative (or more positive) for more difficult items. This was true for reading and problem solving tasks (Goldhammer et al., 2014), Raven items (Goldhammer et al., 2015), lexical decision tasks (Goldhammer et al., 2017), and digital reading (Naumann and Goldhammer, 2017). These results are perfectly in line with the results obtained from local dependency models, and they are also in line with findings by Jeon and De Boeck (2018) that faster than expected response times have a positive covariate effect on the probability of belonging to respondent classes where easy items are even easier, which are interpreted as knowledge retrieval (vs. problem solving) classes in line with the dual-processing hypothesis. The difficulty related dependencies are interpreted from the hypothesis that easy tasks are more amenable to automatization. Because in the studies by Goldhammer and colleagues the relationship between response time and response accuracy was more negative for respondents with high values on the accuracy latent variable, higher levels of skill are also assumed to correspond with higher levels of automatization.

Discussion and Conclusion

We will first discuss the general finding of local dependency, followed by some considerations regarding cognitive process modeling based on response times. For each of the points, conclusions and suggestions for further directions will also be formulated.

The general finding of local dependency between response time and response accuracy is important for at least three reasons. First, the dependency is a violation of measurement invariance because the dependency implies that ability and speed cannot be measured independently. It is important to investigate how large the resulting distortions are. It is possible that the established violations do not cause large measurement distortions. Second, although the local dependency does not give a direct process indication, it can be interpreted as an indirect indication of the main type of processing: automated vs. controlled processing. The distinction, and thus the dual-processing theory, must not necessarily be interpreted as a dichotomy, it can also be interpreted as a continuum. When interpreted as a dichotomy, it corresponds to the class models for response time and response accuracy. When interpreted as a continuum, it corresponds to latent variable models with residual dependencies and to the research line of Goldhammer et al. Third, the dependency seems to have a specific shape indicating that up to a certain point longer response times are associated with an increasing accuracy, after which longer response times become associated with a decreasing accuracy. To be clear, this is not a result based on the relationship between the latent variables; instead it is based on the local dependencies after controlling for latent variables. Following the results from Chen et al. (2018) the turning point comes earlier if the test is more knowledge based and less reasoning based. The shape of the curve may reflect the cost of time and effort on the speed-accuracy tradeoff. Early on in the response process the cost of spending more time is compensated by an increasing chance to find the correct response, but the longer it takes to find the correct response the higher the cost becomes while the perceived chance of finding the correct response may decrease so that the expectation of a correct response does no longer compensate for the cost of effort. This may not play a role for simple cognitive tasks with fast responses, but it seems more likely for problems as presented in a cognitive test, especially when the test has a global time limit. Future research should take the increasing cost of time and effort into account.

Most of the cognitive test research related to response times is focused on measurement and improvement of the quality of measurement, either making use of response times as collateral information for the ability to be measured or to identify and solve issues. One of the major issues is the speed-accuracy trade-off. Working at a slower or faster rate can reflect a natural pace but it may also be induced by a chosen speed-accuracy balance with consequences for the accuracy of responses and thus for ability estimation, and a faster or slower rate can also have consequences for speededness toward the end of the test. Unless an experimental design is used with a manipulation of

the available time, it is not possible to investigate and measure the effects of the SAT. However, experimental manipulations do not inform us about the speed-accuracy balance a respondent chooses when taking a test. The diffusion model seems to give an answer to that important question. It may be a valid answer for the simple two-choice tasks, but it is unclear whether it does for cognitive tests. Further, the assumption of the diffusion model is very similar to Spearman's (1927) assumption that speed and accuracy are governed by cognitive capacity and trading accuracy against speed. Consequently, there is no room for speed as a capacity or as a natural pace variable. Instead there is just one cognitive capacity which determines fast and accurate responses, except for a possibly interfering attitude: the speed-accuracy balance the respondent chooses to work with. To summarize, one cannot simply transpose the diffusion model to cognitive test data and make inferences about the SAT based on that model. Future diffusion model based research should take the nature of cognitive tests into account.

Another major issue is rapid guessing, due to lack of motivation, or due to strategic considerations such as gaining time in order to focus on items with a better perceived chance of success. Rapid guessing is an important practical measurement problem, but it does not inform us about the cognitive processes that play a role when the respondent does work on finding a correct response. It is surprising that response time decomposition models are not used more for cognitive tests, in the line of the cognitive process research by Robert Sternberg. Instead, this more differentiated research is represented in cognitive diagnostic modeling and thus in research and measurement based on response accuracy instead of response time (but see Zhan et al., 2017), whereas response times have a natural relevance for process research. It would be of interest for future research to focus more on response time decomposition models for cognitive test data, beyond the issue of rapid guessing. A combination of response time modeling with cognitive diagnostic model is an alternative and promising avenue for research.

In the future, process research can also come from other types of parallel information, such as eye movement data, recording of actions during the responding process (through clicks and moves on the computer screen), and brain imaging and EEG data. One of the important ongoing trends is the use of data analytics to unravel processes based on recorded actions during the time between the item presentation and the actual response. It is too early for a bet on which approaches will lead to breakthroughs. We should also consider that processes can be so complex and highly variable that it may not pay off to identify what the specific processes are and how they relate, and that it may be more efficient to assess cognitive processes on a higher level of abstraction, for example, how much they are based on automated vs. controlled processes. To summarize, the inclusion of other types of data beyond response times, such as eye tracking data and brain imaging may lead to important novel findings, but, perhaps choices have also to be made regarding the detailed or more general nature of processes one wants to investigate. A good compromise between specificity and generality of processes seems desirable.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Anders, R., Alario, F. X., and Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychol. Methods* 21, 309–327. doi: 10.1037/met0000066

Bolsinova, M., De Boeck, P., and Tijmstra, J. (2017a). Modelling conditional dependence between response time and accuracy. *Psychometrika* 82, 1126–1148. doi: 10.1007/s11336-016-9537-6

Bolsinova, M., and Maris, G. (2016). A test for conditional independence between response time and accuracy. *Br. J. Mathemat. Statist. Psychol.* 69, 62–79. doi: 10.1111/bmsp.12059

Bolsinova, M., and Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Front. Psychol.* 9:1525. doi: 10.3389/fpsyg.2018.01525

Bolsinova, M., Tijmstra, J., and Molenaar, D. (2017b). Response moderation models for conditional dependence between response time and response accuracy. *Br. J. Mathemat. Statist. Psychol.* 70, 257–279. doi: 10.1111/bmsp.12076

Bolsinova, M., Tijmstra, J., Molenaar, D., and De Boeck, P. (2017c). Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. *Front. Psychol.* 8:202. doi: 10.3389/fpsyg.2017.00202

Burbeck, S. L., and Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Percept. Psychophys.* 32, 117–133. doi: 10.3758/BF03204271

Chen, H., De Boeck, P., Grady, M., Yang, C.-L., and Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence* 69, 16–23. doi: 10.1016/j.intell.2018.04.001

Cho, S. J., Brown-Schmidt, S., and Lee, W. Y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: an application to intensive binary time series eye-tracking data. *Psychometrika* 83, 751–771. doi: 10.1007/s11336-018-9604-2

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H. L. J., and Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PLoS ONE* 11; e0155149. doi: 10.1371/journal.pone.0155149

Davison, M. L., Semmes, R., Huang, L., and Close, C. N. (2012). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Edu. Psychol. Measurement* 72, 245–263. doi: 10.1177/0013164411408412

De Boeck, P., Chen, H., and Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *Br. J. Mathemat. Statist. Psychol.* 70, 225–237. doi: 10.1111/bmsp.12094

De Boeck, P., and Wilson, M. (Eds.) (2004). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach.* New York, NY: Springer. doi: 10.1007/978-1-4757-3990-9

DiTrapani, J., Jeon, M., De Boeck, P., and Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence* 56, 82–92. doi: 10.1016/j.intell.2016.02.012

Donders, F. (1869). Over de snelheid van psychische processen. *Nederlands Archief voor Genees- en Natuurkunde*, *4*, 117-145. [Translated and reproduced in 1969 as "On the speed of mental processes, in *Acta Psychologica*, *30*, 412-431).

Embretson, S. E. (Ed.) (1985). *Test Design: Developments in Psychology and Psychometrics*. New York, NJ: Academic Press.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychol.* 3, 359–374. doi: 10.1016/0001-6918(73)90003-6

Fox, J.-P., and Marianti, S. (2016) Joint modeling of ability and differential speed using responses and response times, *Multivariate Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Goldhammer, F., and Kroehne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: experimental time limits, posterior time limits and response time modeling. *Appl. Psychol. Measurement* 38, 255–267. doi: 10.1177/0146621613517164

Goldhammer, F., Naumann, J., and Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *J. Intelligence* 3, 21–40. doi: 10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Toth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *J. Edu. Psychol.* 106, 608–626. doi: 10.1037/a0034716

Goldhammer, F., Steinwascher, M.A., Kroehne, U., and Naumann, J. (2017). Modeling individual response time effects between and within experimental speed conditions: A GLMM approach for speeded tests. *Br. J. Mathemat. Statist. Psychol.* 70, 238–256. doi: 10.1111/bmsp.12099

Hayes, A. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach.* 2nd edn. New York, NY: Guilford Press.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front. Neurosci.* 8:150. doi: 10.3389/fnins.2014.00150

Jeon, M., and De Boeck, P. (2018). An analysis of an item response strategy based on knowledge retrieval. *Behav. Res. Methods*. doi: 10.3758/s13428-018-1064-1

Kang, H. A. (2017). Penalized partial likelihood inference of proportional hazards latent trait models. *Br. J. Mathemat. Statist. Psychol.* 70, 187–208. doi: 10.1111/bmsp.12080

Klein Entink, R.H., Fox, J. P., and van der Linden, W. J. (2009c). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Klein Entink, R.H., Kuhn, J.-T., Hornke, L.F., and Fox, J. P. (2009b). Evaluating cognitive theory: a joint modeling approach using responses and response times. *Psychol. Methods* 14, 54–57. doi: 10.1037/a0014877

Klein Entink, R. H., van der Linden, W.J., and Fox, J.-P. (2009a). A box-cox normal model for response times. *Br. J. Mathemat. Statist. Psychol.* 12, 621–640. doi: 10.1348/000711008X374126

Kyllonen, P. C., and Zu, J. (2016). Use of response time for measuring cognitive ability. *J. Intelligence* 4:14. doi: 10.3390/jintelligence4040014

Lee, Y.-H., and Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychol. Test Assessment Model.* 53, 359–379.

Lo, S., and Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Front. Psychol.* 6:1171. doi: 10.3389/fpsyg.2015.01171

Loeys, T., Rosseel, Y., Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika* 76, 487–503. doi: 10.1007/s11336-011-9211-y

Lohman, D.F. (1989). Individual differences in error and latencies on cognitive tests. *Learn. Individ. Diff.* 1, 179–202. doi: 10.1016/1041-6080(89)90002-2

Luce, R.D. (1986). *Response Times.* New York, NY: Oxford University Press.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables and their application to psychometric models for response times. *Psychometrika* 58, 445–469. doi: 10.1007/BF02294651

Maris, G., and van der Maas, H. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika* 77, 615–633. doi: 10.1007/s11336-012-9288-y

Matzke, D., and Wagenmakers, E. J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: a diffusion model analysis. *Psychon. Bull. Rev.* 16, 798–817. doi: 10.3758/PBR.16.5.798

Meng, X. B., Tao, J., and Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *J. Educ. Meas.* 52, 1–27. doi: 10.1111/jedm.12060

Meyer, J. (2010). A mixture Rasch model with item response time components. *Appl. Psychol. Meas.* 34, 521–538. doi: 10.1177/0146621609355451

Molenaar, D., Bolsinova, M., Rozsa, S., and De Boeck, P. (2016). Response mixture modeling of intraindividual differences in responses and response times to the Hungarian WISC-IV block design test. *J. Intelligence* 4:10. doi: 10.3390/jintelligence4030010

Molenaar, D., Bolsinova, M., and Vermunt, J. (2018). A semi-parametric within-subject mixture approach to the analysis of responses and response times. *Br. J. Mathemat. Statist. Psychol.* 71, 205–228. doi: 10.1111/bmsp.12117

Molenaar, D., and De Boeck, P. (2018). Response mixture modeling: accounting for heterogeneity in item characteristics across response times. *Pychometrika* 83, 279–297. doi: 10.1007/s11336-017-9602-9

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behav. Res.* 50, 56–74. doi: 10.1080/00273171.2014.962684

Molenaar,D., Oberski, D., Vermunt, J., De Boeck, P. (2016). Hidden Markov IRT models for responses and response times. *Multivariate Behav. Res.* 51, 606–626. doi: 10.1080/00273171.2016.1192983

Naumann, J., and Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learn. Individ. Diff.* 53, 1–16. doi: 10.1016/j.lindif.2016.10.002

Novikov, N.A., Nurislamova, Y. M., Zhozhikashvili, N. A., Kalenkovich, E. E., Lapina, A. A., Chernishev, B. V., (2017). Slow and fast responses: Two mechanisms of trial outcome processing revealed by EEG oscillations. *Front. Human Neurosci.* 11:218. doi: 10.3389/fnhum.2017.00218

Partchev, I., and De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence* 40, 23–32. doi: 10.1016/j.intell.2011.11.002

Ranger, J., and Kuhn, J. (2012). A flexible latent trait model for response times in tests. *Psychometrika* 77, 31–47. doi: 10.1007/s11336-011-9231-7

Ranger, J., and Kuhn, J. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *Br. J. Mathemat. Statist. Psychol.* 67, 388–407. doi: 10.1111/bmsp.12025

Ranger, J., and Kuhn, J. (2018). Estimating diffusion-based item response theory models. Exploring the robustness of three old and two new estimators. *J. Edu. Behav. Statist*. doi: 10.3102/1076998618787791

Ranger, J., Kuhn, J. T., and Gavira, J.-L. (2014). A race model for responses and response times in tests. *Psychometrika* 80, 791–810. doi: 10.1007/s11336-014-9427-

Ranger, J., and Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazard model. *Br. J. Mathemat. Statist. Psychol.* 65, 334–349. doi: 10.1111/j.2044-8317.2011.02032.x

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85, 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*. 20, 873–922. doi: 10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends Cogn. Sci.* 20, 260–281. doi: 10.1016/j.tics.2016.01.007

Roskam, E.E. (1987). "Towards a psychometric theory of intelligence," in: *Progress in Mathematical Psychology*, Vol. 1, eds E. E. Roskam and R. Suck (Amsterdam: North-Holland), 151–174.

Roskam, E.E. (1997). "Models for speed and time-limit tests," in *Handbook of Modern Item Response Theory,* eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 187–208. doi: 10.1007/978-1-4757-2691-6_11

Rouder, J., Province, J., Morey, R., Gomez, P., Heathkote, A. (2015). The lognormal race: a cognitive- process model of choice and latency with desirable psychometric properties. *Psychometrika* 80, 491–513. doi: 10.1007/s11336-013-9396-3

Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., and Wittmann, W. W. (2007). Individual differences in components of reaction time distributions

and their relations to working memory and intelligence. *J. Exp. Psychol.* 136, 414-429. doi: 10.1037/0096-3445.136.3.414

Schnipke, D. L., and Scrams, D. J. (2002). "Exploring issues of examinee behavior: insights gained from response-time analyses," in *Computer-Based Testing: Building the Foundation for Future Assessments,* eds C. N. Mills, M. Potenza, J. J. Fremer and W. Ward (Hillsdale, NJ: Lawrence Erlbaum), 237–266.

Semmes, R., Davison, M. L., and Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Appl. Psychol. Measurement* 35, 433–446. doi: 10.1177/0146621611407305

Spearman, C. (1927). *The Abilities of Man*. New York, NY: MacMillan.

Sternberg, R. (1986). Toward a unified theory of human reasoning. *Intelligence* 10, 281–314. doi: 10.1016/0160-2896(86)90001-2

Sternberg, R.J. (1977b). Component processes in analogical reasoning. *Psychol. Rev.* 84, 353–378. doi: 10.1037/0033-295X.84.4.353

Sternberg, R.J. (1980). Representation and process in linear syllogistic reasoning. *J. Exp. Psychol.* 109, 119–159. doi: 10.1037/0096-3445.109.2.119

Sternberg, R.J. (1985). *Beyond IQ. A Triarchic Theory of Human Intelligence*. Cambridge, MA: University Press.

Sternberg, R. J. (1977a). *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, S. (1969). The discovery of processing stages: extensions of donders' method. *Acta Psychol.* 30, 276–315.

Thissen, D. (1983). "Timed testing: an approach using item response theory," in *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, ed D. J. Weiss (New York, NY: Academic Press), 179–203. doi: 10.1016/B978-0-12-742780-5.50019-6

Townsend, J.T., and Ashby, F. G. (1978). "Methods of modeling capacity in simple processing systems," in *Cognitive Theory,* Vol. 3, eds J. Castellan and F. Restle (Hillsdale, NJ: Erlbaum), 200–239.

Tuerlinckx, F., and De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika* 70, 629–650. doi: 10.1007/s11336-000-0810-3

van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika* 70, 359–376. doi: 10.1007/s11336-003-1078-0

van der Linden, W. J. (2006). A lognormal model for response times on test items. *J. Edu. Behav. Statist.* 31, 181–204 doi: 10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *J. Edu. Measurement* 46, 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., and Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika* 75, 120–139. doi: 10.1007/s11336-009-9129-9

van der Maas, H.L.J., Molenaar, D., Maris, G., Kievit, R., and Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychol. Rev.* 118, 339–356. doi: 10.1037/a0022749

van Rijn, P.W., and Ali, U.S. (2018). A generalized speed-accuracy response model for dichotomous data. *Psychometrika* 83, 109–131. doi: 10.1007/s11336-017-9590-9

van Rijn, P. W., and Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *Br. J. Mathemat. Statist. Psychol.* 70, 317–345. doi: 10.1111/bmsp.12101

Van Zandt, T. (2002). "Analysis of response time distributions," in *Stevens' Handbook of Experimental Psychology, 3rd Edn*, Vol. 4*, Methodology in Experimental Psychology,* eds J. T. Wixted and H. Pashler (New York, NY: Wiley), 461–516.

Van Zandt, T., and Ratcliff, R. (1995). Statistical mimicking of reaction time data: single-process models, parameter variability, and mixtures. *Psychonomic Bull. Rev.* 2, 20–54. doi: 10.3758/BF03214411

Verhelst, N. D., Verstralen, H. H. F. M., and Jansen, M. G. (1997). "A logistic model for time-limit tests," in *Handbook of Modern Item Response Theory*, eds W. J. van der Linden, and R. K. Hambleton (New York, NJ: Springer), 169–185.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., and Iversen, G.J. (2004). Assessing model mimicry using parameter bootstrap. *J. Mathemat. Psychol.* 48, 28–50. doi: 10.1016/j.jmp.2003.11.004

Wang, C., Fan, Z., Chang, H. H., and Douglas, J.A. (2013). The linear transformation model with frailties for the analysis of item response times. *Br. J. Mathemat. Statist. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response a. *Br. J. Mathemat. Statist. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Xu, G., and Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–253. doi: 10.1007/s11336-016-9525-x

Wang, T., and Hanson, B.A. (2005). Development and calibration of an item response model that incorporates response time. *Appl. Psychol. Measurement* 29, 323–339. doi: 10.1177/0146621605275984

Whitely, S. E. (1976). Solving verbal analogies: some cognitive components of intelligence test items. *J. Edu. Psychol.* 68, 234–242. doi: 10.1037/0022-0663.68.2.234

Whitely, S. E. (1977). Information processing on intelligence test items: some response components. *Appl. Psychol. Measurement* 1, 465–476. doi: 10.1177/014662167700100402

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol.* 41, 67–85. doi: 10.1016/0001-6918(77)90012-9

Wilding, J. M. (1971). The relation between latency and accuracy in the identification of visual stimuli. I. The effects of task difficulty. *Acta Psychol.* 35, 378–398. doi: 10.1016/0001-6918(71)90012-6

Wise, S.L., and Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Appl. Measurement Educ.* 30, 343–354. doi: 10.1080/08957347.2017.1353992

Wise, S. L., and Kong, X. J. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Zhan, P., Jiao H., and Liao, D. (2017). Cognitive diagnosis modeling incorporating item response times. *Br. J. Mathemat. Statist. Psychol.* 71, 262–286. doi: 10.1111/bmsp.12114

Zhan, P., Jiao, H., Wang, W.-C., and Man, K. (2018a). A multidimensional hierarchical framework for modeling speed and ability in computer-based multidimensional tests. *arXiv:1807.04003*. Available online at: https://arxiv.org/abs/1807.04003

Zhan, P., Liao, M., and Bian, Y. (2018b). Joint testlet cognitive diagnosis modeling for paired local item dependence response accuracy. *Front. Psychol.* 9:607. doi: 10.3389/fpsyg.2018.00607

Zhan, P., Wang, W. C., Jiao, H., Bian, Y. (2018c). Probabilistic-input, noisy conjunctive models for cognitive diagnosis. *Front. Psychol.* 9:997. doi: 10.3389/fpsyg.2018.00997

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 De Boeck and Jeon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 January 2019
doi: 10.3389/fpsyg.2019.00051

Modeling Response Time and Responses in Multidimensional Health Measurement

Chun Wang[1]*, *David J. Weiss*[2] *and Shiyang Su*[3]

[1] College of Education, University of Washington, Seattle, WA, United States, [2] Department of Psychology, University of Minnesota, St. Paul, MN, United States, [3] Department of Psychology, University of Central Florida, Orlando, FL, United States

This study explored calibrating a large item bank for use in multidimensional health measurement with computerized adaptive testing, using both item responses and response time (RT) information. The Activity Measure for Post-Acute Care is a patient-reported outcomes measure comprised of three correlated scales (Applied Cognition, Daily Activities, and Mobility). All items from each scale are Likert type, so that a respondent chooses a response from an ordered set of four response options. The most appropriate item response theory model for analyzing and scoring these items is the multidimensional graded response model (MGRM). During the field testing of the items, an interviewer read each item to a patient and recorded, on a tablet computer, the patient's responses and the software recorded RTs. Due to the large item bank with over 300 items, data collection was conducted in four batches with a common set of anchor items to link the scale. van der Linden's (2007) hierarchical modeling framework was adopted. Several models, with or without interviewer as a covariate and with or without interaction between interviewer and items, were compared for each batch of data. It was found that the model with the interaction between interviewer and item, when the interaction effect was constrained to be proportional, fit the data best. Therefore, the final hierarchical model with a lognormal model for RT and the MGRM for response data was fitted to all batches of data via a concurrent calibration. Evaluation of parameter estimates revealed that (1) adding response time information did not affect the item parameter estimates and their standard errors significantly; (2) adding response time information helped reduce the standard error of patients' multidimensional latent trait estimates, but adding interviewer as a covariate did not result in further improvement. Implications of the findings for follow up adaptive test delivery design are discussed.

Keywords: response time, hierarchical model, health measurement, multidimensional graded response model, item response theory (IRT)

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Lihua Yao,
United States Department of Defense, United States
Fernando Marmolejo-Ramos,
University of Adelaide, Australia

***Correspondence:**
Chun Wang
wang4066@uw.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *22 August 2018*
Accepted: *09 January 2019*
Published: *29 January 2019*

Citation:
Wang C, Weiss DJ and Su S (2019) Modeling Response Time and Responses in Multidimensional Health Measurement. Front. Psychol. 10:51. doi: 10.3389/fpsyg.2019.00051

INTRODUCTION

When assessments are delivered via computer-based devices, collecting persons' response times (RTs) at the item level is straightforward. The analysis of item-level RTs on assessments has attracted substantial interest recently. For example, in personality assessments, RTs have been used to measure attitude strength (Bassili, 1996), to detect social desirability (Holden and Kroner, 1992), and to enhance criterion validity (Siem, 1996). In achievement testing, RTs have been used to

evaluate the speededness of the test (Van Der Linden et al., 1999), to detect aberrant behavior (e.g., Wang and Xu, 2015; Wang et al., 2018b,c), and to design a more efficient test (Bridgeman and Cline, 2004; Van der Linden and Guo, 2008; van der Linden, 2009; Fan et al., 2012). RTs have also been used to evaluate response data quality in web-based surveys (Galesic and Bosnjak, 2009).

In the health measurement domain, response time (sometimes called reaction time) is often used to measure cognitive functioning, particularly in research on aging (e.g., Pearson, 1924; Braver and Barch, 2002; Hultsch et al., 2002; Anstey et al., 2005; Osmon et al., 2018). Similar to the speed test in educational assessments, RTs are usually collected from timed, target stimuli tasks, in which respondents are instructed to respond as quickly as possible. In this case, only RTs, not response accuracy, is of interest. For example, in a study using the United Kingdom Health and Lifestyle Survey (Cox et al., 1987; Der and Deary, 2006), person-level reaction times were examined across different age and gender groups. Another example is using RTs from a stop-signal reaction time task to study response inhibition from patients with Parkinson's disease and other brain disorders (Gauggel et al., 2004; Verbruggen et al., 2013). Despite these widespread applications of RTs, little attention has been paid to the usefulness of item-level response times as collateral information for improving measurement precision. These previous studies have primarily used scale-level, aggregated RTs, such as its mean and standard deviation. However, item-level RTs, routinely collected during computer-based assessment delivery, provide richer information. Only a recent didactic review by Osmon et al. (2018) demonstrated the advantages of examining the entire RT distribution rather than only its mean and standard deviation to understand the efficacy of mental speed assessment in clinical neuropsychology. Therefore, it was of interest to apply advanced psychometric models for item-level RTs in the assessment of reported health behaviors and evaluate if RTs help better estimate the main constructs of interest.

MODELS

Multidimensional Graded Response Model

The most appropriate measurement model for ordered polytomous responses is the graded response model (GRM; Samejima, 1969). The item response function of the unidimensional GRM model is

$$P_{jk}(\theta) = P^{+}_{jk}(\theta) - P^{+}_{j,k+1}(\theta) = \frac{e^{[Da_j(\theta - b_{jk})]}}{1 + e^{[Da_j(\theta - b_{jk})]}} - \frac{e^{[Da_j(\theta - b_{j,k+1})]}}{1 + e^{[Da_j(\theta - b_{j,k+1})]}} \quad (1)$$

where $P_{jk}(\theta)$ is the probability of a randomly selected person with a latent trait θ selecting category k of item j (k−1 … K). $P^{+}_{jk}(\theta)$ is the boundary response function, interpreted as the probability of responding to category k and above for item j given θ. a_j is the item discrimination parameter for item j. b_{jk} is the boundary location parameter for item j in category k ($k = 0, \ldots, K$). D = 1.7 is the normalizing constant. Because by definition, $P^{+}_{j0}(\theta) \equiv 1$ and $P^{+}_{jK+1}(\theta) \equiv 0$, neither b_{j0} nor b_{jK+1} are estimable parameters. Therefore, for an item with four response categories, only three boundary parameters are estimated.

When the instruments include multiple scales measuring different constructs or different aspects of the same construct (e.g., Zickar and Robie, 1999; Fraley et al., 2000; Fletcher and Hattie, 2004; Zagorsek et al., 2006; Pilkonis et al., 2014), the multidimensional extension of the GRM, namely, the MGRM (Hsieh et al., 2010; Jiang et al., 2016), is appropriate. Let $\boldsymbol{\theta}$ be a vector of length H representing the latent traits of interest, and let $h = 1, 2, \ldots, H$. Similar to the unidimensional case, $P^{+}_{j0}(\boldsymbol{\theta}) \equiv 1$ and $P^{+}_{j(K+1)}(\boldsymbol{\theta}) \equiv 0$. When the test displays a simple structure, the boundary response function takes the form of

$$P^{+}_{jk}(\boldsymbol{\theta}) = \frac{1}{1 + \exp\left[-\mathrm{D}a_{jh}\left(\theta_h - b_{jk}\right)\right]} = \frac{1}{1 + \exp\left[-\mathrm{D}\left(a_{jh}\theta_h + c_{jk}\right)\right]}, \quad (2)$$

assuming item j measures dimension h only so that a_{jh} is the item discrimination parameter on the hth dimension of item j. In Equation 2, $c_{jk} = -a_{jh}b_{jk}$ and this a-c parameterization with $D = 1$ is consistent with *flex*MIRT's (Cai, 2013) default parameterization; the c parameter is interpreted as the "intercept." Equation 2 could also be modified to accommodate complex structure; for details, see Reckase (2009).

Bivariate Models of Responses and RTs

Given that RTs carry useful collateral information about both item and person characteristics, the bivariate model of responses and RTs (Molenaar et al., 2015) was considered. The measurement model for responses was as specified in Equation 2, and the measurement model for RTs takes the form

$$\ln t_{ij} = \lambda_j + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \omega_{ij} \quad (3)$$

Here, t_{ij} denotes the RT of patient i on item j, τ_i is the latent speed parameter of patient i, λ_j and φ_j are the time-intensity and time-discrimination parameters of item j, and ω_{ij} is the residual. If the residuals are assumed to be normally distributed, then Equation 3 suggests that the response time t_{ij} follows a log-normal distribution. Other more flexible types of residuals can also be assumed if the data warrants (e.g., Wang et al., 2013a,b).

The term, $\varphi_j\rho_d\theta_{id}$, is called a cross-relation function (Ranger, 2013; Molenaar et al., 2015), and it is assumed that item j measures the dth dimension. Different from van der Linden's (2007) hierarchical model in which a covariance structure is assumed on θ and τ at a second level, this cross-relation term directly models the relationship between the latent ability and observed log-transformed RTs (log-RTs). Certainly, the cross-relation term based on τ_i could alternatively enter into the measurement model of responses; for example, Molenaar et al. (2015) argued that incorporating the cross-relation term in the RT model had unique advantages. That is, when the purpose of including RT information is to improve the measurement

precision of θ, it is preferable to leave the measurement model for the responses unchanged while modeling the information about θ (if any) in the RTs. In this regard, θ accounts for the shared ability variance in the responses and RTs and τ accounts for the additional, unique variance in the RTs. This joint model is termed as Model 0 and its diagram is shown in **Figure 1**.

To ensure model identifiability, several constraints need to be in place. First, regarding the MGRM model, the mean and variance of $\boldsymbol{\theta}$s are restricted to be 0 and 1, respectively. Second, the mean and variance of τ is also constrained to be 0 and 1 such that the residual variance of ω_{ij} is freely estimated[1]. The three θ components are assumed to be correlated, and the correlation matrix is freely estimated. However, all three θs are assumed uncorrelated with τ due to the inclusion of the cross-relation term. The same set of constraints was assumed for all other models introduced hereafter.

Molenaar et al. (2015) suggested identifiability constraints that are similar to those listed, except that $\text{var}(\tau) = 1 - \rho^2$, instead of 1. Both constraints are sufficient, and their choice conveniently allows the interpretation of ρ as a correlation coefficient. Note that in van der Linden's (2007) model the variance of τ is estimable (Equation 22, p. 294). This is because the lognormal model for RT in van der Linden (2007) takes the form

$$f\left(t_{ij};\tau_i,\alpha_j,\beta_j\right) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_j\left(\ln t_{ij} - (\beta_j - \tau_i)\right)\right]\right\} \quad (4)$$

where α_j is interpreted as the *dispersion* parameter that quantifies the variance of the lognormal distribution, rather than the *discrimination* parameter as in Equation 3.

Bivariate Model With Interviewer as a Covariate

Because more than one interviewer was used for data collection, three variations of the bivariate model with interviewer as a covariate were considered. The first model is

$$\ln t_{ij} = \lambda_j + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \sum_{p=1}^{P}\gamma_{jp}x_p + \omega_{ij}, \quad (5)$$

where x_p is a binary indicator variable indicating if interviewer p recorded the RTs for patient i, and P is the total number of interviewers in the data. P equaled 6 for batch 1 and 5 for batches 2–4. Because each patient interacted with only one interviewer, only one non-zero element in the summation $\sum_{p=1}^{P}\gamma_{jp}x_p$ enters into the regression equation for patient i. The model in Equation 5 (Bivariate Model 1) assumes that interviewer effects differed per item, i.e., there is an interaction between interviewer and items.

Model 2 is a slightly restricted version of Model 1, and the measurement model for RT becomes

$$\ln t_{ij} = \lambda_j + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \varphi_j\sum_{p=1}^{P}\gamma_p x_p + \omega_{ij}, \quad (6)$$

where all parameters have the same interpretations as in Equation 5 except τ_i, which can be interpreted as the individual "residual"

[1]By default, M*plus* sets the factor mean to be 0 for both $\boldsymbol{\theta}$ and τ.

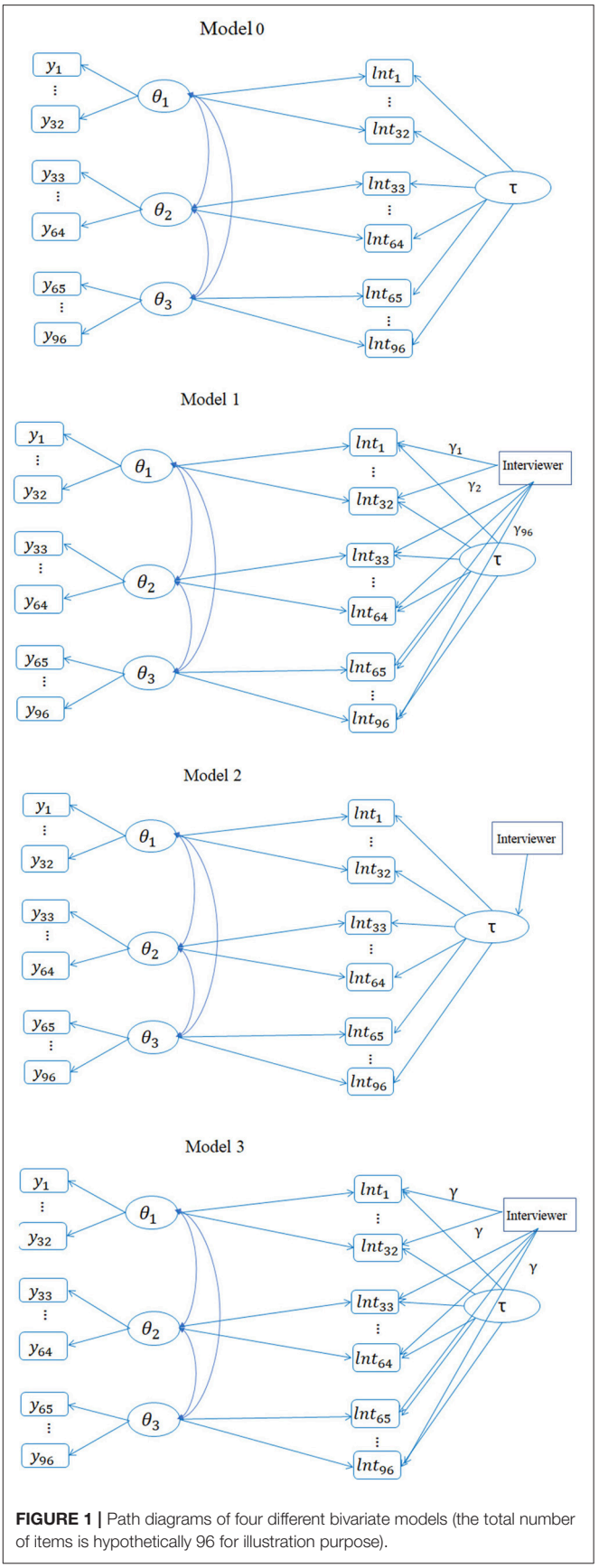


FIGURE 1 | Path diagrams of four different bivariate models (the total number of items is hypothetically 96 for illustration purpose).

speed after removing the interviewer effect. The MGRM model is still used for polytomous responses. In Equation 6, the interviewer effect differs across items but by the same amount, denoted as φ_j. This Model 2 can also be viewed as a hierarchical model in which the interviewer variable predicts the speed at the second level, as follows:

$$\ln t_{ij} = \lambda_j + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \omega_{ij}$$
$$\tau_i = \sum_{p=1}^{P} \gamma_p x_p + \varepsilon_i, \quad (7)$$

where ε_i is the individual residual speed. Compared to Model 1, Model 2 greatly reduces the number of parameters and hence is a more parsimonious model. When fitting the hierarchical model in M*plus* (Muthén and Muthén, 1998-2015), the variance of τ cannot be fixed directly but instead the variance of ε_i is fixed at 1.

Model 3 considers only the interviewer main effect and it assumes that the interviewer effect does not differ across items. Again the MGRM stays the same, and the model for RTs becomes

$$\ln t_{ij} = \lambda_i + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \sum_{p=1}^{P} \gamma_p x_{ip} + \omega_{ij}. \quad (8)$$

Although this Model 3 has essentially the same number of parameters as Model 2, it assumes no interactions between interviewers and items. The path diagrams for the four models are presented in **Figure 1**.

METHODS

Instrument and Subjects

Responses and RTs from the Activity Measure for Post-Acute Care (AM-PAC) were analyzed (Yost et al., 2018). The AM-PAC is the first multi-domain patient reported outcomes measure with the capability to direct care in a hospital rehabilitation environment. The scores from the AM-PAC are intended to be linked to the widely understood stages of the Functional Independence Measure (O'Dell et al., 1998; Huang et al., 2000) such that appropriate rehabilitative care plans can be immediately identified. It is anticipated that the AM-PAC will provide an inexpensive and accurate alternative to clinician assessments. The three domains covered in the AM-PAC include Applied Cognition, Daily Activity, and Mobility. A sample question from the Applied Cognition domain is: "*How much difficulty do you currently have reading a long book (over 100 pages) over a number of days*?", and the four response options are "Unable" (coded as 1), "A lot" (coded as 2), "A little" (coded as 3), and "None" (coded as 4). Items were administered to hospital inpatients via a computer-assisted personal interview using Qualtrics® web survey software. During the field testing of the items, an interviewer read each item to a patient and recorded, on a tablet computer, the patient's responses and the software recorded RTs. A total sample of 2,270 hospitalized patients were recruited to the study; their mean age was 65 years. Roughly 54% were male and 96% were non-Hispanic white, and 78% had two or more comorbidities (Yost et al., 2018).

Questions were grouped into blocks according to domain, and the order of item administration within a block was randomized. Given that there were 324 items in total in the bank, data collection proceeded in four batches to reduce patient burden. The first batch of 109 items was administered to patients, and 24 linking items were selected with eight items in each domain. The number of linking items was determined based on Kolen and Brennan (2004)'s recommendation that at least 20% of the items need to be shared between different test forms to have enough information to link the scale (Wang et al., 2016). These linking items in each domain were selected to produce a composite information function that was closest in shape to the domain information function. Linking items were assembled using the linear programming solver "lp_solve version 5.5" (Diao and van der Linden, 2011). Then, the set of linking items was carried forward in subsequent data collection batches. **Table 1** presents the number of items per domain for the four batches.

Preliminary Data Cleaning

Table 2 presents the summary descriptive statistics for the four batches of data. The cleaned Batch 1 dataset contained 563 respondents after deleting 67 (10.6%) respondents with at least 20 missing items. The cleaned Batch 2 dataset contained 490 respondents after deleting 52 (9.6%) respondents with more than 10 missing items. The cleaned Batch 3 dataset contained 500 respondents after deleting 55 (9.9%) respondents with more than 9 missing items. The cleaned Batch 4 dataset contained 507 respondents after deleting 36 (6.6%) respondents with more than nine missing items. Although each item contained four response categories, for some items, category 1 and/or category 2 received no responses or very few responses. These items were then recoded to ensure that the lowest response category for each item was always 1, but the highest response category could be 4 or less. As shown in **Table 2**, the response time distribution exhibited extreme skewness (ranging from 29.08 to 41.84), and therefore the distribution was truncated by removing the top 2.5% and removing the RTs smaller than 3 s, resulting in skewness from 1.48 to 1.66. The resulting data was entered into modeling analysis. Recent research by Marmolejo-Ramos et al. (2015a) suggested that Box-Cox transformation outperformed the elimination method in normalizing positively skewed data. However, the extremely long and short RTs were trimmed in these data because those RTs were considered as outliers. Extremely long RTs happened when the patient took a break such as "service came in to discuss plans" or "patient lunch came and wanted to stop." The row for the missing proportion of RTs in

TABLE 1 | Number of unique items per domain for the four batches.

Batch	Applied cognition	Daily activity	Mobility	Total	
				Unique	Linking
1	28	27	30	85	24
2	24	24	24	72	24
3	24	24	24	72	24
4	23	23	25	71	24
Linking	8	8	8	24	24
Total	107	106	111	324	—

TABLE 2 | Descriptive statistics of the observed data, by batch.

Variable	Batch 1	Batch 2	Batch 3	Batch 4
SAMPLE SIZE				
Before cleaning	630	542	555	543
After cleaning	563	490	500	507
Trimmed proportion of RTs	6.24%	3.21%	3.16%	4.59%
NUMBER OF ITEMS				
2 categories	1	0	0	0
3 categories	31	26	22	23
4 categories	77	70	94	72
RT BEFORE TRUNCATION				
Mean	9.27	9.79	9.82	8.39
SD	21.28	17.39	25.37	17.62
Skewness	41.84	32.40	35 .85	29.08
RT AFTER TRUNCATION				
Mean	8.21	8.44	8.06	7.18
SD	4.14	4.92	4.80	4.05
Skewness	1.48	1.66	1.65	1.53

Table 2 refers to the proportion of RTs at the person-by-item level, out of the cleaned sample size (e.g., 563 for batch 1), that was deleted either because they were extremely short (<3 s) or extremely long (upper 2.5%).

To further test the normality of item-level RT distributions, the Kolmogorov-Smirnov (K-S) test (Smirnov, 1948) was conducted for all item-level log-RTs. The K-S statistic quantifies the distance between the empirical distribution function of a sample and the cumulative distribution function of a reference distribution, and it is a non-parametric test of the equality of two distributions. For the present purpose, the K-S test was done with response times that were at least 3 s and were below the 97.5% percentile. This item-level K-S test compared the log-RTs of that item to the theoretical normal distribution with the mean and variance computed for the item. The null hypothesis is that the log-RTs follow a normal distribution. Hence, a significant *p*-value (i.e., $p < 0.05$) indicated that the log-RTs distribution was significantly different from normal. Results showed that in Batch 1, 54 out of 110 items exhibited statistically significant *p*-values, but the K-S statistics for those items were very small (ranged from 0.05 to 0.1). In Batches 2, 3, and 4, 30 out of 96, 16 out of 96, 11 out of 95 items, respectively, had significant *p*-values, but again, the K-S statistics were small.

The K-S test was chosen because of its wide popularity. For instance, it was used to evaluate the item RT distributions from computer-based licensure examinations (Qian et al., 2016). However, other tests, such as the Shapiro-Wilk (S-W) test (Royston, 1982) has been found to be more powerful than the K-S test to detect departure from normality (Marmolejo-Ramos and González-Burgos, 2013). Unsurprisingly, using the S-W tests on the same data set showed that 99.1% of Batch 1 items, 90.6% of Batch 2 items, 95.8% of Batch 3 items and 92.6% of Batch 4 items had significant *p*-values. However, the lognormal model was still used as the parametric model for RTs in the following analysis because the skewness (shown in **Table 2**) after truncation was not high, and the lognormal distribution was a convenient choice that most software packages can handle.

Collapsing Response Categories

In the data analysis, response categories for some items were collapsed due to lack of observations in those categories. Specifically, for a given item, if a category received no response or only one response, the response of this option, if any, was combined into the responses of the next higher category. Therefore, as shown in **Table 2**, some items had fewer than four response categories. The treatment of collapsing response categories is legitimate for the graded response model because it does not substantially change the item parameter estimates[2]. For instance, a 4-category GRM item ($k = 1, 2, 3, 4$) item will have four parameters, i.e., a_j, b_{j1}, b_{j2}, b_{j3}. When collapsing the lowest two response categories, the parameters of the same item become $a_j^* \approx a_j$, $b_{j2}^* \approx b_{j2}$, $b_{j3}^* \approx b_{j3}$. This is because the GRM is essentially a difference model (see Equation 1), and the same discrimination parameter is assumed across all boundary response functions [i.e., $P_{jk}^{+}(\theta)$].

Model Fitting and Item Calibration

Bivariate Model Fitting

All four models in **Figure 1** were fit with marginal maximum likelihood estimation (MML) using the Expectation-Maximization (EM) algorithm in M*plus*[3] These models were fitted to each batch of data separately to evaluate global model fit via AIC, BIC and−2Log-likelihood. The M*plus* source code of Batch 4 is provided in the **Appendix**. The same source code was used for other batches, as well. As shown in **Table 3**, Model 2 was the best-fitting model across all four batches of data based on BIC, but Model 1 was preferred based on AIC. In addition, Model 2 and Model 3 are respectively nested within Model 1. The deviance test (i.e., likelihood ratio test) revealed that there was a significant difference between Model 1 and Model 2, Model 1 and Model 3, implying that Model 1 should be preferred. However, Model 2 was used in the following analysis for two reasons: (1) Model 2 is a much more parsimonious model than Model 1 and it is conceptually more reasonable because the interviewer effect should not interact with items, i.e., the interviewer's speed should be relatively static across items; (2) when fitting Model 1 in the concurrent calibration described below, it failed to converge due to complexity and data sparsity.

Concurrent Calibration

When data are collected in different batches, linking items are used to place the items from the different batches onto a common scale. Concurrent calibration has been demonstrated to be more effective than separate calibration plus *post-hoc* linking (Kolen and Brennan, 2004) because the latter approach suffers from linking error. Three models were compared in the concurrent calibration: the MGRM model for responses only, Model 0, and

[2]A separate study (Jiang and Wang, 2019) was conducted that provided analytic and simulation evidence for this claim.

[3]M*plus* was chosen because it is widely used in social science research. M*plus* plotting using R is available via the "rhdf5" package. For details, refer to http://www.statmodel.com/mplus-R/.

Model 2. Models 1, and 3 were not considered because of their poorer fit compared to Model 2. Both the item and person parameters and their standard errors were compared across the three models. The main research question was whether including RTs and interviewer information helped improve the estimation accuracy of both item and person parameters.

When pooling data from the four batches together, the concurrent calibration of Model 0 and Model 2 failed to converge due to the sparsity of data and model complexity. Therefore, a two-stage approach was implemented. In the first stage, data from Batches 2–4 were pooled and a concurrent calibration was conducted on the pooled data. Data from Batch 1 was left out because this batch had the largest number of items flagged under the K-S test. By shrinking the sample size, all models successfully converged. Then in the second stage, Batch 1 data were calibrated using the fixed parameter calibration approach (Kim, 2006). That is, the linking item parameters (i.e., a, b, λ, and φ) were fixed at their estimated values obtained from Stage 1 for each of the three models such that the remaining items were estimated on the same scale as the linking items. Hence, no further linking procedure was needed.

Due to the collapsing of response categories, a side note for the two-stage approach is worth mentioning. Specifically, for the linking items, the threshold parameters of Batch 1 did not always match those in Batches 2–4. For example, an item had four categories (three threshold parameters) in Batches 2–4, but only three categories (two threshold parameters) in Batch 1. The linking items always had the same or fewer categories in Batch 1 as compared to the combined data due to the smaller sample size of Batch 1. In this case, only the corresponding threshold parameters and discrimination parameter for an item were input into the fixed calibration. The rationale is the same as before—collapsing response categories does not substantially change the item parameters.

TABLE 3 | Global fit results (AIC, BIC,−2Log-likelihood) for the four bivariate models, by batch.

Batch and model	Number of free parameters	AIC	BIC	-2Log-likelihood
BATCH 1				
Model 0	736	133566	136755	132094
Model 1	1281	133174	138725	130612
Model 2	**741**	**133316**	**136527**	131834
Model 3	741	133409	136620	131926
BATCH 2				
Model 0	652	102468	105202	101164
Model 1	940	102049	105992	100170
Model 2	**655**	**102235**	**104982**	100924
Model 3	655	102339	105086	101030
BATCH 3				
Model 0	656	111384	114149	110072
Model 1	1040	110613	114996	108532
Model 2	**660**	**111001**	**113783**	109682
Model 3	660	111323	114105	110004
BATCH 4				
Model 0	648	108550	111290	107254
Model 1	1028	107733	112080	105676
Model 2	**652**	**108174**	**110931**	106870
Model 3	652	108364	111121	107060

Bold values highlighted the best-fitting model based on the information criteria.

RESULTS

Global Model Fit

Table 4 presents the global model fit statistics for the three models in both stages. Note that the AIC and BIC from the MGRM are smaller because they are on a different scale compared to Model 0 and Model 2 due to its exclusion of RT information. Consistent with the separate calibration results, Model 2 fit the data better than Model 0, reflected by smaller AIC and BIC values.

Item Parameter Estimates

Figure 2 presents the scatterplots of item discrimination parameters (a_j) across the three models; all points fall along the 45° line, implying a close alignment of item parameter estimates from the three models. This is unsurprising because the variance of θ was fixed at 1 across all models, which fixed the scale of a_j. Means of SEs of estimates of a_j were 0.188 in the MGRM, 0.190 in Model 0, and 0.190 in Model 2. A simple t-test showed no significant differences of SEs between the different models.

The correlations of boundary parameters $\hat{b}_{jk}$ between different models were all 1, and therefore the scatterplots (**Figure 3**) show that the estimates of $\hat{b}_{jk}$ from the different models fall tightly on the 45° line. Moreover, t-tests showed no significant differences of mean SEs of $\hat{b}_{jk}$ between different models. Thus, the results suggest that, in these data, estimation of MGRM item parameters a_j and b_j, and their SEs were not affected by the addition of RT information.

With respect to the item time discrimination parameter, φ_j, the correlation between their estimates from Model 0 and Model 2 was 0.99. The scatterplot (**Figure 4**) shows that these estimates of $\hat{\varphi}_j$ from the two models fell on a line that was not 45°, indicating that there was a linear relationship between $\hat{\varphi}_j$s from the two models. The explanation is as follows: Focusing on the two terms in Model 0 and Model 2, respectively, $\varphi_j(\tau_i - \rho_d\theta_{id})$

TABLE 4 | Global model fit results.

Stage and Model	AIC	BIC
STAGE 1: CONCURRENT CALIBRATION (BATCHES 2-4)		
MGRM	185303.933	190099.963
Model 0	327447.703	336067.811
Model 2	**326589.532**	**335230.884**
STAGE 2: FIXED PARAMETER CALIBRATION		
MGRM	74013.471	75391.453
Model 0	134224.604	136824.572
Model 2	**133924.461**	**136546.095**

Bold values highlighted the best-fitting model based on the information criteria.

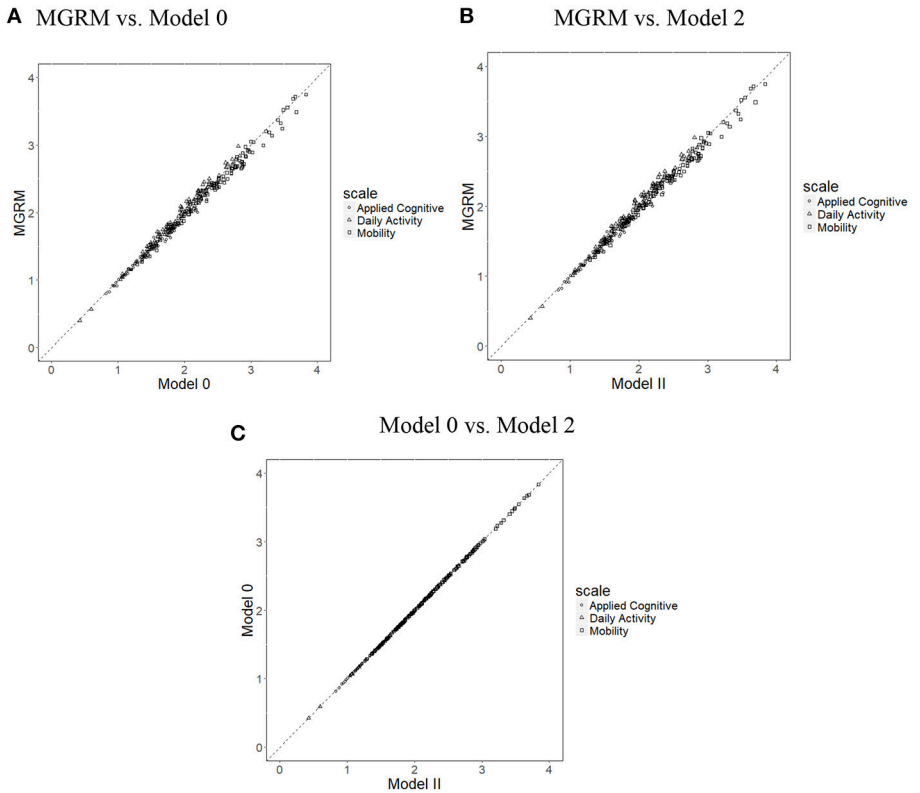


FIGURE 2 | Scatterplots of item discrimination parameters (a_j) across three models. **(A)** MGRM vs. Model 0, **(B)** MGRM vs. Model 2, and **(C)** Model 0 vs. Model 2.

(Equation 3) and $\varphi_j(\sum_{p=1}^{P}\gamma_p x_{ip} + \varepsilon_i - \rho_d\theta_{id})$ (Equation 7) the $(\tau_i - \rho_d\theta_{id})$ and $(\varepsilon_i - \rho_d\theta_{id})$ are the same across the two equations because both τ_i and ε_i are on the 0–1 scale. Due to the data collection design, the same interviewer went through all items in the batch each time, and each interviewer interviewed a portion of the sample. For instance, suppose the sample size is N, and there are n_1, $n_2 - n_1$, $n_3 - n_2$, $n_4 - n_3$, and $N - n_4$ patients interviewed by each interviewer, as shown in **Table 5**. Then, for item j, those patients assigned to Interviewer 1 all carry the same interviewer effect of γ_1, and similarly for the three other groups. Hence, the second and third columns are the same for every item, and also because the mean and variance of these two columns are different, there is a unique linear relationship between $\hat{\varphi}_j$s from the two models. On the other hand, SEs of $\hat{\varphi}_j$ of Model 2 are significantly lower ($p < 0.001$) than those of Model 0: Mean SE was 0.018 in Model 0 and 0.013 in Model 2.

Regarding the time intensity parameter, λ_j, again the correlation between their estimates from Model 0 and Model 2 was larger than 0.99. SEs of $\hat{\lambda}_j$ from Model 2 were significantly higher ($p < 0.001$) than those of Model 0: Mean SE was 0.020 in Model 0 and 0.023 in Model 2. The results suggest that the SE of item response time parameters $\hat{\varphi}_j$ and $\hat{\lambda}_j$ is affected in different directions by the addition of interviewers as covariates. However, the absolute difference in SEs was not too large to be concerning because the difference appeared mostly in the third decimal place.

Person Parameter Estimates

In terms of θ estimation, the $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ from all three models correlated as high as 0.99 to 1. Mean SEs from Models 0 and 2 (**Table 6**) were significantly lower than those from the MGRM ($p < 0.001$), and there was no significant difference of SE between Model 0 and Model 2. This result implies that adding response time decreased the SE of $\hat{\theta}$, which is consistent with prior findings (e.g., van der Linden et al., 2010; Wang et al., 2013a,b), but adding the interviewer variable did not further decrease the SEs.

Table 7 presents the estimated correlation parameters. Consistent with previous findings (e.g., Wang et al., 2018a), there were moderate to high correlations among the three latent traits. Moreover, the speed factor also played a modest role as reflected by the moderate size of ρ_1 to ρ_3. These correlations were higher in Model 2 than in Model 0, which is unsurprising because after removing the interviewer effects on RTs, the individual speed factor should correlate higher with individual latent traits.

The last column in **Table 7** refers to the fixed effects of interviewers. During Batches 2–4 data collection, the same five interviewers were recruited and one of them was randomly selected as the reference for dummy coding. It appears from the estimated $\hat{\gamma}_p$ that interviewers differed substantially and that is why including the interviewer variable in the model helped improve model data fit. For Batch 1 data collection, a different set of six interviewers was recruited; among them, three overlapped with the other set of five. However, because a different reference

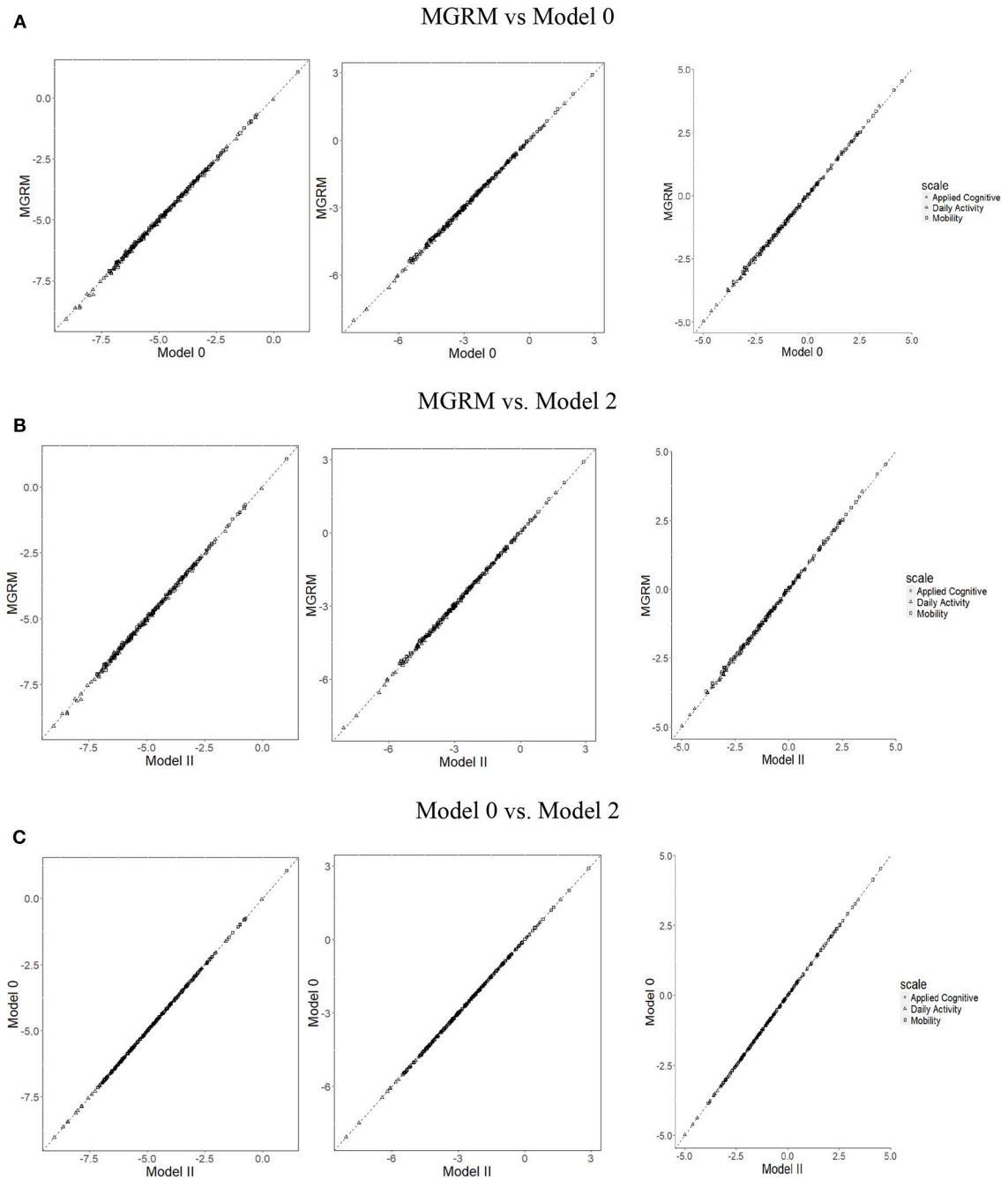


FIGURE 3 | Scatterplots of item boundary parameters (from left to right: $\hat{b}_{j1}$, $\hat{b}_{j2}$, $\hat{b}_{j3}$) across three models. **(A)** MGRM vs. Model 0, **(B)** MGRM vs. Model 2, and **(C)** Model 0 vs. Model 2.

interviewer was selected, the estimated $\hat{\gamma}_p$ from stage 1 and 2 model fitting were not directly comparable. Still, the results show that interviewers operated at different speeds and they contributed to the observed RT variabilities.

DISCUSSION AND CONCLUSIONS

Response time as part of the assessment process data has gained great popularity in recent decades in educational and psychological measurement. This is because collecting RTs has become easy, due to computer-based assessment delivery, and RTs provide an additional source of information for researchers to understand an individual's behavior as well as the characteristics of the items. More than a dozen IRT models have been proposed in the psychometrics literature, with an early focus on modeling the different shapes of RT distributions (e.g., Rouder et al., 2003; van der Linden, 2007; Loeys et al., 2011; Wang et al., 2013a,b) and a later focus on modeling within-subject

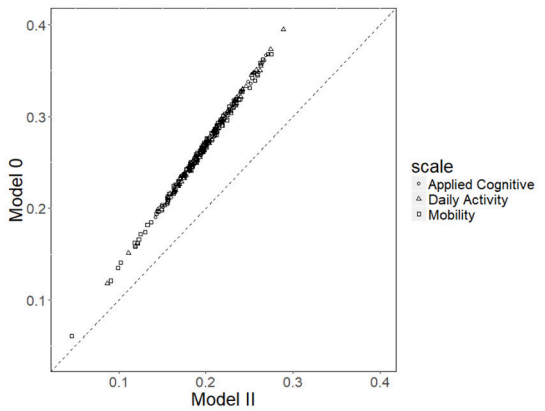


FIGURE 4 | Scatterplot of estimates of $\hat{\varphi}_j$ between Model 0 and Model 2.

TABLE 5 | An illustration of the linear transformation relationship of $\hat{\varphi}_j$ from Model 0 and Model 2.

Patients	Model 0 φ_j^0	Model 2 φ_j^2	Interviewer
1 to n_1	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_1$	1
n_1+1 to n_2	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_2$	2
n_2+1 to n_3	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_3$	3
n_3+1 to n_4	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_4$	4
n_4+1 to N	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id})$	5 (reference)

TABLE 6 | Mean and SD of SE of $\hat{\theta}$ from three models.

θ	MGRM	Model 0	Model 2
SE θ_1			
Mean	0.307	0.280	0.279
SD	0.093	0.076	0.076
SE θ_2			
Mean	0.252	0.242	0.241
SD	0.082	0.071	0.070
SE θ_3			
Mean	0.178	0.171	0.171
SD	0.079	0.074	0.074

variations such as different and changing test-taking behaviors (e.g., Wang and Xu, 2015; Molenaar et al., 2018; Wang et al., 2018b,c). However, the usage of item-level RT information has rarely been explored in health measurement.

This study systematically investigated the application of RTs for improving measurement precision of the target latent traits and the estimation precision of the item parameters. The bivariate joint model discussed in Molenaar et al. (2015) was applied and expanded in two respects: (1) a multivariate θ was considered in the measurement model for responses, and this θ vector was correlated with the latent speed through the cross-relation term; (2) an interviewer covariate was entered into the model to explain the variability of the observed RTs. Patient-reported outcomes obtained from personal interview surveys are widely used in health services research studies (Clancy and Collins, 2010), especially when conducting such surveys among older adults or patients with severe symptoms like the sample used in the present research. Thus, the observed RTs might be contaminated by the interviewer's reaction speed and, hence, the interviewer variable should be included in the model.

Several approaches to including the interviewer variable were explored. Results indicated that Model 2, which is a hierarchical model, consistently best fit the data. In this model, the interviewer's effect on the observed RTs is mediated through patients' latent speed. It is more parsimonious than Model 1 in which the interviewer's effect could differ for different items. Indeed, Model 2 also makes more intuitive sense because the interviewer effect reflects the different interviewers' response styles (i.e., fast or slow responders) that could be considered as the latent speed of an interviewer; hence, it should not change from item to item.

Results from the data analysis revealed that (1) adding response time information did not affect the item parameter estimates and their standard errors significantly; and (2) adding response time information helped reduce the standard error of patients' multidimensional latent trait estimates, but adding interviewer as a covariate did not result in further improvement, although the interviewer effect was significant. Regarding the first point, it is not surprising because Ranger (2013) has proven that the amount of (Fisher) information RTs provide to θ cannot be $> \frac{\rho^2}{1-\rho^2}$ (i.e., an upper bound) regardless of test length and RT distributions. A simple explanation is that RTs only contribute to θ via τ due to the hierarchical structure in van der Linden (2007), and hence the maximum information RTs provide is when τ is "observed," resulting in the information upper bound. As a result, the collateral information provided by RT will be useful when test length is short, but its role diminishes in longer tests when information accrued through responses is already high. That said, it is still worth pointing out that the role of speed as a self-contained construct might be useful for psychological and health assessment. It might be particularly promising to investigate the additional validity of the assessment by including speed in the prediction of external criteria.

An immediate implication for the follow-up adaptive design of the AM-PAC is that RT does not need to be included in interim θ estimation (i.e., selecting items during assessment delivery), but it could be used to improve the final θ estimates. Moreover, to further improve the time efficiency of adaptive testing, the maximum information per time unit (Fan et al., 2012) or its simplified version (Cheng et al., 2017) could be applied. In this case, the interviewer effect could be ignored when estimating an individual patient's speed, as long as item time parameters are provided. This is pragmatically sound because it is likely that different interviewers will be used for adaptive testing data collection in some measurement environments.

Due to the positive skewness of the RT distribution, typical log-transformations were used (van der Linden, 2007; Wang and Xu, 2015; Qian et al., 2016), and the raw RT data was cleaned by trimming the extremely short and long observations. However, recent research by Marmolejo-Ramos et al. (2015a) suggested that the Box-Cox transformation outperformed the elimination methods in normalizing positively skewed data. Vélez et al. (2015) proposed a new approach to estimate the parameter λ in the Box-Cox transformation. In cases in which the

TABLE 7 | Final Pearson correlation parameter estimates for the three models from two calibration stages.

Stage and Model	$\rho_{\theta_1\theta_2}$	$\rho_{\theta_1\theta_3}$	$\rho_{\theta_2\theta_3}$	ρ_1	ρ_2	ρ_3	γ_p
STAGE 1: CONCURRENT CALIBRATION (BATCH 2 TO 4)							
MGRM	0.624	0.468	0.846	–	–	–	–
Model 0	0.625	0.488	0.839	0.425	0.458	0.418	–
Model 2	0.628	0.492	0.840	0.583	0.629	0.578	(1.150, 1.053, 0.725, −0.911)
STAGE 2: FIXED PARAMETER CALIBRATION							
MGRM	0.702	0.545	0.869	–	–	–	–
Model 0	0.707	0.584	0.881	0.400	0.433	0.457	–
Model 2	0.706	0.583	0.880	0.527	0.577	0.605	(1.063, −0.286, 1.315, −0.135, 1.046)

log-transformation is insufficient, the Box-Cox transformation could be a viable alternative. In the present study, the extremely long and short RTs were trimmed because those RTs were considered as outliers. On the other hand, when there is lack of information on the outliers, Ueda's method could be used to automatically detect discordant outliers (Marmolejo-Ramos et al., 2015b). Because observed RTs could exhibit different skewed distributions, a careful decision needs to be made with respect to dealing with outliers, data transformation, and using the mean vs. the median, for making valid inferences (Rousselet and Wilcox, 2018). When the median is used, then quantile regression instead of a mean-based linear model should be considered instead.

AUTHOR CONTRIBUTIONS

CW contributed to constructing the ideas, explaining the results, and drafting the paper. DW contributed to the constructing the ideas and editing the draft. SS contributed to conducting the data cleaning, model fitting/analysis, and constructing tables/figures.

ACKNOWLEDGMENTS

This research was supported by the Eunice Kennedy Shriver National Institutes of Child Health and Human Development of the National Institutes of Health under Award Number R01HD079439 to the Mayo Clinic in Rochester, Minnesota through a subcontract to the University of Minnesota.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00051/full#supplementary-material

REFERENCES

Anstey, K. J., Wood, J., Lord, S., and Walker, J. G. (2005). Cognitive, sensory and physical factors enabling driving safety in older adults. *Clin. Psychol. Rev.* 25, 45–65. doi: 10.1016/j.cpr.2004.07.008

Bassili, J. N. (1996). "The how and why of response latency measurement in telephone surveys," in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, eds N. Schwarz and S. Sudman (San Francisco, CA: Jossey-Bass), 319–346.

Braver, T. S., and Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neurosci. Biobehav. Rev.* 26, 809–817. doi: 10.1016/S0149-7634(02)00067-2

Bridgeman, B., and Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *J. Educ. Measur.* 41, 137–148. doi: 10.1111/j.1745-3984.2004.tb01111.x

Cai, L. (2013). *flexMIRT® Version 2: Flexible Multilevel Multidimensional Item Analysis and Test Scoring [Computer Software]*. Chapel Hill, NC: Vector Psychometric Group.

Cheng, Y., Diao, Q., and Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behav. Res. Methods* 49, 502–512. doi: 10.3758/s13428-016-0712-6

Clancy, C., and Collins, F. S. (2010). Patient-centered outcomes research institute: the intersection of science and health care. *Sci. Trans. Med.* 2, 37cm18-37cm18. doi: 10.1126/scitranslmed.3001235

Cox, B., Blaxter, M., Buckle, A., Fenner, N., Golding, J., Gore, M., et al. (1987). *The Health and Lifestyle Survey. Preliminary Report of a Nationwide Survey of the Physical and Mental Health, Attitudes and Lifestyle of a Random Sample of 9,003 British Adults*. London: Health Promotion Research Trust.

Der, G., and Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: results from the United Kingdom Health and Lifestyle Survey. *Psychol. Aging* 21, 62. doi: 10.1037/0882-7974.21.1.62

Diao, Q., and van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Appl. Psychol. Meas.* 35, 398–409. doi: 10.1177/0146621610392211

Fan, Z., Wang, C., Chang, H.-H., and Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *J. Educ. Behav. Stat.* 37, 655–670. doi: 10.3102/1076998611422912

Fletcher, R. B., and Hattie, J. A. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychol. Sport Exerc.* 5, 423–446. doi: 10.1016/S1469-0292(03)00036-0

Fraley, R. C., Waller, N. G., and Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *J. Pers. Soc. Psychol.* 78, 350. doi: 10.1037/0022-3514.78.2.350

Galesic, M., and Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opin. Q.* 73, 349–360. doi: 10.1093/poq/nfp031

Gauggel, S., Rieger, M., and Feghoff, T. (2004). Inhibition of ongoing responses in patients with Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 75, 539–544.

Holden, R. R., and Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychol. Assess.* 4, 170. doi: 10.1037/1040-3590.4.2.170

Hsieh, C.-A., von Eye, A. A., and Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: the dynamic association between adolescents' social isolation and engagement with delinquent peers in the National Youth Survey. *Multivariate Behav. Res.* 45, 508–552. doi: 10.1080/00273171.2010.483387

Huang, M. E., Cifu, D. X., and Keyser-Marcus, L. (2000). Functional outcomes in patients with brain tumor after inpatient rehabilitation: comparison with traumatic brain injury. *Am. J. Phys. Med. Rehabil.* 79, 327–335. doi: 10.1097/00002060-200007000-00003

Hultsch, D. F., MacDonald, S. W., and Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *J. Gerontol.* 57, 101–115. doi: 10.1093/geronb/57.2.P101

Jiang, S., and Wang, C. (2019). "The different effects of collapsing categories on the graded response model and the generalized partial credit model," in *Paper to be Presented at the 2019 AERA Meeting* (Toronto, ON).

Jiang, S., Wang, C., and Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Front. Psychol.* 7:109. doi: 10.3389/fpsyg.2016.00109

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *J. Educ. Meas.* 43, 355–381. doi: 10.1111/j.1745-3984.2006.00021.x

Kolen, M. J., and Brennan, R. L. (2004). Test *Equating, Scaling, and Linking*. New York, NY: Springer-Verlag.

Loeys, T., Rosseel, Y., and Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika* 76, 487–503. doi: 10.1007/s11336-011-9211-y

Marmolejo-Ramos, F., Cousineau, D., Benites, L., and Maehara, R. (2015a). On the efficacy of procedures to normalize Ex-Gaussian distributions. *Front. Psychol.* 5:1548. doi: 10.3389/fpsyg.2014.01548

Marmolejo-Ramos, F., and González-Burgos, J. (2013). A power comparison of various tests of univariate normality on ex-Gaussian distributions. *Methodology* 9, 137–149. doi: 10.1027/1614-2241/a000059

Marmolejo-Ramos, F., Vélez, J., and Romão, X. (2015b). Automatic detection of discordant outliers via the Ueda's method. *J. Stat. Distrib. Appl.* 2: 8. doi: 10.1186/s40488-015-0031-y

Molenaar, D., Bolsinova, M., and Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *Br. J. Math. Stat. Psychol.* 71, 205–228. doi: 10.1111/bmsp.12117

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *Br. J. Math. Stat. Psychol.* 68, 197–219. doi: 10.1111/bmsp.12042

Muthén, L. K., and Muthén, B. O. (1998-2015). *Mplus User's Guide 7th Edn*. Los Angeles, CA: Muthén & Muthén

O'Dell, M. W., Barr, K., Spanier, D., and Warnick, R. E. (1998). Functional outcome of inpatient rehabilitation in persons with brain tumors. *Arch. Phys. Med. Rehabil.* 79, 1530–1534. doi: 10.1016/S0003-9993(98)90414-2

Osmon, D. C., Kazakov, D., Santos, O., and Kassel, M. (2018). Non-Gaussian distribuitonal analyses of reaction time: improvements that increase efficacy of RT tasks for describing cognitive processes. *Neuropsychol. Rev.* 28, 359–376. doi: 10.1007/s11065-018-9382-8

Pearson, K. (1924). *The Life Letters and Labours of Francis Galton: Volume II. Researches of Middle Life*. Cambridge: Cambridge University Press.

Pilkonis, P. A., Kim, Y., Yu, L., and Morse, J. Q. (2014). Adult Attachment Ratings (AAR): an item response theory analysis. *J. Pers. Assess.* 96, 417–425. doi: 10.1080/00223891.2013.832261

Qian, H., Staniewska, D., Reckase, M., and Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educ. Meas.* 35, 38–47. doi: 10.1111/emip.12102

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika* 78, 538–544. doi: 10.1007/s11336-013-9324-6

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., and Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika* 68, 589–606. doi: 10.1007/BF02295614

Rousselet, G. A., and Wilcox, R. R. (2018). *Reaction Times and Other Skewed Distributions: Problems With the Mean and the Median*. Available online at: https://www.biorxiv.org/content/biorxiv/early/2018/08/10/383935.full.pdf.

Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Appl. Stat.* 31, 115–124. doi: 10.2307/2347973

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr. Suppl.* 34(4 Pt. 2), 100. doi: 10.1007/BF03372160

Siem, F. M. (1996). The use of response latencies to self-report personality measures. *Military Psychol.* 8, 15–27. doi: 10.1207/s15327876mp0801_2

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* 19, 279–281. doi: 10.1214/aoms/1177730256

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *J. Educ. Behav. Stat.* 34, 378–394. doi: 10.3102/1076998609332107

Van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

van der Linden, W. J., Klein Entink, R. H., and Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Appl. Psychol. Meas.* 34, 327–347. doi: 10.1177/0146621609349800

Van Der Linden, W. J., Scrams, D. J., and Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Appl. Psychol. Meas.* 23, 195–210. doi: 10.1177/01466219922031329

Vélez, J. I., Correa, J. C., and Marmolejo-Ramos, F. (2015). A new approach to the Box–Cox transformation. *Front. Appl. Math. Stat.* 1:12. doi: 10.3389/fams.2015.00012

Verbruggen, F., Chambers, C. D., and Logan, G. D. (2013). Fictitious inhibitory differences: how skewness and slowing distort the estimation of stopping latencies. *Psychol. Sci.* 24, 352–362. doi: 10.1177/0956797612457390

Wang, C., Chang, H. H., and Douglas, J. A. (2013a). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

Wang, C., Fan, Z., Chang, H.-H., and Douglas, J. A. (2013b). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J. Educ. Behav. Stat.* 38, 381–417. doi: 10.3102/1076998612461831

Wang, C., Kohli, N., and Henn, L. (2016). A second-order longitudinal model for binary outcomes: item response theory versus structural equation modeling. *Struct. Equat. Model.* 23, 455–465. doi: 10.1080/10705511.2015.1096744

Wang, C., Su, S., and Weiss, D. J. (2018a). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate Behav. Res.* 53, 403–418. doi: 10.1080/00273171.2018.1455572

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Xu, G., and Shang, Z. (2018b). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Wang, C., Xu, G., Shang, Z., and Kuncel, N. (2018c). Detecting aberrant behavior and item preknowledge: a comparison of mixture modeling method and residual method. *J. Educ. Behav. Stat.* 43, 469–501. doi: 10.3102/1076998618767123

Yost, K., Jette, A., Wang, C., Weiss, D., and Cheville, A. (2018). "Computerized adaptive testing to direct delivery of hospital-based rehabilitation: item-bank generation and data collection," in *Poster Prestend at the 25th International Society for Quality of Life Research* (Dublin).

Zagorsek, H., Stough, S. J., and Jaklic, M. (2006). Analysis of the reliability of the leadership practices inventory in the item response theory framework. *Int. J. Select. Assess.* 14, 180–191. doi: 10.1111/j.1468-2389.2006.00343.x

Zickar, M. J., and Robie, C. (1999). Modeling faking good on personality items: an item-level analysis. *J. Appl. Psychol.* 84, 551. doi: 10.1037/0021-9010.84.4.551

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Wang, Weiss and Su. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 December 2018
doi: 10.3389/fpsyg.2018.02339

Modeling Learner Heterogeneity: A Mixture Learning Model With Responses and Response Times

Susu Zhang[1] *and Shiyu Wang*[2]*

[1] Department of Psychology, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [2] Quantitative Methodology Program, Department of Educational Psychology, University of Georgia, Athens, GA, United States

The increased popularity of computer-based testing has enabled researchers to collect various types of process data, including test takers' reaction time to assessment items, also known as response times. In recent studies, the relationship between speed and accuracy in a learning setting was explored to understand students' fluency changes over time in applying the mastered skills in addition to skill mastery. This can be achieved by modeling the changes in response accuracy and response times throughout the learning process. We propose a mixture learning model that utilizes the response times and response accuracy. Such a model accounts for the heterogeneities in learning styles among learners and may provide instructors with valuable information, which can be used to design individualized instructions. A Bayesian modeling framework is developed for parameter estimation and the proposed model is evaluated through a simulation study and is fitted to a real data set collected from a computer-based learning system for spatial rotation skills.

Keywords: response times, learning behaviors, diagnostic classification model, hidden markov model, mixture model

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Yang Liu,
University of Maryland, College Park,
United States
Daniel Bolt,
University of Wisconsin-Madison,
United States

***Correspondence:**
Shiyu Wang
swang44@uga.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *27 August 2018*
Accepted: *08 November 2018*
Published: *05 December 2018*

Citation:
Zhang S and Wang S (2018) Modeling Learner Heterogeneity: A Mixture Learning Model With Responses and Response Times. Front. Psychol. 9:2339. doi: 10.3389/fpsyg.2018.02339

1. INTRODUCTION

Educational researchers have shown long term interests in understanding the heterogeneity among online learners. Learners can differ not only in their initial background and general learning ability, but also in terms of how they learn. For example, learners' affects, that is the attitudes, interests, and values that learners exhibit, can influence their behaviors in the learning process and hence the learning outcomes. Methods were proposed by educational data miners to detect students' affects based on their interactions with the online learning systems (e.g., Baker et al., 2012). By identifying the affects of each student during the learning process, such as boredom, disengagement, confusion, and frustration, educators can provide targeted interventions accordingly to improve learning outcomes. Students can also vary in their preferred mode of instructions. Felder and Silverman (1988) developed the Index of Learning Styles survey, which measured learners' characteristics on the Sensing/Intuiting, Visual/Verbal, Active/Reflective, and Sequential/Global dimensions. A student's learning style can provide indications of possible strengths and difficulties in the learning process.

The increased popularity of computer-based testing has enabled researchers to collect various types of process data, including test takers' reaction time to assessment items, also known as response times. In the field of Psychometrics, extensive research has been conducted on the joint

modeling of response accuracy and response times (e.g., Thissen, 1983; van der Linden, 2006, 2007; Fox and Marianti, 2016). Findings from these studies demonstrated that incorporating the additional information from response times, in addition to response accuracy, can improve the estimation accuracy of item parameters and individuals' latent traits or latent classes, further our understanding of individuals' test-taking behavior and the test items' characteristics, and help differentiate learners using different test-taking strategies (e.g., Meyer, 2010; Wang and Xu, 2015). Most recently, response times have been used to measure students' improvements in skill mastery over time. An example is the work from Wang et al. (2018c), in which response times, together with response accuracy, were incorporated into a higher-order hidden Markov model framework (Wang et al., 2018b) to provide information about learners' mastery of the assessed skills, as well as their fluency of applying the mastered skills.

Wang et al. (2018c) assumed that all learners were engaged in the learning process, that is, they devoted their attention to the learning interventions and answered the assessment questions as correctly as possible. However, as mentioned in the very beginning, learners may have different learning styles. Assuming all learners to have the same learning style may under- or over-predict their learning outcomes. This current study aims to address this limitation with a mixture learning model with response times and response accuracy that can account for the presence of heterogeneities in learning styles among learners.

Response times have shown great potentials in identifying students' learning styles, especially student engagement. As an example, Henrie et al. (2015) provided a comprehensive review of methods for measuring student engagement in technology-based learning environments in the literature, and the time spent on homework, web pages, readings, et cetera were commonly used as an indicator of student engagement. Response times were also used by educational data miners to identify disengaged learners Beck (2004). A statistical approach to identify unobserved subpopulations in the data is by using mixture models. Mixture models have been widely used in psychometrics research, for example, addressing some practical issues in testing, such as identifying rapid-guessing or aberrant behaviors among test-takers (e.g., Wang and Xu, 2015), detecting compromised test items (e.g., McLeod et al., 2003), and modeling test-taking speed in time-constrained testing scenarios (e.g., Bolt et al., 2002). A lot of previous research considered the fit of mixture models to response and response time data collected from educational assessments. For example, Wise and DeMars (2006) proposed an effort-moderated IRT model, under which whether or not the response time of an examinee on a test item exceeds an item-specific threshold is used to infer if the subject has demonstrated efforts on the item, and Wang and Xu (2015) used different underlying response and response time distributions for item responses in different test-taking modes (e.g., solution, pre-knowledge, or rapid-guessing). However, modeling heterogeneity in learning behavior is more challenging than modeling that in testing behavior, as one needs to consider different measurement models as well as the transition models that describe the change of latent constructs over time. The proposed model, which will be described in details in the following section, is more closely related to the literature about Mixture Hidden Markov Models (HMMs). Langeheine and Van de Pol (1990) and Van de Pol and Langeheine (1990) proposed the mixed Markov latent class model, which, in its most general form, is the mixture of several first order hidden Markov models. It assumeed that different subpopulations differed in their initial state distributions, transition probabilities, and the response distributions under a HMM. Vermunt et al. (2008) further extended the mixed Markov latent class model to incorporate time-invariant or time-dependent covariates for each subject at each time point.

The mixture learning model proposed in this study adopts a similar framework for modeling the learners' behaviors in a learning process as that in the mixture HMMs. However, instead of assuming subpopulations of learners throughout the entire learning process, we assume that at each point in time, a learner can be in different learning modes. Furthermore, in addition to the item response data, learners' response times are also used in the measurement model, to measure both the change of learners' latent speed over time and any change in their engagement with the learning process.

The rest of the paper is organized as follows. A motivating example is first presented to demonstrate the utility of response times and response accuracy in the detection of heterogeneous learning behaviors in a computer-based learning program. This is followed by the presentation of the proposed mixture learning model and a Bayesian estimation procedure. We then present the results from fitting the proposed mixture model to the data described in the motivating example. A simulation study is presented to verify the accuracy of proposed estimation algorithm under different conditions and to validate the results from the real data analysis. We further discuss our findings and limitations of this study and propose future research directions in the last section.

2. A MOTIVATING EXAMPLE

This motivating example is presented to illustrate the necessity of using both response times and response accuracy to differentiate learners' behaviors in a learning environment. We start with presenting the results from an exploratory analysis on a data set collected through a spatial rotation learning program (Wang et al., 2018a). This learning program was developed on the basis of a pilot learning program in Wang et al. (2018b) to train four fine-grained mental rotation skills, namely (1) $x90$: 90° rotation along the x-axis; (2) $y90$: 90° y-axis; (3)$x180$: 180° x-axis; and (4) $y180$: 180° y-axis. Test questions in this new learning program were developed based on the ones in Wang et al. (2018b), and these four distinct yet related skills were identified as the set of measured skills by several previous studies (e.g., Maeda et al., 2013; Culpepper, 2015; Wang et al., 2018b). The structure of the learning program is summarized by the flow chart in **Figure 1**. Specifically, the learning program started with a testing module, followed by two consecutive learning modules, and finally ended with a testing module. Each module was composed of 10 different

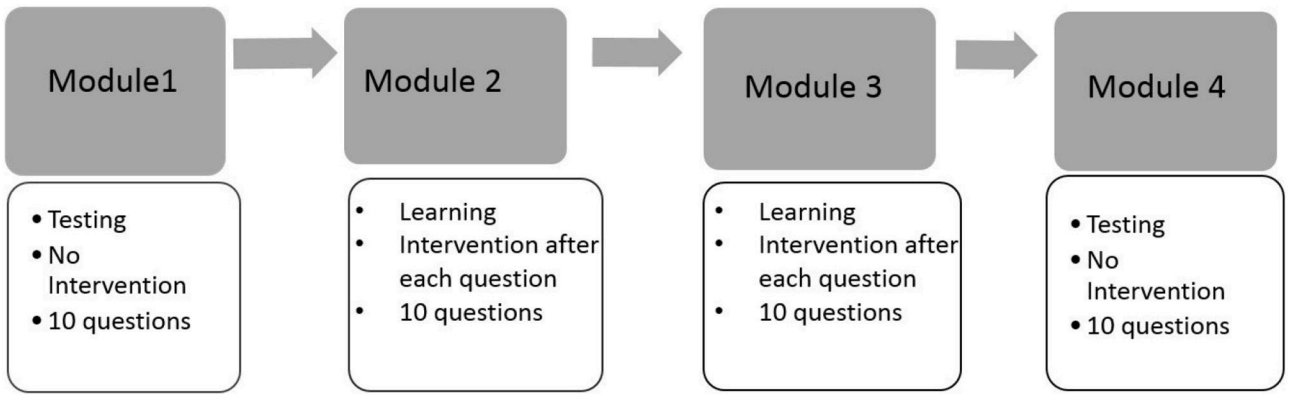


FIGURE 1 | The design of the spatial rotation learning program.

questions that were selected based on various criteria, including item characteristics and how well they assessed or improved the four skills. The main purpose of the two testing modules was to measure accurately the four binary spatial skills at a given point in time, while the two learning modules aimed to improve test-takers' mental rotation skills. Learning interventions were provided only in learning modules, in which participants were provided with learning materials after completing each question. A total of 585 undergraduate students with diverse backgrounds participated in the experiment. Written informed consent was obtained from the participants of this study. These students either received a course credit or a stipend through their participation. For students who received a stipend, their total amount of payment was proportional to the number of correct responses they provided in the experiment. Different incentive strategies may also trigger different learning patterns.

We first explore the data by plotting the log response times of all person-item combinations across four modules in **Figure 2**. It is observed that the response time distributions in modules 1, 2, and 4, especially module 4, have a bimodal structure: the first mode appeared within a short time period, while the second appeared at a later time. The previous studies that had similar observations in a testing environment concluded these two modes represent rapid-guessing and solution behaviors, and this is the evidence for a mixture of two populations with different response behaviors (e.g., Van der Linden and Guo, 2008; Wang et al., 2016). However, in a learning environment, the behavior of fast test-taking does not directly imply random guessing, as there is a confounding factor that the speed, especially in module 4, may be due to the improvement of cognitive skills after receiving learning interventions. To see this, we further identified the faster participants in module 4 and explored their module 4 test scores as well as their testing time and module 1 score. The reason to choose module 1 and 4 is because these two testing modules had similar item characteristics and can be regarded as parallel, thus we can compare the change in response accuracy and response times without worrying about the form effect. **Figure 3** documented the results from four participants. From there we can see that first, Participant 567 and Participant 145 almost had the same speed in modules 1 and 4. However, the former may represent a person with random guessing as he/she had low response accuracy in both modules, and the latter may represent one who mastered or was fluent in the four skills so that he/she can responded quickly while maintaining high accuracy (achieved a full score in each module). The behavior from Participant 576 may indicate this student had a solution behavior in module 1 but switched to random guessing in model 4. The response speed and response accuracy from Participant 383 both increased, and the increased speed may be due to the improvement of the spatial skills. Lastly, participants may switch engagement mode during the learning process. **Figure 4** further documents the examples of learning behaviors of three participants in this experiment. Across all four modules, participant 185 (left) responded to the questions with high speed and low accuracy, indicating he/she was not engaged during the whole experiment. Participant 78 (middle) seemed to be engaged in learning during the first 2 modules, however, his/her response accuracy sharply decreased in module 4 together with the total response time reaching a plateau, indicating he/she started to lose motivation in the last module. Participant 354 (right) presents another pattern, where he/she might not be engaged in the first module, but then switched to be engaged in the following modules. All these findings illustrate the necessity to use response times and response accuracy together to detect different learning behaviors.

3. MIXTURE LEARNING MODEL WITH RESPONSE TIMES AND RESPONSE ACCURACY

3.1. Model Formulation

We introduce the mixture modeling framework using the computer-based learning environment presented in section 2 as an example. It is assumed that N learners are trained to learn K skills at T time points, and that they are assessed with items developed under the Diagnostic Classification Model framework (DCM; also known as Cognitive Diagnosis Model). At time

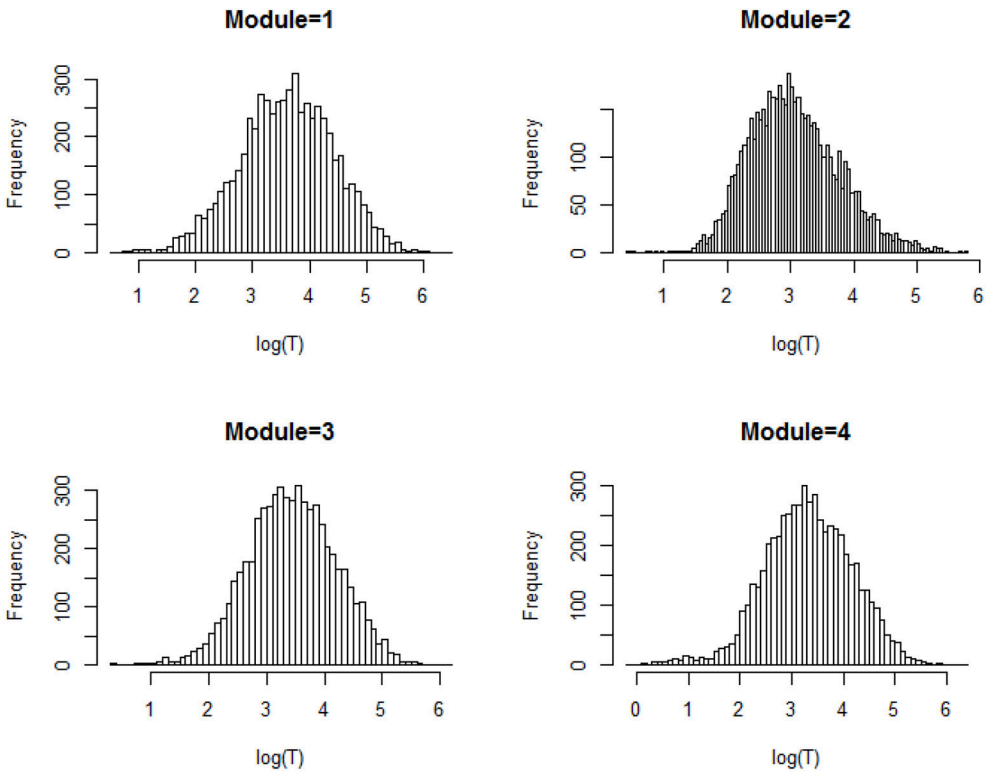


FIGURE 2 | Histogram of the log response time for all person-item combination across four modules.

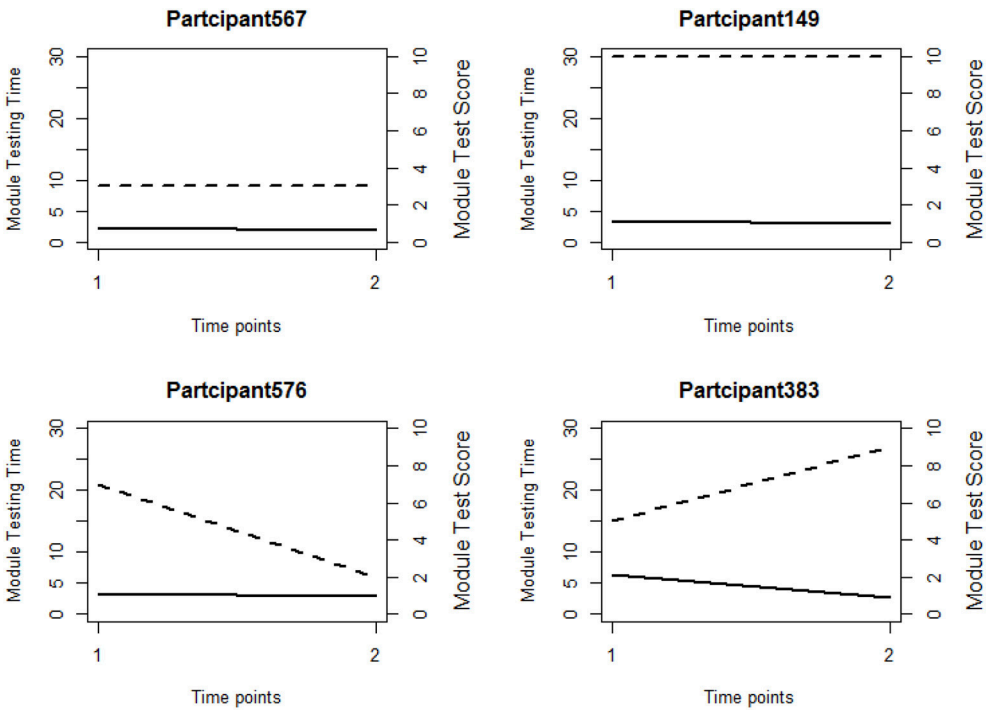


FIGURE 3 | Line plots of four participants' testing time and module scores in module 1 and module 4. The dashed line and solid line represent module score and testing time respectively. Time point 1 represents module 1 and Time point 2 represents module 4.

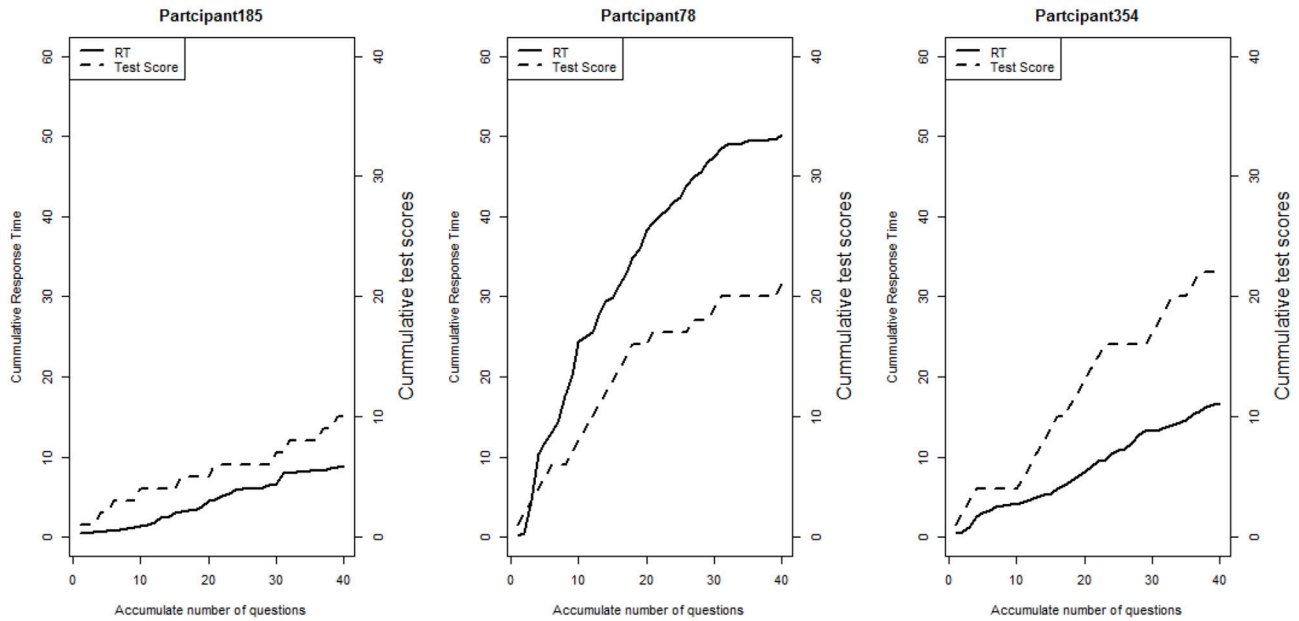


FIGURE 4 | Line plots of three participants' cumulative total scores and total response times at the end of each item.

point t, J_t questions are administered, and the skills measured by each question are documented through a $\mathbf{Q_t}$-matrix, with the j,kth element indexed by q_{jkt}, which equals 1 when item j requires attribute k and 0 otherwise. Let $\mathbf{X}_{it} = (X_{i,1,t}, ..., X_{i,J_t,t})'$ denote responses to the J_t questions from learner i at time t. $X_{i,j,t}$ takes a value of 1 or 0 depending on whether the response is correct or incorrect. The reaction times, or latencies, for these questions are denoted by $\mathbf{L}_{i,t} = (L_{i,1,t}, ..., L_{i,J_t,t})'$. For learner i, the latent skill profile at time t is denoted by $\boldsymbol{\alpha}_{i,t} = [\alpha_{i,1,t}, \ldots, \alpha_{i,K,t}]'$, with $\alpha_{i,k,t} = 1$ indicating mastery of a skill k and $\alpha_{i,k,t} = 0$ indicating non-mastery. Let $D_{i,t}$ be a binary variable that denotes the learning mode of learner i at time point t, with $D_{i,t} = 0$ for an engaged mode and $D_{i,t} = 1$ for a disengaged mode. In this study, we index the time points in the learning process at the module level, that is, each model is regarded as a time point, and a learner is assumed to have the same learning mode and attribute pattern across all items that are administrated at the same time point. We impose this assumption for the consideration of model simplicity, and a generalization of this assumption to item-level time indexing is provided in the discussion section as a future direction. Given the learner's engagement mode at a given time point $D_{i,t}$, the proposed mixture learning model considers the between-mode differences of the learners on the following three sub-models, namely (1) a transition model that captures the change of latent profile between two adjacent time points, (2) a measurement model that describes the distribution of item responses to the assessment questions at a given time point, and (3) a response time model that outlines the distribution of reaction times at a given time point. As the learner is assumed to have only two modes at a given time point, we will address the above three types of models based on whether the learner is in an engaged learning mode or a disengaged learning mode.

First, a learner in an engaged learning mode ($D_{i,t} = 0$) is assumed to employ relevant skills to respond to the assessment questions as accurately as possible. In this case, a reasonable DCM can be chosen as the measurement model. For example, if the deterministic input, noisy-"and"-gate (DINA; e.g., Macready and Dayton, 1977; Junker and Sijtsma, 2001) model is chosen, then the probability of a correct response on item j by learner i at time t is given by

$$P(X_{i,j,t} = 1 \mid \boldsymbol{\alpha}_i, s_j, g_j, \boldsymbol{q}_j) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{i,j,t}}, \qquad (1)$$

where $\eta_{i,j,t} = \prod_{k=1}^{K} \alpha_{i,k,t}^{q_{j,k,t}}$ is the ideal response, indicating whether learner i possesses all required skills to answer item j correctly, and s_j and g_j are the slipping and guessing parameters of item j. Essentially, the DINA model describes the case where a learner needs to master all requisite skills of an item to be able to answer the item correctly with high probability $(1 - s_j)$. Missing any of the item's requisite skills would result in a probability of a correct response of g_j instead. We note that while the DINA model is chosen in the present study, other DCMs can be chosen based on the specific assessment items in hand. This includes, for example, the deterministic input, noisy-"or"-gate model (Templin and Henson, 2006), the reduced reparameterized unified model (DiBello et al., 1995; Hartz, 2002; Roussos et al., 2007), and other general models, such as the log-linear cognitive diagnosis model (Henson et al., 2009), the general diagnostic model (von Davier, 2008), and the generalized-DINA model (de la Torre, 2011).

When a learner engages in solution behavior on an assessment item, we adopt the dynamic response time model proposed by

Wang et al. (2018c) to describe the distribution of the reaction time to this item. Specifically, $L_{i,j,t}$ is assumed to follow a log-normal distribution,

$$\log(L_{i,j,t}) \sim N\left(\gamma_j - (\tau_i + \phi * G_{i,j,t}), \frac{1}{a_j^2}\right). \quad (2)$$

where τ_i is the initial latent speed of learner i, γ_j is the time-intensity parameter of item j, capturing the overall amount of time the item requires, and a_j is the time-discrimination parameter of item j, which captures variance of log-response times at a given τ_i and γ_j. $G_{i,j,t}$ is a covariate defined according to the latent skill profile $\boldsymbol{\alpha}_{i,t}$, and ϕ is the parameter that characterizes the change of the latent speed due to $G_{i,j,t}$. The key part of such a dynamic response time model is the covariate $G_{i,j,t}$, which captures the change in speed of the subject over time as a function of the attribute trajectory of subject i, and here we use the indicator function for G proposed in Wang et al. (2018c), namely

$$G_{i,j,t} = \begin{cases} 1, & \text{if } \boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

With $G_{i,j,t}$ defined this way, a learner can take one of two speed statuses on each item: Depending on whether all the required skills of item j are mastered time t, his or her speed on the item is either τ_i or $\tau_i + \phi$. In terms of the transition probability, we make the assumption that a learner in the engaged mode also has high a engagement level in the learning process and thus may improve in skill mastery over time. In the engaged learning mode, the learner's transitions of attribute pattern from that time point to the next is hence modeled using a simplified version of the higher order hidden Markov DCM (HO-HM DCM) proposed by Wang et al. (2018b), specifically, the logit of the probability of transitioning from non-mastery to mastery on skill k at time $t+1$ is given by

$$\text{logit}[P(\alpha_{i,k,t+1} = 1 \mid \alpha_{i,k,t} = 0, \boldsymbol{\alpha}_{i,t})] = \lambda_0 + \lambda_1\theta_i + \lambda_2 \sum_{\forall k' \neq k} \alpha_{i,k',t}. \quad (4)$$

In this model, θ_i denotes the overall, time-invariant learning ability of learner i. The term $\sum_{\forall k' \neq k} a_{i,k',t}$ represents how many attributes learner i has already acquired other than attribute k at time t. By using a higher order logistic model for the transition probabilities in the hidden Markov model, the effect of different factors on the probability of learning a skill can hence be examined. A monotonicity assumption is also imposed in the current study, where the probability of forgetting a learned skill, $P(\alpha_{i,k,t+1} = 0 \mid \alpha_{i,k,t} = 1, \boldsymbol{\alpha}_{i,t})$, is 0.

On the other hand, if a learner is in a disengaged learning mode at time t, with $D_{i,t} = 1$, we assume this learner takes the rapid-guessing strategy on assessment items and shows low engagement in the learning process. We model their rapid-guessing strategy using similar methods as that in Wang and Xu (2015), where the probability of correctly responding to item j is equal to a parameter $g^* \in (0, 1)$ across all items, and the distribution of response times under the rapid-guessing strategy is also assumed to be the same across items, specifically,

$$\log(L_{i,j,t}) \mid D_{i,t} = 1 \sim N(\mu_1, \sigma_1^2), \quad (5)$$

where μ_1 and σ_1^2 are the mean and variance of the log-response times in the disengaged mode. The disengagement in the learning process is reflected in the transition probabilities from the current stage to the next. In other words, if a learner i is in the disengaged mode at time t, his or her attribute pattern at time $t+1$ is assumed to be unchanged from $\boldsymbol{\alpha}_{i,t}$. As a summary, **Table 1** presents the learning, response, and response time models for the learners under two different learning modes.

TABLE 1 | Components of the mixture learning model under different engagement modes.

Learning mode	Engaged ($D_{i,t} = 0$)	Disengaged ($D_{i,t} = 1$)
$P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}) =$	$\text{logit}[P(\alpha_{i,k,t+1} = 1 \mid \alpha_{i,k,t} = 0, \boldsymbol{\alpha}_{i,t})] = \lambda_0 + \lambda_1\theta_i + \lambda_2 \sum_{\forall k' \neq k} \alpha_{i,k',t}$	$\mathcal{I}(\boldsymbol{\alpha}_{i,t+1} = \boldsymbol{\alpha}_{i,t})$
$P(X_{i,j,t} = 1) =$	$(1 - s_j)^{\prod_{k=1}^{K} \alpha_{i,t,k}^{q_{j,k}}} g_j^{1-\prod_{k=1}^{K} \alpha_{i,t,k}^{q_{j,k}}}$	g^*
$\log(L_{i,j,t}) \sim$	$N\left(\gamma_j - (\tau_i + \phi * G_{i,j,t}), \frac{1}{a_j^2}\right)$	$N(\mu_1, \sigma_1^2)$

3.2. Bayesian Estimation

The proposed mixture learning model with engaged and disengaged modes is fitted under a Bayesian framework. We first outline the prior for each parameter in this modeling framework. Recall that $D_{i,t}$ denotes the membership of learner i at time t in terms of whether one is disengaged, where $D_{i,t} = 1$ if learner i is disengaged at time t, and $D_{i,t} = 0$ otherwise. We assume that

$$D_{i,t} \sim \text{Bernoulli}(\omega), \quad (6)$$

where ω is the probability an arbitrary learner belongs to the disengaged group, and the prior distribution of ω is

$$\omega \sim \text{Beta}(1, 1). \quad (7)$$

The initial attribute pattern of learner i is assumed to be a multinomial sample from all $C = 2^K$ possible classes, with

$$P(\boldsymbol{\alpha}_{i,1} = \alpha_c) = \prod_{c=1}^{C} \pi_c^{\mathcal{I}(\boldsymbol{\alpha}_{i,1} = \boldsymbol{\alpha}_c)}, \quad (8)$$

where a Dirichlet prior distribution for the initial probabilities of each attribute pattern is used,

$$\boldsymbol{\pi} = [\pi_1, \ldots, \pi_C] \sim \text{Dirichlet}(1, \ldots, 1). \quad (9)$$

At time $t \in \{1, \ldots, T-1\}$, if a learner is in the engaged learning mode with $D_{i,t} = 0$, his or her attribute pattern at the next time point, $\boldsymbol{\alpha}_{i,t+1}$, conditioning on the attribute pattern at time t is modeled using the HO-HM DCM in Equation (4). Similar to

Wang et al. (2018c), we used the following prior probabilities for the learning model parameters:

$$\lambda_0 \sim \text{Normal}(0,1), \quad \lambda_1 \sim \text{Log-normal}(0.5,1),$$
$$\lambda_2 \sim \text{Log-normal}(-0.5, 0.6^2). \tag{10}$$

If the learner is disengaged at time t with $D_{i,t} = 1$, $\boldsymbol{\alpha}_{i,t+1}$ is equal to $\boldsymbol{\alpha}_{i,t}$ with probability 1.

The responses of a learner under the engaged mode are assumed to follow the DINA model in Equation (1). A Beta prior was used for the slipping and guessing parameters of all the items, in other words,

$$p(s_j, g_j) \propto s_j^{a_s-1}(1-s_j)^{b_s-1} g_j^{a_g-1}(1-g_j)^{b_g-1}\mathcal{I}(0 \le g_j < 1-s_j \le 1). \tag{11}$$

On the other hand, the response to an item j by a learner in the disengaged mode is assumed to be a Bernoulli sample with success probability g^*, in other words, $P(X_{i,j,t} = 1 \mid D_{i,t} = 1) = g^*$, where g^* is assumed to have a Beta(1, 1) prior distribution.

At each time point $t = 1, \ldots, T$, if $D_{i,t} = 0$, subject i's response time on each item follow the log-normal distribution in Equation (2). Similar to that in Wang et al. (2018c), we use the following priors for the response time model parameters:

$$\gamma_j \sim N(0,1), \phi \sim N(0,1), \text{ and } a_j^2 \sim \text{Gamma}(1,1). \tag{12}$$

If $D_{i,t} = 1$, the reaction time to each item by learner i are assumed to follow the log-normal distribution given in Equation (5), with the following priors for the response time model parameters:

$$\mu_1 \sim N(0,1), \text{ and } \sigma_1^2 \sim \text{Inv-Gamma}(1,1). \tag{13}$$

Lastly, for each learner, his or her latent learning ability θ_i follows a standard normal prior distribution, and his or her initial latent speed τ_i in the engaged mode is assumed to follow a normal distribution with mean 0 and variance σ_τ^2, where the variance, σ_τ^2, has the Inverse-Gamma prior distribution:

$$\sigma_\tau^2 \sim \text{Inv-Gamma}(2.5,1). \tag{14}$$

The conditional distribution for each parameter can be derived based on the specified priors and the likelihood function of the observed data. The details on the full conditional distributions of the model parameters are presented in **Appendix I**. A Metropolis-Hastings within Gibbs sampler is developed to iteratively update the parameters by sampling from their conditional distributions. For θ_i and for $\boldsymbol{\lambda}$, their conditional distributions do not resemble any known families of distributions, and thus, Metropolis-Hastings (MH) steps are used to update these parameters. A special note for the MCMC algorithm is that when $D_{i,t,} = 1$, or in other words when a learner is disengaged, the proposed model assumes that the attribute pattern at the next time point, $\boldsymbol{\alpha}_{i,t+1}$, is the same as $\boldsymbol{\alpha}_{i,t}$. In this case, $\boldsymbol{\alpha}_{i,t}$ and $\boldsymbol{\alpha}_{i,t+1}$ share the same attribute pattern. When updating the $\boldsymbol{\alpha}_{i,t}$s sequentially for each learner, instead of sampling each $\boldsymbol{\alpha}_{i,t}$ separately, sets of consecutive $\boldsymbol{\alpha}_i$s with no transitions in between (e.g., $\boldsymbol{\alpha}_{i,t}$ and $\boldsymbol{\alpha}_{i,t+1}$, if $D_{i,t} = 1$) are sampled together, conditioning on the attribute pattern before the last transition, the learner's attribute pattern after the next transition, and the observed responses and response times at all time points where the underlying attribute pattern is the current one. For example, if student i is disengaged at time 1 and engaged at time 2, then the proposed model predicts that, by the assumptions of "no transition" under the disengaged learning mode, the student should have the same attribute pattern at times 1 and 2. Thus, the algorithm samples $\boldsymbol{\alpha}_{i,1}$ and $\boldsymbol{\alpha}_{i,2}$ together, conditioning on $\boldsymbol{\pi}$, $\boldsymbol{\alpha}_{i,3}$, and the observed item responses and response times at time 2. The detailed description of the MCMC algorithm for parameter estimation is given in **Appendix II**.

4. ANALYZING LEARNING BEHAVIORS IN A SPATIAL ROTATION LEARNING EXPERIMENT

In this section, we apply the proposed mixture learning model to analyze the data in the motivating example. To demonstrate the necessity of fitting this complex model, we in addition fitted two relatively simpler models, one is the model in Wang et al. (2018c), which is a joint model for response accuracy and response time without considering the mixture structure, and the other is an independent model that fit the response accuracy with the HOHM DCM (Wang et al., 2018b) and the response time with a static log-normal model. These three models all converged after 20,000 iterations based on the Gelman-Rubin proportional scale reduction factor (PSRF; Gelman et al., 2014), also known as $\hat{R}$. The last 25,000 iterations were thus used to provide estimates for model parameters. We compared these three models based on the joint Deviance Information Criteria (DIC) and posterior predictive checking. First, the joint DIC for the proposed mixture model is 223104.2, which is the smallest among the three models [joint (224690.7) and independent (226364.1)], indicating a better fit of the proposed mixture model compared with the two simpler models. The testing quantities used in the posterior predictive checking are the minima, maxima, and mean of the change score (total score in Module 4 minus that in Module 1) and change response time (testing time in Module 4 minus that in Module 1). The posterior predictive p-values for these quantities are documented in **Table 2**. In general, an extreme p-value (close to 0 or 1) implies that the model cannot be expected to capture this aspect of the data. Based on the results in **Table 2** we can conclude that the three models had a similar fit in terms of response accuracy. However, the mixture model had the best fit for the response time portion, as the other two models had extreme p-values for the three defined testing quantities. All these results demonstrate that the mixture learning model can improve the data-model fit compared with the two simpler models, and it is necessary to use this model to explore students' learning behaviors.

The average proportion of disengaged participants from the mixture learning model was estimated as $\hat{\omega} = 0.03$ ($SD = 0.004$), indicating on average, about 3% of participants were disengaged at each time point. The following analysis focuses on interpreting

TABLE 2 | Posterior predictive *p*-values for three testing quantities.

Model	Score Change			Time Change		
	Min	Mean	Max	Min	Mean	Max
Mixture	0.375	0.558	0.313	0.706	0.770	0.744
Joint	0.605	0.511	0.311	0.942	0.941	0.930
Independent	0.572	0.514	0.287	0.941	0.949	0.924

the learning behaviors and outcomes in the disengaged learning group and engaged learning group.

4.1. Disengaged Learning Group

Based on the estimated $\hat{D}_{i,t}$ for each participant i, a total of 41 participants were not engaged in at least one of the four time points. There were 11 different disengaged learning patterns, as shown in **Figure 5**. These patterns can be summarized by four types of disengaged learning behaviors. The first is the behavior that participants began as being engaged in answering questions and learning, but they then became disengaged during the learning process. Among participants with this pattern, a relatively large proportion of them were engaged in learning and testing during the first three modules, but switched to disengaged in the very last module. This could possibly explain the exploratory finding in section 2 that the bimodal structure of the log response time distribution in module 4 is more obvious than that in the other three modules. The second type of behavior can be characterized by the participants being disengaged at first and then switching to engagement in later modules. The third type of behavior is characterized by constant switching between disengaged and engaged modes during the learning process. The last type of behavior is complete disengagement throughout the four modules. These different disengaged behaviors may provide feedback on the learning program design. For example, for the participants who were not engaged in the last module, about 70% of them were estimated to have mastered all four skills after the third module. In the last module, participants may become attuned to the nature of the test or bored, which leads to disengagement. This indicates that varieties in testing questions could be enhanced to better attract their attention in the learning program. When participants was not engaged in answering questions, they randomly guessed the item correctly with probability $\hat{g}^* = 0.503(SD = 0.022)$. Their log response time distribution was estimated to follow a normal distribution with mean $\hat{\mu}_1 = 2.528(SD = 0.069)$ and variance $\hat{\sigma}_1^2 = 1.158(SD = 0.038)$. This translates to an expected response time of about 12.5 s per item when a learner is disengaged.

4.2. Engaged Learning Group

The posterior means (EAPs) and standard deviations (SDs) estimated with the MCMC algorithm for the coefficients of the transition model and the speed change rate in the response time model are summarized in **Table 3**. About 52.7% of the participants were estimated as masters of all four skills at the initial time point. In general, when a participant was in an engaged mode, the transition from non-mastery to mastery of a skill at one time point to the next is significantly and positively related to one's general learning ability ($\hat{\lambda}_1 = 2.757$) and the number of mastered skills ($\hat{\lambda}_1 = 0.286$). The speed change rate is estimated as -0.332, indicating participants on average tended to respond more slowly to questions if they mastered the required skills for a question than when they missed some required skills. However, this estimate is the average across all participants; a generalization is to allow each individual to have a different change rate, which could possibly detect the increased speed due to the change of latent skill.

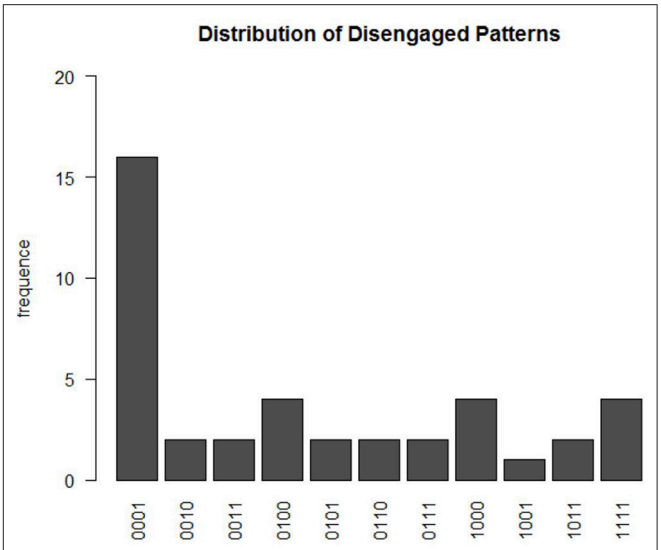


FIGURE 5 | The distribution of disengagement patterns. The x-axis represents the estimated summary pattern of $D_{i,t}$s at four time points, with 1 indicating disengaged and 0 as engaged.

TABLE 3 | The MCMC parameter estimates for the transition model and ϕ from the response time model.

Parameters	λ_0	λ_1	λ_2	ϕ
EAP	−2.214	2.757	0.286	−0.332
SD	0.323	0.781	0.119	0.028

The estimates are the averages across five chains.

The MCMC estimates for the item parameters, including the DINA model item parameters and the response time model item parameters, are documented in **Table 4**. The estimated DINA model item parameters are similar to the findings in Wang et al. (2018c) and Wang et al. (2018b), as these two learning programs share similar test questions. The average of the estimated time intensity parameters is 3.10, indicating participants in the engaged mode spent about 22.2 s answering a test question.

5. SIMULATION STUDY

A simulation study was conducted to achieve three goals. The first was to verify the accuracy of the proposed MCMC algorithm, the second was to provide validation for the real data analysis,

TABLE 4 | The MCMC parameter estimates for item parameters and response time parameters.

Item	s	g	a	γ
1	0.045 (0.014)	0.811 (0.043)	1.410 (0.044)	2.312 (0.040)
2	0.086 (0.018)	0.737 (0.044)	1.776 (0.056)	2.940 (0.037)
3	0.082 (0.019)	0.699 (0.037)	1.865 (0.059)	3.371 (0.035)
4	0.224 (0.026)	0.635 (0.033)	1.679 (0.054)	3.762 (0.035)
5	0.140 (0.023)	0.484 (0.039)	1.781 (0.055)	3.452 (0.034)
6	0.223 (0.025)	0.570 (0.037)	1.702 (0.053)	3.476 (0.035)
7	0.195 (0.026)	0.355 (0.040)	1.869 (0.058)	3.510 (0.033)
8	0.195 (0.025)	0.530 (0.035)	1.737 (0.055)	3.658 (0.034)
9	0.299 (0.029)	0.379 (0.036)	1.735 (0.058)	3.687 (0.035)
10	0.279 (0.029)	0.378 (0.035)	1.533 (0.047)	3.612 (0.037)
11	0.019 (0.008)	0.876 (0.039)	2.103 (0.067)	2.671 (0.034)
12	0.011 (0.006)	0.943 (0.019)	2.271 (0.074)	2.594 (0.033)
13	0.037 (0.010)	0.842 (0.043)	2.113 (0.067)	2.601 (0.033)
14	0.088 (0.015)	0.843 (0.038)	2.150 (0.071)	2.464 (0.034)
15	0.106 (0.014)	0.855 (0.029)	2.155 (0.071)	2.187 (0.035)
16	0.064 (0.015)	0.585 (0.042)	1.820 (0.057)	3.040 (0.035)
17	0.095 (0.019)	0.498 (0.047)	2.011 (0.066)	3.019 (0.035)
18	0.060 (0.013)	0.783 (0.034)	1.975 (0.062)	2.854 (0.034)
19	0.089 (0.016)	0.658 (0.040)	1.723 (0.053)	3.135 (0.036)
20	0.119 (0.019)	0.613 (0.042)	1.655 (0.051)	3.179 (0.037)
21	0.032 (0.010)	0.798 (0.051)	1.848 (0.058)	2.630 (0.035)
22	0.220 (0.022)	0.317 (0.043)	1.769 (0.055)	3.292 (0.035)
23	0.329 (0.025)	0.405 (0.045)	1.947 (0.060)	2.979 (0.034)
24	0.135 (0.019)	0.429 (0.065)	1.500 (0.046)	3.173 (0.040)
25	0.257 (0.024)	0.421 (0.049)	2.099 (0.065)	2.904 (0.034)
26	0.146 (0.020)	0.261 (0.042)	1.817 (0.056)	3.333 (0.035)
27	0.215 (0.023)	0.392 (0.041)	1.732 (0.053)	3.509 (0.036)
28	0.361 (0.026)	0.370 (0.040)	1.810 (0.056)	3.395 (0.034)
29	0.483 (0.026)	0.327 (0.041)	1.749 (0.054)	3.289 (0.035)
30	0.532 (0.026)	0.273 (0.037)	1.743 (0.055)	3.271 (0.036)
31	0.063 (0.013)	0.756 (0.060)	2.108 (0.070)	2.622 (0.034)
32	0.035 (0.009)	0.825 (0.044)	2.106 (0.071)	2.264 (0.035)
33	0.033 (0.009)	0.892 (0.030)	1.867 (0.059)	2.736 (0.036)
34	0.227 (0.022)	0.458 (0.049)	1.701 (0.053)	3.241 (0.036)
35	0.141 (0.019)	0.537 (0.048)	1.727 (0.054)	3.075 (0.036)
36	0.205 (0.022)	0.520 (0.049)	1.780 (0.057)	3.498 (0.036)
37	0.232 (0.023)	0.345 (0.043)	1.546 (0.049)	3.492 (0.038)
38	0.274 (0.024)	0.366 (0.044)	1.662 (0.052)	3.299 (0.037)
39	0.494 (0.027)	0.171 (0.034)	1.373 (0.042)	3.439 (0.041)
40	0.254 (0.024)	0.285 (0.044)	1.368 (0.043)	3.206 (0.041)

Standard errors in parentheses.

and the last was to demonstrate the necessity of modeling the heterogeneity of learning behaviors when they do exist. In order to achieve these goals, the proposed mixture learning model was chosen as the data generation model and the true model parameters were generated according to the estimated parameters from the real data analysis. Two additional factors were considered, one was sample size ($N = 585, 1{,}000, 3{,}000$) and the other was the overall probability of disengagement ($\omega =$ 0.03 or 0.10). Under each simulation condition, 50 data sets were simulated, and the proposed model was refitted through the MCMC algorithm. In addition, under each of the two $N =$ 385 conditions ($\omega = 0.03$ or 0.10), one data set generated from the mixture model was also fitted to the joint learning model of responses and response times under the HO-HM DCM framework proposed by Wang et al. (2018c). This assumes all learners are in the engaged mode across all time points, and the results from this model misspecification scenario can be used to demonstrate the third goal. The estimated parameters were then compared to the ones used to generate the data sets. The details of the simulation procedures and evaluation criteria are presented in the following subsection.

5.1. True Parameters

We simulated the attribute trajectories of $N = 585, 1{,}000$, or 3,000 learners on $K = 4$ skills across $T = 4$ time points. Ten assesment items were administred at each time point ($J_t = 10$). The learners' initial attribute patterns were randomly sampled from the set of all possible attribute profiles ($\{0, 1\}^K$), with probabilities of each profile set to be the expected a posteriori (EAP) estimates from the real data analysis. For each learner, their latent learning ability θ_i was randomly sampled from the standard normal distribution, and their latent speed τ_i was randomly generated from a normal distribution with mean 0 and variance σ_τ^2 estimated from the empirical data.

At each time point $t = 1, \ldots, T$, the learners were randomly assigned to one of two possible learning modes, namely the engaged learning mode ($D_{i,t} = 0$) and the disengaged learning mode ($D_{i,t} = 1$). The true probability of $D_{i,t} = 1$ was set to either $\omega = 0.03$ or $\omega = 0.1$, depending on the simulation condition. Then, conditioning on the learner's mode at time t, the attribute mastery changes, responses, and response times were simulated with different distributions. More specifically:

1. **Transition.** If at time t, learner i is in the engaged learning mode ($D_{i,t} = 0$), the probability that the learner transitions from non-mastery to mastery on a skill is given by the modified HO-HM DCM in Equation (4). Similar to Wang et al. (2018b), we assumed the monotonicity in the growth of attribute mastery, in other words, a mastered skill will not be forgotten. The true intercept (λ_0) and slopes (λ_1, λ_2) of the learning model were set to the EAP estimates from the empirical data analysis presented in **Table 3**. If learner i is disengaged at time t with $D_{i,t} = 1$, the learner's attribute pattern at the next time point, $\boldsymbol{\alpha}_{i,t+1}$, was set to be the same as the current one, $\boldsymbol{\alpha}_{i,t}$.
2. **Response.** When a learner is in the engaged learning mode at time t ($D_{i,t} = 0$), the learner is assumed to engage in the solution behavior, and the responses were simulated under the DINA model in Equation (1). The estimated slipping and guessing probabilities from the empirical data were used as the true parameters of the 40 items (**Table 4**). On the other hand, if the learner is disengaged at time t with $D_{i,t} = 1$, a rapid-guessing strategy is assumed and the learner's responses are generated from Bernoulli(g^*). Similar to the

other parameters, we set g^* equal to the EAP estimate from the real data analysis, which is 0.503.

3. **Response Times.** We assumed that when a learner is in the engaged learning mode, the observed response times follow the log-normal model in Equation (2), with $G_{i,j,t} = \mathcal{I}(\boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j)$, which takes the value 1 if learner i has mastered all requisite skills for item j by time t and 0 otherwise. For each assessment item, the empirically estimated time intensity parameter γ_j and time discrimination parameter a_j in **Table 4** were used as the true parameters in the simulation study and, similarly, the true value of the slope in front of the covariate $G_{i,j,t}$, ϕ was set equal to the EAP obtained from the real data, which is –0.332. If $D_{i,t} = 1$. In other words, learner i is disengaged at time t, the observed reaction time to any item at that time point was simulated from log-normal(μ_1, σ_1), again, the EAPs of μ_1 and σ_1 estimated from the real data were used as the true parameters.

5.2. Parameter Estimation

To start the MCMC, we first generated initial values of all the model parameters, and each of them was sequentially updated given the others from the conditional distributions in the **Appendix**. Specifically, the initial fixed parameters were generated as follows:

$$\begin{array}{lll} \lambda_0 \sim N(0,1), & \lambda_1 \sim U(0,1), & \lambda_2 \sim U(0,1), \\ \boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1}), & \phi \sim U(0,1), & \omega \sim U(0,0.2), \\ g^* \sim U(0,0.5), & s_j \sim U(0,0.3), & g_j \sim U(0,0.3), \\ \mu_1 \sim N(2,1), & \sigma_1 \sim U(0,1), & \gamma_j \sim N(3.45, 0.5^2), \\ a_j \sim U(2,4), & \sigma_\tau^2 \sim \text{Inv-Gamma}(1,1). & \end{array}$$

The random parameters, namely $\boldsymbol{D}, \boldsymbol{\alpha}, \boldsymbol{\theta}$ and $\boldsymbol{\tau}$, were then randomly generated based on the corresponding fixed parameters.

A chain length of 30,000 iterations was used for the MCMC, with the first 5,000 as the burn-in that were excluded for the computation of the point estimates of the parameters. From the post burn-in iterations, we calculated the expected a posteriori (EAP) estimates of each of the model parameters by taking the average of the parameter samples. For the discrete model parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{D}$, the final point estimates were dichotomized depending on whether the associated post burn-in average was $<$ or $>$ 0.5.

5.3. Evaluation Criteria

The performance of the proposed algorithm is evaluated in terms of two aspects. The first is to evaluate the convergence of the MCMC algorithm. Five separate chains with different starting values were run with chain lengths of 30,000 iterations under the $N = 585, \omega = 0.1$ condition, based on one randomly simulated data set. The $\hat{R}$ (Gelman et al., 2014) was calculated for each parameter at different chain lengths, with the first half of the chain as the burn-in, and the progression of the maximum $\hat{R}$ out of all estimated parameters was used to determine an adequate chain length for convergence. The second was to evaluate the ability of the proposed algorithm to accurately recover the true parameters. The following indices were used to evaluate different parameters in the model. Specifically, the recovery of the learners' attribute patterns of at each time point was evaluated using the attribute-wise agreement rate, $AAR = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}\mathcal{I}(\alpha_{ikt}=\hat{\alpha}_{ikt})}{N\times K}$, and the pattern-wise agreement rate, $PAR = \frac{\sum_{i=1}^{N}\mathcal{I}(\boldsymbol{\alpha}_{i,t}=\hat{\boldsymbol{\alpha}}_{i,t})}{N}$, between the true ($\boldsymbol{\alpha}$) and estimated ($\hat{\boldsymbol{\alpha}}$) attribute patterns. Note that the learners who were estimated as not engaged in any of the four time points were excluded from calculating these two indexes, as no information was available to provide estimates for their latent profile at each time point. We further evaluated the recovery of $\phi, \sigma_\tau^2, \boldsymbol{\pi}, \boldsymbol{\lambda}, \omega, \mu_1, \sigma_1$, and g^* by comparing the mean and standard deviation of the posterior parameter samples to the true values. The agreement between true and estimated response time model parameters ($\mathbf{a}$ and $\boldsymbol{\gamma}$), learning ability ($\boldsymbol{\theta}$), and latent speed ($\boldsymbol{\tau}$) was evaluated in terms of the correlation between true and estimated values, and similarly for $\mathbf{a}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{s}$, and $\mathbf{g}$. Note that for each learner, the data used to update θ are the transitions from non-mastery to either non-mastery or mastery at the next time point. Therefore, once a learner becomes a master of all skills, the subsequent $\boldsymbol{\alpha}$s will not provide additional information on θ, and no data on the transitions are available for learners who have mastered all skills at the very beginning. For this reason, the learners whose estimated initial attribute pattern was (1, 1, 1, 1) were excluded from the computation of the correlation between true and estimated learning abilities. The last index is the sensitivity and specificity of the detection of disengagements, that is, the proportion of times that true disengagement is correctly detected, which is defined as $\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{I}(\hat{D}_{i,t}=1, D_{i,t}=1)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{I}(D_{i,t}=1)}$), and the proportion of times that true engagement is correctly identified, which is defined as $\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{I}(\hat{D}_{i,t}=0, D_{i,t}=0)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathcal{I}(D_{i,t}=0)}$.

5.4. Results

5.4.1. Parameter Convergence

Figure 6 presents the change of the maximum univariate $\hat{R}$ among all model parameters as chain length increases. From the figure, we observe that after approximately 2,000 iterations, the maximum $\hat{R}$ fell below 1.2, and that at around 5,000 iterations, $\hat{R}$ has fully stabilized, indicating chain convergence.

5.4.2. Parameter Recovery

Table 5 presents the attribute-wise agreement rates (AARs) and the pattern-wise agreement rates (PARs) between the true and estimated attribute patterns ($\boldsymbol{\alpha}$) at each time point, under different disengagement rate (ω) and sample size (N) conditions. Across all conditions and time points in the learning process, the proposed estimation algorithm achieved over 85% accuracy in measuring the presence/absence of attributes for each participant. The estimation accuracy was the lowest for the initial time point ($t = 1$), and it increased as t increased, achieving over 90% agreement at $t = 4$. We also observed slightly higher accuracy in the $\boldsymbol{\alpha}$ estimates when sample size was larger and when the probability of disengagement was lower.

In **Table 6**, we present the biases and RMSEs of the fixed parameters in the model and the sensitivity and specificity of the learning mode estimates ($D_{i,t}$) averaged across 50 replications. Specifically, these fixed parmeters include the transition model's

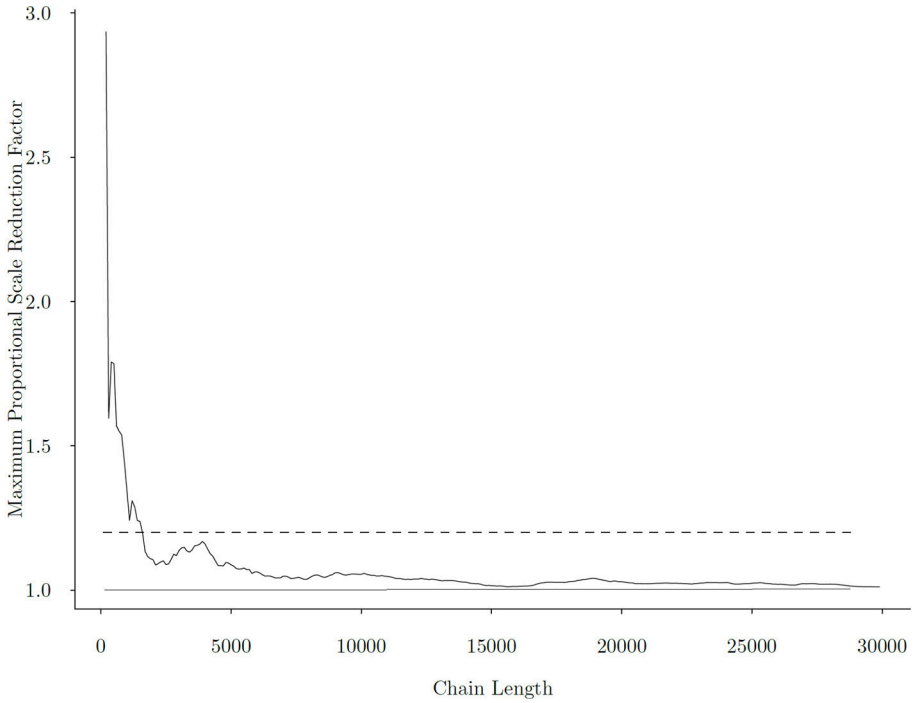


FIGURE 6 | Maximum Gelman-Rubin Proportional Scale Reduction Factor across all parameters with different chain lengths. The $x-$axis is the length of the MCMC chain, and the $y-$axis is the maximum PSRF. The dashed line represents the commonly used threshold of $\hat{R} = 1.2$ for parameter convergence, and the solid line corresponds to $\hat{R} = 1$, the minimum $\hat{R}$ that can be achieved.

TABLE 5 | The averaged attribute-wise and pattern-wise agreement rates (AARs and PARs) between the true and estimated $\boldsymbol{\alpha}$ across 50 repetitions under each simulation condition.

ω	N	Criteria	$t = 1$	$t = 2$	$t = 3$	$t = 4$
0.03	585	AAR	0.872	0.910	0.923	0.923
		PAR	0.683	0.745	0.784	0.793
	1,000	AAR	0.875	0.912	0.926	0.926
		PAR	0.688	0.749	0.789	0.798
	3,000	AAR	0.877	0.913	0.927	0.928
		PAR	0.696	0.752	0.792	0.800
0.10	585	AAR	0.864	0.901	0.916	0.915
		PAR	0.666	0.726	0.769	0.776
	1,000	AAR	0.869	0.903	0.917	0.918
		PAR	0.678	0.732	0.770	0.782
	3,000	AAR	0.872	0.905	0.919	0.919
		PAR	0.688	0.738	0.775	0.784

intercept (λ_0) and slopes (λ_1, λ_2), the correct response probability in the disengaged mode (g^*), the probability of disengagement (ω), the mean (μ_1) and standard deviation (σ_1) of the log response times in the disengaged mode, the coefficient for the increase of latent speed (ϕ) for engaged learners, and the variance of latent speed (σ_τ^2). Across all conditions, the bias of the estimated fixed parameters, except those associated with the transition model ($\boldsymbol{\lambda}$), were relatively small, with small RMSEs. One possible reason for the relatively large bias and RMSE for $\boldsymbol{\lambda}$ is that with $T = 4$, each learner could be observed on at most 3 transitions, and considering that a large proportion of learners started with mastery of all or most of the skills at the initial time point and that some learners might be disengaged at a selection of time points, the actual number of observations for transitions is usually <3 per learner. Thus, the amount of data available for estimating the transition model parameters, as well as the θs, is limited. We further observed that larger sample sizes were associated with slightly lower bias and standard error of the parameter estimates. In addition, a higher rate of disengagement ($\omega = 0.1$) was associated with larger biases and RMSEs of learning model parameter ($\boldsymbol{\lambda}$) and ϕ estimates, but smaller biases and RMSEs of g^* and μ_1, σ_1, the parameters associated with the response and response time distributions in the disengaged mode. This trend is expected, as a higher ω translates to a larger number of observations associated with disengagement and less observations associated with engagement.

Across several repetitions of the simulation study, the estimated learning mode of each learner at each time point, $D_{i,t}$, showed high agreement with the true values, with sensitivity over 95% when $\omega = 0.03$ and over 96% when $\omega = 0.1$, and specificity over 99% across all simulation conditions. This suggests that under the proposed estimation algorithm, whether a learner is disengaged or engaged at a given time point could be detected correctly most of the times based on their response times, responses, and transitions in attribute mastery.

The correlation between true and estimated values of $\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{a}, \boldsymbol{\gamma}, \mathbf{s}$, and $\mathbf{g}$ are presented in **Table 7**. For the items'

TABLE 6 | The bias and RMSE of the fixed parameter estimates under different simulation conditions and the specificity and sensitivity of the $D_{i,t}$ estimates.

True ω	N	λ_0 (RMSE)	λ_1 (RMSE)	λ_2 (RMSE)	g^* (RMSE)
0.030	585	0.562 (0.631)	−0.423 (0.775)	0.043 (0.145)	0.001 (0.026)
0.030	1,000	0.430 (0.499)	−0.408 (0.626)	0.037 (0.135)	0.001 (0.022)
0.030	3,000	0.231 (0.307)	−0.204 (0.352)	−0.004 (0.106)	0.003 (0.013)
0.100	585	0.628 (0.691)	−0.556 (0.837)	0.059 (0.153)	−0.000 (0.014)
0.100	1,000	0.479 (0.546)	−0.359 (0.646)	0.037 (0.137)	0.001 (0.011)
0.100	3,000	0.230 (0.313)	−0.225 (0.388)	0.014 (0.119)	0.001 (0.007)
True ω	**N**	**ω (RMSE)**	**μ_1 (RMSE)**	**σ_1 (RMSE)**	**ϕ_0 (RMSE)**
0.030	585	0.000 (0.005)	−0.000 (0.066)	0.003 (0.042)	−0.017 (0.031)
0.030	1,000	0.000 (0.004)	−0.001 (0.048)	−0.004 (0.033)	−0.010 (0.023)
0.030	3,000	0.000 (0.003)	−0.001 (0.030)	−0.000 (0.022)	−0.003 (0.012)
0.100	585	0.001 (0.008)	−0.000 (0.034)	0.000 (0.023)	−0.023 (0.033)
0.100	1,000	−0.001 (0.007)	−0.001 (0.027)	−0.002 (0.019)	−0.005 (0.022)
0.100	3,000	0.000 (0.004)	−0.003 (0.014)	0.001 (0.010)	−0.005 (0.014)
True ω	**N**	**σ_τ^2 (RMSE)**	***DP*: Sensitivity**	***D*: Specificity**	
0.030	585	0.005 (0.015)	0.952	0.999	
0.030	1,000	0.003 (0.010)	0.954	0.999	
0.030	3,000	0.001 (0.005)	0.954	0.999	
0.100	5,85	0.005 (0.015)	0.967	0.996	
0.100	1000	0.002 (0.010)	0.965	0.996	
0.100	3,000	0.001 (0.005)	0.967	0.996	

Values outside the parenthesis for $\boldsymbol{\lambda}, g^*, \omega, \mu_1, \sigma_1, \phi_0$, and σ_τ^2 are the biases of the parameter estimates averaged across 50 replications. Values in the parenthesis are the RMSEs of the parameter estimates averaged across 50 replications.

TABLE 7 | Correlations between true and estimated latent learning ability ($\boldsymbol{\theta}$) and initial speed ($\boldsymbol{\tau}$) of learners, item response time model parameters ($a, \boldsymbol{\gamma}$), and DINA model item parameters (s, g).

ω	N	ρ_θ	ρ_τ	ρ_a	ρ_γ	ρ_s	ρ_g
0.03	585	0.663	0.968	0.967	0.998	0.989	0.982
	1,000	0.670	0.968	0.982	0.999	0.993	0.989
	3,000	0.669	0.968	0.994	1.000	0.998	0.997
0.10	585	0.643	0.964	0.967	0.998	0.988	0.982
	1,000	0.647	0.964	0.979	0.999	0.993	0.989
	3,000	0.645	0.965	0.993	1.000	0.997	0.996

response time model parameters ($\mathbf{a}, \boldsymbol{\gamma}$), the DINA model parameters ($\mathbf{s}, \mathbf{g}$), and the learners' initial latent speeds ($\boldsymbol{\tau}$), there was a high agreement between the true and estimated values, with correlations over 96%. For the latent learning abilities of the learners ($\boldsymbol{\theta}$), the estimate values demonstrated larger errors with correlations around 0.67 when $\omega = 0.03$ and around 0.64 when $\omega = 0.1$. Similar to the larger errors in the transition model parameter estimates, we think the larger error in the estimation of $\boldsymbol{\theta}$ can potentially be attributed to the paucity of data available to update θ_i for each subject.

5.4.3. Consequences of Misspecification

Finally, we briefly summarize the parameter recovery results when the model is misspecified, that is, when the data generating model is the mixture model but the mixture structure is ignored when refitting data. We note that this is a special case of the proposed mixture model with $D_{i,t} = 0$ for all i and t.

Table 8 presents the summary of the parameter recovery results when the model without mixture is fitted to the data generated from the mixture learning model, with different true disengagement probabilities ($\omega = 0.03$ or 0.10). In both cases, a sample size of $N = 585$ was used. We present the correlations between the true and estimated $\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{a}, \boldsymbol{\gamma}, \mathbf{s}$, and $\mathbf{g}$. In addition, we also present the averaged attribute agreement rate (AAR) between true and estimated $\boldsymbol{\alpha}$ across the four stages.

Compared to when the mixture is explicitly modeled, ignoring the mixture in the data resulted in remarkable decreases in the estimation accuracy of $\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{a}$, and the attribute trajectories of the learners, $\boldsymbol{\alpha}$. The decrease in estimation accuracy is more salient when the proportion of disengagement is higher. Thus, we conclude that when learner disengagement exists in the learning process, assuming that all learners are engaged could greatly sabotage the model parameter estimates, including the estimates of the learner's skill mastery patterns and latent traits.

In addition to the recovery of the true model parameters under model misspecification, we also compare the model-data fit of the missepcified model and that of the mixture model. As a reminder, these two models were fitted to the response and response times data generated under the mixture condition with $N = 585$ and $\omega = 0.03$. On the same data set, the DIC obtained from the mixture learning model and the misspecified model was 223269.1

TABLE 8 | Recovery of the model parameters when the mixture in the data is ignored.

True ω	ρ_θ	ρ_τ	ρ_a	ρ_γ	ρ_s	ρ_g	$\bar{AAR}$
0.03	0.561	0.904	0.926	0.997	0.983	0.958	0.859
0.10	0.426	0.815	0.827	0.993	0.979	0.965	0.802

and 226197.3, respectively. This suggests that when a mixture structure does exist in the observed data, the model without the mixture fits significantly worse than the mixture model.

6. DISCUSSION

In this paper, we propose a mixture learning modeling framework which can address the heterogeneity in learning behaviors. A simple model with two possible learning modes, namely the engaged mode and the disengaged mode, motivated by a real data analysis on a computer-based learning program, is provided as an example. Specifically, with this model, learners are assumed to demonstrate different learning and response behaviors under different modes, leading to differences in the distributions of attribute mastery transitions over time, item responses, and response times. A Bayesian estimation procedure is established to estimate the parameters of the mixture learning model. Different learning behaviors were discovered by applying the proposed model to the real data from the spatial rotation learning program. Simulation studies showed that the model parameters could be accurately estimated, the learners' learning mode could be detected with high accuracy, and the Markov chains stabilized within 5,000 iterations. In addition, the simulation results from the model misspecification scenario suggested the necessity of fitting the proposed mixture learning model instead of a homogeneous learning model when data suggest the existence of a mixed structure of learning modes.

The proposed mixture learning model has the potential to detect learner disengagement in an online learning context. Compared to traditional classroom learning, online learning programs often provide the learners with a significantly more flexible and less controlled environment. Whereas, instructors in traditional classrooms can directly observe the learners' behaviors and their reactions to different interventions, in online learning, the educators do not interact face to face with the learners. This mixture learning model framework provides a way for educators to infer the online learners' learning mode (e.g., engaged or disengaged) and their corresponding latent skills based on the observed responses and reaction times to assessment questions at different time points. This can help the educators to provide different stimuli to different learners through the online learning environment. Furthermore, the proposed model can also help to refine and design individualized learning materials. As demonstrated from the real data analysis, learners may become disengaged at a certain stage of the learning process, and if this can be detected, then different types of learning materials can be delivered so that it does not make the learning tasks boring or transparent. Finally, even though illustrated within a DCM framework, the way to model the engaged and disenaged learning behavior can be generalized to other latent variable models based on specific assessment requirements. For example, if a continuous latent trait is assumed to be measured by the assessment, then a traditional Item Response Theory Model can be used for response accuracy. The latent growth model can be used to describe the change of the continuous latent trait.

Though promising, the proposed mixture learning model has the limitations that it only considers two learning modes and it assumes the learning mode is the same for all items in the same module. These restrictions can all be relaxed in future studies, in which more than two learning modes can be considered to differentiate various types of disengagement or to capture other learning behaviors other than engagement and disengagement, such as a warm-up mode, where students have low familiarity with the learning environment and need some time to adjust before fully engaging. We can also consider the learners' modes and attribute patterns at a finer grain size, such as treating the response to each item as a time point. Another direction is to consider a higher order model that describes the probability that a learner is disengaged at a specific time point, given a set of time dependent or time independent covariates, such as learners' demographic information or other characteristics, the mode of instruction (e.g., video, text, interactive exercise), or the temporal position of the current learning block (e.g., first learning block which may show slow warm-up of the learners, or later learning blocks on which learners may demonstrate fatigue). Lastly, a prior sensitivity analysis needs to be conducted in the future to investigate the sensitivity of the model estimation results to the prior specification.

ETHICS STATEMENT

The experiment related to the spatial rotation learning program was approved by the Institutional Review Board, office of the Vice President for Research at University of Georgia (IRB ID: STUDY00004215). Before the experiment, participants received a consent form that explains the purpose of this research study, the study procedure, Risks and discomforts, benifits, Incentives for participation, Privacy / Confidentiality. They were totally voluntary to participant in this experiment.

DATA AVAILABILITY STATEMENT

The real data set analyzed in this manuscript is not publicly available, because it is part of an ongoing research project from the research team. Requests to access the dataset should be directed to Dr. Shiyu Wang, swang44@uga.edu.

AUTHOR CONTRIBUTIONS

SZ contributed to the development of the methodological framework and the model estimation procedures, the conduction of the simulation studies, and the drafting and revision of the manuscript. SW contributed to providing suggestions on the methodological framework and the model estimation procedures, the conduction of the real data analyses, and the drafting and revision of the manuscript.

FUNDING

This research is partially supported by the Junior Faculty Seed Grant from the Owens Institute for Behavioral Research, The University of Georgia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02339/full#supplementary-material

REFERENCES

Baker, R. S., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., et al. (2012). Towards sensor-free affect detection in cognitive tutor algebra. *Int. Educ. Data Mining Soc.* 126–133.

Beck, J. E. (2004). "Using response times to model student disengagement," in *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments* (Maceio), 13–20.

Bolt, D. M., Cohen, A. S., and Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: application of a mixture rasch model with ordinal constraints. *J. Educ. Meas.* 39, 331–348. doi: 10.1111/j.1745-3984.2002.tb01146.x

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *J. Educ. Behav. Stat.* 40, 454–476. doi: 10.3102/1076998615595403

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995). "Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques," in *Conference on Alternative Diagnostic Assessment,* (Iowa City; IA), 361–389.

Felder, R. M., and Silverman, L. K. (1988). Learning and teaching styles in engineering education. *Eng. Educ.* 78, 674–681.

Fox, J. P., and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, vol. 2. Boca Raton, FL: CRC Press.

Hartz, S. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Henrie, C. R., Halverson, L. R., and Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: a review. *Comput. Educ.* 90, 36–53. doi: 10.1016/j.compedu.2015.09.005

Henson, R., Templin, J., and Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Langeheine, R., and Van de Pol, F. (1990). A unifying framework for markov modeling in discrete space and discrete time. *Sociol. Methods Res.* 18, 416–441. doi: 10.1177/0049124190018004002

Macready, G. B., and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *J. Educ. Stat.* 2, 99–120. doi: 10.3102/10769986002002099

Maeda, Y., Yoon, S. Y., Kim-Kang, G., and Imbrie, P. (2013). Psychometric properties of the revised PSVT:R for measuring first year engineering students' spatial ability. *Int. J. Eng. Educ.* 29, 763–776.

McLeod, L., Lewis, C., and Thissen, D. (2003). A bayesian method for the detection of item preknowledge in computerized adaptive testing. *Appl. Psychol. Meas.* 27, 121–137. doi: 10.1177/0146621602250534

Meyer, J. P. (2010). A mixture rasch model with item response time components. *Appl. Psychol. Meas.* 34, 521–538. doi: 10.1177/0146621609355451

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., and Templin, J. L. (2007). "The fusion model skills diagnosis system," in *Cognitive Diagnostic Assessmentfor Education: Theory and Applications*, eds J. P. Leighton and M. J. Gierl (New York, NY: Cambridge University Press), 275–318.

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287. doi: 10.1037/1082-989X.11.3.287

Thissen, D. (1983). "Timed testing: An approach using item response theory," in *New Horizons in Testing*, ed D. J. Weiss (New York, NY: Academic Press), 179–203.

Van de Pol, F., and Langeheine, R. (1990). Mixed markov latent class models. *Sociol. Methodol.* 20, 213–247. doi: 10.2307/271087

van der Linden, W. J. (2006). A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

Van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

Vermunt, J. K., Tran, B., and Magidson, J. (2008). "Latent class models in longitudinal research," in *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, (ed.) S. Menard (Burlington, MA: Elsevier), 373–385.

Vermunt, J.K., Tran, B., and Magidson, J. (2008). Latent class models in longitudinal research. In: Menard, S. (ed.), Handbook of Longitudinal Research: Design, Measurement, and Analysis, pp. 373-385. Burlington, MA: Elsevier.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* 61, 287–307. doi: 10.1348/000711007X193957

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Xu, G., and Shang, Z. (2016). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Wang, S., Carr, M., and Wang, Q. (2018a). "Diagnostic assessment with learning tools to improve the 3d spatial rotation skills," in *The National Council on Measurement in Education* (New York, NY).

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018b). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

Wang, S., Zhang, S., Douglas, J., and Culpepper, S. (2018c). Using response times to assess learning progress: a joint model for responses and response times. *Meas. Interdiscip. Res. Perspect.* 16, 45–58. doi: 10.1080/15366367.2018.1435105

Wise, S. L., and DeMars, C. E. (2006). An application of item response time: the effort-moderated irt model. *J. Educ. Meas.* 43, 19–38. doi: 10.1111/j.1745-3984.2006.00002.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Zhang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 September 2018
doi: 10.3389/fpsyg.2018.01525

Modeling Nonlinear Conditional Dependence Between Response Time and Accuracy

Maria Bolsinova[1]* *and Dylan Molenaar*[2]

[1] ACTNext, ACT, Inc., Iowa City, IA, United States, [2] University of Amsterdam, Amsterdam, Netherlands

The most common process variable available for analysis due to tests presented in a computerized form is response time. Psychometric models have been developed for joint modeling of response accuracy and response time in which response time is an additional source of information about ability and about the underlying response processes. While traditional models assume conditional independence between response time and accuracy given ability and speed latent variables (van der Linden, 2007), recently multiple studies (De Boeck and Partchev, 2012; Meng et al., 2015; Bolsinova et al., 2017a,b) have shown that violations of conditional independence are not rare and that there is more to learn from the conditional dependence between response time and accuracy. When it comes to conditional dependence between time and accuracy, authors typically focus on positive conditional dependence (i.e., relatively slow responses are more often correct) and negative conditional dependence (i.e., relatively fast responses are more often correct), which implies monotone conditional dependence. Moreover, most existing models specify the relationship to be linear. However, this assumption of monotone and linear conditional dependence does not necessarily hold in practice, and assuming linearity might distort the conclusions about the relationship between time and accuracy. In this paper we develop methods for exploring nonlinear conditional dependence between response time and accuracy. Three different approaches are proposed: (1) A joint model for quadratic conditional dependence is developed as an extension of the response moderation models for time and accuracy (Bolsinova et al., 2017b); (2) A joint model for multiple-category conditional dependence is developed as an extension of the fast-slow model of Partchev and De Boeck (2012); (3) An indicator-level nonparametric moderation method (Bolsinova and Molenaar, in press) is used with residual log-response time as a predictor for the item intercept and item slope. Furthermore, we propose using nonparametric moderation to evaluate the viability of the assumption of linearity of conditional dependence by performing posterior predictive checks for the linear conditional dependence model. The developed methods are illustrated using data from an educational test in which, for the majority of the items, conditional dependence is shown to be nonlinear.

Keywords: response time, conditional dependence, nonlinear relationship, response processes, joint modeling, hierarchical model, response moderation

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Matthew D. Finkelman,
Tufts University School of Medicine, United States
Jean-Paul Fox,
University of Twente, Netherlands

***Correspondence:**
Maria Bolsinova
maria.bolsinova@act.org

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *01 March 2018*
Accepted: *31 July 2018*
Published: *07 September 2018*

Citation:
Bolsinova M and Molenaar D (2018) Modeling Nonlinear Conditional Dependence Between Response Time and Accuracy. Front. Psychol. 9:1525. doi: 10.3389/fpsyg.2018.01525

INTRODUCTION

When psychological and educational tests are presented in a computerized form, it is feasible to not only record the product of the response process (i.e., response accuracy), but also the characteristics of the process itself. The most commonly used process variable is response time. Various psychometric models have been developed to jointly model response accuracy and response time (van der Linden, 2007; Molenaar et al., 2015a,b) which are aimed at (1) better measurement of the ability of interest, and (2) investigating the relationship between the time and accuracy components of the process. The prominent framework for modeling the joint distribution of response time and accuracy is the hierarchical modeling framework (van der Linden, 2007), which specifies separate measurement models for ability and speed and combines them on the higher level through the dependence between speed and ability. In this way, when it comes to the relationship between time and accuracy on the same item, the only thing that the model estimates is the correlation between the latent variables—speed and ability—and all the observed correlations between time and accuracy are assumed to be explained by that correlation. That is, conditional on speed and ability, time and accuracy are assumed to be independent. However, it has been shown in multiple empirical data sets (Meng et al., 2015; Bolsinova and Maris, 2016; Bolsinova and Tijmstra, 2016; Bolsinova et al., 2017a,b) that time and accuracy are in fact not conditionally independent and there is more to learn from the conditional dependence between response time and accuracy.

Several methods have been proposed for testing the assumption of conditional independence (van der Linden and Glas, 2010; Bolsinova and Maris, 2016; Bolsinova and Tijmstra, 2016) and different extensions of the hierarchical model have been proposed to relax this assumption (Ranger and Ortner, 2012; Meng et al., 2015; Bolsinova et al., 2017a,b). From these studies, it appears that a violation of conditional independence is not a rare finding and that substantively interesting phenomena may be revealed by investigating the conditional dependencies (Bolsinova et al., 2017c).

When it comes to conditional dependence between time and accuracy, authors typically focus on positive conditional dependence (i.e., relatively slow responses are more often correct) and negative conditional dependence (i.e., relatively fast responses are more often correct). This implies, that a monotone conditional dependence is assumed for time and accuracy. Moreover, most existing models specify the relationship to be linear. However, this assumption of monotone and linear conditional dependence does not necessarily hold in all situations. It could be that responses which are faster than expected are less often correct than responses with response times close to what is expected, but responses slower than expected are not more often correct than those with response times close to what is expected. Therefore, researchers should be able to test whether linearity of conditional dependence between time and accuracy is plausible and to investigate potential nonlinear conditional dependence.

Nonlinear conditional dependence is interesting from the substantive point of view because by abandoning the assumption of monotonicity and linearity of the conditional relationship between time and accuracy one can get a more complete picture of the response process. Since a linear model can only reveal positive or negative dependence, it may ignore important parts of the response phenomena. Imagine a situation in which an item is solved either using a fast optimal strategy or a slow error-prone strategy (i.e., slow responses are less often correct than relatively fast responses) and, in addition to that, some of the respondents respond to the item by guessing (i.e., very fast responses are rarely correct). If one of these phenomena is much stronger than the other, then a linear effect in one of the directions would be detected (i.e., positive conditional dependence if guessing is the strongest factor, or negative conditional dependence if the difference in strategies is the strongest factor). The linear model might also find no evidence of conditional dependence if the two opposing factors balance each other out. In none of these scenarios, a valid conclusion about the relationship between time and accuracy would be drawn. On the contrary, nonlinear methods would allow one to detect a violation of conditional dependence and to get a better understanding of the response processes.

In this paper we develop methods for exploring nonlinear conditional dependence between response time and accuracy. Three different approaches are proposed. (1) The joint models for conditional dependence between time and accuracy (see e.g., Bolsinova et al., 2017b) are extended to include quadratic effects, which allows one to study nonlinear relationships between residual time and accuracy. (2) Partchev and De Boeck's (2012) model is extended to allow for multiple categories of responses which makes it possible to reveal nonmonotonic relationships between time and accuracy. Moreover, the model is modified in such a way that response time is treated as a continuous variable following a log-normal distribution, and the categories are defined based on the difference between the observed and expected log-transformed response time. This allows one to study the conditional dependence separately from the relationship between speed and ability on the higher-level of the hierarchical model. Bayesian estimation algorithms are developed for the two new joint models for response time and accuracy. (3) We propose using the indicator-level nonparametric moderation method (Bolsinova and Molenaar, in press) with residual log-response time as a predictor for the intercept and the slope of the item characteristic curve (ICC), such that nonparametric relationships between the residual response time and the item parameters can be investigated. Furthermore, we propose using nonparametric moderation to evaluate the viability of the assumption of linearity of conditional dependence. This can be done by performing posterior predictive checks for the linear conditional dependence model.

The remainder of the paper is organized as follows. In section 2 the hierarchical model for response time and accuracy is presented and the assumption of conditional independence is formally defined. In section 3 existing models for conditional dependence are discussed. In section 4 we propose three methods for exploring nonlinear conditional dependence. Section 5 presents an empirical example in which nonlinear conditional

dependence is investigated, and the paper concludes with a discussion.

JOINTLY MODELING RESPONSE TIME AND ACCURACY USING THE HIERARCHICAL MODEL

In the hierarchical model (van der Linden, 2007; Van Der Linden, 2009) the random variables response accuracy and response time of person p on item i, denoted by X_{pi} (with realizations $x_{pi} = 0/1$ for incorrect/correct) and T_{pi} (with realizations t_{pi}), respectively, are assumed to be independent, conditional on the latent variable ability, denoted by θ_p, and speed, denoted by τ_p:

$$f(x_{pi}, t_{pi} \mid \theta_p, \tau_p) = f(x_{pi} \mid \theta_p, \tau_p) f(t_{pi} \mid \theta_p, \tau_p). \quad (1)$$

Furthermore, it is assumed that response accuracy is independent of speed given ability, and that response time is independent of ability given speed. The full specification of the hierarchical model for response times and accuracy requires four model ingredients: (1) a measurement model for response accuracy, typically an item response theory (IRT) model; (2) a measurement model for response times; (3) a model for the relationship between the latent variables; and (4) a model for the relationship between the item parameters. In this section, we will present a simple specification of the model, which we will use as a basis for describing the existing extensions of the hierarchical model allowing for conditional dependence.

For the response accuracy measurement model, we use a two-parameter normal-ogive model (Lord and Novick, 1968) in which the probability of a correct response to the item depends on the ability of the person:

$$\Pr(X_{pi} = 1 \mid \theta_p) = \Phi(\alpha_i \theta_p + \beta_i), \quad (2)$$

where α_i and β_i are the slope and the intercept of the ICC, and $\Phi(\cdot)$ denotes the cumulative standard normal distribution function. Alternatively, the three-parameter normal-ogive model (Klein Entink et al., 2009), logistic IRT models (Bolsinova et al., 2017a), and cognitive diagnostic models (Zhan et al., 2018) have been used as the first ingredient for the hierarchical model.

For the response time measurement model, we use a log-normal model (van der Linden, 2006) in which the response times are assumed to have a log-normal distribution with the mean equal to the difference between the time intensity of the item, denoted by ξ_i, and the speed latent variable:

$$f(t_{pi} \mid \tau_p) = \ln \mathcal{N}(\xi_i - \tau_p; \sigma_i^2) \quad (3)$$

where σ_i^2 is the residual variance of the log-transformed response time. Here, $\frac{1}{\sigma_i^2}$ can be considered a time discrimination parameter since the smaller σ_i^2 is, the larger the proportion of the variance of response times explained by speed is. This model can also be seen as a constrained linear factor model with all factor loadings equal to each other (Molenaar et al., 2015b). Alternatively, one can use an unconstrained linear factor model with additional item-specific factor loadings (Fox et al., 2007). Different choices for the response time model, used as an ingredient for the hierarchical model, include a model based on Box-Cox transformation of response times (Klein Entink et al., 2009), and a Weibull model (Rouder et al., 2003).

For the relationship between the latent variables and for the relationship between the item parameters we use multivariate normal distributions. For identification, the mean vector of the latent variables is constrained to zero, and the variance of θ is constrained to one[1]. For the relationship between the item parameters $(\alpha_i, \beta_i, \xi_i)$ we also use a multivariate normal distribution. Unlike the distribution of the person parameters, here the mean vector and the covariance matrix can be estimated freely.

The conditional independence assumption in Equation (1) means that accuracy and time can be correlated only if ability and speed, which determine their expected values, are correlated. The residual response accuracy and residual log-transformed response time are taken to be noise and the fluctuations on the response accuracy and response time sides of the model are taken to be uncorrelated.

MODELING CONDITIONAL DEPENDENCE BETWEEN TIME AND ACCURACY

The conditional independence assumption can be relaxed and the relationship between residual response time and residual response accuracy can be incorporated into the model. One way to do that is to model the joint distribution of time and accuracy to the same item as a bivariate distribution with a non-zero correlation parameter. Ranger and Ortner (2012) proposed modeling the joint distribution of log-transformed response time (denoted by t^*_{pi}) and augmented continuous response accuracy (denoted by x^*_{pi} defined such that $x_{pi} = \mathcal{I}(x^*_{pi} > 0)$) as a bivariate normal distribution with an item-specific conditional correlation, denoted by ρ_i:

$$f(x^*_{pi}, t^*_{pi} \mid \theta_p, \tau_p) = \mathcal{N}_2\left(\begin{bmatrix} \alpha_i\theta_p + \beta_i \\ \xi_i - \tau_p \end{bmatrix}, \begin{bmatrix} 1 & \rho_i\sigma_i \\ \rho_i\sigma_i & \sigma_i^2 \end{bmatrix}\right). \quad (4)$$

Here, the marginal distribution of response accuracy and response time are the two-parameter normal-ogive model and log-normal model, the same as in the hierarchical model presented in the previous section. Meng et al. (2015) have further extended this model to allow the conditional correlation to vary, not only across persons, but also across items.

Bolsinova et al. (2017b) have shown that the joint model in Equation 4 is equivalent to a model in which the joint distribution of accuracy and time is factorized as a product of the marginal log-normal model for time and a conditional model for accuracy given time, which is a two-parameter normal-ogive model, with the intercept being a linear function of the standardized

[1]Note that if the factor model with item-specific factor loadings is used, then the variance of speed also has to be constrained.

difference between the observed and expected log-transformed response time:

$$\Pr(X_{pi} = 1 \mid t_{pi}, \theta_p, \tau_p) = \Phi\left(\alpha_i\theta_p + \beta_{i0} + \beta_{i1}\frac{\ln t_{pi} - (\xi_i - \tau_p)}{\sigma_i}\right), \quad (5)$$

where β_{i0} is the baseline intercept and β_{1i} is the linear effect of standardized residual log-transformed response time on the intercept of the ICC. In addition to the linear effect on the intercept, the model can be extended with a linear effect on the slope of the ICC (Bolsinova et al., 2017b)[2]:

$$\Pr(X_{pi} = 1 \mid t_{pi}, \theta_p, \tau_p) = \Phi((\alpha_{i0} + \alpha_{i1}z_{pi})\theta_p + \beta_{i0} + \beta_{i1}z_{pi}), \quad (6)$$

where z_{pi} denotes the standardized difference between the observed and expected log-transformed response time $\frac{\ln t_{pi} - (\xi_i - \tau_p)}{\sigma_i}$, and α_{i0} and α_{i1} are the baseline slope and the linear effect of z_{pi} on the slope of the ICC, respectively. The parameters β_{i1} and α_{i1} can be interpreted as the main effect of residual log-transformed response time on response accuracy, and the interaction effect between ability and z_{pi} on accuracy, respectively. Throughout the paper we refer to this model as the linear conditional dependence model.

The approaches discussed above treat the response time as a continuous variable and relate the parameters of the IRT model for accuracy to deviations of the observed log-response time from its expected value. An alternative proposal has been to categorize response time into two classes—fast and slow—and jointly model the dichotomized response time and response accuracy using an IRTree model (De Boeck and Partchev, 2012). In this case, the ICC parameters can differ between the two classes (Partchev and De Boeck, 2012; DiTrapani et al., 2016). If the two-parameter normal-ogive model is used, then the probability of a correct response given response time is:

$$\begin{aligned}\Pr(X_{pi} = 1 | t_{pi}, \theta_p) = \Phi\Big(&\left(\alpha_{iF}\mathcal{I}\left(t_{pi} \le \tilde{t}_i\right) + \alpha_{iS}\left(t_{pi} > \tilde{t}_i\right)\right)\theta_p \\ &+ \beta_{iF}\mathcal{I}\left(t_i \le \tilde{t}_i\right) + \beta_{iS}\mathcal{I}\left(t_i > \tilde{t}_i\right)\Big), \end{aligned} \quad (7)$$

where $\tilde{t}_i$ denotes the median response time to item i, and subscripts F and S denote the fast and the slow class, respectively. Since, only two classes of response time are defined, only a monotonic relationship between response time and accuracy can be explored, for example responses in the slow class being more often correct than responses in the fast class ($\beta_{iS} > \beta_{iF}$), or responses in the slow class being less informative about ability than responses in the fast class ($\alpha_{iS} < \alpha_{iF}$).

It is important to note that separation of the response times into two classes is typically done using an item-level median split. Therefore, this approach is different from the linear models discussed above, since the ICC parameters are related to the categorized *observed* response time, and not to the difference between the expected and observed response time, such that the differences between the fast and slow classes capture not only the conditional dependence, but also the relationship between ability and speed (persons for whom the responses to item i are categorized as fast on average would have a higher speed latent variable in the log-normal model than persons for whom the responses to item i are slow).

[2]Note, that alternatively it has been proposed to include a linear effect on the log-transformed slope of the ICC (Bolsinova et al., 2017a)

MODELING NONLINEAR CONDITIONAL DEPENDENCE

The linear conditional dependence models and the fast-slow model provide quite a simplistic picture of the relationship between response time and accuracy. The residual dependence between time and accuracy is not necessarily monotone and the change of the ICC parameters is not necessarily linear in z_{pi}. To further investigate the relationship between response time and accuracy, we propose two new joint models for conditional dependence between response time and accuracy, and also use a nonparametric moderation method to explore the relationship between the residual log-transformed response time and the parameters of the response accuracy model.

Joint Model for Quadratic Conditional Dependence

To allow for a nonlinear relationship between residual log-transformed response time and the ICC parameters, we extend the conditional model of response accuracy in Equation (6) with quadratic effects. To simplify the notation, we introduce a function $\Psi(\cdot, x) = (\Phi(\cdot))^x(1 - \Phi(\cdot))^{1-x}$. The resulting joint model for time and accuracy is then the following:

$$\begin{aligned} f(x_{pi}, t_{pi} \mid \theta_p, \tau_p) &= f(x_{pi} \mid t_{pi}, \theta_p, \tau_p) f(t_{pi} \mid \tau_p) \\ &= \Psi\Big((\alpha_{i0} + \alpha_{i1}z_{pi} + \alpha_{i2}z_{pi}^2)\theta_p + \beta_{i0} \\ &\quad + \beta_{i1}z_{pi} + \beta_{i2}z_{pi}^2, x_{pi}\Big) \\ &\quad \frac{1}{t_{pi}\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(\ln t_{pi} - \xi_i + \tau_p)^2}{2\sigma_i^2}\right), \end{aligned} \quad (8)$$

where α_{i2} and β_{i2} are the quadratic effects of the residual log-transformed response time on response accuracy. If $\alpha_{2i} < 0$, then the strength of the relationship between ability and the probability of a correct response first increases with residual log-transformed response time and then decreases, and vice versa if $\alpha_{i2} > 0$. Similar interpretations can be given to the sign of β_{i2}. When the quadratic effect is negative, the corresponding parameter of the ICC (i.e., slope or intercept) is the highest when $z_{pi} = -\frac{\alpha_{i1}}{2\alpha_{i2}}$.

Our joint model is an extension of the hierarchical model, therefore in addition to the specification of the joint distribution of the outcome variables, we also need to specify the distribution of the latent variables and the distribution of the item parameters. On the person side we use $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$ where the variance of θ is contrained to be 1. On the item side, we use $\mathcal{N}_7(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ for $\{\alpha_{i0}, \alpha_{i1}, \alpha_{i2}, \beta_{i0}, \beta_{i1}, \beta_{i2}, \xi_i\}$, where $\boldsymbol{\mu}_I$ and $\boldsymbol{\Sigma}_I$ are the mean vector and the covariance matrix of the item parameters, respectively. Note, that while we are including nonlinear effects in modeling the conditional dependence between time and accuracy given

ability and speed, we do not extend the standard hierarchical model with nonlinear effects on the higher level, since it goes beyond the scope of the current paper. However, one may consider more complex models for the joint distribution of the person parameters and for the joint distribution of the item parameters that would allow for a nonlinear relationship on the higher level as well as on the lower level.

This extended joint model for conditional dependence between response time and accuracy can be estimated in a similar way as the linear conditional dependence models (Bolsinova et al., 2017b) using Bayesian estimation. The Appendix contains the full specification of the density of the data, prior and posterior distributions, and the detailed steps of the Gibbs Sampler, in which the parameters are consecutively sampled from their full conditional posteriors.

Multiple-Category Conditional Dependence Model

An alternative to the quadratic conditional dependence model for exploration of nonmonotone dependence is an extension of the slow-fast model. Allowing the ICC parameters to differ not just across two classes of responses, but across multiple classes, makes it possible to uncover nonmonotone relationships between residual response time and the ICC parameters (e.g., an item being most informative for the middle categories and least informative for the extreme categories).

Considering multiple categories is not the only way in which our joint model differs from the existing fast-slow models. Instead of categorizing the response time itself, we are going to use the residual log-transformed response time, since we are interested in the *conditional* dependence between response time and accuracy, taken separately from the relationship between speed and ability.

The joint distribution of response time and accuracy in this model is:

$$f(x_{pi}, t_{pi} \mid \theta_p, \tau_p) = \Psi\left(\left(\alpha_{im} + \sum_{k=1,k\neq m}^{M} \alpha_{ik} z^*_{pik}\right)\theta_p + \beta_{im} + \sum_{k=1,k\neq m}^{M} \beta_{ik} z^*_{pik}, x_{pi}\right) \frac{1}{t_{pi}\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(\ln t_{pi} - \xi_i + \tau_p)^2}{2\sigma_i^2}\right), \quad (9)$$

where M is the number of categories of residual log-transformed response time, m is the baseline category, $z^*_{pik} = \mathcal{I}(q_k \leq z_{pi} \leq q_{k+1})$, and $q_1, \ldots, q_{M+1}$ are the a priori defined thresholds between the categories ($q_1 = -\infty, q_{M+1} = +\infty$). Note, that in this joint model response time is modeled as a continuous variable such that there is no loss of information in the measurement of speed due to categorization.

Given that residual log-transformed response time belongs to the baseline category, the item parameters are equal to $\{\alpha_{im}, \beta_{im}\}$. When z_{pi} belongs to one of the remaining categories $k \neq m$, the parameters are equal to $\{\alpha_{im} + \alpha_{ik}, \beta_{im} + \beta_{ik}\}$. When $M > 2$ the model allows for nonmonotone conditional dependence. For example, if m is the middle category and $\beta_{ik} < 0, \forall k \neq m$, then it means that both responses that are slower than expected and those that are faster than expected are less often correct than responses for which the observed response time is closer to the expected response time. The more categories are used the more flexibly the model can account for different patterns of conditional dependence. However, the more categories there are the smaller the sample size per category is and the less precise the estimates of the item parameters are.

Analogous to the quadratic model, this joint model for time and accuracy can also be estimated using a Gibbs Sampler (see **Appendix** for details). Here we specify the same distribution for the latent variables, and similarly $\mathcal{N}_{2M+1}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ is specified for $\{\alpha_{i1}, \ldots, \alpha_{iM}, \beta_{i1}, \ldots, \beta_{iM}, \xi_i\}$.

Nonparametric Approach

The third approach to exploring nonlinear conditional dependence is in line with the nonparametric indicator-level moderation approach developed by Bolsinova and Molenaar (in press), which is a extension of the local structural equation modeling approach from Hildebrandt et al. (2016). The idea of the method is to explore the nonparametric relationship between the indicator-level covariate and the parameters of the latent variable model. In the case of investigating the conditional dependence between response time and accuracy, this method can be applied by using the residual log-transformed response time as the covariate for the intercept and the slope of the ICCs of the items in the accuracy measurement model. Using residual log-transformed response time instead of the observed response time itself is important because in that way one can investigate the relationship conditional on the latent variables and not the marginal relationship between time and accuracy. By including the residual log-transformed response time as a covariate in the analysis we can look at how the probability of a correct response changes depending on whether the response is shorter than expected or longer than expected (i.e., the intercept being a function of residual log-transformed response time) and how the relationship between ability and the probability of a correct response changes depending on the response being relatively fast or slow (i.e., the slope being a function of the residual log-transformed response time).

Unlike the first two approaches in which the joint distribution of response time and accuracy is modeled, in nonparametric moderation it is not possible to model the two outcome variables jointly since in this approach residual log-transformed response time is treated as an observed covariate. Therefore, we propose using a two-step procedure. First, the measurement model for response times is fitted and the estimates of the standardized residual log-transformed response time are computed:

$$\hat{z}_{pi} = \frac{\ln t_{pi} - \hat{\xi}_i + \hat{\tau}_p}{\hat{\sigma}_i}. \quad (10)$$

Second, the estimates $\hat{z}_{pi}$ are included in the analysis of response accuracy as indicator-level moderators.

For each item, a set of focal points $F_1, \ldots, F_J$ for the value of the standardized residual log-transformed response time are defined for which the slope and intercept of the ICC are estimated. Since for all items the moderator has a mean of zero and a standard deviation of one, it makes sense to have the same focal points for different items. For each focal point j and for each item, the estimates of α_{ji} and β_{ji} are obtained by weighting the responses to the item from each person p using the distance between the value $\hat{z}_{pi}$ and the focal point. For each combination of an item i and a focal point j a vector of weights $\mathbf{w}_{ji}$ is defined with each element corresponding to a particular person p:

$$w_{pji} = \exp\left(-\frac{(\hat{z}_{pi} - F_j)^2}{2\left(hN^{-\frac{1}{5}}\right)^2}\right), \tag{11}$$

there h is the bandwidth factor which serves as a smoothing parameter and determines how far from the focal point $\hat{z}_{pi}$ has to be to have a relatively large impact on the estimates of the parameters α_{ji} and β_{ji}. We will use the vale of 1.1 for h, which has been proposed in the nonparametric literature (Silverman, 1986) and has been successfully used for indicator-level moderation (Bolsinova and Molenaar, in press).

The item slopes and intercepts of the K items in the test are estimated in an iterative procedure. To start, the values of the slope and intercept are initialized for each combination of a person and an indicator, that is $N \times K$ matrices of response-specific slopes and intercepts, denoted by $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ respectively, are defined. The estimates of the item slopes and intercepts from the conditional independence hierarchical model can be used as starting values. After initialization, repeatedly for each item the estimates of α_{ji} and β_{ji} are obtained for each focal point j by maximizing the weighted log-likelihood:

$$\begin{aligned}&\ln \mathcal{L}(\alpha_{ji}, \beta_{ji} \mid \mathbf{X}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, \mathbf{w}_{ji}) \\ &= \sum_p \ln \int \left(\Psi(\alpha_{ji}\theta + \beta_{ji}, x_{pi})\right)^{w_{pji}} \\ &\quad \prod_{k \neq i} \Psi(\alpha^*_{pk}\theta + \beta^*_{pk}, x_{pk}) \mathcal{N}(\theta; 0, 1) d\theta,\end{aligned} \tag{12}$$

where the responses to item i are weighted with $\mathbf{w}_{ji}$, while for the rest of the items $k \neq i$ the current values of response-specific slopes and intercepts contained in $\boldsymbol{\alpha}^*_{\cdot k}$ and $\boldsymbol{\beta}^*_{\cdot k}$ are used.

After α_{ji} and β_{ji} are obtained, we update the values of $\boldsymbol{\alpha}^*_{\cdot i}$ and $\boldsymbol{\beta}^*_{\cdot i}$ as follows:

$$\alpha^*_{pi} = \begin{cases} \alpha_{1i} & \text{if } \hat{z}_{pi} < F_1, \\ \alpha_{ji} + (\hat{z}_{pi} - F_j)\frac{\alpha_{(j+1)i} - \alpha_{ji}}{F_{j+1} - F_j} & \text{if } F_j \leq \hat{z}_{pi} \leq F_{j+1}, \forall j \in [1, J-1], \\ \alpha_{Ji} & \text{if } \hat{z}_{pi} > F_J; \end{cases} \tag{13}$$

with a similar specification for β^*_{pi}. That is, if $\hat{z}_{pi}$ is outside of the range of the focal points, then the parameters are set equal to the parameters at the nearest focal point, and otherwise α^*_{pi} and β^*_{pi} are computed using piece-wise linear regression.

Under this nonparametric approach the significance of conditional dependence can be tested using permutation tests. To perform these tests, one needs to repeatedly estimate the nonparametric relationship between the residual log-transformed response time and the parameters of the ICCs in permuted data sets, that is, data sets in which the response accuracy data points are kept intact but the residual log-transformed response times are randomly assigned to different persons in the sample. As a first tool to draw inferences about the significance of the relationship between the residual log-transformed response time and the ICC parameter, one can use graphical checks of deviations of the observed relationship and the relationship in the permuted data sets. However, a more rigorous test is to use the variance of the parameters across focal points as a statistic and compare the observed value to its distribution in the permuted data sets. The proportion of permuted data sets in which the variance is larger than in the observed data can be used to approximate the p-value for testing the hypothesis of conditional independence.

Furthermore, nonparametric moderation can be used to evaluate the viability of the assumption of linearity of conditional dependence. This can be done by performing posterior predictive checks (Meng, 1994; Gelman et al., 1996) for the linear dependence model. The idea of posterior predictive checks is to compare the observed relationship between the residual log-transformed response time and accuracy (as estimated using the nonparametric method) with its posterior predictive distribution under the linear conditional dependence model. To do so one needs to (1) sample from the posterior distribution of the model parameters of the linear conditional dependence model, (2) using the values of the parameters sampled from this posterior generate replicated data under the model, and (3) evaluate the relationship between residual log-transformed response time and the parameters of the ICCs in each of the replicated data sets using the nonparametric method. In addition to the visual comparison of the estimated relationship in the observed data set and multiple replicated data sets, one can also use some measure quantifying a deviation from linearity and compare the observed measure with its posterior predictive distribution in the replicated data sets. To obtain such a measure one can first compute residuals in a simple linear regression model with the estimates of the ICC parameter at focal points ($\hat{\alpha}_{1i}, \ldots, \hat{\alpha}_{Ji}$ or $\hat{\beta}_{1i}, \ldots, \hat{\beta}_{Ji}$) as an outcome variable and the focal points as a predictor, and then compute the maximum of the absolute value of the cumulative sum of these residuals. The higher this value, the larger the deviation from linearity is. The proportion of replicated data sets in which the deviation from linearity is larger than in the observed data approximates the posterior predictive p-value. Small posterior predictive p-values (i.e., below 0.05) indicate that the deviation from linearity in the observed data is too large to conclude that the assumption of linearity of conditional dependence is viable.

EMPIRICAL EXAMPLE

Method

To illustrate how the nonlinear conditional dependence between response time and accuracy can be investigated, the proposed methods were applied to a data set of a high-stakes arithmetic

test[3]. One of the test versions with 38 items answered by 4,632 persons was available for analysis. For this data set several models were fitted: (1) the conditional independence model, (2) the linear conditional dependence model, (3) the quadratic conditional dependence model, and (4) the multiple-category conditional dependence model. In Model 4 we considered 5 categories for residual log-transformed response time and the thresholds between the categories were set equal to -1.5, -0.5, 0.5, and 1.5 (i.e., the thresholds are symmetric around zero and each two neighboring thresholds are one standard deviation away from each other), the middle category (i.e., the category where the response times are the closest to their expected values) was used as a baseline.

The four models were fitted using Gibbs Samplers with 10,000 iterations including 5,000 iterations of burn-in. For the details of the estimation algorithm for the conditional independence model and the linear conditional dependence model see Bolsinova et al. (2017b). Gibbs Samplers for Models 3 and 4 are described in the **Appendix**. The fitted models were compared using the modified Bayesian information criterion (BIC) which has been previously used for comparing and selecting joint models for response time and accuracy (Bolsinova et al., 2017b)[4]. The criterion is modified in the sense that posterior means of the model parameters are used instead of the maximum likelihood estimates of the parameters. The models allowing for nonlinear conditional dependence have a larger penalty term based on their larger number of parameters (i.e., quadratic effects in addition to the baseline ICC parameters and the linear effects in the quadratic model, and category-specific ICC parameters for the multiple category model).

In addition to fitting the joint models for response time and accuracy, the nonparametric moderation method was applied to the data. To do so the standardized residuals of log-transformed response time in the one-factor model with equal factor loadings (i.e., which is equivalent to the log-normal model in Equation 3) were computed using "lavPredict" function from the R-package "lavaan" (Rosseel, 2012). As focal points we used [-2, -1.5, -1,-0.5, 0, 0.5, 1, 1.5, 2], that are the points where the observed log-transformed response time is equal to the expected value, and where the deviation from the expected value are equal to 0.5, 1, 1.5, and 2 residual standard deviations. To test the significance of the effect of residual log-response time on the slopes and the intercepts of the ICCs, permutation tests with 500 replications were performed.

Finally, to test the linearity of conditional dependence, posterior predictive checks were performed for the linear conditional dependence model. Given each 10th sample of the model parameters after the burn-in a replicated data set was generated under the linear conditional dependence model (i.e., 500 replicated data sets were generated). The nonparametric moderation method was applied for each of the replicated data sets in the same way as for the observed data. The relationship between standardized residual log-transformed response time and the ICC parameters in the replicated data sets and the observed data were compared graphically. Furthermore, in each data set for each effect, the maximum of the absolute value of the cumulative sum of the residuals in the simple linear regression model with the focal points as a predictor and the ICC parameter as an outcome variable was computed. For each effect, the proportion of replicated data sets in which the deviation from linearity (quantified by the maximum of the absolute value of the cumulative sum of the residuals) was larger than in the observed data was computed to approximate the posterior predictive *p*-value for the linearity check.

TABLE 1 | Information criteria for the four joint models for time and accuracy.

Model	-2LL	*P*	Modified BIC
Conditional independence	2046864	–	2126317
Linear conditional dependence	2042274	76	2122368
Quadratic conditional dependence	2040146	152	2120882
Multiple-category model	2039572	304	2121591

P is the number of additional parameters compared to the conditional independence model.

Results

Table 1 shows the information criteria for the fitted joint models. The conditional independence model has the worst values compared to all models which take conditional dependence into account. This result shows that the conditional independence assumption does not hold for this test. Furthermore, models allowing conditional dependence to be nonlinear have lower information criteria values than the linear conditional dependence model, which shows that the assumption of linearity of conditional dependence also does not hold. The quadratic model was better than the multiple-category model, which points in the direction that the ICC parameters are not homogeneous within each category.

It is important to investigate whether the main inferences that are made based on the linear conditional dependence model would also hold for the nonlinear conditional dependence models and for the nonparametric moderation method. The first question is about the presence of the effects on the intercept and the slope of the ICCs of the separate items. In the linear conditional dependence model for 24 and 30 items, the 95% credible intervals of α_{1i} and β_{1i} respectively, did not include zero, which can be seen as evidence of the presence of the effects. In the quadratic model for 33 and 37 items the 97.5% credible intervals[5] of either α_{i1} or α_{i2}, and of either β_{i1} or β_{i2} did not contain zero, which can be seen as evidence of the presence of

[3]We would like to thank Dutch National Institute for Measurement in Education (CITO) for making this data set available to us. For confidentiality reasons we cannot disclose the content of the test items analyzed in this paper, but example items can be found at http://www.cito.nl/onderwijs/voortgezet%20onderwijs/rekentoets_vo/voorbeeldtoetsen. IRB approval was not needed for the study, since the data were collected previous to the study within high-stakes testing, and only response time and accuracy data and no information identifying the respondents was available for analysis.

[4]We are only using the modified BIC and not the modified Akaike information criterion (AIC) which has also been evaluated by the authors because they have shown that AIC tends to be too liberal.

[5]We decided to use a wider credible interval for the quadratic model because here two parameters are evaluated for each ICC parameter to make a conclusion about the presence of the effect instead of one, that is the area outside of the credible interval was divided by the number of parameters which were evaluated.

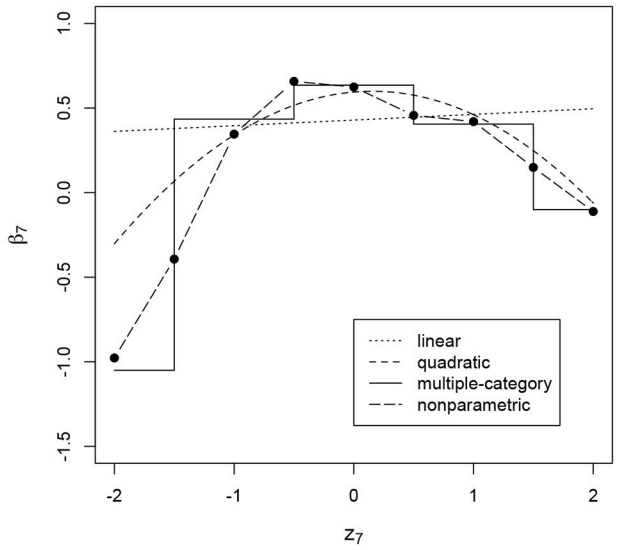


FIGURE 1 | Intercept of the item characteristic curve of item 7 (β_7, on the *y*-axis) as a function of residual log-transformed response time (z_7, on the *x*-axis) estimated with different methods. The item intercept is modeled as a function of residual log-transformed response time, such that the intercept is different depending on the value of that residual.

conditional dependence for these items. In the multiple-category conditional dependence model for 29 and 37 items the 98.75% credible intervals[6] of at least one of $\alpha_{ik}, k \neq m$ and one of $\beta_{ik}, k \neq m$ did not contain zero. For 25 and 35 items the permutation test had *p*-values below 0.05 for the effects on the slopes and the intercept respectively, pointing to the presence of main and interaction effects of residual log-transformed response time on response accuracy.

We note that the nonlinear methods are more flexible and complex and therefore provide noisier results and have less power for detecting the effects, so it would not be surprising if a linear effect is detected by the simpler linear method, but not by more complex nonlinear methods. On the contrary, having items for which the linear conditional dependence model does not detect the effect, while it is detected by the nonlinear models should be worrying, since it would mean that the effect is not detected due to its nonlinear nature. This is the case, for example, for the effect on the intercept of item 7: **Figure 1** shows the estimated relationship between the residual log-transformed response time and the intercept of the ICC for this item under the linear conditional dependence model and under the three nonlinear methods. It can be seen that when we allow the effect to be nonlinear and nonmonotone there is a clear relationship, while with the linear model the resulting relationship is close to a horizontal line.

The second kind of conclusion that is typically made based on the linear conditional dependence model is about the correlation between the baseline intercept of the items and the effect of residual log-transformed response time on the intercept. In multiple data sets previously this correlation was found to be negative (Bolsinova et al., 2017a,b). In our data set we found the same relationship. **Figure 2** (top left) shows the relationship between the estimates of β_{i0} and β_{i1} in the linear conditional dependence model. For easier items the effects are more often negative, and for more difficult items the effects are more often positive. To check whether a similar conclusion would be made using the nonlinear methods we performed the following analyses: (1) For the quadratic model for the items with negative β_{i2} (i.e., items for which there exists a value of z_{pi} which maximizes the intercept of the ICC) we plotted the points at which the intercept is maximized ($-\frac{\beta_{i1}}{2\beta_{i2}}$) against the baseline intercept (see **Figure 2**, top right); (2) For the multiple-category model we plotted the category for which the item intercept is the highest against the intercept in the baseline category (see **Figure 2**, bottom left); (3) For the nonparametric method we plotted the focal points for which β_{ji} is the highest against the overall proportion of correct responses to the item (see **Figure 2**, bottom right). In all three additional plots we see a similar relationship as for the linear conditional dependence model: For easier items relatively fast responses tend to be most often correct, while for difficult items relatively slow responses tend to be most often correct.

The comparison of the information criteria shows that linearity of conditional dependence does not hold for the test as a whole. Additionally, we examined the estimates of the item hyper-parameters specifying the mean and the variance of the quadratic effects. The means of the quadratic effects across items were estimated to be -0.02 [-0.07, 0.04] for α_{i2}s, and -0.09 [-0.15, -0.03] for β_{i2}s. The variances of the quadratic effects were 0.03 [0.02, 0.05] for α_{i2}s and 0.03 [0.02, 0.05] for β_{i2}s. For the effects of the item intercepts there is a clear pattern of the intercept first increasing and then decreasing with residual log-transformed response time since the mean of β_{i2} is negative, but for the effects on the item slopes the pattern is not so clear.

In addition to the overall conclusions about the presence of nonlinear effects, at least for some of the items, it is also important to look at each item separately and evaluate the results of the posterior predictive checks for linearity. For 27 and 30 items the posterior predictive *p*-value for linearity was below 0.05 for the effects on the slope and the intercept of the ICC, respectively. **Figures 3**–**6** give examples of some of the items with the largest deviations of conditional dependence from linearity. For item 1 the intercept of the ICC increases very steeply when the response is faster than expected, while positive deviations from the expected response time hardly result in further increase of the probability of a correct response (see **Figure 3**). From this figure, one can also see that the strength of the effect is underestimated in the linear conditional dependence model since the effect is averaged across the ranges of z in which there is an effect and where there is no effect. The slope of item 2 first increases and then decreases, for which the quadratic model gives quite a good approximation, while the linear conditional dependence severely misrepresents the relationship between residual log-transformed response time and the item slope (see **Figure 4**). For items 28 and 30 (see **Figures 5**, **6**), the direction of the effect changes

[6]We decided to use a wider credible interval for the multiple-category model because here four parameters are evaluated for each ICC parameter to make a conclusion about the presence of the effect instead of one parameter, that is the area outside of the credible interval was divided by the number of parameters which were evaluated.

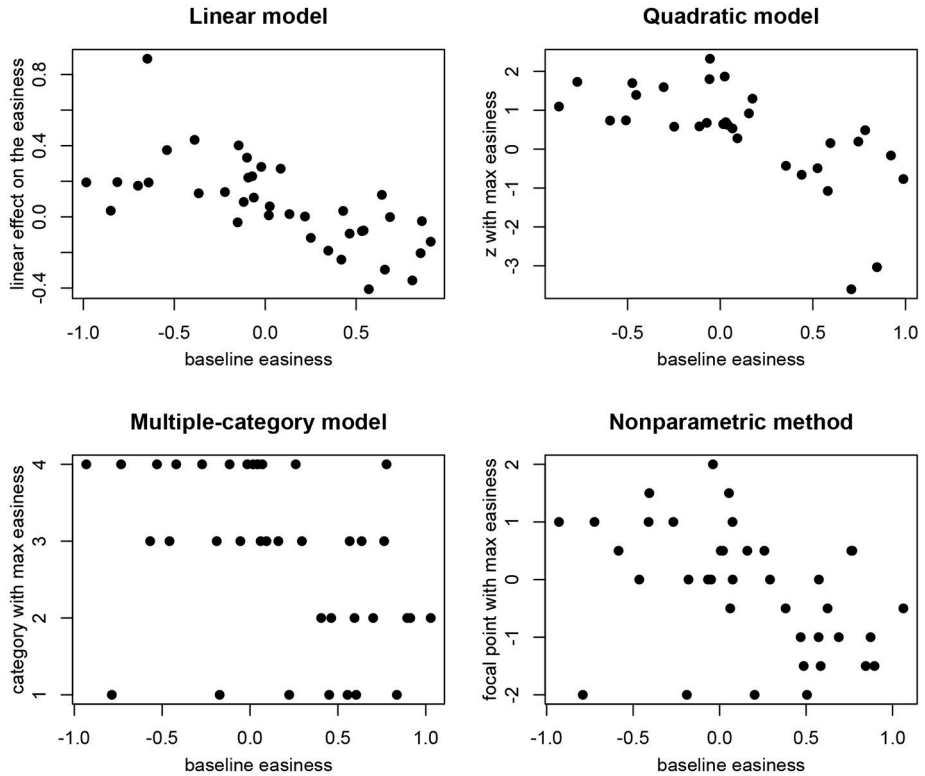


FIGURE 2 | Differences in the effect of residual log-transformed response time (z) on item easiness depending on the baseline easiness.

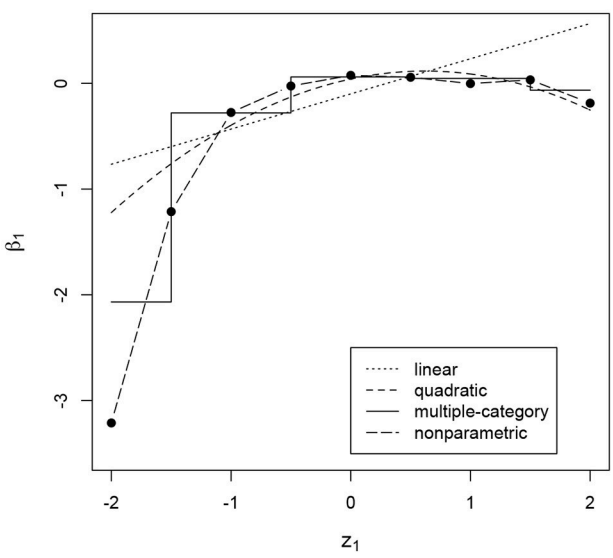


FIGURE 3 | Intercept of the item characteristic curve of item 1 (β_1, on the *y*-axis) as a function of residual log-transformed response time (z_1, on the *x*-axis) estimated with different methods.

in the area where the observed response time is close to its expected value: Responses both faster than expected and slower than expected are less often correct than the responses with response times close to their expected values. **Figures 5**B, **6B** illustrate the posterior predictive check for the intercepts of items 28 and 30. Here, the relationships in the observed data (black lines) clearly deviate from what would be expected if the data were generated under the linear conditional dependence model (gray lines). Note, that for the first of these two items the linear conditional dependence model reports a positive conditional dependence between response time and accuracy and for the second one it reports a negative conditional dependence, which is correct only for a part of the scale of the residual log-transformed response time and does not adequately represent the pattern of conditional dependence as a whole.

Additionally, we compared the estimates of ability under the conditional independence model, the linear conditional dependence model and the two nonlinear conditional dependence models (quadratic and multiple-category models) to check how the inclusion of conditional dependence in a model (and the exact way in which it is modeled) influences the inferences about the respondents. The correlations between the estimates of θ under each pair of models was very high, the lowest value of the correlation was above 0.988, and the highest value of the correlation was above 0.999. Therefore, in this example modeling conditional dependence does not change the measured construct, while it does allow learning more about the relationship between time and accuracy compared to the standard conditional independence model.

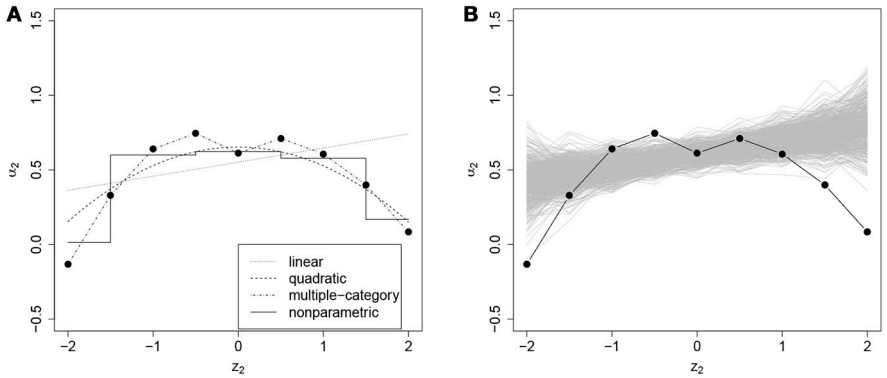


FIGURE 4 | (A) Slope of the ICC of item 2 (α_2, on the *y*-axis) as a function of residual log-transformed response time (z_2, on the *x*-axis) estimated with different methods; **(B)** posterior predictive check for linearity of conditional dependence: each gray line represents the relationship between the residual log-transformed response time of item 2 (z_2) and the slope of the ICC of item 2 (α_2) estimated in the replicated data generated under the linear model, and the black line represents the relationship in the observed data.

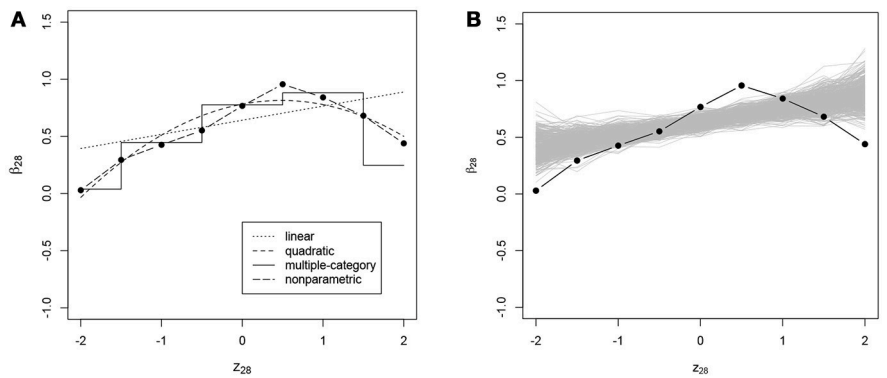


FIGURE 5 | (A) Intercept of the ICC of item 28 (β_{28}, on the *y*-axis) as a function of residual log-transformed response time (z_{28}, on the *x*-axis) estimated with different methods; **(B)** posterior predictive check for linearity of conditional dependence: each gray line represents the relationship between the residual log-transformed response time of item 28 (z_{28}) and the intercept of the ICC of item 28 (β_{28}) estimated in the replicated data generated under the linear model, and the black line represents the relationship in the observed data.

DISCUSSION

Our empirical example shows that conditional dependence between response time and accuracy can be nonlinear: in this example models allowing for nonlinear dependence are preferred over the linear dependence model, and for the majority of the items the posterior predictive checks indicate violations of linearity of the relationship between residual log-transformed response time and the ICC parameters. Using a linear conditional dependence model may in some situations lead to incorrect conclusions about the relationship between response time and accuracy: (1) One may conclude that conditional independence holds, when conditional independence is violated in a nonmonotone way such that the positive dependence in one range of the z-values and the negative dependence on another range cancel each other out; (2) The strength of the effect may be underestimated, when the effect is strong in some range of z-values and is either very weak or is absent in another range; (3) One may conclude that the dependence is, for example, negative while in fact it is both positive and negative depending on the range of z-values. In such situations, by modeling nonlinear conditional dependence one can get a better picture of the relationship between response time and accuracy in the data and get closer to understanding the response processes behind this relationship.

The approaches proposed in this paper make use of the difference between the observed and expected log-transformed response times, z_{pi}, as a predictor variable to account for unobserved heterogeneity in the responses. In the model, we do not explicitly separate the unobserved heterogeneity by means of additional latent variables. As a result, z_{pi}, which contains noise, is fully incorporated in the response model which decreases the power to detect an effect as the parameter estimates will have increased sampling fluctuations due to the noise in the residual log-transformed response time. However, we did not want to further complicate the model by introducing additional latent variables. In addition, introducing more latent variables may also decrease the power to detect an effect due to increased estimation error. Another aspect of the conditional dependence models is that false positives may arise if the response time model

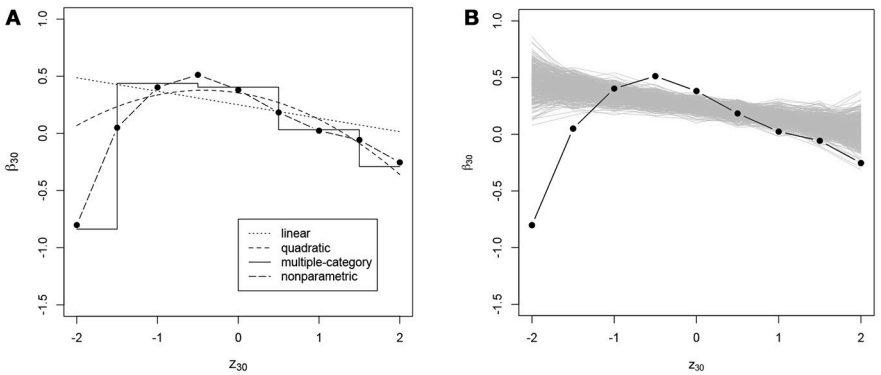


FIGURE 6 | (A) Intercept of the ICC of item 30 (β_{30}, on the *y*-axis) as a function of residual log-transformed response time (z_{30}, on the *x*-axis) estimated with different methods; **(B)** posterior predictive check for linearity of conditional dependence: each gray line represents the relationship between the residual log-transformed response time of item 30 (z_{30}) and the intercept of the ICC of item 30 (β_{30}) estimated in the replicated data generated under the linear model, and the black line represents the relationship in the observed data.

is misspecified. That is, such misspecifications will be absorbed in z_{pi} which in turn may be detected as a linear or non-linear conditional dependence effect if the misspecification is large enough. As a result, ideally one should carefully consider model fit of the response time measurement model before interpreting the results of the present parametric approach.

The conclusion about the negative relationship between the baseline intercept of the items and the effects of residual log-transformed response time on the intercept, previously found in other datasets (see e.g., Bolsinova et al., 2017b) and also found in our empirical example, seems to be robust regarding the violation of the linearity of the effect. With all three methods allowing for nonlinear dependence, we observed a relationship between the overall easiness of the item and the pattern of conditional dependence. When nonlinear conditional dependence is considered, we can no longer talk about the single effect on the intercept, instead we are considering the range of values of z_{pi} for which the intercept (and therefore response accuracy) is the highest. For easier items, the optimal values of z_{pi} tend to be more negative (responses faster than expected), while for difficult items, the optimal z_{pi} is positive (responses slower than expected).

In this paper we used three different approaches to modeling nonlinear conditional dependence: (1) the quadratic conditional dependence model, (2) the multiple-category conditional dependence model, and (3) the nonparametric modeling approach. These three approaches all have their comparative advantages and disadvantages. An important difference between the first two methods and the third one is that the first two methods allow modeling response time and accuracy jointly, while the third method requires a two-step procedure in which the estimates $\hat{z}_{pi}$ are treated as observed covariates for the distribution of response accuracy. This can be seen as a disadvantage of the nonparametric approach. At the same time, the nonparametric approach allows for more flexibility in the relationship between residual log-transformed response time and the ICC parameters. A limitation of the quadratic approach is that it restricts the possible relationship between the residual log-transformed response time and the ICC parameters to having a particular parametric shape and does not allow exploration of the shape of the conditional dependence. One way in which the quadratic shape of the relationship between z_{pi} and the ICC parameters is restrictive is that the function is symmetric, whereas it could be that the decrease of the parameter when moving away from the maximum point (given that the quadratic effect is negative and there is a maximum) is stronger when z_{pi} is becomes smaller that its optimal value than when it becomes larger. The nonparametric method allows us to more closely follow the shape of the relationship, however due to its flexibility the method requires larger sample sizes. A limitation of the multiple-category approach is that it assumes that within each category of residual log-transformed response time the item parameters are constant, which might not necessarily be the case in practice.

While the empirical example considered an application from educational measurement, the developed methodology can be expected to be relevant for applications relating to ability measurement in general, in cases where both response time and accuracy are recorded. Like the traditional hierarchical model, the models proposed in this paper make it possible to obtain additional information about ability based on the observed response times, but the methods also allow one to further study and model the complex relationship that may exist between response time and accuracy. This can, for example, be considered relevant in the context of developing and applying intelligence tests or other complex cognitive tests, where one might expect that items display relevant patterns of conditional dependence. For example, it may be that response time is indicative of the particular problem solving strategy that a respondent employs, which may also affect how likely one is to provide a correct response. Additionally, it may be that long response times are indicative of aberrant test taking behavior, such as inattention or distraction, which makes it plausible that such responses should be seen as less informative of ability than responses for which the response times do not indicate aberrant behavior. Our methods allow one to take this into account, by allowing the discrimination parameter of the item to be influenced by residual response time. In this way, the proposed methods allow researchers to work with models for ability measurement that take both response time and accuracy into account and that are

highly flexible with regard to the relationship between these two outcome variables that can be dealt with, and can accommodate a variety of deviations from conditional independence that can be expected in both high- and low-stakes psychological testing.

AUTHOR CONTRIBUTIONS

MB and DM designed the study, MB wrote software, performed the analysis and wrote the paper, DM provided feedback on the manuscript.

FUNDING

The research by DM was made possible by a grant from the Netherlands Organization for Scientific Research (NWO VENI-451-15-008).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01525/full#supplementary-material

REFERENCES

Bolsinova, M., De Boeck, P., and Tijmstra, J. (2017a). Modeling conditional dependence between response time and accuracy. *Psychometrika*. 82, 1126–1148. doi: 10.1007/s11336-016-9537-6

Bolsinova, M., and Maris, G. (2016). A test for conditional independence between response time and accuracy. *Brit. J. Math. Stat. Psychol.* 69, 62–79. doi: 10.1111/bmsp.12059

Bolsinova, M., and Molenaar, D. (in press). Nonlinear indicator-level moderation in latent variable models. *Multivar. Behav. Res*.

Bolsinova, M., and Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *J. Educ. Behav. Stat.* 41, 123–145. doi: 10.3102/1076998616631746

Bolsinova, M., Tijmstra, J., and Molenaar, D. (2017b). Response moderation models for conditional dependence between response time and response accuracy. *Brit. J. Math. Stat. Psychol*. 70, 257–279. doi: 10.1111/bmsp.12076

Bolsinova, M., Tijmstra, J., Molenaar, D., and De Boeck, P. (2017c). Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. *Front. Psychol.* 8:202. doi: 10.3389/fpsyg.2017.00202

De Boeck, P., and Partchev, I. (2012). IRTrees: tree-based item response models of the GLMM family. *J. Stat. Softw*. 48, 1–28. doi: 10.18637/jss.v048.c01

DiTrapani, J., Jeon, M., De Boeck, P., and Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence* 56, 82–92. doi: 10.1016/j.intell.2016.02.012

Fox, J.-P., Klein Entink, R., and van der Linden, W. (2007). Modeling of responses and response times with the package cirt. *J. Stat. Softw*. 20, 1–14.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin*. 4, 733–807.

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., and Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivar. Behav. Res.* 51, 257–258. doi: 10.1080/00273171.2016.1142856

Klein Entink, R. H., Fox, J. P., and van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Klein Entink, R. H., van der Linden, W. J., and Fox, J. P. (2009). A box–cox normal model for response times. *Brit. J. Math. Stat. Psychol.* 62, 621–640. doi: 10.1348/000711008X374126

Lord, F., and Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Meng, X.-L. (1994). Posterior predictive p-values. *Ann. Stat.* 22, 1142–1160. doi: 10.1214/aos/1176325622

Meng, X. B., Tao, J., and Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *J. Educ. Meas.* 52, 1–27. doi: 10.1111/jedm.12060

Molenaar, D., Tuerlinckx, F., and van der Maas, H. (2015a). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivar. Behav. Res.* 50, 56–74. doi: 10.1080/00273171.2014.962684

Molenaar, D., Tuerlinckx, F., and van der Maas, H. (2015b). A generalized linear factor model approach to the hierarchical framework for responses and response times. *Brit. J. Math. Stat. Psychol.* 68, 197–219. doi: 10.1111/bmsp.12042

Partchev, I., and De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence* 40, 23–32. doi: 10.1016/j.intell.2011.11.002

Ranger, J., and Ortner, T. (2012). The case of dependency of responses and response times: a modeling approach based on standard latent trait models. *Psychol. Test Assess. Model.* 54, 128–148.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., and Zhou, D. (2003). A hierarchical bayesian statistical framework for response time distributions. *Psychometrika* 68, 589–606. doi: 10.1007/BF02295614

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis, Vol. 26*. London: CRC Press.

van der Linden, W. (2006). A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

Van Der Linden, W. J. (2009). Conceptual issues in response-time modeling. *J. Educ. Meas.* 46, 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., and Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika* 75, 120–139. doi: 10.1007/s11336-009-9129-9

Zhan, P., Jiao, H., and Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *Brit. J. Math. Stat. Psychol*. 71, 262–286. doi: 10.1111/bmsp.12114

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Bolsinova and Molenaar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 May 2019
doi: 10.3389/fpsyg.2019.01186

Understanding Therapeutic Change Process Research Through Multilevel Modeling and Text Mining

Wouter A. C. Smink[1,2], Jean-Paul Fox[2], Erik Tjong Kim Sang[3], Anneke M. Sools[1], Gerben J. Westerhof[1] and Bernard P. Veldkamp[2]*

[1] Psychology, Health & Technology, University of Twente, Enschede, Netherlands, [2] Research Methodology, Measurement & Data Analysis, University of Twente, Enschede, Netherlands, [3] Netherlands eScience Centre, Amsterdam, Netherlands

Online interventions hold great potential for Therapeutic Change Process Research (TCPR), a field that aims to relate in-therapeutic change processes to the outcomes of interventions. Online a client is treated essentially through the language their counsellor uses, therefore the verbal interaction contains many important ingredients that bring about change. TCPR faces two challenges: how to derive meaningful change processes from texts, and secondly, how to assess these complex, varied, and multi-layered processes? We advocate the use text mining and multi-level models (MLMs): the former offers tools and methods to discovers patterns in texts; the latter can analyse these change processes as outcomes that vary at multiple levels. We (re-)used the data from Lamers et al. (2015) because it includes outcomes and the complete online intervention for clients with mild depressive symptoms. We used text mining to obtain basic text-variables from e-mails, that we analyzed through MLMs. We found that we could relate outcomes of interventions to variables containing text-information. We conclude that we can indeed bridge text mining and MLMs for TCPR as it was possible to relate text-information (obtained through text mining) to multi-leveled TCPR outcomes (using a MLM). Text mining can be helpful to obtain change processes, which is also the main challenge for TCPR. We showed how MLMs and text mining can be combined, but our proposition leaves open how to obtain the most relevant textual operationalization of TCPR concepts. That requires interdisciplinary collaboration and discussion. The future does look bright: based on our proof-of-concept study we conclude that MLMs and text mining can indeed advance TCPR.

Keywords: therapeutic change processes research (TCPR), multilevel models (MLMs), text mining, process data, online interventions, text variables

OPEN ACCESS

Edited by:
Prathiba Natesan, University of North Texas, United States

Reviewed by:
Timothy R. Brick, Pennsylvania State University, United States
Shianghau Wu, Macau University of Science and Technology, Macau

****Correspondence:***
Wouter A. C. Smink
w.a.c.smink@utwente.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *14 September 2018*
Accepted: *06 May 2019*
Published: *29 May 2019*

Citation:
Smink WAC, Fox J-P, Tjong Kim Sang E, Sools AM, Westerhof GJ and Veldkamp BP (2019) Understanding Therapeutic Change Process Research Through Multilevel Modeling and Text Mining. Front. Psychol. 10:1186. doi: 10.3389/fpsyg.2019.01186

1. INTRODUCTION

Traditional forms of psychotherapy are nowadays increasingly supplemented by online interactions: it is not uncommon that a counsellor seeks contact with a client through e-mail, text, chat, or other text-bearing messages. As the contact between counsellor and client becomes increasingly digitally mediated, it should be possible to trace the factors that contributed to the beneficial outcome of treatment back to these textual interactions.

In this light, the field of Therapeutic Change Process Research (TCPR) re-establishes its importance. TCPR aims to identify the mechanisms through which psychological treatments bring about positive and therapeutic change (Greenberg, 1986; Orlinsky et al., 2004; Elliott, 2010, 2012). TCPR has a long-standing tradition of studying the linguistic "products" of therapy (e.g., homework exercises, diaries, transcripts) in order to understand therapeutic change (Kazdin and Nock, 2003; Imel et al., 2015).

The rising popularity of Internet-based interventions (cf. Hoogendoorn et al., 2017) allow researchers to ask new TCPR research questions and re-establish the relevance of several known questions. Questions pertaining to the change processes that are beneficial to clients necessitate investigation of the "active ingredients" of therapy, of which many are linguistic (Muntigl and Horvath, 2005; Imel et al., 2015). TCPR has thus the potential to reveal the fundamental processes that are related to change. Aside from insight in what helps patients improve their functioning and reduce (clinical) symptoms, the importance of TCPR is also related to the rising number of people diagnosed with mental health disorders (see e.g., Andrade et al., 2013; Whiteford et al., 2013).

Over many decades, researchers attempted to answer TCPR questions; Orlinsky et al. (2004) estimated that there are more than 2000 published process-outcome studies of psychotherapy. Crits-Christoph et al. (2013) discuss several (methodological) issues related to TCPR, and express that "individual psychotherapy is not based just on an individual: it is a dyadic relationship consisting of a patient and therapist." Similar to Kenny and Hoyt (2009) and Crits-Christoph et al. (2013) argued that—from a statistical point of view—patients are nested within their therapist, hence, TCPR is concerned with *multi-level models* (MLMs; also known as hierarchical linear models, mixed models, random coefficient, or random effects models).

Yet, we found few studies that applied MLMs specifically to study therapeutic language. In the current work, we will present an approach for the study of therapeutic change processes based on text mining and MLMs by (re-)analyzing e-mails send between counsellor and client (Lamers et al., 2015). We do so by first making a comprehensive argument for the importance of understanding multi-layered change processes (Knobloch-Fedders et al., 2015), and argue for the use of text mining to study TCPR.

1.1. TCPR: Therapeutic Change Process Research

Progress in psychotherapy research is not made by only demonstrating the (average) effectiveness of a treatment; the history of psychotherapy research is marked by a gradual increase in the understanding of psychotherapeutic change processes (Orlinsky et al., 2004; Braakmann, 2015). Hence, psychotherapy benefits from a greater understanding of TCPR[1], which is defined as the scientific investigation of what occurs during psychotherapy, with regard to its clinical meaningfulness; in other words, it investigates the process through which clinically relevant changes occur within psychotherapy (Gelo and Manzo, 2015, p. 248).

Questions concerning the underlying processes that benefit the client also align with the interests of many clinical practitioners (Norcross and Wampold, 2011): *what* treatment, by *whom*, is most effective for *this* individual with *that* specific problem, and under *which* set of *circumstances* (Paul, 1967, p. 111)? Studies aimed at demonstrating average effects at group level fail to show what aspects of the intervention are related to the change the intervention realized (Barkham et al., 1993; Nock, 2007). Still, more effort is devoted to the analysis of the outcomes of psychotherapeutic interventions.

1.1.1. TCPR and the Study of the Therapeutic Conversation

As early as Freud's talking cure, the importance of looking at language to understand the therapeutic process has been recognized. Conversation is still the interactive medium central to most forms of psychotherapy (Muntigl and Horvath, 2005). The idea that the verbal exchange between counsellor and client contains important ingredients of therapy fueled TCPR (Greenberg, 1986; Hill and Lambert, 2004; Elliott, 2010), which is known for its a long-standing tradition of studying the linguistic "products" of therapy (e.g., homework exercises, diaries, transcripts) in order to understand therapeutic change (Gelo et al., 2015, p. 303, 392). For example, the *Narrative Processes Coding System* is "focused on the strategies and processes by which a client and counsellor transform the events of everyday life into a meaningful story that both organizes and represents the client's sense of self and others in the world" (Angus et al., 1999).

Another reason to specifically choose text over other types of TCPR data is that a valid understanding of psychotherapeutic processes require measurements collected from multiple perspectives, including that of the client, counsellor, and (possibly) external observers (Knobloch-Fedders et al., 2015). A good way to do so is to study the text-based representation of the therapeutic interaction. Because these transcripts are a direct observation of the therapeutic process, they reflect what actually happened in therapy. Transcripts can thus provide the basis to obtain the perspective of the client, counsellor, and (or independent) observer on the therapeutic process. Such interpretations are usually measured by questionnaires or interviews, and are retrospective reflections. Transcripts come with the additional benefit that they are relatively

[1]It should be noted that various terminologies are used in the literature, e.g., Change Process Research (CPR: Elliott, 2010; Greenberg, 2007), Psychotherapy Process Research (PPR: Gelo et al., 2012), and some of the early works simply refer to "change" (Braakmann, 2015; Hill and Corbett, 1993). The term "process-outcome research" is also often used, for example by Orlinsky et al. (2004), who defined it as "(primarily) the actions, experiences, and relatedness of patient and therapist in therapy session when they are physically together, and (secondarily) the actions and experiences of participants specifically referring to one another that occur outside of therapy sessions when they are not physically together" (Crits-Christoph et al., 2013, p. 311). To emphasize that we are dealing with change resulting from *therapy*, we propose to describe change processes as TCPR: *Therapeutic Change Process Research*.

straightforward to obtain after providing transcripts of the therapeutic conversation.

In this light, it is not unsurprising to see TCPR moving toward online interventions. Aside from being cost-efficient, web-based self-help interventions directly produce the textual interaction between therapist and counsellor, and come with the additional benefits that they are effective (see e.g., Andrews et al., 2004, 2010; Andersson et al., 2014), and easily accessible by large groups of people (Wang et al., 2007; Hoogendoorn et al., 2017). Just like transcripts, assessment of the interaction between counsellor and client in a web-based intervention has the potential of being a direct observation of the therapy process (Pennebaker et al., 2003; Schegloff, 2007; Elliott, 2012; Gelo et al., 2012).

Transcription and manual analyses mark the labor-intensive nature of TCPR, which is also the main reason why the field did not yet reach its full potential (Smink et al., under review). Traditional research methods start with the recording and transcription of a psychotherapeutic intervention so that human raters can (manually) code and analyse these transcripts (Atkins et al., 2014). Because the understanding of change processes mainly relies on qualitative analysis, these methods are only as fast as the researcher(s) conducting the research, which in practice limits their use to small scale studies (Atkins et al., 2012; Imel et al., 2015).

To strike a balance between TCPR's ambition to unravel the black-box through which therapy attains its effects and the labor-intensity of the TCPR methods, we propose to use automated text analysis methods. Text mining, a computational approach to text analysis, can be used to automatically extracted text features that can contribute to the understanding of the active ingredients of therapy. We are observant of the criticisms that algorithms have yet to achieve the same depth of analysis as humans. However, in our view, it would be a shortcoming to TCPR's ambitions if the insights that basic text features can offer remain unused. In the next session we will discuss how text mining can scale up TCPR by finding text-based predictors—also known as input variables or independent variables—from therapy related texts. We will do so making use of *multi-level models* (MLMs), an advanced statistical model that is able to capitalize on the hierarchical structure of text data.

1.2. Text Mining: Scaling Up TCPR

As language is an important mediator of psychotherapeutic processes, obtaining information about these processes through texts is one of the first applications of text mining. Mergenthaler (1996) compared five computer-assisted measures for the analysis of textual data of two psychotherapies, and was among the first to apply text mining for psychology. He used text mining, which he then called "computer assisted analysis of textual data," to identify turning points in sessions, which could then be explored more deeply by humans through (qualitative) analyses methods. Anderson et al. (1999) developed *Computer Assisted Language Analysis System* (CALAS) to examine the relationship of various linguistic measures to outcome measures in high and low verbalized affect segments. Many applications of text mining are still centered around finding key moments in the therapeutic process (cf. Lepper and Mergenthaler, 2005; Pfäfflin et al., 2005; Fontao and Mergenthaler, 2008), which is also a common approach in TCPR (e.g., the "Significant Events Approach" in Elliott, 2010).

Practically, typical text mining approaches in psychology include counting words, identifying topics, and coupling the terms to a domain-specific ontology (Hoogendoorn et al., 2017). Text mining[2] refers to a general methodological framework that includes several automated methods to analyse large corpora of texts (cf. Jurafsky and Martin, 2017). As text mining is a methodological framework that combines and includes numerous techniques and methods from many disciplines, it is not surprising that terms referring to the automatic extraction of information from text are used sometimes interchangeably, such as text mining and NLP.

1.2.1. Text Mining Emotions

The Linguistic Inquiry and Word Count (LIWC) software by Pennebaker et al. (2015a) is used by many researchers, and has showed to be effective in predicting therapist empathy (Gibson et al., 2015), counsellor behavior (Pérez-Rosas et al., 2017), and identifying emotional and cognitive process in psychotherapy (McCarthy et al., 2017). LIWC categorizes word usage by counting the percentage of words that reflect—among other categories—thinking styles, emotional states, and social concerns (Pennebaker et al., 1997; Hirsh and Peterson, 2009; Tausczik and Pennebaker, 2010). LIWC taps into the underlying idea that word use is one of the most direct means of expressing thoughts and feelings (Fast and Funder, 2008), as the way individuals talk and write provides a window into their emotional and cognitive worlds psychological characteristics (Pennebaker et al., 2003, 2015b).

The writing intervention by Lamers et al. (2015) focused on different life themes, with one theme central to each of the seven modules. By asking clients to describe specific positive and several difficult memories, clients adjusted their life stories step-by-step by integration of these memories. Lamers et al. (2015) did not study the content of the e-mails.

Previous studies showed that positive therapeutic outcomes from writing interventions are associated relatively high rate of positive emotion words, few negative emotion words, and with an increasing number of "cognitive"[3] words throughout the intervention (e.g., Campbell and Pennebaker, 2003; Pennebaker et al., 1997, 2015b). As the intervention of Lamers et al. (2015) focuses specifically on positive and difficult memories and emotions with the aim of integrating these two, we study words the reflective of these aspects in e-mails. As the intervention by Lamers et al. (2015) aims to improve integration of positive and negative memories, we expect that LIWC's "cause" and "insight" categories are mostly reflective of that process. We aim to find further evidence for these findings in data from Lamers et al. (2015), by relying on text mining and multi-level models.

[2]We recommend Manning and Schütze (1999); Feldman and Sanger (2007); Jurafsky and Martin (2017) for a detailed overview of text mining. For aspiring text mining practitioners, we recommend the NLTK library available in the programming language Python (Bird et al., 2009).

[3]Cognitive words are a word category from the LIWC program.

1.2.2. Text Mining and MLMs

Although the idea to relate words or textual aspects (in psychotherapeutic texts) to outcomes is well-established in TCPR, there are methodological issues that are specifically relevant when analyzing text data. Studying change processes in e-mails mandates accounting for the *dyadic* relation (Crits-Christoph et al., 2013, p. 301), and is therefore dependent on both the counsellor and client.

While the assumption of *independence of observations* is the basis for traditional statistical models, such as the ANOVA or regression model, some text mining models relax this assumption. For example, the naive Bayes classifier *assumes* independence assumptions between observations. The model classifies units to the category that has the highest probability; a common application of the model is the spam-filter, where e-mails are classified as either spam or "ham" (no-spam). He et al. (2012) used naive Bayes to find words that could discriminate between texts written by soldiers with or without PTSD.

Naive Bayes is a family of algorithms based on the assumption that the value of a particular (text)feature is independent of the value of any other feature. This independence assumption is too strong ("naive"); in reality, independence does not hold for texts that are written by the same person. In doing so, the model "naively' neglects the nesting of e-mails within person, ignoring the assumption of independence. In the next section work, we will argue for the importance of applying MLMs to analyse textual data for correct statistical inference, as MLMs do not violate the non-independence in e-mail data (Kenny et al., 2002). A consequence of failing to recognize the nested and hierarchical structures in e-mails is that standard errors of the estimated coefficients are underestimated, leading to an overstatement of statistical significance. MLMs recognize the existence of hierarchies in data by allowing for residual components at each level of the hierarchy.

1.2.3. Psychotherapy As a Multi-Leveled Procedure

Because MLMs offer the possibility to include predictors at the level of the individual, the group and at any other level of organization, the model arises quite naturally for TCPR (Raudenbush and Bryk, 2002). Many individual change phenomena can be represented through a two-level hierarchical model. The first level represents each clients' development by an individual growth trajectory that depends on the repeated measures for each client. The second level unit represents variables that are not repeatedly measured, such as gender, income, or depressive symptoms. The first level consists out of—for example—experienced pain at the beginning, middle, and at the end of therapy. The second level consists of the clients themselves, who could be (at a third level) nested within their therapist, for examples see Baldwin et al. (2007) and (Baldwin and Imel, 2013).

From a statistical viewpoint, TCPR practically equates to research questions concerning either a (longitudinal) development over time (Crowder and Hand, 1990; Baldwin et al., 2007; Nissen-Lie et al., 2010; Fitzmaurice et al., 2011; Adler, 2012), an (dyadic) interaction between a counsellor and its client (Tasca and Gallop, 2009; Kenny and Hoyt, 2009; Crits-Christoph et al., 2013), or to both. MLMs are—compared to traditional statistical methods—particularly useful to both of these situations as they capitalize on hierarchically organized data. Many kinds of data, including observational data collected in the human and biological sciences, have a hierarchical or clustered structure.

Considering that the psychotherapeutic practice is a multi-leveled procedure, it becomes apparent that client and counsellor are the two pre-eminent levels of organization. As counselors (almost) always treat more clients, clients could be viewed as grouped within their counsellor, similar to the students being *nested* within their class (Kenny and Hoyt, 2009; Crits-Christoph et al., 2013). Crucial to any MLM is that the unit of analysis at the lowest level (the students or clients) are nested within higher level units (classes or counsellor), that itself could also be nested within (higher) even higher units (schools, therapeutic practices, or clinical institutions).

Many of the applications of MLMs in psychotherapy resolve around the question of how to assess psychotherapeutic effectiveness. Adelson and Owen (2012) examined the influence of psychotherapists on clients' clinical outcomes. Baldwin et al. (2007) and Marcus et al. (2009) both showed that higher rates of therapeutic alliance could be relate to better therapeutic outcomes through MLMs (Crits-Christoph et al., 2013). Baldwin and Imel (2013) searched the literature for studies comparing outcomes of therapists. Nissen-Lie et al. (2010) accounted for variation in early patient-rated alliance by means of various self-reports of therapists providing treatment in a naturalistic outpatient setting.

1.3. Research Questions

Online a client is treated essentially through the language their counsellor uses, therefore the verbal interaction contains many important ingredients that bring about change. TCPR faces two challenges, first, how to derive meaningful change processes from (the) large bodies of texts (that online interventions produce)? Second, how to assess these complex, varied, and multi-layered processes? These two questions are intimately linked: insight in complex change processes gives an indication of how to derive other meaningful processes, and visa-versa.

We therefore advocate the combination of text mining and MLMs: the former offers tools and methods to discover patterns and trends in texts; the latter can analyse processes that vary at multiple levels. As the study by Lamers et al. (2015) is a writing intervention of which the writing assignment, the e-mails themselves, and the outcomes of the intervention are available, we give a proof-of-concept based on data from this study.

2. METHODS

2.1. Participants

The dataset derived from 174 clients who were recruited by Lamers et al. (2015) through advertisements in Dutch newspapers and websites. Only participants who felt depressed and were interested in writing about their life were included by Lamers et al. (2015). The sample was thus a self-selected group of individuals who had expressed interest in the program.

All participants had moderate depressive symptomatology and were randomly allocated to either the life-review "the stories we live by" (auto-biographic writing; AW), or the "expressive writing" (EW) intervention, or a waiting list condition. The mean age of the participants in the AW condition was 57.7 ($SD = 10.3$) years old, and the majority was female (75.9%). The mean age in the EW condition was 56.8 ($SD = 7.9$), and the majority was female (77.6%). In both conditions, the majority of the participants received a higher form of education (i.e., universities or colleges; AW: 48.3%, EW: 37.9%). For more details see Lamers et al. (2015).

2.2. Design

2.2.1. Study by Lamers et al. (2015)

2.2.1.1. Auto-biographic writing (AW)

The AW condition was a life-review self-help intervention that consisted of homework assignments, divided over modules that had to be completed over the course of 10 weeks. Clients communicated about their progress with trained counselors through a weekly e-mail interaction. According to Lamers et al. (2015) the self-help model program was based on insights from the autobiographical memory (Serrano et al., 2004; Brewin, 2006; Williams et al., 2007), narrative therapy (White and Epston, 1990; White, 2007), and life-review (Butler, 1963; Birren and Deutchman, 1991; Haight and Webster, 1995; Bluck and Levine, 1998; Westerhof et al., 2010b), and has been shown effective in previous studies (Korte, 2012; Westerhof et al., 2017).

2.2.1.2. Expressive writing (EW)

According to Lamers et al. (2015) the EW intervention was based on the method of expressive writing (Pennebaker et al., 1997). The method consisted of daily writing about emotional experiences, for $15 - 30$ min on $3 - 4$ consecutive days during 1 week. Lamers et al. (2015) extended and adapted this method to an intervention with seven modules, to make it a comparable with the life-review intervention.

2.2.2. Current Study

Our first intention was to demonstrate how text mining can be used to obtain change processes from e-mails. Lamers et al. (2015) concentrated their efforts on the analysis of the outcomes of the interventions but did not analyse the content of textual characteristics of the e-mails. After pre-processing, we obtained the insight, cause, positive, and negative emotion words from the LIWC program.

Our second intention was to demonstrate how multi-level models (MLMs) can be used to assess text-based measures of e-mails to aid understanding of the change processes. Similar to Lamers et al. (2015) we used the post-treatment measurement of the CES-D scale as the main outcome variable.

2.3. Materials

2.3.1. Questionnaires

The data available to us included the pre- and post-therapeutic measurements of the CES-D. The Center for Epidemiologic Studies Depression Scale (CES-D) is a brief self-report questionnaire to measure severity of depressive symptoms in the general population (Radloff, 1977). Lamers et al. (2015) used the Dutch version of the CES-D (Beekman et al., 1997); higher CES-D scores indicated more depressive symptoms (20 items, range $0 - 60$, $\alpha = 0.78$).

The intervention of Bohlmeijer and Westerhof (2010) teaches participants about autobiographical reasoning by specifically improving the ability to reason about the autobiographical self (Lamers et al., 2015). This form of reasoning describes the process of relating episodic memories to the conceptual self (Pasupathi and Carstensen, 2003; Thorne et al., 2004). By making the moral of an individual's life-story explicit, (s)he obtains insight in what the particular memory could reveal, explain, cause, give insight, or provide a (life) lesson learned about the (autobiographical) self. These processes are extensively researched by—for example—Pennebaker and Chung (2011), mainly in the context of showing how analog experiences, such as emotions, are translate to digital forms that bear meaning, such as of stories.

This process is operationalized by phrases that LIWC analyses can detect from the *insight* (e.g., "I now realize that…") and *cause* (e.g., "I understand why…"; Pennebaker and Chung, 2011). As the increase in insight and cause words are intractly related to emotional writing, we also study the (increase in) positive words, and (decrease in) negative words from LIWC (Westerhof et al., 2010a; Pennebaker and Chung, 2011).

2.3.2. Software

We used the LIWC software of Pennebaker et al. (2015b) to analyse the e-mails for the emotion and insight categories. We used the NLTK library of Bird et al. (2009) in the programming language Python (Python Software Foundation, 2018, version 3.6), for pre-processing and dividing the e-mail texts in words and sentences.

For our statistical analyses, we relied on the programming language `R` (R Core Team, 2019, version 3.5.1). We used package `lme4` for estimation and evaluation of our MLMs (Bates et al., 2015), and package `psych` for making descriptions of our variables (Revelle, 2018).

2.4. Data

2.4.1. Available Data

The data included the pre- and post-therapeutic measurements of the CES-D scale, and the e-mails exchanged between counselors and clients (2079 e-mails in total).

2.4.2. Complete Cases

In total, data of 174 clients was available to us from Lamers et al. (2015). We only used clients with no missing data. 166 of the 174 clients (95.4%) had a complete CES-D score. Not all e-mails were available, so we could only analyse the e-mails of 104 clients (59.8%). After removing duplicates, we included 97 clients in our analyses (55.7%, all percentages calculated against the original total of 174 clients).

2.4.3. Anonymization

Identifying information has been removed from the dataset that contained the outcomes ("structured" data), we identified

clients based on a unique four digits number. The e-mails ("unstructured" data) have been anonymized by removing all (e-mail) addresses, phone numbers, names of persons, organizations, and locations. Client names and counsellor names have been replaced by the previously mentioned unique four digits number so that it remained possible to identify which mails were written by the same person and which clients were treated by the same counsellor. The counselors were also anonymized.

2.4.4. Process Data

The e-mails of Lamers et al. (2015) should include the whole therapeutic process because they are the only form of interaction between counsellor and client. The e-mail procedure is explained (in Dutch) in detail in Bohlmeijer and Westerhof (2010). We will give some quotes that we translated from Dutch to English to give an impression of process data in a therapeutic context.

The first quote comes from a female participant: "*My trust in people is damaged pretty badly, I'm no longer in such good faith as I was in the past*." In response, the counsellor asks: "*Can you tell us a bit more about this? How did this happen? Are there times when you feel that you can trust people?*"

In the second week a male participant writes: "*By writing about myself, and especially naming the nice aspects about my life, I notice that writing is already paying off*." In the sixth week he writes: "*I feel that I am coming back to who I am*." He also expresses his graduate toward the counsellor: "*I do not have a specific question for you, a reaction from you based on my writing already is already enough. However, if you do ask questions, that would help me even further*."

The third example comes from a (different) female participant: "*How should I continue with my life? Is it okay? Almost thirty years ago I lost my brother and my sister-in-law. I lost my 10-year-old daughter... Losing a child is pretty much the worst thing that can happen to you*." In week seven she wrote (about her daughter): "*The tears are rolling down my cheeks as I think about you intensively. Over the duration of the course I have learned to balance between positive and negative emotions by means of communication or through writing. I succeeded, because I know that you knew that I am still an optimist in life. You and dad have a share in this. You were both never judgemental, but always stimulating*."

2.5. Procedure

2.5.1. Selection of the Text Variables From LIWC

We chose to use the number of *insight* and *cause* words from the cognitive process category, and the number of *positive* and *negative* words form the LIWC program (Pennebaker et al., 2015b). We had several reasons for doing so, first of all, past studies showed that positive therapeutic outcomes are associated with writing assignments of individuals that include relatively high rates of positive emotion words, few negative emotion words, and with an increasing number of cognitive words throughout the intervention (Pennebaker et al., 1997, 2015b; Campbell and Pennebaker, 2003; Campbell et al., 2013). Secondly, these basic text features are—as the name implies—relatively straightforward to obtain from an e-mail. Third, it is our ambition to show how textual information can be *obtained* through text mining and *analyzed* with MLMs. We do not aim to advance TCPR theory in our current paper: determining which textual predictors are meaningful is beyond the scope of our work. We intend to show how TCPR can be modeled in e-mails. Lastly, by bridging text mining and MLMs other TCPR researchers are enabled to advance TCPR theory using these two methodologies.

2.5.2. Pre-processing

We used the NLTK library to preprocess the e-mails. NLTK pounts sentences by counting word-terminal end-of-sentence punctuation like the period, question mark and / or exclamation mark. NLTK has a limited list of abbreviations, which are not included in the punctuation/sentence count. Word-internal punctuation, like the first period in e.g., is ignored. Handling of interjections depends on their punctuation, for example, "Oh?" is a separate sentence while "Oh," is part of the following sentence. Sentence fragments and quotes with end-of-sentence punctuation are counted as separate sentences.

NLTK is an often used Python library for text pre-processing, as it provides detailed documentation in Bird et al. (2009) on the order and content of the preprocessing steps.

2.5.3. Pre- and Post-therapeutic Measurements of the Text-Variables

We calculated the pre- and post-therapeutic scores of the text-variables (*insight*, *cause*, *positive*, and *negative* words form LIWC program) by averaging over the number of these words as counted by Pennebaker et al. (2015b) in the first and last three e-mails of the intervention by (Lamers et al., 2015). The original intervention also included a third time-point (T_0 a depression measure at the onset of the writing treatment, T_1 a measure at the end of the treatment, and T_2 a follow-up measure). However, only for the first two measurements (those at the beginning and end of therapy) we had e-mail data available. Hence, we dropped the follow-up measure (T_2) from our dataset, as we could not use in our text mining models.

2.6. Analyses

In total, we estimated five MLMs, see **Figure 1** for an overview and the R code. The regression equations below will give an indication of how the R code and equations are related. The data we used were the pre- and post-therapeutic measurements of the CES-D and the *insight*, *cause*, and the *positive* and *negative emotion* words of the LIWC Pennebaker et al. (2015b).

The pre- and post-therapeutic measurements of the CES-D scale were considered to be an outcome variable of the MLMs. Each MLM had a random intercept for the client to describe the variability in outcome scores across clients. An index i is used to refer to a pre-therapeutic score ($i = 1$) or post-therapeutic score ($i = 2$), and an index j is used to refer to the j^{th} client. Then, the outcomes can be described with a MLM, which is represented by

$$\text{CES-D}_{ij} = \mu + u_{0j} + \mathbf{X}_{ij}\boldsymbol{\beta}_1 + e_{ij}. \quad (1)$$

The errors e_{ij} are assumed to be normally distributed with a mean of zero and variance σ_e^2, and the random intercepts u_{0j} is also assumed to be normally distributed with mean

```
# Modelling the null / basic model.
M0 <- lmer(cesd ~ post + (1 | id),
 data)

# Modelling of the positive and
 negative words categories.
ME1 <- lmer(cesd ~ post*posemo
 + (1 | id), data)
ME2 <- lmer(cesd ~ post*negemo
+ (1 | id), data)

# Modelling of the insight and cause
 categories.
MCP1 <- lmer(cesd ~ post*insight +
 (1 | id), data)
MCP2 <- lmer(cesd ~ post*cause   +
 (1 | id), data)
```

FIGURE 1 | R code of the five multi-level models (M0, ME1, ME2, MCP1, and MCP2) using package `lme4`. In all models, we estimated the post-therapeutic measurement of CES-D (`cesd`) based on a random intercept for each client (`id`). In ME1 we estimated the post-therapeutic effect of the number of positive emotion words as the interaction effect between the number of positive emotion words (`posemo`) and an indicator variable (`post`). The other models are similar, in ME2 we estimated the effect of the number of negative emotion words (`negemo`), in MCP1 we estimated the effect of the number of `insight` words, and in the MCP2 we estimated the effect of the `cause` words. M0 is nested under each of these models.

zero and variance τ. The parameter μ is the general mean across scores. The predictor variables are stored in a matrix $\mathbf{X}$. The common effects, $\boldsymbol{\beta}_1$, represent the effects of the predictor variables on the outcomes CES-D. The predictor variables $\mathbf{X}$ explain variance in scores across the pre- and post-therapeutic measurement, and do not explain any change between the pre- and the post-therapeutic scores. To assess change, an indicator variable is used for the post-therapeutic measurement with $D_{1j} = 0$ for all the pre-measurements, and $D_{2j} = 1$ for the post-measurements. A significant interaction between the post-therapeutic measurement scores and a predictor variable would identify a change.

The MLM described in Equation (1) can be recognized as a repeated measures model, where the model describes the profile of two measurements for each subject. The well-known models for pre- post-therapeutic measurements are the change-score model (the difference in outcomes is regressed on the predictor variables) and the regressor variable method (ANCOVA; the post-therapeutic measurement is regressed on the pre-therapeutic measurement and predictor variables, e.g., Allison, 1990). Allison (1990) and Kutner and Brogan (1982) showed that the repeated measures model is more general than the change score model, which is more restrictive and provides less information about the data. Furthermore, it is possible to control for additional group differences at the pre-therapeutic measurement by including additional predictor variables (Schmidt et al., 2016). This can be beneficial for instances when different groups have not been randomly assigned to different treatments and pre-therapeutic measurement differences between groups need to be accounted for to measure treatment effects. According to the repeated measures model, the MLM for the CES-D scores using the post-therapeutic measurement indicator D is given by,

$$\begin{aligned}\text{CES-D}_{1j} &= \mu_0 + u_{0j} + \mathbf{X}_j\boldsymbol{\beta}_1 + e_{1j}\\ \text{CES-D}_{2j} &= \mu_0 + u_{0j} + \mathbf{X}_j\boldsymbol{\beta}_1 + \mu_1 I(D_{2j} = 1)\\ &\quad + \mathbf{X}_j I(D_{2j} = 1)\boldsymbol{\beta}_2 + e_{2j}.\end{aligned}$$

The parameters $\boldsymbol{\beta}_1$ represent the common effects of the predictor variables $\mathbf{X}$ on the outcomes CES-D and explain part of the common variance in the pre- and post-therapeutic measurements. The intercept μ_0 represents the average score level at the pre-therapeutic measurement, and the μ_1 the average change in scores between the pre- and post-therapeutic measurements. Given the effects of the predictor variables, the μ_1 represents the assessed average change in measurements that is not explained by any predictor variable. The parameters $\boldsymbol{\beta}_2$ represents the contribution of the predictor variables in explaining unique variance in the post-therapeutic measurement scores. Significant interaction $\boldsymbol{\beta}_2$ effects identify and explain a change in scoring between the pre- and post-therapeutic measurements.

The first model, our "null" model, acted as a baseline, hence the name M0. In M0, we test whether a random intercept for each client explains variability in outcome scores across clients. In ME1, we test whether the text-predictor variable positive emotion words contributes to explaining the unique variance in post-therapeutic scores. In ME2, MCP1, and MCP1 we test similar hypotheses, but then with the number of the number of negative emotion, insight and cause words.

3. RESULTS

We intended this section as a guideline for TCPR researchers who aspire to use text mining for multilevel modeling. We start with a statistical summarization of the variables that we used in our five multilevel models. Then we present and interpret the fixed and random effects of these models, and the corresponding goodness of fit measures. In doing so, we hope to give guidance of how these two frameworks should be combined, without presenting results of statistical significance.

3.1. Variable Descriptions

In total, we used five variables, one from the intervention from Lamers et al. (2015); we obtained the other four (text) variables from the LIWC program by Pennebaker et al. (2015b). The CES-D score ($M = 19.42$, $SD = 8.75$), the number of positive emotion words ($M = 40.25$, $SD = 27.15$), the number of negative emotion words ($M = 21.62$, $SD = 11.73$), the number of insight words ($M = 48.19$, $SD = 28.49$), and the number of cause words ($M = 20.93$, $SD = 14.1$) are summarized in **Table 1** (mean and standard deviations in the text are combinations of the pre- and post-therapeutic measurements).

3.2. Multilevel Models

In total, we estimated five multilevel models (see **Figure 1**). The post-therapeutic measurement of CES-D was the main outcome. In M0, model 0, we estimated the post-therapeutic measurement based on a random intercept for each client. M0 is nested under the other four models. In ME1 ("Model *Emotion*"), we estimated the post-therapeutic effect of the number of positive emotions words and a random intercept for each client. ME2 was similar to ME1, but instead of positive emotion words, we estimated the effect of (the number of) negative emotions words. MCP1 ("Model *Cognitive Process*") was similar in the same respect: we estimated the effect of insight words (instead of positive or negative words), and in MCP2 we estimated the effect of cause words.

TABLE 1 | Descriptive statistics of the CES-D score, insight, cause, positive, and negative emotion words from the e-mails of the clients on the pre- (T_0) and post-therapeutic (T_1) measurement.

Variable	Time	*M*	*SD*	Median	Min.	Max.
CES-D	T_0	23.41	7.51	23	10	49
	T_1	15.42	8.07	14	1	37
Positive emotion	T_0	36.78	20.73	35	2	110
	T_1	43.71	32.07	34	0	162
Negative emotion	T_0	25.47	16.17	22	1	77
	T_1	17.76	13.76	14	0	62
Insight	T_0	50.52	27.03	50	1	142
	T_1	45.86	29.84	41	2	173
Cause	T_0	21.32	12.92	18	2	59
	T_1	20.54	15.25	18	0	85

3.3. Interpretation

The data do not support our hypotheses that the writing intervention improves the number of positive, insight and cause words, while decreasing the number of negative words. Rather than using the data of Lamers et al. (2015) as a case to obtain new insights about TCPR, we present it as a use case for process researchers who wish to investigate e-mail data through multilevel models. Accordingly, we assessed the results in **Table 2** in four steps.

3.3.1. 1. Fixed Effects: Intercept and Post-therapeutic Indicator

The post-therapeutic effect of the writing intervention is estimated as the interaction ("interaction" in **Table 2**) between the model specific variable ("variable," with a varying meaning between the models, variable indicates the number of positive emotion words in ME1, negative emotion words in ME2, insight words in MCP1, and cause words in MCP2) and the post-therapeutic indicator ("post. indi.") in **Table 2**. As we are specifically interested in the post-therapeutic interaction effect, we do not interpret the effect of the model specific variable and post-therapeutic indicator in **Table 2**. The fixed effect of M0 is the grand mean (μ), which is interpretable as the positive effect of the writing treatment, without specific change effects of the

TABLE 2 | Model fit, parameter estimates and corresponding standard errors of the fixed and random effects of the five multilevel models.

	Baseline		Emotion				Cognitive processes			
	M0		ME1		ME2		MCP1		MCP2	
Fixed	23.41 (0.792)	*	22.83 (1.516)	**	22.61 (1.397)	**	21.62 (1.588)	**	21.69 (1.447)	**
Intercept	-7.99 (0.858)	*	0.02 (0.035)		0.03 (0.045)		0.04 (0.027)		0.08 (0.057)	
Post-indicator			-5.60 (1.735)	*	-6.33 (1.649)	**	-5.33 (1.820)	*	-5.50 (1.663)	**
Variable interaction			-0.06 (0.039)		-0.08 (0.066)		-0.05 (0.033)		-0.12 (0.068)	
Random										
σ_e^2	35.74		35.37		35.80		35.46		35.66	
τ	25.06		24.89		25.21		25.21		24.72	
Model fit										
deviance	1327.38		1323.48		1325.89		1324.58		1324.24	
AIC	1335.38		1335.48		1337.89		1336.58		1336.24	
BIC	1348.45		1355.09		1357.50		1356.19		1355.85	
LogLik	-663.69		-661.74		-662.95		-662.29		-662.12	
χ^2			3.89		1.48		2.80		3.13	
χ^2 df			2		2		2		2	
Effect size										
Ω_0^2 df	0.67		0.68		0.67		0.68		0.67	

The mean of the text-variable, indicated by "variable" in the table, changes between the five models: in ME1 it is the number of positive words, in ME2 it is the number of negative words, in MCP1 it is the number of insight words, and in MCP2 it is the number of cause words. The "interaction" variable is the interaction between the text variable and the post-therapeutic indicator ("post. indi.").

Coefficients (and standard errors).

**p < 0.01.*

***p < 0.001.*

word categories we included. We also estimated the effect of the post-therapeutic indicator. However, this effect should not be interpreted, as it merely acts as a dummy variable in our model.

3.3.2. 2. Assess Post-treatment Effects

There are two ways to evaluate the model(s). The first is based on values of the post-therapeutic interactions. **Table 2** does not give an indication that models ME1, ME2, MCP1, and MCP2 have significant post-treatment effects at the $p < 0.05$ level. Because all the relevant information lies in the interaction effect, the effect of the (text-)"variable" should also not be interpreted.

The second way to evaluate models is based on model fit. Of the all the model fit information in **Table 2**, the χ^2-test is perhaps the most straightforward to interpret, as it comes with a significance test. As none of the χ^2-tests are significant, the model fit information in **Table 2** does not indicate that one of the four models (ME1, ME2, MCP1, and MCP2) is a (significant) improvement over the baseline model M0. The other fit criteria should be seen as measures that indicate good model fit if they are closer to zero (there are several good sources, we suggest Burnham and Anderson, 2004, as a starting point).

3.3.3. 3. Random Effects

The variance of the random effect τ express the variation in post-therapeutic depression scores for individuals. The variance of the residual error σ_e^2 expresses the variance of the measurement errors, conditional on the individuals (the random effects). **Table 2** shows that the main effect of the text variables are—relative to the interaction effects—quite large. This is an indication that the sample (and population) are quite heterogeneous, making it difficult to estimate the effect of the writing intervention, as homogeneous treatment effect are simpler to estimate.

3.3.4. 4. Effect Size

For the calculation of the effect sizes, we followed the suggestions of Xu (2003). Ω_0^2 in **Table 2** is a generalization of the well-known R^2 measure, which can be interpreted as a measure for explained variance in multilevel models. Overall, **Table 2** shows that all models have a relative large proportion of explained variance. However, as model fit is (decimally) similar for all models, we cannot conclude that one model should be preferred over the others.

4. DISCUSSION

Key questions of Therapeutic Change Process Research (TCPR) usually adhere to obtaining a thorough understanding of the change processes that are (most) beneficial to the client. For TPCR, the pertinent question is not whether psychotherapy is effective, but how change occurs. It is common for TCPR to study the language used in the (therapeutic) interaction between client and counsellor in order to obtain answers to this question. Two challenges arise, how to obtain text-measures that relate to change processes, and how to analyse these change processes. We argued that text mining could be used for the first challenge, and multi-leveled models (MLMs) to overcome the second.

4.1. Conclusion

The complete-data subset from Lamers et al. (2015) does not suggest that the writing intervention contributes to change in the (number of) insight, cause, positive, and negative emotion words. The analyses show that the intervention does decrease post-therapeutic depression, however, the data did not indicate that this decrease could be associated with one of the text variables.

We aimed to make a case for the correct analyses of e-mail data, by obtaining text variables from large bodies of text, not to obtain theoretical insights. We showed that text mining is an appropriate tool to model change processes, as it can answer questions related to change processes.

The second goal of our paper was to show how complex and multi-layered change processes should be assessed. We presented a straightforward re-parametrization of multi-level models, that allowed for assessing post-therapeutic change. The way we parametrize our MLMs allows for modeling a baseline (pre-therapeutic score) and change (post-therapeutic score) over time, while accounting for the dependency between pre- and post-therapeutic score of each client. This also corresponds to growth modeling of multilevel data, where measurements are nested within subjects (Muthén, 1997). The association of specific text variables to the outcomes of the intervention was illustrative for these two points. Based on this proof-of-concept, we conclude that obtaining and analyses of textual information through text mining and MLMs can indeed advance TCPR.

4.1.1. Relevance

The main advantage of these models is that it opens up the possibility to engage more with clients in therapeutical settings. With online interventions on the rise, there is clear room to do so. The information from texts, which is directly accessible and does not require intensive transcription procedures, and can then be used to steer the therapeutic process in the desirable direction. Text mining can thus be used as a form of "direct feedback," as MLMs allow for correct modeling of the relations between variables.

4.2. Open Challenges

We proposed that text mining can be used to identify the important change processes within therapy related texts, and MLMs can be used to explain the relations between processes and outcomes. Full demonstration of the capabilities of this framework requires multiple datasets, and many of the problems that we faced require the attention of more researchers. We start the discussion session by describing these (open) challenges. Then, in the next section, we cover the limitations specific to our study.

4.2.1. Operationalization

Operationalization is one of the first challenges that users of text mining for TCPR face. Many of the TCPR constructs are theoretical, and need to be operationalized into linguistic features so that they are clearly distinguishable, measurable, and understandable in terms of empirical observations. Examples of these variables include emotional ventilation, dramatic relief,

tension release, abreaction, or catharsis (for more examples, see Grencavage and Norcross, 1990). Operationalization is not only an important aspect for TCPR, nor is it limited to psychology, the whole social and life sciences require good operationalizations.

The linguistic products of therapy (diaries, psychotherapeutic assignments, or transcripts of the therapeutic interaction) provide rich source of research material, provided that the variables of interest are adjustable to texts. In our current work, we used a basic text features from LIWC. We justified our use of these basic text features because we aimed to give a proof-of-concept with the intend of showing how TCPR and MLM can be bridged.

However, our choice for such a basic text variable leaves one of the largest challenges open: what to (text) mine? Traditionally, the text mining community was more concerned with collecting, storing and managing large bodies of unstructured text rather than applying theoretical models from other fields. Advances in the field of computer science made technical issues less insurmountable than they were a decade ago (Mayer-Schönberger and Cukier, 2013, p. 8). As a results, text mining is no longer reserved for those with a computer science degree.

The increase in solved technical issues did not lead to insights in "what to mine." We did not aim to advance TCPR theory with our current paper; we intended our work as a method paper, because with the current state of the literature, it is difficult to determine which textual predictors are meaningful. Also, we feel that our proposition to bridge text mining and MLMs itself allows for advancing TCPR theory. Constructs as described by Grencavage and Norcross (1990), Orlinsky et al. (2004), Elliott (2010), and Elliott (2012) require a 'translation', or adjustment, before text mining is applicable to these data types. Domain experts in the TCPR field are well-equipped to face this question, but this requires an interdisciplinary approach.

We showed how MLMs and text mining can be combined, but our proposition leaves open how TCPR concepts should be operationalized for text mining metrics. That would require an interdisciplinary collaboration and discussion. However, the future does look bright: based on our proof-of-concept study we conclude that MLMs and text mining can indeed advance TCPR.

The next step in that direction, would be to—aside from LIWC—incorporate other existing text mining software, such as TCM (*Therapeutic Cycles Model*; Mergenthaler, 1996), or CALAS (*Computer Assisted Language Analysis System*; Anderson et al., 1999).

4.2.2. Measurement Error

Elliott (2010) argued that TCPR is plagued by measurement error. Although the term "error" is often used, in our experience, it can refer to two different concepts depending on the field of study. With the risk of over-generalization, in the machine learning community and other fields that rely heavily on predictive analytics, error often refers to the error or confusion matrix. The table of confusion reports the number of false positives and negatives, and the true positives, and negatives. These measurement represent the performance of an algorithm. Error then refers to measures of predictive error, the difference between the observed values and the values predicted by the model.

In statistics, error is related to measurement error, which represents the difference between a measured value of a quantity and its true value. Measurement error is often used to indicate whether or not measurement is reliable. Reliability expresses how repeatable measurements are when remeasured. The reliability of a measure is then a direct function of the amount of error is present in the measurement. Because no behavioral measure is perfectly reliable, some degree of measurement error will always occur. Therefore, reliability is low when there is a abundance of error, and vice versa. The underlying idea is that every observation is a combination of the hypothetical true score plus some measurement error.

Although nowadays ideas appear to be floating freely between machine learning and statistics (Wasserman, 2010, p. 8), some concepts—such as measurement error—are traditionally more associated with one branch rather than the other (see for example Donoho, 2017). Measurement error is well-established in statistics, and has potential for machine learning disciplines such as text mining. Variables are simply an operationalization of the process, behavior or item that we are trying to measure. Estimation of the measurement error reflects the uncertainty present in the estimate. Consistency of the research measures benefits when accounting for measurement error.

In fact, with respect to measurement error, MLMs are the way forward. MLMs recognize the existence of several levels, nesting and hierarchies in data. MLMs capitalize on this concept by allowing for the inclusion of residual components at each level of the hierarchy. Hence, the precision of the estimation of measurement error increases, as the residual variance is partitioned.

4.2.3. Sample Size

TCPR is rooted in qualitative research methods; MLMs come from the quantitative sciences. Intensive case-studies are not uncommon for qualitative scientists, but will lead to statistical power issues for MLMs. As MLMs introduce multiple levels, the total number of units observed for each level become the sample size. The relevant sample size for power issues depends on the parameters that are being tested. Unlike the traditional regression, there is a difference between testing a regression coefficient or a variance parameters in a MLM.

The main limit is the sample size at the highest level of organization. Naturally, having multiple measures (at the first level) for one client (second level) is less informative then having these same multiple measures for multiple clients. The number of clients will therefore be one of the main issues for using MLMs for TCPR, but it will limit the wide scale application of MLMs for TCPR.

4.3. Limitations

We already gave an impression of some overarching open challenges that—in their current form—limit the applicability and wide-scale impact of the ideas we presented in the current work.

4.3.1. Excluded Therapy

Based on the design of Lamers et al. (2015), it would also have been straightforward to model the effect of treatment. Modeling treatment as a random effect could have provided an insight in the efficacy of the treatment for each individual client. The fixed effect of treatment would have given some insight in the average efficacy of the treatment groups in comparison with each other.

We however, as Lamers et al. (2015), we could not differentiate between the two conditions of the treatment. They found both writing conditions to be helpful in comparison to the control group, but could not differentiate between the expressive writing and autobiographical writing conditions.

We justified our exclusion furthermore because we only intended to show that text mining can be used to obtain additional predictors for multilevel models. Our intend was not to offer new theoretical insights for psychological writing interventions; we intended to offer methodological rather theoretical insights.

4.3.2. Complete Cases

We only included clients with complete cases and did not attempt to account for the missing data. First of all, it was difficult to determine why certain measurements where missing for an individual. Lamers et al. (2015) gave an overview of drop-out and missing data: it was challenging for us to determine *post-hoc* what the exact reason for missing data or drop out was for an individual based on general information.

Because we did not understand the underlying reason for the occurrence of missing data, we were hesitant in choosing an imputation technique. Also, because we did not intend to draw theoretical conclusions from our work, we felt that the issues with generalization and validity associated with ignoring missing data were less relevant for our proof-of-concept.

4.4. Future Research

MLMs come with the well-known advantage that the model can incorporate the hierarchical structure of the data. This is idea holds potential for TCPR, as change processes are often multifaceted and multi-layered. For example, an interesting analyses would be to see the effect counselors have on their clients. As a counsellor almost always treats multiple clients, it is possible to estimate the effect of a counsellor on its clients. Combining this form of nesting with other forms of nesting, such as the treatment effect itself, it would then be possible to estimate counsellor efficacy in different arms of the treatment. Accounting for clustering influences the estimation of the treatment effect as these influences are expressed as parameters in the model.

TCPR would also receive an enormous boost when change processes could be automatically detected through text mining. Some methods, such as the *Innovative Moments Coding Scale* (perhaps better known under its abbreviated name ICMS, see Gonçalves et al., 2009, 2010), already provided an avenue for doing so.

We are optimistic about TCPR's future through the happy marriage between text mining and MLMs. Especially in the social sciences, many phenomena can considered to be leveled, and the usage of text mining is already picking up. Social scientists in general often intend to learn about relations between variables in the population. In our view, in comparison with machine learning models, MLMs are of use to social scientists because they can provide theoretical insights in the relationships between, rather than building a black box model with the goal of attaining good predictive qualitative. MLMs can thus be used to explain relations between variables, whereas text mining can thus be used to obtain important therapy related variables, given that other TCPR research point in the direction of which important constructs are present in texts.

ETHICS STATEMENT

We reused the data of Lamers et al. (2015). Westerhof is also a project member of the current team, so we reused his data, that already received medical ethical approval and - as can be read in the paper - was anonymized before shared with the current authors.

AUTHOR CONTRIBUTIONS

WS conducted the research and wrote the paper. J-PF contributed to the multilevel analyses. ET contributed to data handling in Python and pre-processing of the data. GW contributed the data. J-PF, ET, GW, AS, and BV helped with setting up project and gave feedback throughout.

FUNDING

The *What Works When for Whom* project was supported by the Accelerating Scientific Discovery Call grant from the Netherlands eScience Centre (Amsterdam, the Netherlands): 027.015.G04. The eScience centre is funded through the Netherlands Organization for Scientific Research (NWO *Dutch*: Nederlandse organisatie voor Wetenschappelijk Onderzoek). We received no additional forms of funding.

REFERENCES

Adelson, J. L., and Owen, J. (2012). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy* 49, 152–162. doi: 10.1037/a0023990

Adler, J. M. (2012). Living into the story: agency and coherence in a longitudinal study of narrative identity development and mental health over the course of psychotherapy. *J. Pers. Soc. Psychol.* 102, 367–389. doi: 10.1037/a0025289

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociol. Methodol.* 20, 93–114.

Anderson, T., Bein, E., Pinnell, B. J., and Strupp, H. H. (1999). Linguistic analysis of affective speech in psychotherapy: a case grammar approach. *Psychother. Res.* 9, 88–99.

Andersson, G., Cuijpers, P., Carlbring, P., Riper, H., and Hedman, E. (2014). Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. *World Psychiatry* 13, 288–295. doi: 10.1080/16506070903318960

Andrade, L., Caraveo-Anduaga, J. J., Berglund, P., Bijl, R., Kessler, R. C., Demler, O., et al. (2013). Cross-national comparisons of the prevalences and correlates of mental disorders. *Bull. World Health Organ.* 78, 413–426. doi: 10.1590/S0042-96862000000400003

Andrews, G., Cuijpers, P., Craske, M. G., McEvoy, P., and Titov, N. (2010). Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. *PLOS ONE* 5:e13196. doi: 10.1371/journal.pone.0013196

Andrews, G., Issakidis, C., Sanderson, K., Corry, J., and Lapsley, H. (2004). Utilising survey data to inform public policy: comparison of the cost-effectiveness of treatment comparison of the cost-effectiveness of treatment of ten mental disorders. *Br. J. Psychiatry* 184, 526–533.

Angus, L. E., Levitt, H., and Hardtke, K. (1999). The narrative processes coding system: research applications and implications for psychotherapy practice. *J. Clin. Psychol.* 55, 1255–1270.

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., and Christensen, A. (2012). Topic models: a novel method for modeling couple and family text data. *J. Fam. Psychol.* 26, 816–827. doi: 10.1037/a0029607

Atkins, D. C., Steyvers, M., Imel, Z. E., and Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implement. Sci.* 9:49. doi: 10.1186/1748-5908-9-49

Baldwin, S. A., and Imel, Z. E. (2013). "Chapter 8: Therapist effects," in *Handbook of Psychotherapy and Behavior Change, 6th Edn*, eds M. J. Lambert (Hoboken, NJ: John Wiley & Sons, Ltd.), 258–297.

Baldwin, S. A., Wampold, B. E., and Imel, Z. E. (2007). Untangling the alliance-outcome correlation: exploring the relative importance of therapist and patient variability in the alliance. *J. Consult. Clin. Psychol.* 75, 842–852. doi: 10.1037/0022-006X.75.6.842

Barkham, M., Stiles, W. B., and Shapiro, D. A. (1993). The shape of change in psychotherapy: longitudinal assessment of personal problems. *J. Consult. Clin. Psychol.* 61:667.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Beekman, A. T. F., Deeg, D. J. H., van Limbeek, J., Braam, A. W., de Vries, M. Z., and van Tilburg, W. (1997). Criterion validity of the Center for Epidemiologic Studies Depression scale (CES-D): results from a community-based sample of older subjects in the Netherlands. *Psychol. Med.* 27, 231–235.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python, 1st Edn*. Sebastopol, CA:O'Reilly Media.

Birren, J. E., and Deutchman, D. E. (1991). *Guiding Autobiography Groups for Older Adults: Exploring the Fabric of Life*. John Hopkins University Press.

Bluck, S., and Levine, L. J. (1998). Reminiscence as autobiographical memory: a catalyst for reminiscence theory development. *Ageing Soc.* 18, 185–208.

Bohlmeijer, E. T., and Westerhof, G. J. (2010). *Op Verhaal Komen. Je Autobiografie als Bron van Wijsheid [The Stories We Live by]*. Amsterdam: Uitgeverij Boom.

Braakmann, D. (2015). "Chapter 3: Historical paths in psychotherapy research," in *Psychotherapy Research: Foundations, Process, and Outcome, 1 Edn*, eds O. C. G. Gelo, A. Pritz, and B. Rieken (Wien: Springer), 39–65.

Brewin, C. R. (2006). Understanding cognitive behaviour therapy: a retrieval competition account. *Behav. Res. Ther.* 44, 765–784. doi: 10.1016/J.BRAT.2006.02.005

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33:261304.

Butler, R. N. (1963). The life review: an interpretation of reminiscence in the aged. *Psychiatry* 26, 65–76.

Campbell, L. F., Norcross, J. C., Vasquez, M. J. T., and Kaslow, N. J. (2013). Recognition of psychotherapy effectiveness: the APA resolution. *Psychotherapy* 50:98. doi: 10.1521/psyc.2009.72.1.32

Campbell, R., and Pennebaker, J. W. (2003). The secret life of pronouns: flexibility in writing style and. *Sour. Psychol. Sci.* 14, 60–65.

Crits-Christoph, P., Gibbons, M. B. C., and Mukherjee, D. (2013). "Chapter 9: Psychotherapy process-outcome research," in *Handbook of Psychotherapy and Behavior Change, 6th Edn*, ed M. J. Lambert (Hoboken, NJ: John Wiley & Sons, Ltd.), 298–340. , .

Crowder, M. J., and Hand, D. J. (1990). *Analysis of Repeated Measures, 1st Edn*. New York, NY: Chapman & Hall/CRC.

Donoho, D. (2017). 50 years of data science. *J. Comput. Graph. Stat.* 26, 745–766. doi: 10.1080/10618600.2017.1384734

Elliott, R. (2010). Psychotherapy change process research: realizing the promise. *Psychother. Res.* 20, 123–135. doi: 10.1080/10503300903470743

Elliott, R. (2012). "Chapter 6: Qualitative methods for studying psychotherapy change processes," in *Qualitative Research Methods in Mental Health and Psychotherapy: A Guide for Students and Practitioners, 1 Edn*, eds A. R. Thompson and D. Harper (Chichester: Wiley-Blackwells), 69–111.

Fast, L. A., and Funder, D. C. (2008). Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *J. Pers. Soc. Psychol.* 94, 334–346. doi: 10.1037/0022-3514.94.2.334

Feldman, R., and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis, 2nd Edn*. Hoboken, NJ: John Wiley & Sons.

Fontao, M. I., and Mergenthaler, E. (2008). Therapeutic factors and language patterns in group therapy application of computer-assisted text analysis to the examination of microprocesses in group therapy: preliminary findings. *Psychother. Res.* 18, 345–354. doi: 10.1080/10503300701576352

Gelo, O. C. G., and Manzo, S. (2015). "Chapter 13: Quantitative approaches to treatment process, change process, and process-outcome research," in *Psychotherapy Research*, eds O. C. G. Gelo, A. Pritz, and B. Rieken (Vienna: Springer), 247–277.

Gelo, O. C. G., Rieken, B., and Pritz, A. (2015). *Psychotherapy Research: Foundations, Process, and Outcome*. Wien: Springer.

Gelo, O. C. G., Salcuni, S., and Colli, A. (2012). Text analysis within quantitative and qualitative psychotherapy process research: introduction to special issue. *Res. Psychother. Psychopathol. Process Outcome* 15, 45–53. doi: 10.7411/RP.2012.005

Gibson, J., Malandrakis, N., Romero, F., Atkins, D. C., and Narayanan, S. S. (2015). Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. in *Sixteenth Annual Conference of the International Speech Communication Association* (Dresden), 1947–1951.

Gonçalves, M. M., Matos, M., and Santos, A. (2009). Narrative therapy and the nature of "innovative moments" in the construction of change. *J. Construct. Psychol.* 22, 1–23.

Gonçalves, M. M., Ribeiro, A. P., Matos, M., Santos, A., and Mendes, I. (2010). Innovative moments coding system: a methodological procedure for tracking changes in psychotherapy. *YIS Yearbook Idiogr. Sci.* 2, 107–130.

Greenberg, L. S. (1986). Change process research. *J. Consult. Clin. Psychol.* 54:4.

Greenberg, L. S. (2007). A guide to conducting a task analysis of psychotherapeutic change. *Psychother. Res.* 17, 15–30. doi: 10.1080/10503300600720390

Grencavage, L. M., and Norcross, J. C. (1990). Where are the commonalities among the therapeutic common factors? *Profess. Psychol. Res. Pract.* 21, 372–378.

Haight, B. K., and Webster, J. D. (1995). *The Art and Science of Reminiscing: Theory, Research, Methods, and Applications*. Washington, DC: Taylor & Francis.

He, Q., Veldkamp, B. P., and de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: a text mining approach. *Psychiatry Res.* 198, 441–447. doi: 10.1016/j.psychres.2012.01.032

Hill, C. E., and Corbett, M. M. (1993). A perspective on the history of process and outcome research in counseling psychology. *J. Counsel. Psychol.* 40, 3. doi: 10.1037/0022-0167.40.1.3

Hill, C. E., and Lambert, M. J. (2004). "Methodological issues in studying psychotherapy processes and outcomes," in *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*, Vol. 5, ed M. J. Lambert (New York, NY: John Wiley & Sons), 87–135.

Hirsh, J. B., and Peterson, J. B. (2009). Personality and language use in self-narratives. *J. Res. Pers.* 43, 524–527. doi: 10.1016/j.jrp.2009.01.006

Hoogendoorn, M., Berger, T., Schulz, A., Stolz, T., and Szolovits, P. (2017). Predicting social anxiety treatment outcome based on therapeutic

email conversations. *IEEE J. Biomed. Health Inform.* 21, 1449–1459. doi: 10.1109/JBHI.2016.2601123

Imel, Z. E., Steyvers, M., and Atkins, D. C. (2015). Computational psychotherapy research: scaling up the evaluation of patient–provider interactions. *Psychotherapy* 52:19. doi: 10.1037/a0036841

Jurafsky, D., and Martin, J. H. (2017). *Speech and Language Processing: An Introduction to Natural Language Processing, 3 Edn*. New Jersey, NJ: Pearson Prentice Hall.

Kazdin, A. E., and Nock, M. K. (2003). Delineating mechanisms of change in child and adolescent therapy: Methodological issues and research recommendations. *J. Child Psychol. Psychiatry* 44, 1116–1129.

Kenny, D. A., and Hoyt, W. T. (2009). Multiple levels of analysis in psychotherapy research. *Psychother. Res.* 19, 462–468. doi: 10.1080/10503300902806681

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., and Kashy, D. A. (2002). The statistical analysis of data from small groups. *J. Pers. Soc. Psychol.* 83, 126–137. doi: 10.1037/0022-3514.83.1.126

Knobloch-Fedders, L. M., Elkin, I., and Kiesler, D. J. (2015). Looking back, looking forward: a historical reflection on psychotherapy process research. *Psychother. Res.* 25, 383–395.

Korte, J. (2012). *The Stories We Live by : The Adaptive Role of Reminiscence in Later Life* (PhD thesis). University of Twente, Enschede.

Kutner, M. H. and Brogan, D. R. (1982). Comparative analyses of pretest-posttest research designs. *Am. Stat.* 34, 229–232.

Lamers, S. M., Bohlmeijer, E. T., Korte, J., and Westerhof, G. J. (2015). The efficacy of life-review as online-guided self-help for adults: a randomized trial. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* 70, 24–34. doi: 10.1093/geronb/gbu030

Lepper, G., and Mergenthaler, E. (2005). Psychotherapy research exploring group process. *Psychother. Res.* 15, 433–444. doi: 10.1080/10503300500091587

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing, 1st Edn*. Cambridge, MA: The MIT Press.

Marcus, D. K., Kashy, D. A., and Baldwin, S. A. (2009). Studying psychotherapy using the one-with-many design: the therapeutic alliance as an exemplar. *J. Counsel. Psychol.* 56, 537–548. doi: 10.1037/a0017291

Mayer-Schönberger, V., and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: Houghton Mifflin Harcourt.

McCarthy, K. L., Caputi, P., and Grenyer, B. F. (2017). Significant change events in psychodynamic psychotherapy: is cognition or emotion more important? *Psychol. Psychother. Theor. Res. Pract.* 90, 377–388. doi: 10.1111/papt.12116

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: a new way of describing psychotherapeutic processes. *J. Consult. Clin. Psychol.* 64, 1306–1315.

Muntigl, P., and Horvath, A. (2005). "Chapter 10: Language, psychotherapy and client change: an interdisciplinary perspective," in *A New Agenda in (Critical) Discourse Analysis*, eds R. Wodak and P. A. Chilton (Amsterdam: John Benjamins), 213–240.

Muthén, B. (1997). Latent variable growth modeling with multilevel data. *Sociol. Methodol.* 27, 453–480.

Nissen-Lie, H. A., Monsen, J. T., Rønnestad, M. H., Trygve Monsen, J., and Rønnestad, M. H. (2010). Therapist predictors of early patient-rated working alliance: a multilevel approach. *Psychother. Res.* 20, 627–646. doi: 10.1080/10503307.2010.497633

Nock, M. K. (2007). Conceptual and design essentials for evaluating mechanisms of change. *Alcohol. Clin. Exp. Res.* 31, 4s–12s.

Norcross, J. C., and Wampold, B. E. (2011). What works for whom: tailoring psychotherapy to the person. *J. Clin. Psychol.* 67, 127–132.

Orlinsky, D. E., Ronnestad, M. H., and Willutzki, U. (2004). Fifty years of psychotherapy process-outcome research: Continuity and change. *Handb. Psychother. Behav. Change* 5, 307–389.

Pasupathi, M., and Carstensen, L. L. (2003). Age and emotional experience during mutual reminiscing. *Psychol. Aging* 42, 798–808. doi: 10.1037/0882-7974.18.3.430

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *J. Consult. Psychol.* 31, 109–118.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015a). *The Development and Psychometric Properties of LIWC2015*. UT Faculty/Researcher Works.

Pennebaker, J. W., and Chung, C. (2011). "Chapter 18: Expressive writing and its links to mental and physical health," in *The Oxford Handbook of Health Psychology*, ed H. S. Friedman (Oxford, UK: Oxford University Press), 417–437. doi: 10.1093/oxfordhb/9780195342819.001.0001

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2015b). *Linguistic Inquiry and Word Count: LIWC 2015*. Mahwah, NJ: Lawrence Erlbaum Associates, 71.

Pennebaker, J. W., Mayne, T. J., and Francis, M. E. (1997). Linguistic Predictors of Adaptive Bereavement. *J. Pers. Soc. Psychol.* 72, 863–871.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annu. Rev. Psychol.* 54, 547–577. doi: 10.1146/annurev.psych.54.101601.145041

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., and An, L. (2017). "Understanding and predicting empathic behavior in counseling therapy," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vancouver, BC), 1426–1435.

Pfäfflin, F., Böhmer, M., Cornehl, S., and Mergenthaler, E. (2005). What happens in therapy with sexual offenders? A model of process research. *Sex. Abuse J. Res. Treat.* 17, 141–151. doi: 10.1007/s11194-005-4601-2

Python Software Foundation (2018). *Python Language Reference, version 3.6*. Wilmington. Available online at: http://www.python.org.

R Core Team (2019). *R: A Language and Environment for Statistical Computing [Version 3.5.3]*. Vienna. doi: 10.1007/978-3-540-74686-7

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Measure.* 1, 385–401. doi: 10.1177/014662167700100306

Raudenbush, S. W., and Bryk, A. S. (eds.). (2002). "Chapter 6: Applications in the study of individual change," in *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn* (Los Angeles, CA: SAGE Publications Ltd), 160–204.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. [R package version 1.8.12]. Evanston, IL. Available online at: https://cran.r-project.org/package=psych.

Schegloff, E. (2007). *Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis*. New York, NY: Cambridge University Press.

Schmidt, S., Zlatkin-Troitschanskaia, O., and Fox, J.-P. (2016). Pretest-posttest-posttest multilevel IRT modeling of competence growth of students in higher education in Germany. *J. Educ. Meas.* 53, 332–351.

Serrano, J. P., Latorre, J. M., Gatz, M., and Montanes, J. (2004). Life review therapy using autobiographical retrieval practice for older adults with depressive symptomatology. *Psychol. Aging* 19, 272–277. doi: 10.1037/0882-7974.19.2.272

Tasca, G. A., and Gallop, R. (2009). Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics. *Psychother. Res.* 19, 429–437. doi: 10.1080/10503300802641444

Tausczik, Y. R., and Pennebaker, J. W. (2010). The Psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 24–54. doi: 10.1177/0261927X09351676

Thorne, A., McLean, K. C., and Lawrence, A. M. (2004). When remembering is not enough: reflecting on self-defining memories in late adolescence. *J. Pers.* 72, 513–541. doi: 10.1111/j.0022-3506.2004.00271.x

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., et al. (2007). Worldwide use of mental health services for anxiety, mood, and substance disorders: results from 17 countries in the WHO World Mental Health (WMH) Surveys. *Lancet* 370, 841–850. doi: 10.1016/S0140-6736(07)61414-7

Wasserman, L. (2010). *All of Statistics*, Vol. C. Berlin: Springer Texts in Statistics. doi: 10.1007/978-0-387-21736-9

Westerhof, G. J., Bohlmeijer, E. T., and McAdams, D. P. (2017). The relation of ego integrity and despair to personality traits and mental health. *J. Gerontol. Ser. B* 72, 400–407. doi: 10.1093/geronb/gbv062

Westerhof, G. J., Bohlmeijer, E. T., Van Beljouw, I. M., and Pot, A. M. (2010a). Improvement in personal meaning mediates the effects of a life review intervention on depressive symptoms in a randomized controlled trial. *Gerontologist* 50, 541–549. doi: 10.1093/geront/gnp168

Westerhof, G. J., Bohlmeijer, E. T., and Webster, J. D. (2010b). Reminiscence and mental health: a review of recent progress in theory, research and interventions. *Ageing Soc.* 30, 697–721. doi: 10.1017/S0144686X09990328

White, M. (2007). *Maps of Narrative Practice*. New York, NY: WW Norton & Company.

White, M., and Epston, D. (1990). *Narrative Means to Therapeutic Ends*. New York, NY: WW Norton & Company.

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., et al. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* 382, 1575–1586. doi: 10.1016/S0140-6736(13)61611-6

Williams, J. M. G., Barnhofer, T., Crane, C., Hermans, D., Raes, F., Watkins, E., et al. (2007). Autobiographical memory specificity and emotional disorder. *Psychol. Bull.* 133, 122–148. doi: 10.1037/0033-2909.133.1.122

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Stat. Med.* 22, 3527–3541. doi: 10.1002/sim.1572

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Smink, Fox, Tjong Kim Sang, Sools, Westerhof and Veldkamp. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

METHODS
published: 24 April 2019
doi: 10.3389/fpsyg.2019.00887

Scaling the Variance of a Latent Variable While Assuring Constancy of the Model

Karl Schweizer[1]*, *Stefan J. Troche*[2] and *Christine DiStefano*[3]

[1] Institute of Psychology, Goethe University Frankfurt, Frankfurt, Germany, [2] Department of Psychology, University of Bern, Bern, Switzerland, [3] Department of Educational Studies, University of South Carolina, Columbia, SC, United States

This paper investigates how the major outcome of a confirmatory factor investigation is preserved when scaling the variance of a latent variable by the various scaling methods. A constancy framework, based upon the underlying factor analysis formula that enables scaling by modifying components through scalar multiplication, is described; a proof is included to demonstrate the constancy property of the framework. It provides the basis for a scaling method that enables the comparison of the contribution of different latent variables of the same confirmatory factor model to observed scores, as for example, the contributions of trait and method latent variables. Furthermore, it is shown that available scaling methods are in line with this constancy framework and that the criterion number included in some scaling methods enables modifications. The impact of the number of manifest variables on the scaled variance parameter can be modified and the range of possible values. It enables the adaptation of scaling methods to the requirements of the field of application.

Keywords: scaling, variance parameter, variance of latent variable, confirmatory factor analysis, structural equation modeling, scaling methods, constancy framework

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Lixiong Gu,
Educational Testing Service,
United States
Ben Kelcey,
University of Cincinnati, United States

***Correspondence:**
Karl Schweizer
K.Schweizer@psych.uni-frankfurt.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *04 September 2018*
Accepted: *03 April 2019*
Published: *24 April 2019*

Citation:
Schweizer K, Troche SJ and DiStefano C (2019) Scaling the Variance of a Latent Variable While Assuring Constancy of the Model. Front. Psychol. 10:887. doi: 10.3389/fpsyg.2019.00887

INTRODUCTION

In evaluating the results of factor analysis, the focus is usually on the factor loadings as related to the magnitude and the direction of the relationship to the latent variable. While also a parameter of the model, under factor analysis, the variance of the latent variable is largely ignored as a source of information for evaluation. A reason for ignoring the variance as a source of information is its dependency on the indicator selected for scaling in order to achieve model identification. It is well-known that modifying scaling by replacing one indicator by another one changes the value of the variance among other consequences (e.g., Gonzalez and Griffin, 2001; Steiger, 2002). Such dependency does not endorse the variance of the latent variable as a reliable source of information.

Despite the dependency on indicator selection, factor variance can be an important piece of information for evaluation. Even though it is commonly ignored, the variance of a latent variable has been recognized as a useful source of information for some specific areas, in particular, longitudinal research and invariance analyses (McArdle and Cattell, 1994; Schmitt and Kuljanin, 2008; McArdle, 2009). For example, the variance of a latent variable is used for evaluating development across time and for gaining insight about differences between groups. Besides these statistical approaches, there are further analysis strategies that may profit from comparisons of the variances of latent variables, such as the multitrait-multimethod approach (Marsh and Grayson, 1995) and the bifactor approach (Reise, 2012). Especially when using a multitrait-multimethod

design, it may be important to know how large the trait variance is in comparison to the method variance. This information reveals the relative contributions of different measures to the representation of a construct.

The particular interest in the scaling of latent variables has given rise to several specific methods that satisfy the needs of the corresponding areas of research (Little et al., 2006). For example, in longitudinal research (McArdle and Cattell, 1994) it is useful to scale the variance in such a way that it is set equal to one at the first measurement occasion. This approach establishes a baseline, and changes from the baseline to successive measurement occasions are more readily interpretable. Thus, different scaling methods may be of interest to achieve specific goals based upon the design under consideration. However, despite the different goals giving rise to different scaling methods, all methods must be able to preserve the major outcome of a confirmatory factor investigation while scaling transforms a statistic into a new reference system. Therefore, it should be possible to relate the various scaling methods to each other and to integrate them into a common framework.

The available methods for scaling variances (either implicitly or explicitly) include a definition of the relationship between the factor loadings and the variance of the corresponding latent variable (Little et al., 2006). Such a definition is also required in confirmatory factor analysis for specifying the model of the covariance matrix[1] (Jöreskog, 1970). Therefore, this model is considered as the framework that may preserve the major outcome of an investigation and is suitable for investigating scaling methods. We discuss this point in greater detail in the following sections where different scaling methods are considered and consequences of possible modifications are demonstrated.

SCALED VARIANCES AS SOURCES OF INFORMATION

In order to be regarded as an important source of information, the variance of the latent variable must be scaled; i.e., it must be adapted to the reference system of interest. This kind of adaptation requires that a content area is identified that potentially profits from the availability of scaled variances. Some content areas for scaling are already mentioned. In this section the perspectives of models of measurement are used for considering areas that may profit from scaling the variances of latent variables. Furthermore, scaling in confirmatory factor analysis (CFA) is compared with standardization for obtaining meaningful weights in linear regression analysis. Standardized regression weights enable the comparison of the contributions of independent variables to the explanation of the dependent variable.

Before detailing the process, the specificity of the variance characterizing a latent variable needs to be addressed. Both the latent variable and the variance are parts of a tested model and, therefore, to some degree are shaped by the characteristics of this model. The variance of the latent variable is assigned the role of a parameter of a model that is thought to reflect dispersion, but is not equivalent to the variance defined as the sum of squared deviations (Verbeke and Molenberghs, 2003; Stoel et al., 2006). For ease in communication, we stay with the term variance.

At first, the possible advantage of scaling the variance of the latent variable of a one-factor confirmatory factor model is considered. This model relates the $p\times1$ vector $\mathbf{x}$ representing the centered observations to the product composed of the $p\times1$ vector $\boldsymbol{\lambda}$ representing the factor loadings and of the latent variable ξ and to the $p\times1$ vector $\boldsymbol{\delta}$ representing the error influences (Graham, 2006):

$$\mathbf{x} = \boldsymbol{\lambda}\xi + \boldsymbol{\delta}. \tag{1}$$

There is also an extended version of this model (Miller, 1995; Raykov, 1997). It additionally includes the $p\times1$ vector $\boldsymbol{\mu}$ of latent intercepts and applies to the non-centered observations $\mathbf{X}$ instead of the centered observation $\mathbf{x}$:

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\lambda}\xi + \boldsymbol{\delta}. \tag{2}$$

This unidimensional model mainly serves the investigation of the structural validity and also of the convergent and discriminant validity of scales. Examinations are expected to provide information on the correctness of this model with respect to the given data. If the information suggests correctness (as shown by acceptable fit), it is retained; otherwise it is rejected. No further information requiring scaling is necessary unless there is a repeated application of the model.

The model of measurement of Equation (2) is designed according to the assumption that there is only one systematic source of responding. It ignores, for example, the well-known impurity problem that was observed in cognitive measures (Jensen, 1982; Miyake et al., 2000). It states that it is virtually impossible to complete many cognitive items without stimulating auxiliary processes besides the intended cognitive process. In other words, it is quite likely that there is at least one other process influencing the responses besides the process reflecting the construct, which is in the focus of the scale. This second process needs to be represented in the model of measurement as another source of responding by an additional component. Enlarging the model of measurement of Equation (2) gives the following:

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\lambda}_{\text{first source}}\,\xi_{\text{ first source}} + \boldsymbol{\lambda}_{\text{second source}}\,\xi_{\text{ second source}} + \boldsymbol{\delta} \tag{3}$$

where the labels *first source* and *second source* distinguish as subscripts between the construct source reflecting the intended cognitive process and the other source, the auxiliary process.

In the case of the two-factor confirmatory factor model, it may not be sufficient to know that the model is correct because there are two different sources showing different qualities. In the case of the second source being unrelated to the source captured by the scale, the two sources are a "good" source (related to the

[1]The model of the covariance matrix refers to the form of the Σ matrix of relationships included in the general factor analysis formula, typically written as: $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \Theta$ and defined in Equation (6).

construct of interest) and a "bad" source (clouding measurement of the construct) and, therefore, it is at least important to know whether the good one dominates the response, and it is even better to be able to show that the influence of the bad source is a minor influence on the responses. This means that the two latent variables constitute a reference system for scaling.

Distinguishing between good and bad sources is not just an idea but a real problem of substantive research. There are, for example, impure measures of working memory capacity showing this characteristic. We mention one major case of controversy that highlights the importance of quantifying the contributions of the additional sources to responding: there are now a number of studies reporting very high correlations between working memory capacity and intelligence suggesting more or less equivalence of working memory capacity and fluid intelligence. However, there is also good reason for suspecting that measures of working memory capacity do not only tap working memory capacity but also processing speed (Chuderski, 2013, 2015). Using a very large sample in an investigation focused on this issue, it was possible to demonstrate that minimizing the possible influence of processing speed lowered the correlation substantially. That processing speed is a threat to the validity of a measure is not only a problem of cognitive research but also of assessment in general. If there is a time limit in testing, processing speed is likely to contribute to performance (Oshima, 1994). The combination of a time limit in testing and processing speed impairs the validity of measurement (Lu and Sireci, 2007).

A similar situation is noted in linear regression analysis with two or more independent variables. The dependent variable is explained/predicted by the independent variables, and it is of interest to know about the relative contributions of the individual independent variables. These contributions are reflected by the regression weights. For demonstrating the structural similarity with the model of Equation (3), assume the dependent variable Y, the independent variables X_1 and X_2, the intercept b_0 and the error e (notation according to Osborne, 2017) that relate to each other according to the following equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + e \tag{4}$$

where b_1 and b_2 are the regression weights. Standardized regression weights signify the contributions of independent variables to the explanation of the dependent variable. These regression weights can be compared. For example, the weights can be used for evaluating contributions of independent variables that, for example, may be considered as variables reflecting good and bad sources.

The confirmatory factor model of Equation (3) includes equations showing a structure similar to Equation (4), as is obvious when using a more detailed way of presenting the vectors:

$$\begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ X_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_p \end{bmatrix} + \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ \lambda_p \end{bmatrix}_{\text{first source}} \times \xi_{\text{first source}} + \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ \lambda_p \end{bmatrix}_{\text{second source}} \times \xi_{\text{second source}} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ . \\ . \\ \delta_p \end{bmatrix} \tag{5}$$

There are factor loadings serving more or less the same purpose as the regression weights in regression analysis (λ_i instead of b_i). Although the estimation methods used in confirmatory factor analysis and linear regression analysis may differ from each other and lead to somewhat differing estimates, factor loadings, and regressions weights show some functional similarity.

However, in confirmatory factor analysis, the two sources that are to be compared with each other show not only one factor loadings, but p of them. This means that the factor loadings need to be integrated into one statistic. The variance can be this statistic since factor loadings and the variance of the latent variable depend on each other, as is demonstrated in the next section. The dependency is established by a framework. By means of this framework it becomes possible to relate variances scaled with respect to multiple indicators to the initially mentioned scaling by fixing one indicator (e.g., Gonzalez and Griffin, 2001; Steiger, 2002). Given this framework, it is shown in one of the following sections that it possible to achieve scaled variances, which can serve for comparisons like those by standardized regression weights, by one of the scaling methods.

CONSTANCY DUE TO SCALAR MULTIPLICATION

This section addresses the issue of constancy regarding the reproduction of the empirical covariance matrix by the model of the covariance matrix, despite scaling variance parameters. It is argued that constancy despite scaling by means of the various methods is accomplished by means of scalar multiplication. Scalar multiplication denotes the multiplication of a scalar and a matrix. The usefulness of scalar multiplication is detailed below.

Constancy is considered with respect to the model of the covariance matrix (Jöreskog, 1970) that is often symbolized by $\boldsymbol{\Sigma}$. This matrix (i.e., model of the covariance matrix) is denoted as the $p \times p$ model-implied covariance matrix for p manifest variables ($\boldsymbol{\Sigma} \in \Re^{p \times p}$) and is specified to reproduce the $p \times p$ empirical covariance matrix $\mathbf{S}$ ($\mathbf{S} \in \Re^{p \times p}$). Under CFA, the definition of the model $\boldsymbol{\Sigma}$, is given by the following equation:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{6}$$

where $\boldsymbol{\Sigma}$ is defined as the sum of $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$ and Θ. The product $\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'$ is composed of the $p \times q$ matrix of factor loadings $\boldsymbol{\Lambda}$ ($\boldsymbol{\Lambda} \in \Re^{p \times q}$) (and its transpose $\boldsymbol{\Lambda}'$) and the $q \times q$ matrix $\boldsymbol{\Phi}$

($\Phi \in \Re^{q \times q}$) consists of the variances and covariances of q latent variables. The second component in the equation is the $p \times p$ diagonal matrix of error components $\mathbf{\Theta}$ ($\mathbf{\Theta} \in \Re^{p \times p}$), which is linked additively to the first component.

The reasoning regarding constancy concentrates on $\mathbf{\Lambda \Phi \Lambda'}$ since constancy of this part of the model with respect to a specific empirical covariance matrix $\mathbf{S}$ implies that $\mathbf{\Theta}$ is also constant. Scaling the variance parameters of $\mathbf{\Lambda \Phi \Lambda'}$ in a manner that assures constancy means that the product (as a whole) is constant, although the factor loadings and the variance and covariance parameters may change.

A constancy framework for scaling. Assume the $p \times q$ matrices of factor loadings denoted $\mathbf{\Lambda}$ and $\mathbf{\Lambda^*}$ and the $q \times q$ matrices of the variances and covariances of latent variables denoted as $\mathbf{\Phi}$ and $\mathbf{\Phi^*}$. Constancy in the sense of equality of $\mathbf{\Lambda \Phi \Lambda}$and $\mathbf{\Lambda^* \Phi^* \Lambda^*}$,

$$\mathbf{\Lambda \Phi \Lambda'} = \mathbf{\Lambda^* \Phi^* \Lambda^{*\prime}}, \tag{7}$$

is given if there is a scaling constant c $\left(c \in \Re^{>0}\right)$ such that

$\mathbf{\Lambda^*} = c\ \mathbf{\Lambda}$

and

$\mathbf{\Phi^*} = 1/\ c^2\ \mathbf{\Phi}$.

Scaling is achieved by multiplying both $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ with c respectively the inverse of its square. In the following section it is demonstrated that the available scaling methods can be described in terms of this framework.

In order to ensure that the stated equality is correct, a proof is provided. The proof consists of three parts:

a) transformation of the left-hand side of Equation 7 to the right-hand side to illustrate equivalence (Proof 1)
b) demonstration that the products of matrices included in Equation 7 produce matrices of the same size (Proof 2), and
c) demonstration that all entries of the two products of matrices are the same (Proof 3).

Proof 1. First, $c \in \Re^{>0}$ is introduced into the left-hand side of Equation 7:

$$\mathbf{\Lambda \Phi \Lambda'} = 1 \times \mathbf{\Lambda \Phi \Lambda'} = \frac{c \times c}{c \times c} \times \mathbf{\Lambda \Phi \Lambda'} = c \times \frac{1}{c^2} \times c \times \mathbf{\Lambda \Phi \Lambda'}$$

The $\times$ symbol is used for explicitly emphasizing some cases of multiplication. As c is a scalar, $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ are matrices and the entries of the matrices are real numbers. Thus, the commutative and associative properties of scalar multiplication enable reordering of the scalars:

$$\begin{aligned} c \times \frac{1}{c^2} \times c \times \mathbf{\Lambda \Phi \Lambda'} &= c \times \mathbf{\Lambda} \times \frac{1}{c^2} \times \mathbf{\Phi} \times c \times \mathbf{\Lambda'} \\ &= (c \times \mathbf{\Lambda}) \times \left(\frac{1}{c^2} \times \mathbf{\Phi}\right) \times \left(c \times \mathbf{\Lambda'}\right) \\ &= (c\mathbf{\Lambda}) \left(\frac{1}{c^2}\mathbf{\Phi}\right) \left(c\mathbf{\Lambda'}\right) \end{aligned}$$

Finally, a product term is achieved that includes components that are in line with the replacement rules introduced in combination with Equation (2), $\mathbf{\Lambda^*} = c\ \mathbf{\Lambda}$ and $\mathbf{\Phi^*} = 1/\ c^2\ \mathbf{\Phi}$:

$$(c\mathbf{\Lambda}) \left(\frac{1}{c^2}\mathbf{\Phi}\right) \left(c\mathbf{\Lambda'}\right) = \mathbf{\Lambda^* \Phi^* \Lambda^{*\prime}}.$$

Proof 2. Since the product of the matrix of factor loadings and of the matrix of variances and covariances (and also the transpose of the matrix of factor loadings) is an additive component of the sum that constitutes the model of the covariance matrix according to Equation 6, the size of $\mathbf{\Lambda \Phi \Lambda'}$ is the same as the size of $\mathbf{\Sigma}$ that is, a $p \times p$ matrix. It remains to demonstrate that $\mathbf{\Lambda^* \Phi^* \Lambda^{*\prime}}$ is also a $p \times p$ matrix. Since c is a scalar, it does not change the size of the matrix to which it serves as multiplier. This means that the size of $c\ \mathbf{\Lambda}$ is the same as the size of $\mathbf{\Lambda}$, the size of $1/\ c^2\ \mathbf{\Phi}$ the same as the size of $\mathbf{\Phi}$, and the size of $c\ \mathbf{\Lambda'}$ the same as the size of $\mathbf{\Lambda'}$. Consequently, for $\mathbf{\Lambda^*} = \mathrm{c}\ \mathbf{\Lambda}$ and $\mathbf{\Phi^*} = 1/\ \mathrm{c}^2\ \mathbf{\Phi}$, the size of $\mathbf{\Lambda^* \Phi^* \Lambda^{*\prime}}$ is the same as the size of $\mathbf{\Lambda \Phi \Lambda'}$.

Proof 3. This proof requires the demonstration that the entries of $\mathbf{\Lambda \Phi \Lambda'}$ are the same as the entries of $\mathbf{\Lambda^* \Phi^* \Lambda^{*\prime}}$. Both products of matrices are considered as the true part of a $p \times p$ model-implied covariance matrix (i.e., the summand, excluding error of Equation 6); therefore, the entries of the ith row and jth column are represented by $\sigma_{\tau ij}$ and $\sigma^*_{\tau ij}$ across the two matrices, respectively. Given the research interest in investigating the variance at the latent level, $\mathbf{\Phi}$ is assumed to be a diagonal matrix[2].

In the case of q latent variables and diagonal $\mathbf{\Phi}$, the true (i.e., population) part of the ith row and jth column $\sigma_{\tau ij}$ is given by:

$$\sigma_{\tau ij} = \lambda_{i1}\Phi_{11}\lambda_{j1} + ... + \lambda_{iq}\Phi_{qq}\lambda_{jq}. \tag{8}$$

Analogically, the true part of the entry of the ith row and jth column ${\sigma_{ij}}^*$ is described by the following equation:

$$\sigma^*_{\tau ij} = \lambda^*_{i1}\phi^*_{11}\lambda^*_{i1} + ... + \lambda^*_{iq}\phi^*_{qq}\lambda^*_{iq}. \tag{9}$$

The next steps make use of scaling constant c as introduced in combination with Equation (7). Since $\mathbf{\Lambda^*}$ is set equal to $c\mathbf{\Lambda}$, the entry of the ith row and jth column of $\mathbf{\Lambda^*}$ (i.e., ${\lambda_{ij}}^*$) can be replaced by the entry of the ith row and jth column of $c\mathbf{\Lambda}$ (i.e., $c\lambda_{ij}$). Furthermore, as $\mathbf{\Phi^*}$ corresponds by definition to $1/c^2\ \mathbf{\Phi}$, the entry of the ith row and jth column of $\mathbf{\Phi^*}$ that is ${\phi_{ij}}^*$ can be replaced by the entry of the ith row and jth column of $1/c^2\ \mathbf{\Phi}$ that is $1/c^2\phi_{ij}$ such that:

$$\sigma^{*}_{\tau ij} = c\lambda_{i1}\frac{1}{c^2}\phi_{11}c\lambda_{j1} + ... + c\lambda_{iq}\frac{1}{c^2}\phi_{qq}c\lambda_{jq}.$$

Because scalar multiplication is also distributive, the equation can be further transformed into:

$$\sigma^{*}_{\tau ij} = c \times \frac{1}{c^2} \times c \times \left(\lambda_{i1}\phi_{11}\lambda_{j1} + ... + \lambda_{iq}\phi_{qq}\lambda_{jq}\right).$$

Since the sum given in parentheses of the right-hand side of this equation corresponds to the right-hand side of Equation (8), it can be replaced by the left-hand side of Equation 8:

$$\sigma^{*}_{\tau ij} = c \times \frac{1}{c^2} \times c \times \sigma_{\tau ij}.$$

[2]In this investigation, we are assuming orthogonal factors to place the focus on the decomposition of the latent variance in non-overlapping parts. This assumption is in line with the majority of models of measurement employed in assessment research (Graham, 2006). The omission of the interaction term helps to keep the illustration succinct and of manageable size.

In the final step, coefficients are arranged to provide a ratio that amounts to one:

$$\sigma^{*}_{\tau ij} = \frac{c^2}{c^2} \times \sigma_{\tau ij} = \sigma_{\tau ij} \quad \blacksquare$$

THE INTEGRATION OF THE SCALING METHODS INTO THE CONSTANCY FRAMEWORK

Given that the proof applies to all $p \times q$ matrices of factor loadings $\mathbf{\Lambda}$, it also applies to all $p \times 1$ matrices of factor loadings, referred to as $p \times 1$ vectors of factor loadings ($\boldsymbol{\lambda}$). In this case, the matrix of variances and covariances, $\mathbf{\Phi}$, reduces to the scalar, ϕ. This scalar is the *variance parameter* which represents the variance of the latent variable in the model of the covariance matrix. The status of this parameter as variance has been questioned since it can be assigned a negative value in the process of parameter estimation (Verbeke and Molenberghs, 2003; Stoel et al., 2006).

In this case of a $p \times 1$ vector of factor loadings, Equation (7) reduces to:

$$\boldsymbol{\lambda\phi\lambda}' = \boldsymbol{\lambda}^{*}\boldsymbol{\phi}^{*}\boldsymbol{\lambda}^{*\prime} \tag{10}$$

if there is a scaling constant $c \in \Re^{>0}$such that $\boldsymbol{\lambda}^{*} = c\,\boldsymbol{\lambda}$ and $\phi^{*} = 1/c^2\phi$.

The one-factor version of the constancy framework, as described by Equation (10) in combination with the two replacement rules, provides the basis for the following equation that related the scaled variance parameter ϕ_{sc} to the scaling constant c and to the original variance parameter ϕ:

$$\phi_{\mathrm{sc}} = \frac{1}{c^2}\phi. \tag{11}$$

Scaling the variance parameter through use of Equation (11) is a general scaling method, as c may be selected to represent different scaling methods. Furthermore, this equation can be used to investigate the properties of specific scaling methods and to compare their effects. The following subsections relate this approach to available scaling methods, including the marker-variable method, the reference-group method and criterion-based methods (e.g., effect-coding method; Little et al., 2006; Little, 2013). In the following subsections, each method is described.

The Marker-Variable Method

This frequently used method for scaling the variance parameter states that a value of one is assigned to one of the factor loadings (i.e., a marker variable) while the other factor loadings and the variance parameter of the latent variable are freely estimated. Such a configuration of free and fixed factor loadings is illustrated by **Figure 1**. A double circle identifies the factor loading selected for serving as indicator.

However, the influence of the marker variable is incorporated into the variance of the latent variable. Integrating this specific method into the constancy framework requires the choice of c with respect to the originally selected factor loading λ_i such that:

$$1 = \boldsymbol{\lambda}^{*}_{\mathbf{i}} = c\boldsymbol{\lambda}_{\mathbf{i}} \tag{12}$$

where λ_i refers to the left-hand part of Equation (10) and λ^{*}_{i} to the right-hand part. If $\lambda_i > 0$ then $c \in \Re^{>0}$. Given the original variance parameter, ϕ, the scaled variance parameter ϕ_{sc} is obtainable by means of Equation (11).

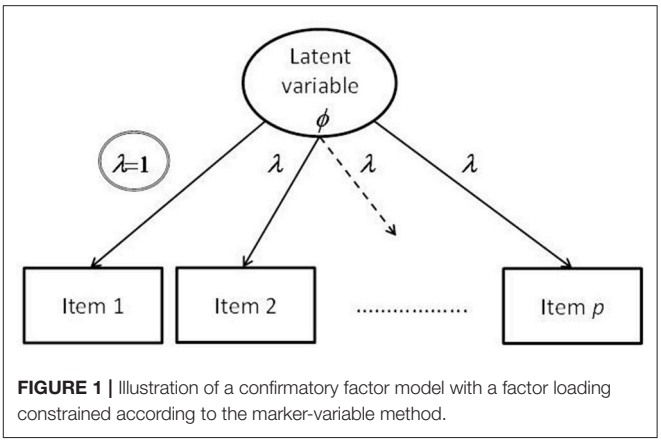


FIGURE 1 | Illustration of a confirmatory factor model with a factor loading constrained according to the marker-variable method.

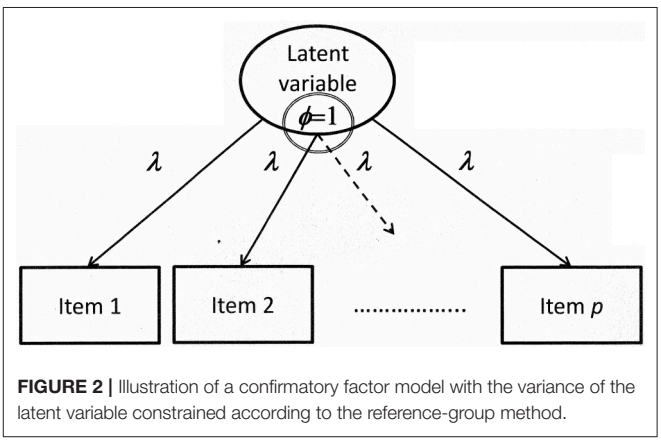


FIGURE 2 | Illustration of a confirmatory factor model with the variance of the latent variable constrained according to the reference-group method.

The Reference-Group Method

The reference-group method requires that the value of one is assigned to the variance parameter (i.e., standardized latent variables) while all factor loadings are freely estimated. This means that

$$1 = \phi_{\mathrm{sc}} = \frac{1}{c^2}\phi. \tag{13}$$

Figure 2 includes the graphical illustration of major parts of a model of measurement with the variance parameter ϕ set equal to one.

If ϕ_{sc} corresponds to the original variance parameter ϕ, c is equal to one. Otherwise, if ϕ is given, c is obtainable by means of

a reordered version of Equation (11):

$$c = \sqrt{\frac{\phi}{\phi_{sc}}}.$$

Criterion-Based Methods

Methods including a criterion number, p_c, are referred to as criterion-based methods. The number selected as criterion is related to the sum of factor loadings or the sum of squared factor loadings. Criterion-based methods differ from each other in the number selected as the criterion and the way of summing the factor loadings. First, there is the effect-coding method (Little et al., 2006) that is equivalent to effect-coding used in analysis of variance where factor loadings are replaced by numbers that represent the coding of the effect. These numbers must be adjusted in such a way that their sum equals the number of manifest variables (p) and the adjusted numbers are used in the estimation of the variance parameter. In an example provided by Little et al. (2006), each one of the factor loading is set equal to one. It is highlighted that the estimate of the latent variance corresponds to the average of the indicator variables' variances (p. 63).

Equation (14) gives the formal representation of the basic characteristic of this method; that is, the selection of constraints such that the sum corresponds to p_c. In considering the scaling constant, c, the method is related to the outlined constancy framework:

$$p_c = \mathbf{1}'\boldsymbol{\lambda}^*_{coding_constraints} = \mathbf{1}'c\boldsymbol{\lambda}_{coding_constraints} \quad (14)$$

where $\mathbf{1}$ is a $p \times 1$ vector of ones, $\boldsymbol{\lambda}^*_{coding_constraints}$ the vector of adjusted numbers serving as factor loadings and $\boldsymbol{\lambda}_{coding_constraints}$ the vector of original numbers selected for coding the effect. The scaling constant c is necessary whenever the numbers selected for coding the effect do not directly sum to p_c.

A second criterion-based method relates the criterion number to the sum of squared factor loadings that was suggested for investigations focusing on variances and covariances (Schweizer, 2011). The number of manifest variables p is set equal to the product of the $p \times 1$ vectors of adjusted factor loadings $\boldsymbol{\lambda}^*$, respectively the vectors of original factor loadings $\boldsymbol{\lambda}$ with multiplier c:

$$p_c = \boldsymbol{\lambda} * '\boldsymbol{\lambda} * = c\boldsymbol{\lambda}'c\boldsymbol{\lambda}. \quad (15)$$

Using principles of scalar multiplication, the cs can be put in front of the product of vectors so that:

$$p_c = c^2\boldsymbol{\lambda}'\boldsymbol{\lambda}. \quad (16)$$

The product of vectors reveals that in this case the scaling aims at the variance explained by the factor. Given p_c and $\boldsymbol{\lambda}$, c is obtainable by means of a reordered version of Equation (16).

The graphical illustration for demonstrating the criterion-based methods includes products of the scaling constant c and λ (see **Figure 3**).

If λ originates from parameter estimation and not from effect coding, it may be necessary to estimate the value in the first step and fix it for scaling in the second step.

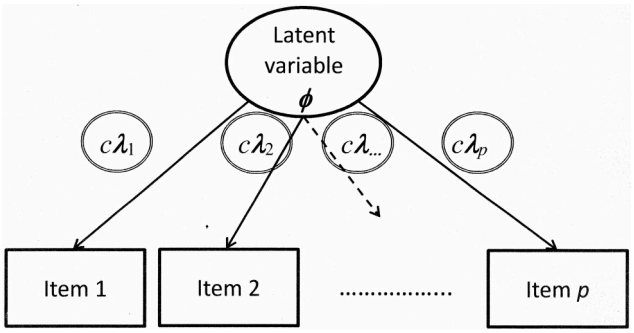


FIGURE 3 | Illustration of a confirmatory factor model with all factor loadings constrained according to the criterion-based methods.

THE EFFECT OF THE CRITERION NUMBER ON THE OUTCOME OF SCALING

While the marker-variable method and the reference-group method are rather restricted, the criterion-based methods include a criterion number that enables the adjustment to special expectations regarding the size of scaled variance parameters. This adjustment does not violate the constancy property. Although this criterion number is set equal to the number of manifest variables for good reasons in the version provided by Little et al. (2006), the number is changeable and may be changed to achieving variance values that vary within a smaller or larger range of possible values for the scaled variance parameter.

To demonstrate the effect of different choices of p_c, let p_{cA} and p_{cB} (where $p_{cA} > p_{cB}$) be two criterion numbers selected for the scaling of the variance parameter. Given the product $\boldsymbol{\lambda}\boldsymbol{\lambda}'$ and the initial inequity of p_{cA} and p_{cB}, Equation 16 suggests that

$$c_A^2\boldsymbol{\lambda}'\boldsymbol{\lambda} > c_B^2\boldsymbol{\lambda}'\boldsymbol{\lambda}.$$

Because both sides of the inequity include the product $\boldsymbol{\lambda}\boldsymbol{\lambda}'$, the inequity can be reduced to

$$c_A^2 > c_B^2.$$

The consequence of this inequity for the scaled variance parameters ϕ_{scA} and ϕ_{scB} when computed according Equation (11) is described by the next inequity:

$$\phi_{scA} = \frac{1}{c_A^2}\phi < \phi_{scB} = \frac{1}{c_B^2}\phi \quad (17)$$

The scaled variance parameter ϕ_{scA} is smaller than the scaled variance parameter ϕ_{scB} since ϕ_{scA} includes the larger scaling constant c as divisor. This inequity reveals that the larger p_c, the smaller the scaled variance parameter.

To demonstrate the practical consequences of selecting different values for p_c, the empirical consequences of changing p_c are reported in the following section for a number of different conditions. The computations are conducted according to Equations (11, 15). The outset is given by setting the original

TABLE 1 | Sizes of scaled variance parameters for criterion numbers set equal to the number of manifest variables (p) or proportions of it (r) in combination with different sizes of the factor loadings and numbers of manifest variables.

Proportion of r	Number of manifest variables p	Sizes of scaled variances		
		Factor loading of 0.2	Factor loading of 0.4	Factor loading of 0.6
1/1	4	0.04	0.16	0.36
1/1	8	0.04	0.16	0.36
1/1	12	0.04	0.16	0.36
1/2	4	0.08	0.32	0.72
1/2	8	0.08	0.32	0.72
1/2	12	0.08	0.32	0.72
1/4	4	0.16	0.64	1.44
1/4	8	0.16	0.64	1.44
1/4	12	0.16	0.64	1.44

TABLE 2 | Sizes of scaled variance parameters for criterion numbers (p_c) independent of the number of manifest variables combined with different sizes of the factor loadings and numbers of manifest variables.

Criterion number	Number of manifest variables	Sizes of scaled variances		
		Factor loading of 0.2	Factor loading of 0.4	Factor loading of 0.6
1	4	0.160	0.640	1.440
1	8	0.320	1.280	2.880
1	12	0.480	1.920	4.320
5	4	0.032	0.128	0.288
5	8	0.064	0.256	0.576
5	12	0.096	0.384	0.864
10	4	0.016	0.064	0.144
10	8	0.032	0.128	0.288
10	12	0.048	0.192	0.432

variance parameter equal to one and the factor loadings to 0.2, 0.4, or 0.6. Furthermore, the number of manifest variables is set to 4, 8, or 12. In the first step, it is investigated how p_c as proportion of $p_,$ that is defined to correspond to the number of manifest variables, influences the size of the scaled variance parameter. Three proportions are considered: 1, 1/2, and 1/4. The proportion of 1 requires the consideration of p_cs of 4, 8 and 12, the proportion of 1/2 p_cs of 2, 4, and 6, and the proportion of 1/4 p_cs of 1, 2, and 3.

The results are reported in **Table 1**. The first to third rows give the results for the original size of the criterion number, the fourth to sixth rows for the half of the original size and the seventh to ninth rows for the quarter of the original size. The inspection of the individual sections of **Table 1** reveals that the number of manifest variables has no influence on the size of the scaled variance parameter, whereas the increase of factor loadings leads to an increase of the scaled variance parameter. The results suggest that the larger the factor loadings, the larger the scaled variance parameter. In contrast, the comparison of the sections shows that the smaller the proportion of p_c, the larger the scaled variance parameter. This increase is predicted by the inequity of Equation (17). In the smallest proportion the factor loadings of 0.6 even lead to scaled variance parameters larger than one.

Furthermore, there is the opportunity to define the criterion number p_c independent of the number of manifest variables. In order to explore this possibility, the criterion number is set equal to 1, 5 and 10. Additionally the numbers of manifest variables (4, 8, 12) and the sizes of factor loadings (0.2, 0.4, 0.6) are varied.

The results are reported in **Table 2**. This table shows the same structure as **Table 1**. The comparisons of the three sections display an overall decrease of the scaled variance parameter from the first to the last one. This decrease is in line with the inequity of Equation (17). Furthermore, within the sections there is an increase of the scaled variance parameter from four manifest variables to 12 manifest variables. As also observed in **Table 1**, there is an increase of the scaled variance parameter associated with the increase of factor loadings.

Taken together, the results show that the increase of the factor loadings leads to an increase of the scaled variance parameter and that an increase of the criterion number leads to a decrease of the scaled variance parameter. Furthermore, the comparison of the results of **Tables 1**, **2** reveals that linking the criterion number to the number of manifest variables leads to constancy of the scaled variance parameter whereas otherwise, (i.e., when there is independency of the number of manifest variables) an increase of the number of manifest variables leads to an increase of the scaled variance parameter.

SCALING FOR ACHIEVING VARIANCES FOR COMPARISONS

The achievement of scaled variances for comparing the influences of latent variables on responding like standardized regressions weights in regressions analysis is presented as a major aim in the second section of the paper. For reaching this aim we resort to a basic method of factor analysis for estimating the variance explained by a factor. This method suggests the computation of the sum of squared factor loading $\boldsymbol{\lambda}'\boldsymbol{\lambda}$. It can alternatively be achieved by the trace of the corresponding matrix:

$$\boldsymbol{\lambda}'\boldsymbol{\lambda} = \text{trace}\left(\boldsymbol{\lambda}\boldsymbol{\lambda}'\right).$$

Although a variance parameter is not considered, it can be assumed being set equal to one ($\phi = 1$) and being omitted for convenience. In order to achieve similarity of the right-hand part of this Equation and the left-hand part of Equation (10) and also Equation (7), ϕ (= 1) is inserted in the right-hand part of this Equation:

$$\boldsymbol{\lambda}'\boldsymbol{\lambda} = \text{trace}\left(\boldsymbol{\lambda}\phi\boldsymbol{\lambda}'\right). \tag{18}$$

In the next step the matrix included in the parentheses is transformed by making use of the second criterion-based method (Equation 15). The criterion number p_c is set to 1:

$$1 = \boldsymbol{\lambda} * '\boldsymbol{\lambda} * = c\boldsymbol{\lambda}' c\boldsymbol{\lambda}. \tag{19}$$

The scaling framework of Equation (10) respectively Equation (7) enables the replacement of the vectors in the parentheses of Equation (18) and the assignment of the scaling constant as numerator to the variance parameter:

$$\boldsymbol{\lambda}'\boldsymbol{\lambda} = \text{trace}\left(\boldsymbol{\lambda} * \frac{\phi}{c^2}\boldsymbol{\lambda} * '\right)$$

Since the ratio of ϕ and c^2 is a scalar, it can be removed from the parentheses and is replaced by the scaled variance parameter ϕ^*:

$$\boldsymbol{\lambda}'\boldsymbol{\lambda} = \phi * \text{trace}\left(\boldsymbol{\lambda} * \boldsymbol{\lambda} * '\right).$$

Because of setting the criterion number p_c to 1, the trace must be 1 so that

$$\boldsymbol{\lambda}'\boldsymbol{\lambda} = \phi * \times 1 = \phi^*.$$

The contributions of all factor loadings are transferred to the scaled variance parameter. If this method is applied to the variances of two latent variables of the same model, as for example to the latent variables of Equation (3), there are two scaled variances that incorporate the contributions of all factor loadings on the corresponding latent variables. It enables the comparison of the influences of these latent variables on responding.

EXAMPLE: SCALING TRAIT AND METHOD LATENT VARIABLES WITH MTMM

We demonstrate consequences of employing different criterion numbers for scaling the variance of the latent variable through an investigation of a Multitrait-Multimethod (MTMM) design. For illustration, the MTMM matrix from the classic article by Campbell and Fiske (1959) was used; however, we recognize that the original matrix was a synthetic example, and thus, may not demonstrate optimal fit. Using the original MTMM matrix as correlation matrix input for CFA and specifying the model of measurement according to the correlated trait-correlated method model (Widaman, 1985) revealed two problems: (1) two negative error variances and (2) relationships among standardized error variances did not reflect expected relationships for the complements of reliability estimates provided along the main diagonal (0.89, 0.89, 0.76, 0.93, 0.94, 0.84, 0.94, 0.92, 0.85). In order to assure positive values of the error variances and to establish the expected relationship, the main diagonal of the matrix was changed from (1, 1, 1, 1, 1, 1, 1, 1, 1) to (1.145, 1.140, 1.145, 1.005, 0.965, 0.965, 0.940, 1.010, 0.980). Following the argument in justifying the use of the ridge option (Yuan et al., 2011), it was assumed the modification would affect error components of variances but not the systematic components themselves.

Furthermore, the insignificant correlations among the trait and method latent variables were eliminated from the full correlated trait-correlated method model. Only the correlations of the second and third method latent variables ($r = 0.52$) and the first and second trait latent variables ($r = 0.31$) remained. The revised correlated trait-correlated method model yielded good model fit, $\chi^2(16) = 17.63$, normed $\chi^2 = 1.10$, RMSEA = 0.014, SRMR = 0.065, CFI = 1.00, GFI = 0.99. This model estimated factor loadings, while the variance parameters of the model were set equal to one for identification.

Various methods for scaling are investigated[3]. At first, the results of criterion-based scaling are reported. Since Equation (14) was proposed for coding effects, Equation (15) guided the computation. Setting the criterion number to 3, that is, to the number of manifest variables for each construct and method led to the following variance parameter estimates: $\phi_{\text{method 1}} = 0.286$; $\phi_{\text{method 2}} = 0.528$; $\phi_{\text{method 3}} = 0.527$; $\phi_{\text{trait 1}} = 0.472$; $\phi_{\text{trait 2}} = 0.481$; $\phi_{\text{trait 3}} = 0.359$. No reported variance estimate was larger than one.

After setting the criterion number to 1, the following estimates of the variance parameter were observed: $\phi_{\text{method 1}} = 0.859$; $\phi_{\text{method 2}} = 1.585$; $\phi_{\text{method 3}} = 1.582$; $\phi_{\text{trait 1}} = 1.415$; $\phi_{\text{trait 2}} = 1.444$; $\phi_{\text{trait 3}} = 1.076$. All estimates of the variances of the trait latent variables were larger than one and two method latent variable variances were larger than one. While not reported, the t values for parameter significance testing were independent of the criterion number.

Next, the marker-variable method was used. One of the three factor loadings on each one of these latent variables was set equal to one whereas the remaining factor loadings and the variance parameter were free for estimation. Setting the first factor loading on each factor to one led to the following estimates of the variance parameter: $\phi_{\text{method 1}} = 0.291$; $\phi_{\text{method 2}} = 0.534$; $\phi_{\text{method 3}} = 0.500$; $\phi_{\text{trait 1}}$, $= 0.711$; $\phi_{\text{trait 2}} = 0.745$; $\phi_{\text{trait 3}} = 0.517$. Results for setting the second factor loading on each factor to one were: $\phi_{\text{method 1}} = 0.245$; $\phi_{\text{method 2}} = 0.522$; $\phi_{\text{method 3}} = 0.543$; $\phi_{\text{trait 1}} = 0.361$; $\phi_{\text{trait 2}} = 0.348$; $\phi_{\text{trait 3}} = 0.299$, respectively. Finally, the selection of the third factor loading on each latent variable for constraining values to one provided the following estimates: $\phi_{\text{method 1}} = 0.322$; $\phi_{\text{method 2}} = 0.529$; $\phi_{\text{method 3}} = 0.539$; $\phi_{\text{trait 1}} = 0.342$; $\phi_{\text{trait 2}} = 0.351$; $\phi_{\text{trait 3}} = 0.260$. In sum, different selections led to different estimates of the variance parameters. For example, selecting the first and second manifest variables as markers revealed the variance of the first method latent variable as the smallest one whereas in selecting the third manifest variables as marker the variance of the third trait latent variable was smallest.

The criterion-based method and the marker-variable method were considered for scaling the variance parameters obtained for Campbell and Fiske's MTMM. Different properties of these methods became apparent. The largest estimates were observed for the criterion-based method when the criterion number was one. Setting the criterion number to three led to overall smaller estimates. The marker-variable method led to different rank-orders of the variance estimates for different selections of marker variables. A unique set of variance estimates was not obtainable by means of this method. The reference-group method was not considered since this method only makes sense if dependences

[3]All t values for parameter significance were large, with $p < 0.05$ for all tests. We have eliminated t values to keep the focus on the variance estimates; however, t values are available upon request.

among the latent variables can be assumed as in longitudinal research. In contrast, trait and method latent variables are independent of each other.

DISCUSSION AND CONCLUSIONS

Although a variance parameter is a necessary component of factor analysis models, researchers often do not consider the effect that the scaling of this parameter has on the variance of the latent variable. One major issue addressed in this paper is the preservation of information when changing from one reference system to another through scaling. Scaling the variance of a latent variable must preserve the result regarding the structure of the data while simultaneously improving interpretability and comparability of the result. The consistency framework presented in this paper reveals how the preservation occurs, and we provide insight into the crucial role of scalar multiplication. Scalar multiplication enables the change of parts of the model of the covariance matrix that is basic to the confirmatory investigation while exhibiting constancy of the product of these parts.

The investigation of the available methods for scaling the variance parameter reveals that the available methods fit to the constancy framework; however, methods present different degrees of flexibility. Whereas, the reference group method is totally fixed, the marker-variable method allows some adaptability to the data in that the method enables the selection of the indicator variable from the set of all manifest variables. We understand that different methods of setting a marker variable for identification may lead to different standard error terms for parameters, and subsequently, different significance test (i.e., Z) values (Gonzalez and Griffin, 2001), the method of scaling latent variables is consistent in terms of fit and parameter estimates.

Criterion-based methods, however, are potentially adaptable to specific needs as the criterion number may be changed to meet a specific situation. The use of a criterion number provides the opportunity to design the method for scaling the variance parameter in such a way that it is possible to: (1) choose between dependency and independency on the number of manifest variables, and (2) opt for lower or larger values of the variance parameter, i.e. different ranges of the possible sizes of the scaled variance parameter, starting with zero.

The application of the scaling methods concentrated on the MTMM provided by Campbell and Fiske (1959). All scaling methods were considered; however, not all of them were able to fit the MTMM matrix. The reference-group method does not apply if there is only one sample; however, it provides a starting point for scaling according to other methods as estimates of the factor loadings are obtained by setting the variances of the latent variables equal to one. The application of the marker-variable methods requires the selection of marker variables; results revealed that different marker-variables lead to different values as the result of scaling. This is not a good property if unique statistics (e.g., means, standard deviations) are expected. Uniqueness of scaled variance estimates are achieved by the criterion-based method.

The criterion-based method also provides an opportunity to achieve scaled variances similar to eigenvalues. Using positive integers as criterion numbers, the largest scaled variance parameters are obtainable for one as criterion number. According to the results of an empirical study, use of the value of one as a criterion number leads to estimates of the variance parameter that correspond to eigenvalues if the model for investigating the data is unidimensional and specific procedural properties are considered (Schweizer et al., 2017). This property enlarges the range of possible applications of scaled variances. Whereas, variance parameters scaled in another way can only be compared with each other, the scaling in using one as criterion number additionally enables comparisons of scaled variances with eigenvalues.

AUTHOR CONTRIBUTIONS

KS conceptualized the study and contributed to the writing. ST and CD contribute substantially to the writing of the whole article.

REFERENCES

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not? *Intelligence* 41, 244–262. doi: 10.1016/j.intell.2013.04.003

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Pers. Individ. Diff.* 85, 98–104. doi: 10.1016/j.paid.2015.04.046

Gonzalez, R., and Griffin, D. (2001). Testing parameters in structural equation modeling: every "one" matters. *Psychol. Methods* 6, 258–269. doi: 10.1037/1082-989X.6.3.258

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educ. Psychol. Meas.* 66, 930–944. doi: 10.1177/0013164406288165

Jensen, A. R. (1982). "The chronometry of intelligence" in *Advances in Research on Intelligence, Vol. 1*, ed R. J. Sternberg (Hillsdale, NJ: Erlbaum), 93–132.

Jöreskog, K. G. (1970). A general method for analysis of covariance structure. *Biometrika* 57, 239–257. doi: 10.1093/biomet/57.2.239

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York, NY: Guilford Press.

Little, T. D., Slegers, D. W., and Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Struct. Equat. Model.* 13, 59–72. doi: 10.1207/s15328007sem1301_3

Lu, Y., and Sireci, S. G. (2007). Validity issues in test speededness. *Educ. Meas.* 26, 29–37. doi: 10.1111/j.1745-3992.2007.00106.x

Marsh, H. W., and Grayson, D. (1995). "Latent variable models of multitrait–multimethod data" in *Structural Equation Modeling: Concepts, Issues and Application*, ed R. H. Hoyle (Thousand Oaks, CA: Sage), 177–198.

McArdle, J. J. (2009). Latent variable modeling of differences and change with longitudinal data. *Annu. Rev. Psychol.* 60, 577–605. doi: 10.1146/annurev.psych.60.110707.163612

McArdle, J. J., and Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivar. Behav. Res.* 29, 63–113. doi: 10.1207/s15327906mbr2901_3

Miller, M. B. (1995). Coefficient α: a basic introduction from the perspectives of classical test theory and structural equation modeling. *Struct. Equat. Model.* 2, 255–273. doi: 10.1080/10705519509540013

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Osborne, J. W. (2017). *Regression and Linear Modeling*. Thousand Oaks, CA: Sage Publishing.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *J. Educ. Meas.* 31, 200–219. doi: 10.1111/j.1745-3984.1994.tb00443.x

Raykov, T. (1997). Estimation of composite reliability for congeneric models. *Appl. Psychol. Meas.* 21, 173–184. doi: 10.1177/014662169702 12006

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivar. Behav. Res.* 47, 667–696. doi: 10.1080/00273171.2012.715555

Schmitt, N., and Kuljanin, G. (2008). Measurement invariance: review of practice and implications. *Hum. Resour. Manage.* 18, 210–222. doi: 10.1016/j.hrmr.2008.03.003

Schweizer, K. (2011). Scaling variances of latent variables by standardizing loadings: applications to working memory and the position effect. *Multivar. Behav. Res.* 46, 938–955. doi: 10.1080/00273171.2011. 625312

Schweizer, K., Troche, S., and Reiß, S. (2017). Can variances of latent variables be scaled in such a way that they correspond to eigenvalues? *Int. J. Stat. Probab.* 6, 35–42. doi: 10.5539/ijsp.v6n6p35

Steiger, J. H. (2002). When constraints interact: a caution about reference variables, identication constraints, and scale dependencies in structural equation modeling. *Psychol. Methods* 7, 210–227. doi: 10.1037/1082-989X.7.2.210

Stoel, R. D., Garre, F. G., Dolan, C., and van de Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychol. Methods* 11, 439–455. doi: 10.1037/1082-989X.11.4.439

Verbeke, G., and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* 59, 254–262. doi: 10.1111/15 41-0420.00032

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl. Psychol. Meas.* 9, 1–26.

Yuan, K.-H., Wu, R., and Bentler, P. M. (2011). Ridge structural equation modeling with correlation matrices. *Br. J. Math. Stat. Psychol.* 64, 107–133. doi: 10.1348/000711010X497442

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Schweizer, Troche and DiStefano. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

METHODS
published: 17 April 2019
doi: 10.3389/fpsyg.2019.00829

Compromised Item Detection for Computerized Adaptive Testing

Cheng Liu[1,2,3], *Kyung T. Han*[4] *and Jun Li*[1]*

[1] Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN, United States, [2] Center for Research Computing, University of Notre Dame, Notre Dame, IN, United States, [3] Department of Psychology, University of Notre Dame, Notre Dame, IN, United States, [4] Graduate Management Admission Council, Reston, VA, United States

Item leakage has been a serious issue in continuous, computer-based testing, especially computerized adaptive testing (CAT), as compromised items jeopardize the fairness and validity of the test. Strategies to detect and address the problem of compromised items have been proposed and investigated, but many solutions are computationally intensive and thus difficult to apply in real-time monitoring. Recently, researchers have proposed several sequential methods aimed at fast detection of compromised items, but applications of these methods have not considered various scenarios of item leakage. In this paper, we introduce a model with a leakage parameter to better characterize the item leaking process and develop a more generalized detection method on its basis. The new model achieves a high level of detection accuracy while maintaining the type-I error at the nominal level, for both fast and slow leakage scenarios. The proposed model also estimates the time point at which an item becomes compromised, thus providing additional useful information for testing practitioners.

Keywords: computerized adaptive testing, CAT, compromised item detection, generalized linear model, test security

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Okan Bulut,
University of Alberta, Canada
Yoav Cohen,
National Institute for Testing and Evaluation (NITE), Israel

***Correspondence:**
Jun Li
jun.li@nd.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *12 September 2018*
Accepted: *28 March 2019*
Published: *17 April 2019*

Citation:
Liu C, Han KT and Li J (2019) Compromised Item Detection for Computerized Adaptive Testing. Front. Psychol. 10:829. doi: 10.3389/fpsyg.2019.00829

1. INTRODUCTION

Due to advances in information technology, continuous testing has been offered for many large-scale testing programs, and test takers can take such exams nearly any time during the year. Although continuous testing provides test takers with considerable flexibility and convenience, it also raises serious security concerns. Individuals who take the test earlier in a testing window could share the items orally or online (e.g., via social media platforms), which would benefit subsequent test takers, jeopardizing the validity and fairness of the test. Studies have shown the severe and negative impact of compromised items (Chang and Zhang, 2002, 2003; Davey and Nering, 2002; McLeod et al., 2003; Yi et al., 2008; Guo et al., 2009; Zhang et al., 2012). Items administered frequently are vulnerable to leakage, and many methods have been proposed to control item exposure to protect test security (Sympson and Hetter, 1985; Stocking, 1994; Stocking and Lewis, 1995; Mills and Stocking, 1996; Hetter and Sympson, 1997; Way, 1998; Chang and Zhang, 2002, 2003; Davey and Nering, 2002; Chen et al., 2003). Sympson-Hetter (SH) method (Sympson and Hetter, 1985) is one of the widely used applications of this strategy. SH method needs an upper cutoff proportion (e.g., 20%) as a parameter. Only those items that are exposed to <20% of all test takers can be selected. This way, SH method is able to prevent the over exposure of an item to the public, which in return reduces the potential damage caused by the item compromise. Since then, researchers have developed many exposure-control strategies following the same direction.

Although these methods are generally effective in keeping items from being over exposed, they are only preventive measures and do not directly address the problem of items that have been compromised.

Alternatively, many methods have been developed to proactively detect item preknowledge (McLeod et al., 2003; Belov et al., 2007; Belov and Armstrong, 2010, 2011; Obregon, 2013; Belov, 2014). These methods allow testing practitioners to determine whether an examinee has preknowledge of a set of suspicious items by comparing the estimates of the examinee's ability with and without suspicious items. As examples, Drasgow et al. and Armstrong et al. proposed detection methods using likelihood-based person fit statistics (Drasgow et al., 1985; Armstrong et al., 2007), Levine and Dragsgow proposed another method based on Neyman-Pearson lemma (Levine and Drasgow, 1988), and Belov et al. proposed to use Kullback-Leibler divergence for detection (Belov and Armstrong, 2011). However, in practice, there are two major limitations with the application of these methods. First, it is difficult to identify a set of suspicious items without context or prior information, especially when considering that the item set varies from examinee to examinee. Second, these methods rely heavily on the estimation of the test-taker's ability. When the true ability of the test taker is known, the method usually performs well. In practice, however, the test-taker's true ability is unknown and needs to be estimated. In the case of severe item leakage, the estimation of an individual's ability can become biased, which in turn can lead to inefficiency in detecting item preknowledge.

The above-mentioned proactive methods focus on individual-level test statistics, but in recent years, several item-level sequential methods have been proposed to detect compromised items in computerized adaptive testing (CAT) (Zhang, 2014; Zhang and Li, 2016; Choe et al., 2018). These methods focus on monitoring the change of the expected probability of getting a correct response for an item. To enhance the sensitivity of the detection procedure, they suggest imposing a moving window to select a group of responses from the n^{th} response to the $(n + m)^{th}$, where n is the starting point of the window and m is the size of window. Then a hypothesis test is performed to tell whether the expected probabilities of getting an item correct are the same before and within this window. The item will be flagged as compromised if the change is significant. One advantage of this sequential algorithm is that it is computationally fast and hence can be used for real-time detection. Existing studies, however, are limited in several ways and warrant further study. First, these methods require specification of the best window size, which may be challenging for test professionals. Second, the simulation considers only the scenario where the expected probability of a correct response has a sharp increase after an item is compromised. The utility of these methods in the face of a gradual change of the expected probability is unknown. Third, the current sequential detection method can only tell when the leakage is detected but cannot estimate when the item is compromised, from which test practitioners can review the impact of the leakage and re-evaluate test-takers' ability estimation. For example, an item compromised at day t_1 can be detected as an compromised item at day t_2. There is a $t_2 - t_1$ lag in between before a significant conclusion could be drawn. In this case, t_2 is the detection day, which is known. And t_1 is the compromised day which is not known.

Therefore, there is need for a new, flexible method to account for various item-leaking processes in real life, where compromised items can spread at different rates and item leakage can result from many causes. The new method should be able to detect leakage under different scenarios, and provide an estimate of when an item is leaked.

First, compromised items may spread at different speeds, and the expected probability of correctly response to an item may not jump abruptly to a fixed, high value. For example, a posting on a popular social media website could quickly spread preknowledge of an item, whereas sharing within a small group of acquaintances might result in slower spreading. Therefore, to make the sequential detection approach more robust, it is important to develop a flexible method that takes these underlying dynamics into consideration.

Second, there are many probable causes of item leakage. A common scenario as detailed above could involve a test taker who posts the items received on a website, where future test takers could gain preknowledge on those items. A more severe case is organized item theft, which has been discussed in Yi et al. (2008). In this case, profit-driven organizations may send thieves to take the exam at the early stage in a testing window. The thieves will intentionally memorize the items they receive, aiming to profit from disclosing the items to future test takers. In the "random item leakage" scenario, the time when an item becomes compromised is random. In the "organized item theft" scenario, on the other hand, the leakage usually happens at the very beginning of a testing window. When investigating the performance of detection methods, these different scenarios should be considered.

We therefore propose a new method for proactive detection of compromised items that largely addresses the stated limitations of existing approaches. Our method uses generalized linear modeling with complementary log-log transformation (cloglog) as the link function, and it takes the potential leaking mechanism into consideration. Compared with existing methods, it has the following advantages: (1) It can handle more complicated item leakage mechanisms, both fast and slow; (2) Unlike existing sequential approaches, it does not need a moving window to boost the detection sensitivity, and thus saves the trouble of determining the best window size; instead, it improves the detection accuracy by utilizing complete testing information. (3) It enables the estimation of the "compromise time," i.e., the time point at which the item was compromised. (4) It is computationally more efficient compared with those item preknowledge detection methods since it does not depend on the selection of suspicious items.

The model is validated by both simulation data and real data in practice. For simulation, the test is performed with simulations under different scenarios and parameters. The simulated datasets are generated as diverse as possible. First, the model we use for simulating data is purposefully designed to differ from our model for leakage detection, in order to test the robustness of our leakage detection method when

the underlying leaking mechanism is unknown. Second, our simulation covers two distinct leakage scenarios, organized item theft and random item leakage. Third, for each simulation scenario, we investigated the values of the leakage rate in a wide range, in order to mimic different spread speeds in practice. In addition of all simulation studies above, we also showed an application of our proposed method to a real large-scale testing dataset. Both studies, i.e., simulation and real data, perform well. In our study, an application based on the estimation of the compromised day, t_1, is also proposed, which successfully links the compromised item detection with the person-level preknowledge detection. Simulation results show that t_1 can provide important information for the preknowledge detection in CAT and significantly improve the accuracy of the person's ability estimation.

2. METHODS

We detect compromised items by monitoring the responses of test takers. When an item is compromised, the expected probability for test takers to answer it correctly will increase. Instead of assuming all responses to always be correct (Yi et al., 2008) or to be a constant probability (Zhang, 2014) immediately after an item becomes compromised in simulation, we propose a gradual change model as a function of time, hereafter referred to as the *leakage model*. The leakage model acknowledges the fact that responses to a compromised item may not always be correct right after its compromise. Instead, as more people are exposed to this compromised item over time, the probability for the item to be correctly answered will gradually increase to 1. This increase can be slow or fast, depending on the rate parameter. When this rate is large, our model will degrade to the previous models mentioned above.

2.1. Generalized Linear Model for Detection

In computerized adaptive testing, the probability for a test taker to give a correct response to an uncompromised item can be modeled by a three-parameter logistic (3PL) item response theory (IRT) model (Lord, 1980):

$$P(U=1|\theta) = c + (1-c)\frac{1}{1+e^{-1.7a(\theta-b)}}, \tag{1}$$

where θ is the latent ability of a test taker, a is the discrimination parameter, b is the difficulty parameter, and c is the pseudo-guessing parameter. It has been shown (Birnbaum, 1968) that the item will be assigned to test takers whose provisional ability estimate is close to

$$\theta_0 = b + \frac{\ln(1+\frac{\sqrt{1+8c}}{2})}{1.7a}, \tag{2}$$

when the maximum item information method is used to select the next item. The expected probability for test takers to answer the item correctly is $(1+\sqrt{1+8c})/4$. Thus, the probability to answer the target item correctly should fluctuate roughly around this expected probability. When an item is compromised, the expected probability will increase accordingly. In practice, since the ability estimate may not be sufficiently accurate at the beginning of the test, the expected probability to correctly answer the item might not be exactly $(1+\sqrt{1+8c})/4$ initially. On the other hand, to control the potential damage from item thieves during high stakes exams, an item exposure control component will be implemented, which is a random factor on top of the item selection criterion. Therefore, it is rare that the very item expected to exhibit the largest Fisher information would actually be selected and administered. One of the items with higher information will, though. As the test progresses, however, and if the item pool is sufficiently large, the expected probability should hold, a property that could be used to detect the compromised item. A similar idea was also discussed in Zhang (2014).

In this study, the proposed detection algorithm concerns only the time series of responses of a single item, and all items are treated independently. Unless stated otherwise, we will use a representative item to hereafter illustrate the detection model.

Suppose the expected probability for a test taker to answer this item correctly is $1-\pi_t$ on day t (i.e., the probability to get an incorrect answer is π_t). Therefore, the number of incorrect answers y_t should approximately follow a binomial distribution, $y_t \sim \text{Bin}(n_t, \pi_t)$, where n_t is the total number of examinees taking this item on day t. Thus, the overall log-likelihood for all T testing days for the item of interest is

$$l = \log L = \sum_{t=1}^{T} \left[y_t \log \pi_t + (n_t - y_t)\log(1-\pi_t) \right], \tag{3}$$

where $t = 1, 2, \cdots, T$. Please note that although we are using days as the unit of t for illustration, t actually can be any time units. For example, t can be hours instead as long as there are enough samples within the time interval.

In order to design an effective model to detect the leakage pattern in real data, we worked with the researchers in the large-scale testing company in this study. **Figure 1** shows four typical curves from the empirical data analysis. These curves are selected from a large-scale operational CAT program that has 2905 items records. The item pool was rotated every 10 days in order to secure the test from item compromise (For more information of this dataset, please see section REAL DATA APPLICATION). The error bars are the 95% confidence interval of the probability of incorrect response of that day, which is calculated by $1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$ (Agresti, 2013). **Figure 1A** is an item without leakage. **Figures 1B,C** represent the leakage with two different leakage rates: slow and fast. **Figure 1D** shows a scenario where the expected probability goes back up after a significant decrease. For both **Figures 1B,C**, a sigmoid-shaped pattern curve could be used to model the probability change. For scenario d, although the probability goes back after a significant dip, this scenarios should also be flagged out as well, since: (1) in a continuous test, we can only make our decision based on the data we have at hand. (2) a significant decrease of the expected probability should always be alarmed and carefully investigated,

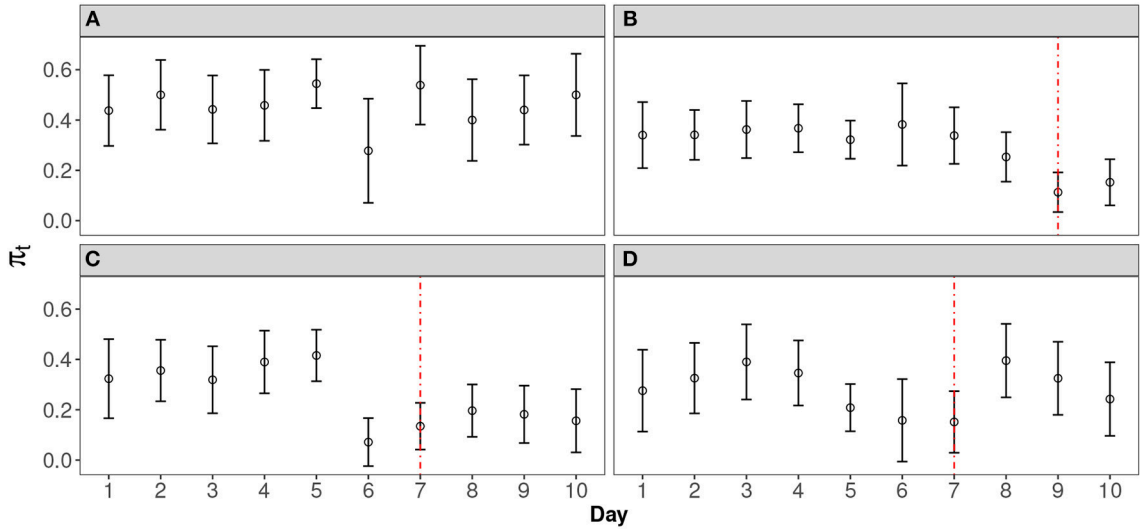


FIGURE 1 | Representative Curves for Different Scenarios. **(A)** Item without any leakage; **(B)** Item with slow leakage; **(C)** Item with fast leakage; **(D)** Item with leakage that goes back thereafter.

to enhance the security of the test. In this case, a sigmoid-shaped pattern curve can also be used to model the part before it goes back.

To detect the gradual change of the expected probability, two possible methods could be used to model the probability π_t as a function of time: logit

$$\text{logit}\frac{\pi_t}{\pi_0} = \beta(t - t_0), \tag{4}$$

or cloglog

$$\text{cloglog}\frac{\pi_t}{\pi_0} = \beta(t - t_0), \tag{5}$$

where π_0 is the expected probability before leakage and β is a coefficient that controls the speed of the leakage. Here t_0 is the point at which the item is compromised. **Figure 2** illustrates the shape of π_t under different combinations of π_0 and β for both logit and cloglog functions. In general, π_t decreases in a sigmoid manner when β is negative, and a larger absolute value of β corresponds to a faster decrease, suggesting a faster leakage of the compromised item. In the beginning, π_t presumably changes relatively faster than later in the test cycle. This is when some test takers who are eager to obtain preknowledge of the compromised item would like to take the test, since the compromised item likely is still available. In such case, it will induce a faster drop of probability of incorrect response when leakage starts. For this reason, the asymmetry of the cloglog function is favored in this study and will be selected to model π_t. When $t = t_0$, $\pi_t = \pi_0(1 - e^{-1})$, which is around 0.63 of the expected probability before leakage. Note that Equation 5 is actually equivalent to the following model

$$\text{cloglog}\frac{\pi_t}{\pi_0} = \beta t + \alpha. \tag{6}$$

That is, $\log(-\log(1 - \frac{\pi_t}{\pi_0})) = \beta t + \alpha$, which gives

$$\pi_t = \pi_0(1 - e^{-e^{\beta t + \alpha}}). \tag{7}$$

For a compromised item, a negative β is expected. Therefore, the problem of detecting a compromised item is converted to performing the following hypothesis test:

$$H_0 : \beta = 0 \quad vs. \quad H_a : \beta < 0. \tag{8}$$

Note that our test is one-sided, since a positive beta corresponds to an increasing π_t, which is not a desired pattern we want to flag out.

In order to perform the hypothesis test, we need $\hat{\beta}$, as well as the estimate of its variance or standard error. $\hat{\beta}$ and $\hat{\sigma}_\beta$ are obtained via maximum likelihood estimation. Since there is no closed form analytical solution, we use the coordinate-wise Newton-Raphson method to obtain a numerical solution. Compared with the conventional Newton-Raphson that updates all model parameters at the same time, the proposed method successfully avoids the calculation of the inverse of the Hessian matrix, which can be near-singular and cause numerical instability. Our approach has proved to be efficient and stable in all our simulation studies.

Let $\mathbf{\Psi}$ be the coefficients in our model that need to be estimated. We have $\mathbf{\Psi} = (\psi_1, \psi_2, \psi_3)^T$, where $\psi_1 = \pi_0$, $\psi_2 = \beta$ and $\psi_3 = \alpha$. In this way, we can use one general symbol $\mathbf{\Psi}$ to represent all three parameters. The steps of the coordinate-wise Newton-Raphson algorithm are as follows:

1. Initialize model parameters with random starting values. We use $\pi_0^{(0)} = 0.5$, $\beta^{(0)} = 0$ and $\alpha^{(0)} = 0$ for all our simulation studies.
2. Update $\mathbf{\Psi}$ by updating each of its elements consecutively. That is,

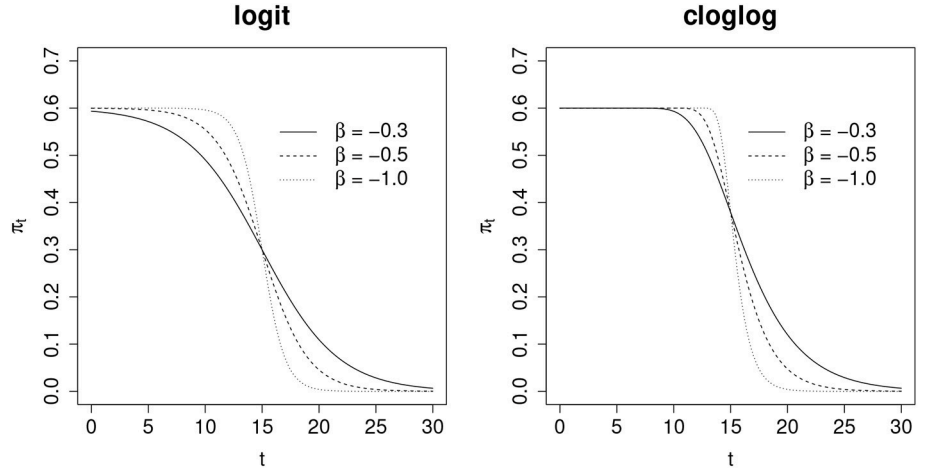


FIGURE 2 | Comparison of link functions with logit and cloglog transformations.

(a) First, keep $\hat{\beta}$ and $\hat{\alpha}$ unchanged, update $\hat{\pi}_0$;
(b) Then, keep $\hat{\pi}_0$ and $\hat{\alpha}$ unchanged, update $\hat{\beta}$; and
(c) Third, keep $\hat{\pi}_0$ and $\hat{\beta}$ unchanged, update $\hat{\alpha}$.

Each of the above updates is given by

$$\hat{\psi}_k \leftarrow \hat{\psi}_k - \frac{\frac{\partial l(\Psi)}{\partial \psi_k}\Big|_{\Psi=\hat{\Psi}}}{\frac{\partial^2 l(\Psi)}{\partial \psi_k^2}\Big|_{\Psi=\hat{\Psi}}}. \tag{9}$$

(See the **Appendix** for more details about this equation.)

3. Repeat Step 2 until convergence. The convergence is checked by calculating the change of the log-likelihood after each iteration. If the change is less than a threshold, e.g., 0.001, the model has converged. Then the element of the Fisher information matrix (see **Appendix** for detail) is

$$[I_{k_1 k_2}(\pi_0, \beta, \alpha)] = -\mathbb{E}\Big[\frac{\partial^2 l(\pi_0, \beta, \alpha)}{\partial \psi_{k_1} \partial \psi_{k_2}}\Big], \tag{10}$$

where $k_1, k_2 = 1, 2, 3$. According to the co-factor method of getting the inverse matrix of $\boldsymbol{I}$, we have

$$\hat{\sigma}_\beta = \sqrt{\frac{I_{11} \cdot I_{33} - I_{13} \cdot I_{31}}{det(\boldsymbol{I})}}. \tag{11}$$

Given $\hat{\beta}$ and $\hat{\sigma}_\beta$, the Wald statistic is given by

$$z = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_\beta} \sim N(0, 1). \tag{12}$$

When the null hypothesis is rejected (one-sided test), the item will be flagged as compromised. The time from when the item starts to leak, i.e., the "compromised time," to when a leaked item is flagged, is defined as "detection lag." This definition of detection lag is the same as that in Zhang (2014) and Shao et al. (2015). Note that the compromised time, denoted as t_c, is unknown in real applications. We propose an estimate of it, $\hat{t}_c$, as the time when π_t drops to a certain percentage, say ϵ, of π_0. Based on our model in Equation 7, it is easy to show that

$$\hat{t}_c = \frac{\ln(\ln\frac{1}{1-\epsilon}) - \hat{\alpha}}{\hat{\beta}}. \tag{13}$$

Especially, we use $\epsilon = 90\%$. The bias of this estimate is defined as the "estimation lag".

Further, the variance of $\hat{t}_c$ is given by

$$\text{var}(\hat{t}_c) = \Big(\frac{\partial \hat{t}_c}{\partial \hat{\alpha}}, \frac{\partial \hat{t}_c}{\partial \hat{\beta}}\Big) \cdot \hat{\Sigma} \cdot \begin{pmatrix} \frac{\partial \hat{t}_c}{\partial \hat{\alpha}} \\ \frac{\partial \hat{t}_c}{\partial \hat{\beta}} \end{pmatrix}, \tag{14}$$

where $\hat{\Sigma}$ is the variance-covariance matrix of $(\hat{\alpha}, \hat{\beta})$, and

$$\frac{\partial \hat{t}_c}{\partial \hat{\alpha}} = -\frac{1}{\hat{\beta}},$$
$$\frac{\partial \hat{t}_c}{\partial \hat{\beta}} = \frac{\alpha - \ln(\ln\frac{1}{1-\epsilon})}{\hat{\beta}^2}.$$

The elements of $\hat{\Sigma}$ can be easily estimated by the inverse matrix of $\boldsymbol{I}$, similar to how the variance of β is derived in Equation 11.

2.2. Leakage Simulation Model

Our primary goal for introducing a different leakage model is to test the effectiveness of the proposed detection method with unknown underlying leakage rates. The leakage simulation model should have these two features: (1) After the item is compromised, the expected probability to get a correct response will increase; (2) The spread rates of the compromised item may differ across items. In this study, the leaking process is simulated using an exponential function as follows,

$$P(\text{a test taker already knows the answer}|\lambda, t_0) = 1 - e^{-\lambda(t-t_0)}, \tag{15}$$

where λ is the leakage parameter that regulates how fast the item will be exposed to the public, t_0 is the time point at which the item is compromised, and $t - t_0$ is the time interval since the item was first compromised. The probability for any test taker to have item preknowledge is a function of t, or $P(t)$. Therefore, after integration with the 3PL IRT model, the overall probability for a test taker to correctly answer an item can be captured in a mixture model as follows:

$$P(U = 1|\theta, \lambda, t_0) = (1 - e^{-\lambda(t-t_0)}) + e^{-\lambda(t-t_0)} \cdot \left[c + (1 - c)\frac{1}{1 + e^{-1.7a(\theta-b)}}\right]. \quad (16)$$

If the test taker already knows the answer to the item due to item preknowledge, the response process is described by the first component of Equation 16, which is $1 - e^{-\lambda(t-t_0)}$. Otherwise, the process follows the 3PL IRT model with a probability of $e^{-\lambda(t-t_0)}$. Therefore, the total expected probability for a test taker to correctly answer the item is given by Equation 16. Again, the first component of Equation 16 is a function of time and therefore captures the leakage process, where λ controls the speed of the leakage. For example, given a moderate leakage parameter λ and a compromise time point t_0, responses to the compromised item will contain increasingly more 1 s (i.e., correct responses), as time t increases. With a large λ, the responses will almost always be 1 after item compromise, as assumed by previous studies (Yi et al., 2008). Thus, the gradual change model is more flexible and general.

Note that, in this study, simulation model is only used to test the detection model, not to detect the leakage. Compared with the detection model, simulation model is more complex with extra parameters including a person's ability θ. Although we can also use the leakage model to fit the curve and run the hypothesis test thereafter, a simultaneous estimation of person's ability will make the fitting less efficient than the detection model. Since we only care about the detection of probability curve's leakage pattern, the proposed detection model is more straightforward and easier to converge.

3. SIMULATION DESIGN

Simulation studies are conducted to investigate the performance of the proposed detection method. The parameters in our simulation were chosen according to previous publications (du Toit, 2003; Yi et al., 2008; Zhang, 2014). A total of 400 randomly generated items serve as the item pool. The underlying IRT model is 3PL with item parameters generated as follows:

$$a \sim \text{lognormal}(0, 0.5), \quad b \sim \text{N}(0, 1), \quad c \sim \text{U}(0, 0.25). \quad (17)$$

The discrimination parameters are generated by lognormal distribution. An exposure control procedure is implemented to prevent items from being over-exposed and to protect test security. The exposure rate for an item is defined as

$$\rho = \frac{\#\text{ of times an item has been administered}}{\text{total number of test takers}}. \quad (18)$$

In this study, the exposure control parameter is set to be 0.2, meaning only items with exposure rate lower than 0.2 are eligible for administration. Items in the bank belong to three content areas with percentages 40, 30, and 30%, respectively. Test length is set at 40. A content control procedure is implemented in the simulation to ensure that 40, 30, and 30% of items are selected from each content area for every test taker (i.e., 16, 12, and 12, respectively). The item with the lowest exposure rate in the desired content area will be selected as the first item for the incoming test taker. A sample of 500 test takers (θs) are generated each day to take the exam, whose abilities follow standard normal distribution. The simulation is replicated 10 times and all the distribution figures presented in the remainder of this paper are generated based on results aggregated over replications.

A test item could become compromised for a variety of reasons. The interest of this study is to investigate the effectiveness of the detection algorithm in general. In order to achieve this goal, we studied two common scenarios, which form the core of this paper:

1. Organized item theft. Organized item theft is one of the most severe item leakage scenarios in computer-based testing (Yi et al., 2008). Since organized theft usually occurs early in one testing window for maximal gain, 20 item thieves are randomly generated in the first 4 days of the exam cycle. A simple assumption that each thief can randomly remember 10 items is used here, although professional item thieves could remember more. The items will be treated as compromised when they are remembered. Leakage simulation model will be applied thereafter.
2. Random item leakage. Some test takers simply share the items that they have memorized with the public. In this instance, the leakage could occur any time. For the purpose of this study, 20 such item sharers are randomly selected during one testing window. A testing window is so defined that no item pool maintenance such as rotation or replenishment occurs within that window. In other words, the item bank remains the same throughout the window. In this study, the testing window is set to be 30 days (one month). In practice, this number highly depends on the operation of testing company. It might not be a fixed value even for the same test. For simulation, we use monthly rotation to demonstrate the methodology. On average, we assume each item sharer could remember at random 10 out of the 40 items and share these with the public. Usually the motivation to share items is weak near the end of a testing window. For this reason, this simulation study assumes that such random sharing behavior happens only in the first 25 days.

For each test taker, the first item is selected from the item bank that has the lowest exposure rate at that time from the desired content areas. The probability of the test taker to give a correct answer to the target item is calculated based on the mixture leakage model (Equation 16). Then a uniform distributed random number will be generated within (0, 1). If its value is less than the mixture probability, the response will be 1 (i.e., a correct answer). Otherwise, the response will be

0. The expected a posterior (EAP) method (Bock and Mislevy, 1982) is used to estimate an individual's ability given this person's previous responses. After the first item, the standard CAT procedure using maximum item information method (Lord, 1980) is adopted to select the next item according to the estimated $\hat{\theta}$.

In some extreme cases, the probability of getting an item correctly after it is compromised is 1 for all test takers. But, in practice, item leakage could be a gradual change process. In this study, leakage parameter λ (see Equation 15) are set to be 0.05, 0.1, 0.3, 0.5, 0.7, 1, and 1.5 to regulate the differential speed of item leakage. When λ is large, e.g., $\lambda = 1.5$, the simulation represents a severe leakage scenario, in which nearly all responses will be correct once an item has been compromised.

4. RESULTS

As illustrated in the Method section, the proposed leakage detection model intentionally uses Equation 7, which differs from the true underlying model (Equation 16) that is used to generate the item responses. Parameter λ controls the speed of leakage. The days to reach the probability's half-drop can be approximately estimated by $-\frac{1}{\lambda}\log(0.5)$, which are around 14, 1.4, and 0.5 when λ is 0.05, 0.5, 1.5, respectively.

4.1. Organized Item Theft

As mentioned earlier, this study assumes that all item thieves have taken the test in the first 4 days within a testing window. **Table 1** shows the results of detection accuracy and corresponding type-I error in this case. Detection accuracy is defined as the proportion of compromised items correctly identified as such. Type-I error is the proportion of uncompromised items that are incorrectly identified as compromised items.

$$\text{accuracy} = \frac{\#\text{ of compromised items that are correctly flagged}}{\#\text{ of total compromised items}} \tag{19}$$

$$\text{type-I error} = \frac{\#\text{ of items that are incorrectly flagged as compromised items}}{\#\text{ of total uncompromised items}} \tag{20}$$

For a desired 95% confidence interval, the detection accuracy is about 99% for those λs larger than 0.05. When $\lambda = 0.05$, the detection accuracy drops to 93.70%. This is because $\lambda = 0.05$ represents a very slow leakage process, which is hard to detect within the 30-day window. On the other hand, the type-I errors for all λs are well controlled at ∼5%, consistent with the desired 95% confidence interval. **Figure 3** represents the distribution of the detected date of item compromise for different λs within the 30-day window. Overall, when λ is small, the distribution shows large variability. When λ is large, the detection is rather accurate, i.e., pinpointing compromise within the first 4 days. In addition, when $\lambda = 0.05$, the distribution of detected dates for compromised items shows a significant portion of items being truncated by the end of the 30-day testing window. **Figure 3** provides a direct explanation why the detection accuracy is only 93.70% when λ is small. It is expected that, given more time, more compromised items would be detected and the detection accuracy would be higher.

TABLE 1 | Detection accuracy and Type-I error for organized item theft (standard error is given in parenthesis).

Leakage rate (λ)	Accuracy (%)	Type-I error (%)
0.05	93.70 (0.54)	4.49 (0.37)
0.10	99.86 (0.09)	6.56 (0.56)
0.30	99.93 (0.06)	7.67 (0.70)
0.50	99.43 (0.27)	4.09 (0.76)
0.70	99.61 (0.13)	4.89 (0.74)
1.00	99.04 (0.16)	4.99 (0.34)
1.50	98.85 (0.26)	4.32 (0.83)

Table 2 shows the detection lag and the estimation lag. According to **Table 2**, the mean detection lag is more than 10 days when λ is small (0.05 and 0.10 in our study). When $\lambda \geq 0.3$, the detection lag drops to ∼4 days. The probabilities for a coming test taker to have preknowledge of the item are estimated using Equation 15 with λ and average detection lag. Although the detection lag for small λ is large, the impact of the large lag is actually smaller than the cases with large λs. On the other hand, the estimation lag is about 1 day for all λs. All the above results are obtained using $\epsilon = 90\%$ in Equation 13. When $\epsilon = 85\%$ or $\epsilon = 95\%$ is used, the estimation lag is slightly worse yet still quite comparable.

Figure 4 shows the distribution of items that are incorrectly flagged as compromised (type-I error) as a function of item difficulty, at different leakage rates. It suggests that, in general, easier items are much more prone to type-I error. Since most of the test takers could correctly answer an easy item without any preknowledge, the majority of the responses will be 1 s regardless of item leakage. In this case, the detection algorithm will capitalize on the randomness of item responses, which in turn triggers more false positives.

Further, given the estimation of an item compromise point, test practitioners could re-evaluate a test-taker's ability by removing the responses to the suspicious items from ability estimation. Suspect items are defined as those compromised items administered to test takers who take the test after the item compromise point. For example, if an item is flagged as being compromised on day 3 and it was assigned to a test taker on day 4, this item will be classified as a suspicious item for that test taker. **Figure 5** compares the ability estimation with and without suspicious items. The results indicate that, after removing the suspicious items, the ability estimation is significantly better than the one in which all items are used, as evidenced by higher correlation between true and estimated ability, and smaller RMSE in ability estimates. **Figure 5C** shows the effective number of items for ability estimation, meaning the number of items left after removing suspicious items. Under the organized item theft scenario, the effective test length could drop to as low as 22 items, which is about half of the original test length

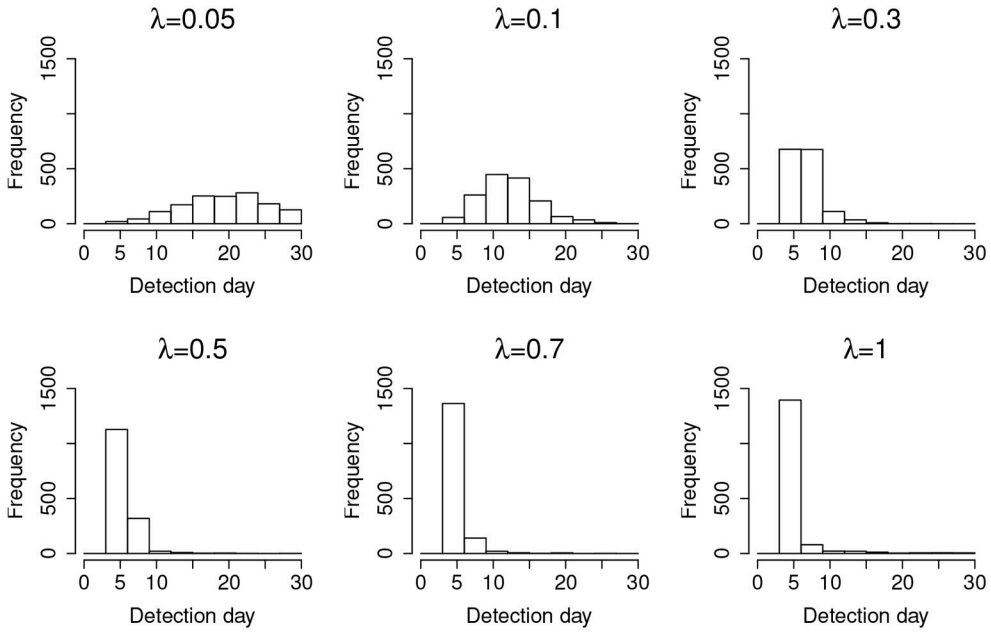


FIGURE 3 | Distribution of the detection day for organized item theft.

TABLE 2 | Detection lag and estimation lag for organized item theft (standard error is given in parenthesis).

Leakage rate (λ)	Detection lag (days)	Estimation lag (days)	Probability of preknowledge
0.05	17.47 (0.16)	0.61 (0.19)	0.58
0.10	10.61 (0.11)	–0.46 (0.11)	0.65
0.30	4.69 (0.06)	–0.95 (0.04)	0.76
0.50	3.47 (0.05)	–1.07 (0.03)	0.82
0.70	3.03 (0.05)	–1.13 (0.02)	0.88
1.00	3.11 (0.08)	–1.07 (0.02)	0.96
1.50	4.96 (0.14)	–1.00 (0.01)	1.00

40. Since the number of effective items will affect the accuracy of the ability estimation, it is expected that the estimation should be more accurate when λ is small (when there are more effective items left, as shown in **Figure 5C**), corresponding to an increase in correlation and decrease in RMSE as shown in **Figures 5A,B**.

4.2. Random Item Leakage

Results from studying the random item leakage conditions show common patterns with those of the organized item theft conditions. However, unlike the scenario of organized item theft, random item leakage does not always start at the beginning of the item bank rotation. The leakage can occur any time before the rotation of the item pool. Therefore, more data are available before the leakage. This part of the study examines how the model performs under such a scenario. **Table 3** shows the detection power when random item leakage happens in the first 25 days. As with organized item theft, the detection accuracy is very close to 100% when $\lambda \geq 0.3$. Due to the shortage of detection time when λ is small, the detection accuracy drops significantly given the 30-day simulation window. Therefore, it is difficult to effectively detect slow leaking items when the compromise date is close to the end of the test cycle. For example, if a test taker decides to share the test items assigned to him/her at day 25, π_t will not change much from day 25 to 30 when λ is small. When λ is large, however, a significant change of π_t could still be observed within 5 days. **Figure 6** shows the distribution of the detection days under the random leakage conditions. Compared with **Figure 3**, the distribution has large variability. The truncation of the detection day is severe in this case when λ is small.

Table 4 shows the detection lag and the estimation lag of the compromise time for random item leakage. The detection lags are about 1 day shorter than in the case of organized item theft, which suggests the model works better for the random leakage scenario. This is because, for the organized item theft scenario, the detection could not start until day 4, since all the item theft is assumed to happen in the first 4 days. As a consequence, for those items compromised at day 1, the earliest detection day is day 4 (i.e., the lowest possible lag is 3 days). On the other hand, although the assumption of organized item theft affects the detection lag, it does not significantly affect the estimated compromise time very much. The method used to estimate the item compromise time shows similar results in both scenarios, which is about 1 day. Similar to **Figures 4**, **7** also shows that most of the type-I errors are related to those easy items under random leakage scenario as well.

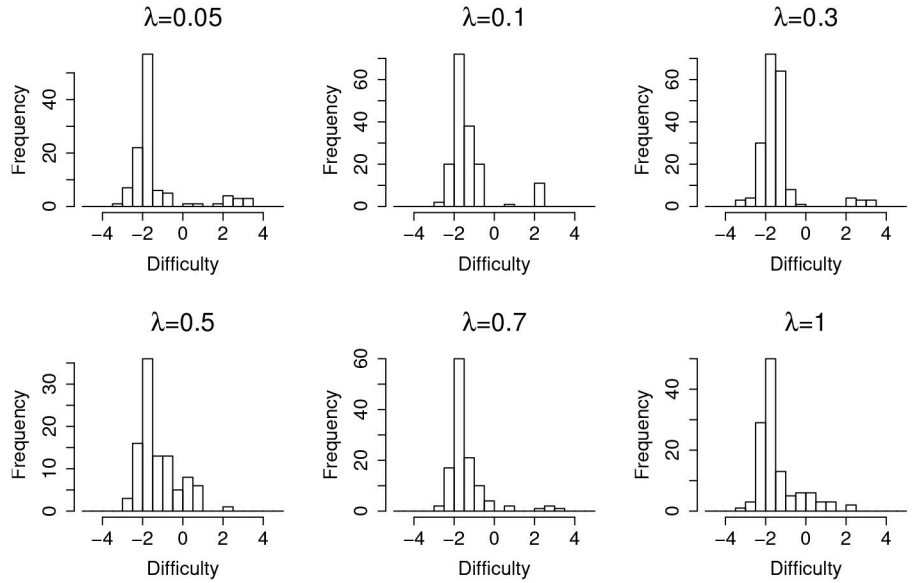


FIGURE 4 | Item distribution of Type-I error items for organized item theft.

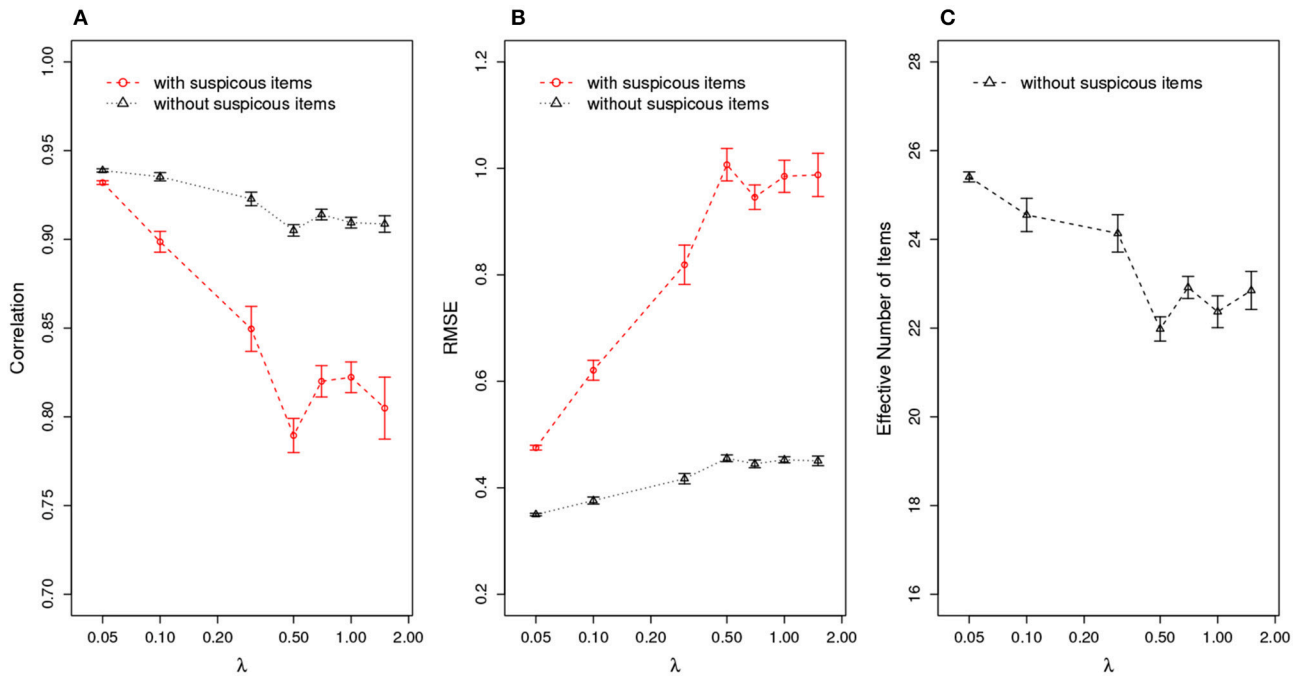


FIGURE 5 | Ability Estimation with/without Suspicious Items for Organized Item Theft. **(A)** correlation of estimated $\hat{\theta}$ with true θ; **(B)** RMSE of estimated $\hat{\theta}$; **(C)** effective number of items after removing suspicious items. (X axis is log scale).

Figure 8 compares the ability estimation with and without those suspicious items. Similar to the scenario of organized item theft, the ability estimation is significantly improved after removing suspicious items. **Figure 8C** shows that the effective test length is about four items longer than the other scenario above.

5. REAL DATA APPLICATION

In this study, we demonstrate the use of the proposed methods with real data from a large-scale operational CAT program that offers continuous testing. Item response data for about 10 days from two operational item pools are used for the analysis.

There are 2905 items in total and only 32 items are flagged as being compromised, with nominal alpha level at 0.05. This result indicates that this operational testing program is rather secure, with only slightly over 1% (32 out of 2905) of potential leakage detected. Although the nominal Type-I error is 0.05, the empirical alpha level may be different due to many factors, e.g., the short testing interval (10 days). For all four typical curves illustrated in **Figure 1**, dashed lines indicate when the leakage is detected by our proposed method. There are 14, 4, and 8 flagged items, respectively, in Types b, c and d. Since the method is designed to monitor the probability change sequentially, information after the detection (dashed line) is not used for fitting the model. In contrast to Types b and c, Type d items are challenging to interpret. They may not necessarily be compromised but the large fluctuation that triggered the flag for these items suggests testing practitioners should investigate further these items closely in case there is a leakage. The Type d scenario might indicate group preknowledge of the item of interest. One conjecture is that those who cheat often also attempt to time the item pool rotation. For example, they try to schedule and take tests as soon as they have certain amount of preknowledge of items after the pool rotation, to improve their chance of seeing some of the leaked items before the pool rotates again. Since future responses are not foreseen and we can only draw conclusions based on the response data currently at hand, in practice, once an item is flagged (no matter if it is Type b, c, or d), it should be removed from the item pool at least temporarily. When an item is flagged, one cannot be sure if its probability curve will eventually go back up or not. Test practitioners need to balance between being conservative and liberal. Given the importance of test security, if only a small number of items are flagged as being potentially compromised, the cost to exclude those items from test administration is limited so the choice is obvious.

TABLE 3 | Detection accuracy and Type-I error for random item leakage (standard error is given in parenthesis).

Leakage rate (λ)	Accuracy (%)	Type-I error (%)
0.05	67.63 (2.61)	2.02 (0.44)
0.10	87.00 (1.98)	4.54 (0.43)
0.30	96.64 (1.03)	5.14 (0.92)
0.50	98.80 (0.33)	4.67 (1.27)
0.70	99.13 (0.50)	4.85 (0.50)
1.00	99.79 (0.10)	5.22 (0.90)
1.50	99.74 (0.14)	4.01 (0.83)

6. MODEL COMPARISON

We compare our proposed detection method with the existing method (Zhang, 2014), using both simulation data and real data. Zhang's sequential model requires setting of two parameters, the length of burn-in period and the size of the moving window. We follow Zhang's simulation study and set them as 150 and 50, respectively. Our proposed cloglog detection model, on the other hand, contains no tuning parameters and the detection starts automatically at day 4 since the model has three coefficients to fit.

First, we apply Zhang's detection model to our simulation data with a leakage process taken into consideration. **Tables 5**, **6** summarize the results on the random leakage scenario and the

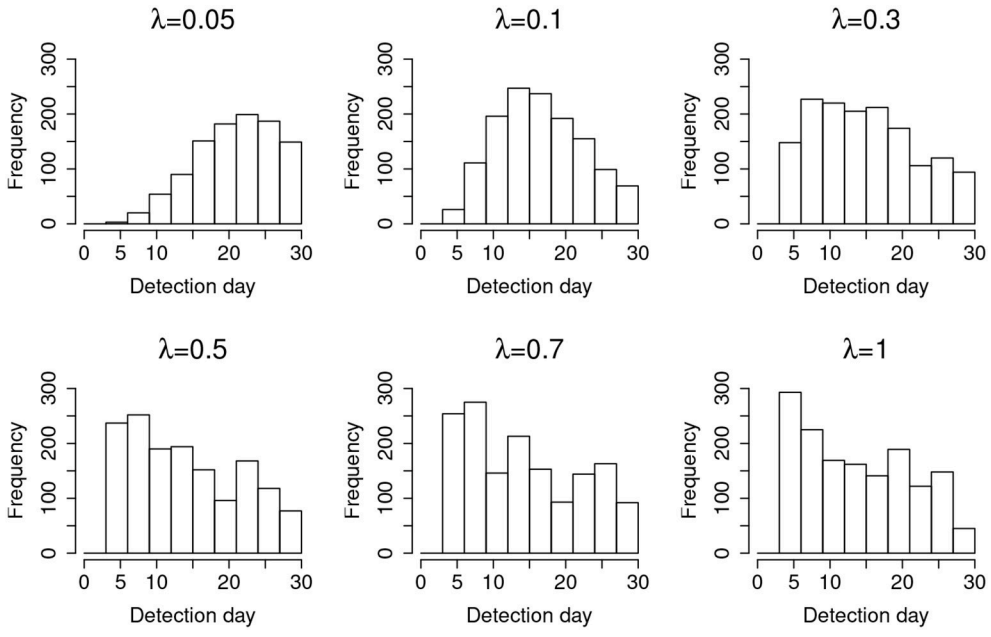


FIGURE 6 | Distribution of the detection day for random item leakage.

organized theft scenario, respectively. Firstly, we notice that in all scenarios, the type-I error is much larger than the α value, the nominal type-I error. For example, under $\alpha = 0.05$, the type-I error is larger than 70% under every simulation scenario. This means that Zhang's method did a poor job in controlling the type-I error. Secondly, even if we ignore the inflation of type-I error, Zhang's method still has a lower power than our method. The difference in power of the two methods is especially large when the leakage rate is small. For example, under the random leakage scenario with a low leakage rate ($\lambda = 0.05$), the power of our method is 67.63%, while Zhang's method is 37.44% when the type-I error is reasonably low ($< 2\%$ achieved under $\alpha = 0.0001$). This agrees with our expectation: sliding-window-based methods are not as efficient in capturing slow leakage as methods that describe and utilize the shape of probability change.

TABLE 4 | Detection lag and estimation lag for random item leakage (standard error is given in parenthesis).

Leakage rate (λ)	Detection lag (days)	Estimation lag (days)	Probability of preknowledge
0.05	12.42 (0.15)	0.66 (0.16)	0.46
0.10	7.76 (0.08)	0.18 (0.08)	0.54
0.30	3.88 (0.05)	–0.90 (0.05)	0.69
0.50	3.00 (0.04)	–1.00 (0.04)	0.78
0.70	2.66 (0.04)	–1.16 (0.04)	0.84
1.00	2.38 (0.04)	–1.14 (0.03)	0.91
1.50	2.64 (0.07)	–1.17 (0.03)	0.98

We also apply Zhang's sequencial method to the real dataset we used in the section of "Real Data Application". **Figure 9** shows how the number of items flagged as compromised increases as the nominal type-I error level increases. Strange enough, while the number of leaked items flagged by our method shows a roughly linear increase as the nominal α value increases, the number of leaked items flagged by Zhang's method shows a dramatic increase when the nominal type-I error is in the range of 0.02 and 0.05. When α is 0.05, Zhang's method flagged over 600 items as compromised, which is more than 20% of the entire item pool. Although it is hard to make any conclusive statement on a real dataset with no knowledge about which items are truly leaked, based on results from our simulation study it is not completely unreasonable to suspect that this high rate of detection may be due to severe inflation of the type-I error.

7. CONCLUSION AND DISCUSSION

In this study, we have proposed a general detection model that considers the practical dynamics of the item leaking process. The method shows, through all our simulation studies, a strong detection power for various leakage rates with well-controlled type-I error. The model also provides a way to estimate the time point at which an item is compromised, which may be helpful for testing practitioners to better secure the testing process.

The goal of our method is to detect the item leakage for various leakage rates with unknown underlying leaking processes. Therefore, the simulation model of the leakage is purposefully designed to differ from the compromised item detection model. The results show that the proposed model for detection performs very well under such scenarios, which

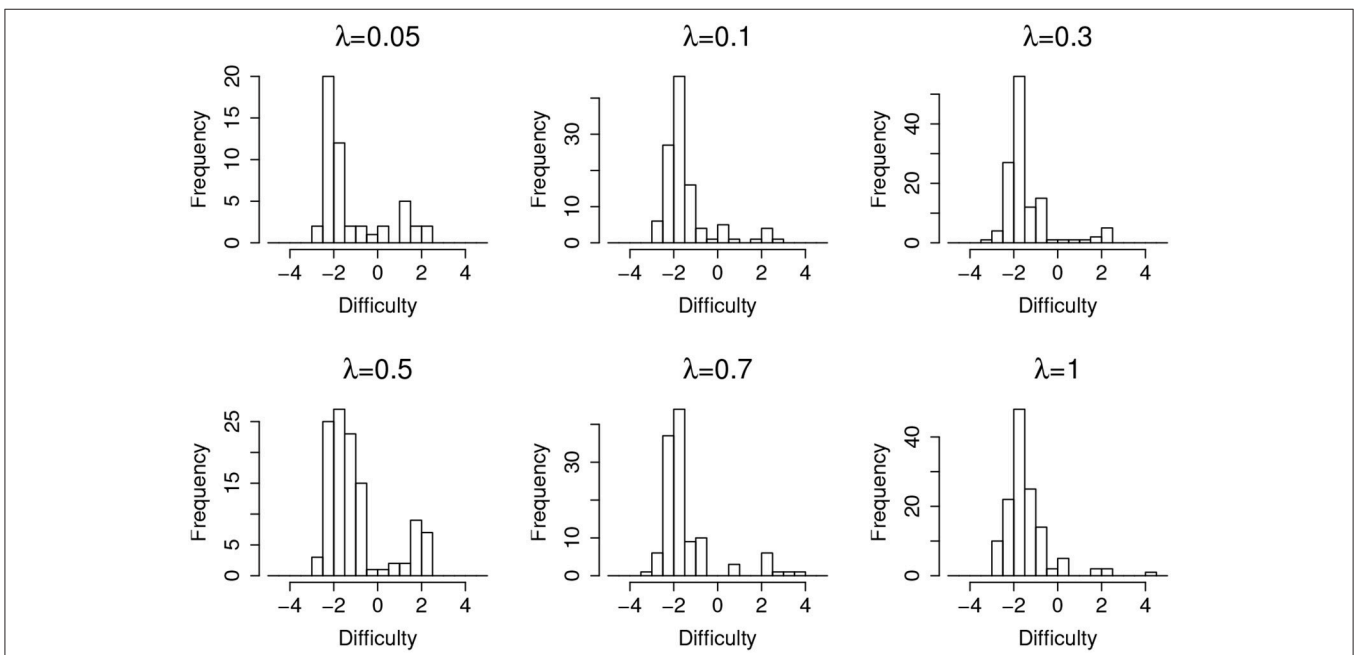


FIGURE 7 | Item distribution of Type-I error items for random item leakage.

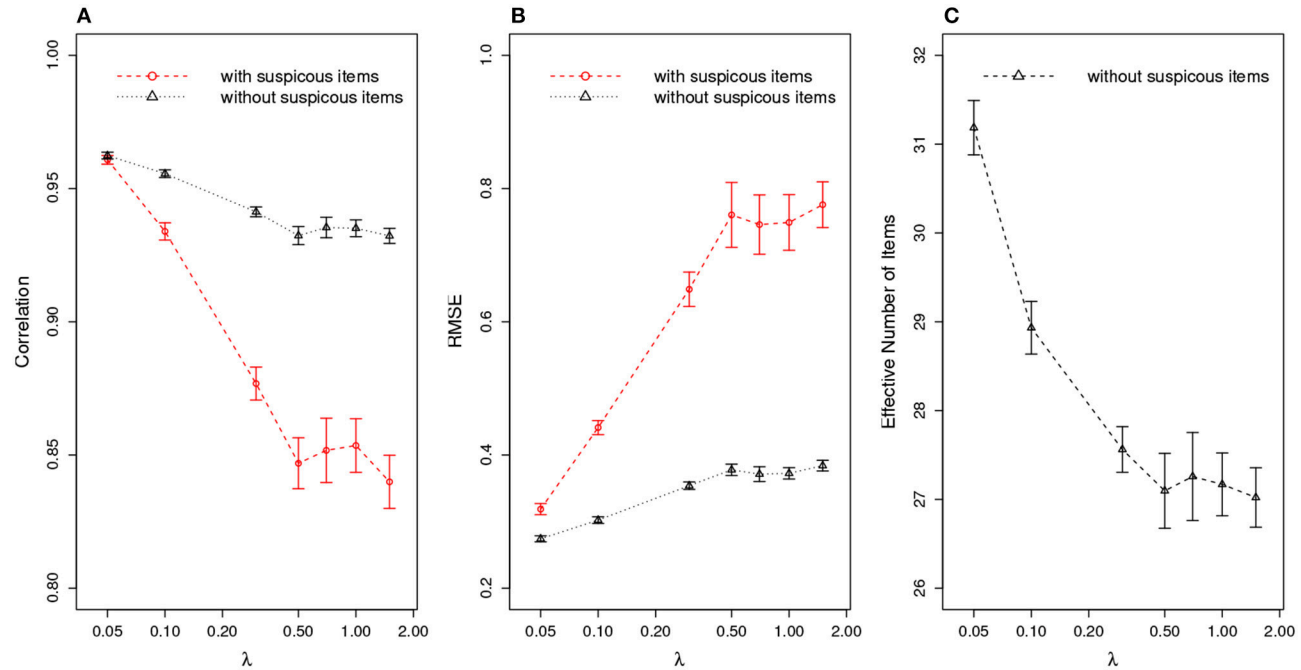


FIGURE 8 | Ability Estimation with/without Suspicious Items for Random Item Leakage. **(A)** correlation of estimated $\hat{\theta}$ with true θ; **(B)** RMSE of estimated $\hat{\theta}$; **(C)** effective number of items after removing suspicious items. (X axis is log scale).

TABLE 5 | Application of Zhang's sequential method to random leakage scenario.

Leakage rate	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.001$		$\alpha = 0.0001$	
	Accuracy(%)	Type-I(%)	Accuracy(%)	Type-I(%)	Accuracy(%)	Type-I(%)	Accuracy(%)	Type-I(%)
0.05	99.68	89.87	94.50	48.36	73.95	8.90	37.44	1.58
0.1	99.87	84.29	98.08	42.37	91.78	9.75	71.31	2.95
0.3	99.40	77.97	99.31	39.00	97.18	9.30	90.51	2.98
0.5	99.93	77.89	99.74	37.83	98.01	8.56	92.05	2.75
0.7	99.94	74.96	99.68	32.80	97.78	6.79	93.06	2.46
1.0	100.00	76.40	99.41	35.60	96.81	8.11	91.46	2.46
1.5	99.93	74.40	99.01	32.38	95.74	7.47	91.36	1.98

TABLE 6 | Application of Zhang's sequential method to organized theft scenario.

Leakage rate	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.001$		$\alpha = 0.0001$	
	Accuracy(%)	Type-I(%)	Accuracy(%)	Type-I(%)	Accuracy(%)	Type-I(%)	Accuracy(%)	Type-I(%)
0.05	100.00	81.81	99.81	36.89	91.75	9.04	44.80	2.63
0.1	99.93	72.07	98.86	36.41	95.32	9.66	56.41	3.89
0.3	99.93	75.37	99.93	40.17	92.22	11.56	64.05	4.51
0.5	99.93	76.72	99.63	41.27	90.39	11.92	61.03	3.10
0.7	99.68	79.17	99.65	42.74	83.88	10.96	60.33	2.80
1.0	99.54	76.60	96.63	40.86	80.73	9.96	61.54	2.56
1.5	99.19	77.48	93.23	43.83	74.96	10.34	56.59	1.88

is a strong indicator of the generality and powerfulness of our detection method. Estimates of both detection accuracy and detection lag are close to the expected value when the leakage rate is not too small. When the leakage is very slow, we have observed a longer detection lag time. The impact of this lag, however, can be quite mild in real data applications: When λ is small, the

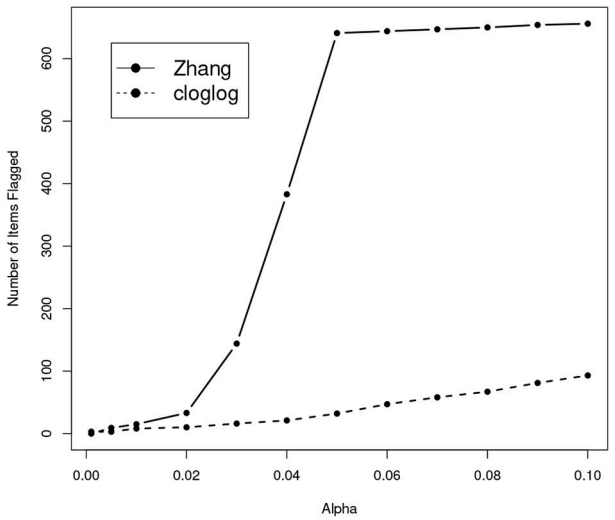


FIGURE 9 | Number of items that are flagged as compromised with different α for two models.

change in probability of getting a correct answer is not large even with a relatively large detection lag. Further, this lag is inevitable: Determining whether an item is compromised when the leakage is slow is intrinsically difficult, no matter what method is used.

The assumption of the detection model is that, given an infinitely long testing window, all test takers eventually will be aware of the compromised item and hence be able to respond correctly, which is implicit in Equation 7. In practice, it may be the case that some portion of test takers will not gain any preknowledge of the items, no matter the length of long the testing window. Therefore, the probability of correctly answering a compromised item ultimately may never reach 100%. In that case, we can generalize our method to cover such a scenario as follows:

$$\text{cloglog}\frac{\pi_t - \pi_e}{\pi_0 - \pi_e} = \beta t + \alpha \Rightarrow \pi_t = \pi_e + (\pi_0 - \pi_e)(1 - e^{-e^{\beta t + \alpha}}), \tag{21}$$

where one more parameter π_e is introduced to represent the expected upper asymptote after the item has been compromised. The π_e could be any value between $[0, \pi_0]$. When π_e=0, the model reduces to the simplified model in Equation 7. It will be our future work to implement this more general model.

The validation of the model was performed both by simulation and real data. Through the simulation study we were able to generate different leakage dynamics and test the effectiveness of our proposed method in these scenarios. Note that, although both models control the leakage speed, the parameter β in the detection model is not mathematically related to the λ in the simulation model. Actually, our proposed method essentially focuses on detecting the leakage pattern. As long as the overall pattern of the expected probability curve is similar to what we proposed, the method should work. We also applied the method to real data to demonstrate its utility in practice.

The simulation study shows that our proposed method is powerful and reliable when applied to CAT using the maximum item information method for item selection. But the method is not limited to a particular item selection method. Letting $g(\theta)$ represent the distribution of an individual's ability assigned to an item, the expected probability for a correct response of this item is:

$$E(P(\theta|a,b,c)) = \int_{-\infty}^{\infty} P(\theta|a,b,c)g(\theta)d\theta = F(a,b,c). \tag{22}$$

The expected probability, therefore, does not depend on the distribution of θ. Different item selection algorithms provide different $g(\theta)$, but will still lead to a constant expected probability of a given item. Furthermore, this method can also be applied to a non-CAT scenario. Compared with the CAT scenario, where individual's abilities fluctuate around the item difficulty, the distribution of non-CAT is expected to be more spread out. Since, for the CAT scenario, only test takers whose estimated abilities are close to a certain value (see Equation 2) will be assigned to this item, the distribution of their abilities is less variable than the original $g(\theta)$. Therefore, it is expected that more data are required if this detection model is applied, in order to draw a statistically significant conclusion.

Although the time unit in this study is set at the day level, its selection is very flexible and can be set at finer levels if necessary. The best way to select a time unit depends on the property of the test of interest and expert judgment of experienced testing practitioners. For example, given a large number of scheduled test takers per day, the time unit could be further divided by hourly increments. This would allow for more time points to be used for model fitting, subsequently leading to higher detection sensitivity. On the other hand, instead of aggregating the data by time, one could also choose to aggregate the data by a fixed number of item responses, e.g., every 20 responses. In addition, the type-I error for the hypothesis test is set to be 0.05 in this study, following convention. In practice, the cutoff could be chosen per test practitioners' preference as well.

Our study shows that the ability estimation $\hat{\theta}$ can be significantly improved by removing the responses of suspicious items. A potential future study is to apply our method to determine whether or not an examinee has preknowledge of some test items. This could be accomplished by comparing the ability estimates derived with and without the suspicious items (i.e., items that are flagged by our method). As mentioned in the introduction, many individual-level preknowledge detection methods essentially compare the ability estimates obtained from the secure vs. suspicious test items (Belov, 2016). Our method allows practitioners to identify a set of suspicious items, which is critical to the success of those individual-level detection methods. A retest may be necessary for those test takers whose ability estimates significantly differ with and without suspicious items.

Comparing with the existing sequential method (Zhang, 2014), our method shows large performance boosts in all our simulation data with a variety of leakage rates. On real data, although it is impossible to evaluate and compare the true performance of different methods, our method does not show the apparently erroneous shape of the curve of how the number of flagged items changes according to the nominal Type-I error.

Further, Zhang's method asks the user to set the window size parameter, which can be almost unfeasible, our method does not have such tuning parameters.

AUTHOR CONTRIBUTIONS

CL and JL developed the method. CL did the simulation study and analysis on real data. KH provided the real data and did the real data analysis with CL and JL. CL, JL, and KH wrote the manuscript.

ACKNOWLEDGMENTS

The authors have relied upon data supplied by the Graduate Management Admission Council (GMAC) to conduct the independent research that forms the basis for the findings and conclusions stated by the authors in this article. These findings and conclusions are the opinion of authors only, and do not necessarily reflect the opinion of GMAC. The authors also would like to thank the comments and suggestions from editor and three reviewers.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis*, 3rd Edn. New Jersey, NJ: John Wiley & Sons, Inc.

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., and Shi, M. (2007). On the performance of the lz person-fit statistic. *Pract. Assess. Res. Eval.* 12:16. Available online at: http://pareonline.net/getvn.asp?v=12&n=16

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *J. Comput. Adapt. Test.* 2, 37–58. doi: 10.7333/1410-0203037

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Appl. Psychol. Meas.* 40, 83–97. doi: 10.1177/0146621615603327

Belov, D. I., and Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback Leibler divergence and K-index. *Appl. Psychol. Meas.* 34, 379–392. doi: 10.1177/0146621610370453

Belov, D. I., and Armstrong, R. D. (2011). Distributions of the kullback leibler divergence with applications. *Br. J. Math. Stat. Psychol.* 64, 291–309. doi: 10.1348/000711010X522227

Belov, D. I., Pashley, P. J., Lewis, C., and Armstrong, R. D. (2007). "Detecting aberrant responses with Kullback Leibler distance," in *New Trends in Psychometrics*, eds K. Shigemasu, A. Okada, T. Imaizumi, and T. Hoshino (Tokyo: Universal Academy Press), 7–14.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley), 397–479.

Bock, R. D., and Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Appl. Psychol. Meas.* 6, 431–444. doi: 10.1177/014662168200600405

Chang, H. H., and Zhang, J. (2002). Hypergeometric family and test overlap rates in computerized adaptive testing. *Psychometrika* 67, 387–398. doi: 10.1007/BF02294991

Chang, H. H., and Zhang, J. (2003). "Assessing CAT security breaches by the item pooling index," in *Paper Presented at the Annual Meeting of National Council on Measurement in Education* (Chicago, IL).

Chen, S., Ankenmann, R. D., and Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *J. Educ. Meas.* 40, 129–145. doi: 10.1111/j.1745-3984.2003.tb01100.x

Choe, E. M., Zhang, J. amnd Chang, H. H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika* 83, 650–673. doi: 10.1007/s11336-017-9596-3

Davey, T., and Nering, N. (2002). "Controlling item exposure and maintaining item security," in *Computer-Based Testing: Building the Foundation for Future Assessments*, eds C. N. Mills, M. T. Potenza, J. J. Fremer, and W. C. Ward (Mahwah, NJ: Erlbaum), 165–191.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *Br. J. Math. Stat. Psychol.* 38, 67–86. doi: 10.1111/j.2044-8317.1985.tb00817.x

du Toit, M. (ed.). (2003). *IRT from SSI: BILOGMG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International, Inc.

Guo, J., Tay, L., and Drasgow, F. (2009). Conspiracies and test compromise: an evaluation of the resistance of test systems to small-scale cheating. *Int. J. Test.* 9, 283–309. doi: 10.1080/15305050903351901

Hetter, R., and Sympson, B. (1997). "Item exposure control in CAT-ASVAB," in *Computerized Adaptive Testing: From Inquiry to Operation*, eds W. Sands, B. Waters, and J. McBride (Washington, DC: American Psychological Association), 141–144.

Levine, M. V., and Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika* 53, 161–176. doi: 10.1007/BF02294130

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

McLeod, L., Lewis, C., and Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Appl. Psychol. Meas.* 27, 121–137. doi: 10.1177/0146621602250534

Mills, C. N., and Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Appl. Meas. Educ.* 9, 287–304. doi: 10.1207/s15324818ame0904_1

Obregon, P. (2013). "A Bayesian approach to detecting compromised items," In *Paper Presented at the Annual Meeting of the National Council on Measurement in Education* (San Francisco, CA).

Shao, C., Li, J., and Cheng, Y. (2015). Detection of test speededness using change-point analysis. *Psychometrika*. 81, 1118–1141. doi: 10.1007/s11336-015-9476-7

Stocking, M. L. (1994). *Three Practical Issues for Modern Adaptive Testing Item Pools (ETS RR-94-95)*. Princeton, NJ: ETS.

Stocking, M. L., and Lewis, C. (1995). *A New Method of Controlling Item Exposure in Computerized Adaptive Testing (ETS RR-95-25)*. Princeton, NJ: ETS.

Sympson, J. B., and Hetter, R. D. (1985). "Controlling item-exposure rates in computerized adaptive testing," in *Proceedings of the 27th Annual Meeting of the Military Testing Association* (San Diego, CA: Navy Personnel Research and Development Center), 973–977.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educ. Meas.* 17, 17–27. doi: 10.1111/j.1745-3992.1998.tb00632.x

Yi, Q., Zhang, J., and Chang, H. (2008). Severity of organized item theft in computerized adaptive testing: a simulation study. *Appl. Psychol. Meas.* 32, 543–558. doi: 10.1177/0146621607311336

Zhang, J. (2014). A sequential procedure for detecting compromised items in the item pool of a CAT system. *Appl. Psychol. Meas.* 38, 87–104. doi: 10.1177/0146621613510062

Zhang, J., Chang, H., and Yi, Q. (2012). Comparing single-pool and multiple-pool designs regarding test security in computerized testing. *Behav. Res. Methods* 44, 742–752. doi: 10.3758/s13428-011-0178-5

Zhang, J., and Li, J. (2016). Monitoring items in real time to enhance CAT security. *J. Educ. Meas.* 53, 131–151. doi: 10.1111/jedm.12104

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Liu, Han and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

According to the chain rule of differentiation, the deduction could be divided into two parts: First derive the log-likelihood toward π_t; then calculate the derivative of π_t toward model parameters π_0, β and α. Therefore, for convenience, let $\mathbf{\Psi} = (\pi_0, \beta, \alpha)^T$ and let

$$l = \log L = \sum_{t=1}^{T} \left[y_t \log \pi_t + (n_t - y_t) \log(1 - \pi_t) \right] = \sum_{t=1}^{T} f(\pi_t, n_t, y_t), \tag{A1}$$

where $f(\pi_t, n_t, y_t)$ is function of π_t, n_t, y_t. Then we have,

$$\begin{cases} \dfrac{\partial l}{\partial \psi_k} = \displaystyle\sum_{t=1}^{T} \frac{\partial f(\pi_t, n_t, y_t)}{\partial \pi_t} \frac{\partial \pi_t}{\partial \psi_k} \\ \dfrac{\partial^2 l}{\partial \psi_k^2} = \displaystyle\sum_{t=1}^{T} \left[\frac{\partial^2 f(\pi_t, n_t, y_t)}{\partial \pi_t^2} \left(\frac{\partial \pi_t}{\partial \psi_k} \right)^2 + \frac{\partial f(\pi_t, n_t, y_t)}{\partial \pi_t} \frac{\partial^2 \pi_t}{\partial \psi_k^2} \right] \\ \dfrac{\partial^2 l}{\partial \psi_{k_1} \partial \psi_{k_2}} = \displaystyle\sum_{t=1}^{T} \left[\frac{\partial^2 f(\pi_t, n_t, y_t)}{\partial \pi_t^2} \left(\frac{\partial \pi_t}{\partial \psi_{k_1}} \frac{\partial \pi_t}{\partial \psi_{k_2}} \right) + \frac{\partial f(\pi_t, n_t, y_t)}{\partial \pi_t} \frac{\partial^2 \pi_t}{\partial \psi_{k_1} \partial \psi_{k_2}} \right], k_1 \neq k_2 \end{cases} \tag{A2}$$

where $k = 1, 2, 3$ and $t = 1, 2, \ldots, T$. In addition, from equation 3 we get,

$$\begin{cases} \dfrac{\partial f(\pi_t, n_t, y_t)}{\partial \pi_t} = \dfrac{y_t}{\pi_t} - \dfrac{n_t - y_t}{1 - \pi_t} \\ \dfrac{\partial^2 f(\pi_t, n_t, y_t)}{\partial \pi_t^2} = -\dfrac{y_t}{\pi_t^2} - \dfrac{n_t - y_t}{(1 - \pi_t)^2}. \end{cases} \tag{A3}$$

For convenience, let

$$\Omega = \frac{\pi_0 - \pi_t}{\pi_0} \Rightarrow \pi_t = \pi_0(1 - \Omega), \tag{A4}$$

from both equations A2 and A3, we finally get, for π_0

$$\begin{cases} \dfrac{\partial \pi_t}{\partial \pi_0} = 1 - e^{-e^{\beta t + \alpha}} = \dfrac{\pi_t}{\pi_0} = 1 - \Omega \\ \dfrac{\partial^2 \pi_t}{\partial \pi_0^2} = 0 \end{cases} \tag{A5}$$

and for β,

$$\begin{cases} \dfrac{\partial \pi_t}{\partial \beta} = \pi_0 t \cdot e^{\beta t + \alpha} \cdot e^{-e^{\beta t + \alpha}} = t(\pi_0 - \pi_t) \cdot \log \dfrac{\pi_0}{\pi_0 - \pi_t} = -t\pi_0 \Omega \cdot \log \Omega \\ \dfrac{\partial^2 \pi_t}{\partial \beta^2} = \pi_0 t^2 \cdot e^{\beta t + \alpha} \cdot e^{-e^{\beta t + \alpha}} \cdot (1 - e^{\beta t + \alpha}) = t^2(\pi_0 - \pi_t) \cdot \log \dfrac{\pi_0}{\pi_0 - \pi_t} \cdot (1 - \log \dfrac{\pi_0}{\pi_0 - \pi_t}) \\ \qquad\qquad = -t^2 \pi_0 \Omega \cdot \log \Omega \cdot (1 + \log \Omega) \end{cases} \tag{A6}$$

and for α,

$$\begin{cases} \dfrac{\partial \pi_t}{\partial \alpha} = \pi_0 \cdot e^{\beta t + \alpha} \cdot e^{-e^{\beta t + \alpha}} = (\pi_0 - \pi_t) \cdot \log \dfrac{\pi_0}{\pi_0 - \pi_t} = -\pi_0 \Omega \cdot \log \Omega \\ \dfrac{\partial^2 \pi_t}{\partial \alpha^2} = \pi_0 \cdot e^{\beta t + \alpha} \cdot e^{-e^{\beta t + \alpha}} \cdot (1 - e^{\beta t + \alpha}) = (\pi_0 - \pi_t) \cdot \log \dfrac{\pi_0}{\pi_0 - \pi_t} \cdot (1 - \log \dfrac{\pi_0}{\pi_0 - \pi_t}) \\ \qquad\qquad = -\pi_0 \Omega \cdot \log \Omega \cdot (1 + \log \Omega) \end{cases} \tag{A7}$$

and the derivatives for cross terms are,

$$\begin{cases} \dfrac{\partial^2 \pi_t}{\partial \pi_0 \partial \beta} = t \cdot e^{\beta t + \alpha} \cdot e^{-e^{\beta t + \alpha}} = \dfrac{t(\pi_0 - \pi_t)}{\pi_0} \cdot \log \dfrac{\pi_0}{\pi_0 - \pi_t} = -t\Omega \cdot \log \Omega \\ \dfrac{\partial^2 \pi_t}{\partial \pi_0 \partial \alpha} = e^{\beta t + \alpha} \cdot e^{-e^{\beta t + \alpha}} = \dfrac{\pi_0 - \pi_t}{\pi_0} \cdot \log \dfrac{\pi_0}{\pi_0 - \pi_t} = -\Omega \cdot \log \Omega \\ \dfrac{\partial^2 \pi_t}{\partial \beta \partial \alpha} = \pi_0 t \cdot e^{\beta t + \alpha} \cdot e^{-e^{\beta t + \alpha}} \cdot (1 - e^{\beta t + \alpha}) = t(\pi_0 - \pi_t) \cdot \log \dfrac{\pi_0}{\pi_0 - \pi_t} \cdot (1 - \log \dfrac{\pi_0}{\pi_0 - \pi_t}) \\ \qquad\qquad = -t\pi_0 \Omega \cdot \log \Omega \cdot (1 + \log \Omega) \end{cases} \tag{A8}$$

Therefore, the Fisher information matrix is,

$$[I_{jk}(\pi_0, \beta, \alpha)] = -E\Big[\frac{\partial^2 l(\pi_0, \beta, \alpha)}{\partial \psi_j \partial \psi_k}\Big]. \tag{A9}$$

Take equation A2 into equation A9, and we have,

$$\begin{aligned}[I_{jk}(\pi_0, \beta, \alpha)] &= -E\Bigg[\sum_{t=1}^{T}\Bigg[\frac{\partial^2 f(\pi_t, n_t, y_t)}{\partial \pi_t^2}\Big(\frac{\partial \pi_t}{\partial \psi_{k_1}}\frac{\partial \pi_t}{\partial \psi_{k_2}}\Big) + \frac{\partial f(\pi_t, n_t, y_t)}{\partial \pi_t}\frac{\partial^2 \pi_t}{\partial \psi_{k_1}\partial \psi_{k_2}}\Bigg]\Bigg] \\ &= -\sum_{t=1}^{T}\Bigg[E\Big[\frac{\partial^2 f(\pi_t, n_t, y_t)}{\partial \pi_t^2}\Big]\Big(\frac{\partial \pi_t}{\partial \psi_{k_1}}\frac{\partial \pi_t}{\partial \psi_{k_2}}\Big) + E\Big[\frac{\partial f(\pi_t, n_t, y_t)}{\partial \pi_t}\Big]\frac{\partial^2 \pi_t}{\partial \psi_{k_1}\partial \psi_{k_2}}\Bigg].\end{aligned} \tag{A10}$$

Since

$$\begin{cases} E\Big[\frac{\partial f(\pi_t, n_t, y_t)}{\partial \pi_t}\Big] = E\Big[\frac{y_t}{\pi_t} - \frac{n_t - y_t}{1 - \pi_t}\Big] \\ \qquad\qquad = \frac{n_t \cdot \pi_t}{\pi_t} - \frac{n_t - n_t \cdot \pi_t}{1 - \pi_t} = 0 \\ E\Big[\frac{\partial^2 f(\pi_t, n_t, y_t)}{\partial \pi_t^2}\Big] = E\Big[-\frac{y_t}{\pi_t^2} - \frac{n_t - y_t}{(1 - \pi_t)^2}\Big] \\ \qquad\qquad = -\frac{n_t \cdot \pi_t}{\pi_t^2} - \frac{n_t - n_t \cdot \pi_t}{(1 - \pi_t)^2} = -\frac{n_t}{\pi_t(1 - \pi_t)}, \end{cases} \tag{A11}$$

the Fisher information matrix could be simplified to

$$[I_{jk}(\pi_0, \beta, \alpha)] = \sum_{t=1}^{T}\frac{n_t}{\pi_t(1 - \pi_t)} \cdot \Big(\frac{\partial \pi_t}{\partial \psi_{k_1}}\frac{\partial \pi_t}{\partial \psi_{k_2}}\Big). \tag{A12}$$

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 March 2019
doi: 10.3389/fpsyg.2019.00620

A Multidimensional IRT Approach for Dynamically Monitoring Ability Growth in Computerized Practice Environments

Jung Yeon Park[1]*, *Frederik Cornillie*[1], *Han L. J. van der Maas*[2] and *Wim Van Den Noortgate*[1]

[1] Faculty of Psychology and Educational Sciences, imec–ITEC, KU Leuven, Leuven, Belgium, [2] Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

Adaptive learning systems have received an increasing attention as they enable to provide personalized instructions tailored to the behaviors and needs of individual learners. In order to reach this goal, it is desired to have an assessment system, monitoring each learner's ability change in real time. The Elo Rating System (ERS), a popular scoring algorithm for paired competitions, has recently been considered as a fast and flexible method that can assess learning progress in online learning environments. However, it has been argued that a standard ERS may be problematic due to the multidimensional nature of the abilities embedded in learning materials. In order to handle this issue, we propose a system that incorporates a multidimensional item response theory model (MIRT) in the ERS. The basic idea is that instead of updating a single ability parameter from the Rasch model, our method allows a simultaneous update of multiple ability parameters based on a compensatory MIRT model, resulting in a multidimensional extension of the ERS ("M-ERS"). To evaluate the approach, three simulation studies were conducted. Results suggest that the ERS that incorrectly assumes unidimensionality has a seriously lower prediction accuracy compared to the M-ERS. Accounting for both speed and accuracy in M-ERS is shown to perform better than using accuracy data only. An application further illustrates the method using real-life data from a popular educational platform for exercising math skills.

Keywords: multidimensional IRT, Elo rating system, adaptive practice, speed-accuracy trade-off, e-learning

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Xiaoou Li,
University of Minnesota Twin Cities,
United States
Chun Wang,
University of Washington,
United States

***Correspondence:**
Jung Yeon Park
ellie.park@kuleuven.be

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *18 November 2018*
Accepted: *06 March 2019*
Published: *29 March 2019*

Citation:
Park JY, Cornillie F, van der Maas HLJ and Van Den Noortgate W (2019) A Multidimensional IRT Approach for Dynamically Monitoring Ability Growth in Computerized Practice Environments. Front. Psychol. 10:620. doi: 10.3389/fpsyg.2019.00620

INTRODUCTION

Over the past decade adaptive learning systems have received an increasing attention as they enable to provide instructions tailored to the behaviors, needs, and learning pace of individual learners. In this way the learners can benefit from more personalized learning items. Therefore, it is desired for the systems to have a learner modeling method that keeps track of the learner's cognitive states and its evolution in a timely and flexible manner. In the context of computerized adaptive testing (CAT; van der Linden, 2000) the use of item response theory (IRT) is a common method to model the relationship between the learner's ability level and their responses to different measurable items.

As originally intended for high-stakes standardized tests, studies related to CAT primarily zooms in on how to increase the precision of the examinee's ability level estimate by successively rendering

most informative items, or how to decrease the number of test items while maintaining a high level of precision in estimating the ability level. Because typically no feedback is given during the test, the true ability level is not expected to evolve. The idea of CAT also can be applied to learning environments in which learners interact with items, toward computerized adaptive practice (CAP; Klinkenberg et al., 2011). An a-priori expectation is that the learners in a learning environment, unlike in a testing environment, tend to develop their knowledge by interacting with the items rendered (and by getting feedback on their responses), and their true ability levels consequently evolve in real time. Therefore, a first step toward the goal of the adaptive learning system of optimizing the learning gain is tracing the learners' ability evolution in a fast and accurate manner.

In the context of intelligent tutoring systems (ITS), there are specialized approaches for tracing the learner's mastery of knowledge. A representative example is Bayesian knowledge tracing (BKT; Corbett and Anderson, 1994). In BKT, the learner's knowledge state is represented by a set of multiple binary latent variables that indicate mastery or non-mastery of the skills. The probability of having mastered each skill is estimated by binary measurement outcomes (correct or incorrect responses to items) and iteratively updated by using the rule of Bayes. Similar to CAT, however, the methods require a calibration on large samples using some nontrivial estimation techniques (expectation-maximization algorithm, or exhaustive search) that require high computational power (Papousek et al., 2014; Pálanek, 2016).

To that extent, an interesting alternative that can be considered for tracking the learner's ability evolution is the Elo rating system (ERS; Elo, 1978). The ERS was originally developed for calculating relative skill levels of players in chess performances, and the method also has widely been used in sport statistics for paired competitions (e.g., major league baseball). More recently, the ERS has been applied to various contexts of educational and psychological studies (e.g., Attali, 2014; Brinkhuis et al., 2015). In regard of its application to online-learning environment, the paired competition can be thought of as an interaction between the learner and the item. In general, the ERS algorithm is formulated to update the learner ability and item difficulty parameters from the Rasch model. To be specific, once a learner has responded to an item, the ERS updates the individual learner's ability level estimate that was based on his or her previous trajectory. Given the learner's current ability level, the next item is chosen by its difficulty level. A practical strength of this approach is that the method is conceptually fast and readily implementable in any software.

Several articles compared the performance of ERS with that of traditional IRT modeling to explore whether its parameter estimation is as accurate as the traditional approach. Maris and Van der Maas (2012) showed that the ability estimates updated from ERS method is highly correlated with the expected a posteriori (EAP) estimates from an IRT model when a speed-accuracy trade-off scoring rule was used. Studies also compared the performance of ERS with alternative methods for estimating item difficulties. For example, Wauters et al. (2012) compared the quality of the ERS-based item difficulty estimates with those based on maximum likelihood procedures, proportion correct, and human judgement methods, and found that the ERS provides reliable results with a sample size of 200 learners. Similarly, Pálanek (2016) provided evidence that there is a high correlation between ERS-based item difficulty estimates and joint maximum likelihood-based estimates.

Researchers (e.g., Klinkenberg et al., 2011; Savi et al., 2015; Braithwaite et al., 2016; Coomans et al., 2016; Hofman et al., 2018) also provided empirical evidence in favor of the ERS, by means of massive log data from Maths Garden, a CAP system where the learner ability and item difficulty levels are updated on the fly. Park et al. (2018) proposed a method to alleviate the cold-start in adaptive learning systems—the problem that for new learners we do not have an idea of their ability and therefore the adaptive learning environment might not perform well until the learner made a substantial number of items. The authors proposed using an explanatory IRT model based on learner-item interaction data and learner features (e.g., age, gender, or learning disability) and estimate the learners initial ability levels and their ability changes while not engaged in the learning environment.

Despite the increasing number of studies applying the ERS in adaptive learning systems, in the majority of these studies, the ERS is intended to track just a single broad ability. In contrast, monitoring multiple abilities not only forms the basis of learners' understanding of the material, but also provides direct information to educational researchers and instructors as to the areas that learners need to improve upon (Ferrini-Mundy and Schmidt, 2005). Therefore, identifying his or her progress on more fine-grained ability dimensions would imply an important advancement of the adaptive learning system, because of the sheer amount of information about the learner's learning state. Doebler et al. (2015) and Pálanek (2016) proposed an improved ERS algorithm for tracking multiple dimensions of ability. Yet, their methodological focus is still on situations where items are allowed to load on only one of the multiple ability dimensions in the answering process. More recently, Chen et al. (2018) and Tang et al. (2018) used a Markov decision process to track multiple dimensions of ability. In these studies, the learner's ability was modeled by a set of multiple binary latent variables that indicate mastery or non-mastery of the skills while a reinforcement learning approach was proposed to recommend personalized items.

In the current article, therefore, we propose to address these issues by using a multidimensional IRT (MIRT) model to track the (continuous) ability parameter estimates within ERS. The basic idea is that instead of assuming a unidimensional trait of item responses, our approach will assume that a single item may involve more than one ability parameters. Therefore, we extend the standard ERS that updates a single ability parameter based on the Rasch model, and will allow to have a simultaneous update on multiple ability parameters based on a compensatory MIRT model ("M-ERS").

In the next section we give more details on the methodological framework of the ERS and its application to educational settings. We then propose our method ("M-ERS") that is formulated to update multiple abilities. Next, we will evaluate the performance of our method through three simulation studies. Furthermore,

the method will be demonstrated using a real application of learning data obtained from an educational platform for children's math ability development. We end with conclusions and implications.

ELO RATING SYSTEM

The ERS is originally rooted in the Bradley Terry Luce (BTL; Bradley and Terry, 1952) model, a probabilistic model that predicts the outcome of players in a type of paired competitions. Specifically, the expected outcome that one player defeats his or her opponent is formulated as follows:

$$P_{ij} = P\left(i \text{ defeats } j\right) = \frac{\theta_i}{\theta_i + \theta_j}, \quad (1)$$

where θ_i and θ_j represent the ratings (e.g., latent traits) of players i and j, respectively. By setting up $\theta_i = e^{\theta_i}$, Equation (1) can be transformed to a logistic function of the difference between θ_i and θ_j, which comprises of the expected outcome of the ERS. That is,

$$P_{ij} = P\left(i \text{ defeats } j\right) = \frac{exp(\theta_i - \theta_j)}{1 + exp(\theta_i - \theta_j)} \quad (2)$$

Likewise, both the BTL model and the ERS are based on the probability of winning a competition; however, the latter method is additionally intended to supply easy-to-compute updates as new outcomes are observed. In other words, the ERS takes an algorithmic heuristic to easily update the expected outcome for the next iteration, based upon the estimated latent trait (i.e., θ_i and θ_j) at the current iteration. Kiraly and Qian (2017) showed that the derivative of a likelihood function for Equation (2) based on a single data point produces the following updating component for the ERS algorithm:

$$\frac{\partial l\left(\theta_i, \theta_j \middle| Y_{ij}\right)}{\partial \theta_i} = Y_{ij}\left(1 - P_{ij}\right) - \left(1 - Y_{ij}\right) P_{ij} = Y_{ij} - P_{ij}, \quad (3)$$

where $l\left(\theta_i, \theta_j \middle| Y_{ij}\right) = Y_{ij} \log P_{ij} + (1 - Y_{ij})\log(1 - P_{ij})$.

In sum, given the observations for a competition between players i and j, the estimates of θ_i and θ_j are updated simultaneously. Specifically,

$$\begin{aligned} \hat{\theta}_i &= \hat{\theta}_i + K\left\{Y_{ij} - P_{ij}\right\} \text{ for a player i,} \\ \hat{\theta}_j &= \hat{\theta}_j - K\left\{Y_{ij} - P_{ij}\right\} \text{for a player j.} \end{aligned} \quad (4)$$

In the equation above, the term $\left\{Y_{ij} - P_{ij}\right\}$ can be viewed as the discrepancy between what is expected and what is observed. In fact, the ERS can be viewed as a type of the stochastic gradient descent (SGD; Robbins and Monro, 1951) algorithm where the updating rule in the system corresponds to the update of parameters along the error gradient (Pálanek, 2016). The update will be larger if the current parameter setting produces a large discrepancy. Note that K is a step size that defines to what extent the ability estimate can be affected by the difference between the current and expected responses for the student p.

APPLICATION TO ADAPTIVE LEARNING SYSTEMS

In adaptive learning environments, the paired competition occurs when the learner interacts with the learning material (=item). The ERS process can be applied as follows. Consider $\theta_{i(t)}$ be an ability of a learner i (unidimensional continuous variable) after solving an item at measurement occasion t. Also, suppose $Y_{ij(t)}$ be the learner i outcome for item j measured at measurement occasion t, where the outcome is dichotomously scored (0 = incorrect; 1 = correct answer to the item). Then the ERS for updating the ability parameter takes the following sequence:

$$\begin{aligned} \hat{\theta}_{i(t)} &= \hat{\theta}_{i(t-1)} + K\left\{Y_{ij(t)} - P_{ij(t)}\right\} \text{ for a learner i} \\ \hat{\beta}_{j(t)} &= \hat{\beta}_{j(t-1)} - K\left\{Y_{ij(t)} - P_{ij(t)}\right\} \text{ for an item j,} \end{aligned} \quad (5)$$

where $\hat{\theta}_{i(t-1)}$ is the ability estimate at the previous measurement time $t-1$ for the learner i, $\hat{\beta}_{j(t-1)}$ is the item difficulty estimate at the previous measurement time $t-1$ for the item j, and $P_{ij(t)}$ is the expected response for the current measurement occasion t. Consequently, a learner interacting with a very difficult item risks losing a little bit of ability level in case of failure, with the possibility of gaining much greater ability level in case of success. Several studies have explored the optimal step size K for the ERS for student modeling. Wauters et al. (2012) suggested using a constant step size, $K = 0.4$ in the context of educational data. On the other hand, other studies (e.g., Glickman, 1999; Klinkenberg et al., 2011; Papousek et al., 2014; NiŽnan et al., 2015) proposed that the step size needs to decrease as a function of a total number of item answered and therefore the system gains more information about the learner's true ability level.

In Equation (5), it is possible that the outcome $Y_{ij(t)}$ can be scored by considering whether the learner completed the item within the allotted limit. Maris and Van der Maas (2012) derived a scoring rule that accounts for response time and accuracy, and applied it within ERS. While the ERS can be used to gradually obtain reliable estimates of both student's abilities and item difficulties, adaptive item sequencing can be more efficient if we could start from a pre-calibrated item bank, including information on item difficulty and possibly other characteristics of items, and from which items with undesired characteristics are excluded (van Groen et al., 2014). In this case, $\hat{\beta}_{j(t)}$ in Equation (5) needs not be updated.

MULTIDIMENSIONAL EXTENSION OF THE ERS (M-ERS)

In this section we propose an extended version of the ERS that enables the system to track multidimensional abilities in real time. Specifically, the proposed algorithm can handle two types of dimensional structures in the item bank–(a) "within-item dimensionality" where a single item can be associated with more than one task ability; as well as (b) "between-item dimensionality" where a set of items is associated with multiple abilities, while each item measures only one of those abilities.

Suppose an adaptive learning environment contains an item bank that is designed to measure a total of M-dimensional abilities i.e., $\theta_i = (\theta_{i1}, \ldots, \theta_{iM})'$ for a student i. The multidimensional dichotomous logistic model (Reckase, 1985) that describes the probability of a correct answer to item j can be formulated by either conjunctive or compensatory assumptions about how the latent abilities are combined. In a conjunctive model assumption, it is assumed that the learner should have each of the relevant abilities in order to answer an item correctly. The probability of a correct response therefore is a joint product of the inverse logit function of the difference between each of the abilities and the corresponding item difficulty:

$$P_{ij} = P\left(Y_{ij} = 1\right) = \prod_{m=1}^{M} \frac{exp(\alpha_{jm}[\theta_{im} - \beta_j])}{1 + exp(\alpha_{jm}[\theta_{im} - \beta_j])}, \quad (6)$$

where $P\left(Y_{ij} = 1\right)$ indicates the probability of a correct answer, θ_{im} is the mth ability parameter of the learner i ($m = 1, \ldots, M$), α_{jm} is the item discrimination of the item j corresponding to mth ability dimension, and β_j denotes the overall difficulty level of the item j. On the other hand, in a compensatory model, on the other hand, it is assumed that the lack of one ability can be compensated by greater level of another ability, as follows:

$$P_{ij} = P\left(Y_{ij} = 1\right) = \frac{exp(\sum_{m=1}^{M} \alpha_{jm}\theta_{im} - \beta_j)}{1 + exp(\sum_{m=1}^{M} \alpha_{jm}\theta_{im} - \beta_j)}. \quad (7)$$

The difference between the observed and the expected performance P_{ij} based on the multidimensional IRT models is used to update the ability parameters after each item response. Specifically, the P_{ij} within ERS for the m-th ability for person i on measurement occasion t is updated as follows:

$$\begin{aligned} \hat{\theta}_{im(t)} &= \hat{\theta}_{im(t-1)} + D_{m(t)}\, K\, \left\{Y_{ij(t)} - P_{ij(t)}\right\} \\ \hat{\beta}_{j(t)} &= \hat{\beta}_{j(t-1)} - D_{m(t)}\, K\, \left\{Y_{ij(t)} - P_{ij(t)}\right\}, \end{aligned} \quad (8)$$

where $D_{m(t)}$ is a weight to specify whether the mth ability is indicated by the item given at t-th step. For the ability that is indicated by the item, $D_{m(t)}$ equals 1. For the ability that is not indicated by the item, the weight takes values between zero and one.

SIMULATION STUDY

To explore the performance of the M-ERS method in terms of estimating the real-time evolution of multidimensional ability parameters for individual learners, we apply the method to data generated under a variety of conditions. In accordance with our research questions, the simulation study consists of three parts. In Study 1, we examine the result of a standard ERS that naively assumes the unidimensionality of ability parameter, where in fact data involve a multidimensional ability. In Study 2, we explore the performance of the M-ERS in relation to the total number of items answered. In Study 3, we investigate the performance of the modified M-ERS in which both response time and accuracy data are incorporated.

TABLE 1 | Patterns of multidimensionality (a sample of 15 items from two item banks).

Item ID	β_j	Item bank 1			Item bank 2		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
1	−2.534	1	.	.	1	.	.
2	−2.21	1	.	.	1	.	.
3	1.326	1	.	.	1	.	.
4	0.253	1	.	.	1	.	.
5	1.275	1	.	.	1	.	.
6	0.089	1	.	.	1	.	.
7	−0.001	1	1	.	.	1	.
8	−1.256	1	1	.	.	1	.
9	2.242	1	1	.	.	1	.
10	−1.556	1	1	.	.	1	.
11	2.213	1	1	.	.	1	.
12	−3.3	1	.	1	.	.	1
13	0.753	1	.	1	.	.	1
14	−2.246	1	.	1	.	.	1
15	−1.156	1	.	1	.	.	1

"1" indicates that the item loads on the dimension.

Item Bank

Following the literature on the MIRT (e.g., Adams et al., 1997; Hartig and Höhler, 2008), we consider two loading structures to determine patterns of the multidimensionality. Specifically, two types of item banks are created– (a) when items are allowed to load on more than a single ability dimension ("Item bank 1"); and (b) when items are allowed to load on only one of the multiple ability dimensions ("Item bank 2"). Each of them includes 200 operational items, measuring a total of three dimensions. Item bank 1 consists of a primary dimension θ_1 indicated by all the items, and two auxiliary dimensions, θ_2 and θ_3, indicated by 35% of items and the 25% of the remaining items. In Item bank 2, each item involves only one out of the three dimensions. Specifically, 40% of items involves the 1st dimension, 35% of them involves the 2nd dimension, and the remaining 25% of them involves the 3rd dimension (as an illustration, **Table 1** gives a sample of 15 items from two item banks). Based on each item bank, data are generated under a compensatory IRT model with difficulty and discrimination parameters, mimicking realistic test items:

$$P_{ij} = P\left(Y_{ij} = 1\right) = \frac{\exp\left(\sum_{m=1}^{3} \alpha_{mj}\theta_{jm} - \beta_j\right)}{1 + \exp\left(\sum_{m=1}^{3} \alpha_{mj}\theta_{jm} - \beta_j\right)}, \quad (9)$$

where the generating parameter values for the difficulty and the slope parameters are drawn from $\beta_j \sim N(0, 1)$ and $\alpha_{mj} \sim U(0.5, 2)$, where $j = 1, \ldots, 200$ (items) and $m = 1, 2, 3$ (dimensions of ability).

Persons

A total of $n = 250$ learners are considered in the simulation studies. The population distribution of ability parameters is taken to be $N(\mu, \Sigma)$, where $\mu = (1, 1, 1)'$. In Σ, all variances are equal to

1 and the three bivariate correlations are equal: dimensions were independent ($\rho_{mm'} = 0.0$), weakly correlated ($\rho_{mm'} = 0.2$), or moderately correlated ($\rho_{mm'} = 0.5$).

A total of 6 data sets were generated by following 6 scenarios (2 patterns of dimensionality × 3 correlations among ability dimensions). In each condition, the M-ERS method will be used to update the three ability parameter estimates for each learner as he or she attempts on a sequence of items. We assume that each learner is assigned a sequence of 200 items that are randomly selected. That is, item sequences are varied across learners. In M-ERS, in Equation (7), the expected response is estimated by using a compensatory IRT model with a constraint that the slope parameters α_{mj}'s are equal to 1 for simplicity (the inclusion of different loadings is a logical further extension). That is,

$$P\left(Y_{ij(t)} = 1\right) = \frac{\exp\left(\sum_{m=1}^{3} \theta_{jm} - \beta_j\right)}{1 + \exp\left(\sum_{m=1}^{3} \theta_{jm} - \beta_j\right)}.$$

In the equation, we use step size K that linearly decreases as a function of a total number of items answered between the maximum value of 0.4 and the minimum value of 0.1. Note that we assume that item difficulty parameters are considered as known (based on a calibration study), and the difficulty estimates therefore will not be further updated within the M-ERS.

STUDY 1: PREDICTION ACCURACY OF A STANDARD ERS AND M-ERS

In the first study, we explore the extent to which the unidimensional ability assumption embedded in a standard ERS has an impact on the prediction accuracy (in terms of the learners' future responses), when the truth is that the responses to the learning items involve three-dimensional ability parameters. As seen in Equation (8), predictions of the responses of learners can be achieved by using the ERS algorithm based on the known item parameters and the learner ability estimate(s) predicted by the preceding step. The prediction accuracy is calculated by classifying the expected response ($= P_{ij(t)}$) with a certain cut-point into the observed response ($= Y_{ij(t)}$) on measurement occasion t. To evaluate the quality of predictions, we use a Receiver Operating Characteristic curve (ROC). The ROC curve represents the relation between true positive rates [=TP/(TP + FN)] and false positive rates [FP/(FP + FN)] at various probability cut-off points. In case of totally random predictions, the Area Under Receiver Operating Characteristic curve (AUROC) is approximately equal to 0.5.

Figure 1 visualizes the ROC curves, comparing the performances of a standard ERS and the M-ERS. Each panel in the figure includes 6 curves, representing a combination of simulation conditions (2 patterns of dimensionality × 3 correlations among ability parameters). The x-axis and y-axis indicate the false positive rate and the true positive rate. Note that the best possible prediction method would yield a point in the upper left corner or coordinate (0, 1) of the ROC space, representing 100% sensitivity and specificity. Therefore, results from the two panels suggest that M-ERS outperforms the standard ERS in all six simulated scenarios. **Table 2** summarizes the area under the ROC curves (AUROC). In case of M-ERS, the AUROCs are much higher than the expected values using random predictions (i.e., 0.5 for AUROC). It is seen that the AUROCs for M-ERS are 0.8088 and 0.7881 for Item banks 1 and 2, respectively. However, the standard ERS generally reveals only performs marginally better than the random predictions (i.e., 0.5299 for Item bank 1 and 0.5345 for Item bank 2).

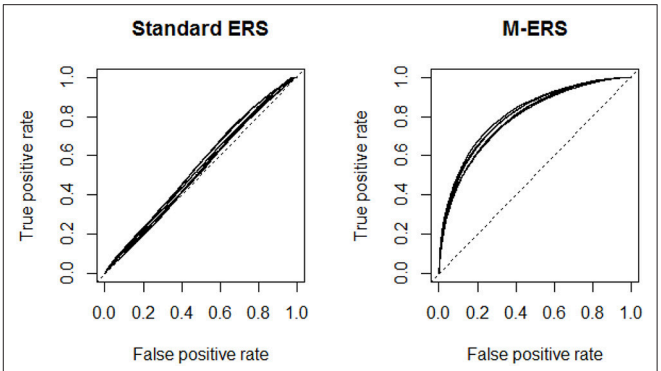


FIGURE 1 | Receiver operating characteristic (ROC) curves for a standard ERS and M-ERS. *Note*. Each panel includes six ROC curves representing a total of 6 simulation conditions (2 patterns of dimensionality × 3 correlations among ability parameters).

TABLE 2 | Area Under ROC curve (AUROC) for a standard ERS and M-ERS.

	Item bank 1		Item bank 2	
	Standard ERS	**M-ERS**	**Standard ERS**	**M-ERS**
$\rho = 0.0$	0.5197	0.8038	0.5216	0.7850
$\rho = 0.2$	0.5204	0.8070	0.5324	0.7900
$\rho = 0.5$	0.5496	0.8157	0.5494	0.7893
Average	0.5299	0.8088	0.5345	0.7881

STUDY 2: ABILITY PARAMETER ESTIMATION AS A FUNCTION OF NUMBER OF ITEMS ANSWERED

In the second study, we investigate the performance of M-ERS as a function of the total number of items answered. We also examine the effect of different simulation conditions (i.e., dimensionality patterns and correlations among true three ability parameters) on the ability parameter estimation. To evaluate the quality of the ability estimation, the estimated ability parameters are summarized by mean squared error (MSE). In particular, the differences between the true and estimated abilities at measurement occasion t are squared, and averaged over the entire sample size of new students. That is, for the learner i at the measurement occasion t:

$$MSE(\hat{\theta}_{m(t)}) = \frac{\sum_{i=1}^{N} \left(\hat{\theta}_{im(t)} - \theta_{im(t)}\right)^2}{N} \quad (10)$$

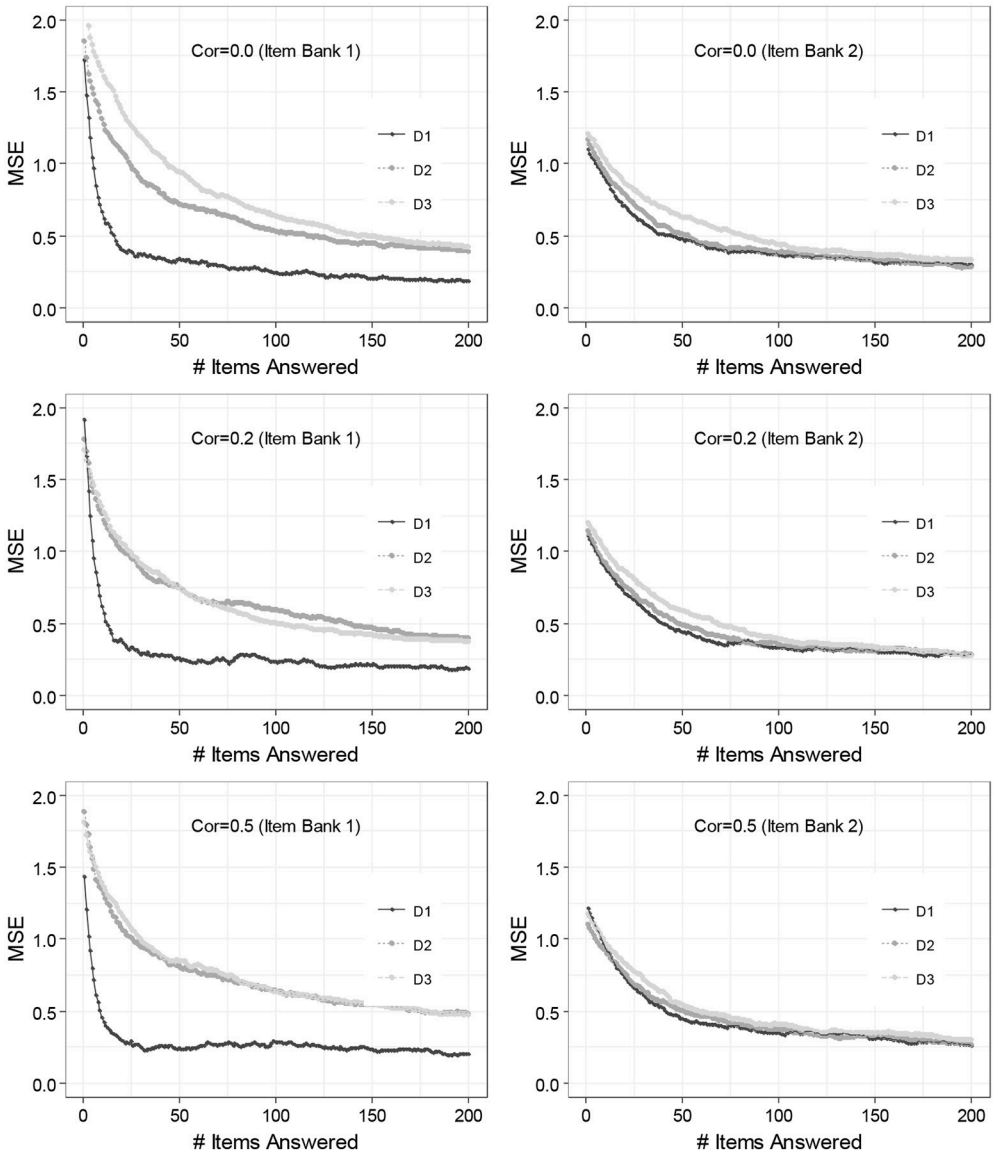


FIGURE 2 | Result of 3-dimensional ability estimation across the number of items answered. Cor, correlations between dimensions; D1, 1st dimension; D2, 2nd dimension; D3, 3rd dimension.

Figure 2 includes line plots demonstrating the performances of the M-ERS for two different patterns of dimensionality. Each panel in the figure represents a different simulation condition with the patterns of dimensionality (Item banks 1 and 2) and the true correlations among ability parameters ($\rho_{mm'} = 0.0$, 0.2, and 0.5), with a total number of items answered being on the x-axis and the MSEs on the y-axis. The three lines in each panel comprise the squared difference between the $\hat{\theta}_{im(t)}$ and $\theta_{im(t)}$ averaged across $n = 250$ individual learners for the three dimensions ("D1," "D2," and "D3" in the legend). Remind that the three panels on the left-hand side summarize the performance of the M-ERS method for Item bank 1 that exhibits one primary dimensions plus two auxiliary dimensions (Item bank 1). The other three on the right-hand side are based on the data where each item involves only a single dimension. See the section of "Study Design" for more details.

Overall, results suggest that that the MSEs tend to decrease as the total number of items answered increases. The finding is common to all three dimensions ("D1," "D2," and "D3"), but the speed of decrease varies by the number of items the ability dimension involves. For Item bank 1 (column left), it is seen that MSE for the 1st dimension (on which all items load) reveals a dramatic decrease while the first 20 items are answered (to around 0.2). Similarly, though more gradually, MSEs for 2nd and 3rd dimensions (on which only 35 and 25% of the items load) also tend to decrease. In these two auxiliary

dimensions, however, the MSE does not reach 0.2, even up to 200 items.

For Item bank 2 (column right), on the other hand, the degrees of the decreasing trends are extremely similar among the three dimensions. This can be explained by the fact that the three dimensions involve a similar amount of items i.e., 40, 35, and 25% of items load on the D1, D2, and D3, respectively. For moderate scenarios, in particular, the difference becomes extremely tiny. It is seen that the true correlations among ability parameters do not have an impact on MSEs across any measurement occasions.

As an alternative way to check the performance of M-ERS, **Figure 3** compares the ability estimates of 250 learners after 200 Elo-updates with the expected a posteriori (EAP) ability estimates obtained by fitting a compensatory IRT approach. Overall, ability estimates from EAP and M-ERS are highly correlated, regardless of any simulation conditions and the ability dimensions. The correlation coefficients (3 dimensions × 2 item banks × 3 correlation among true abilities) range from 0.967 to 0.990. Note that the EAP estimates are the results of an analysis that requires responses of many persons on many items, and is computing-intensive and therefore cannot be used on the fly. Therefore, the EAP estimates are used here as a benchmark, but they cannot be considered as a viable alternative for the ERS approach.

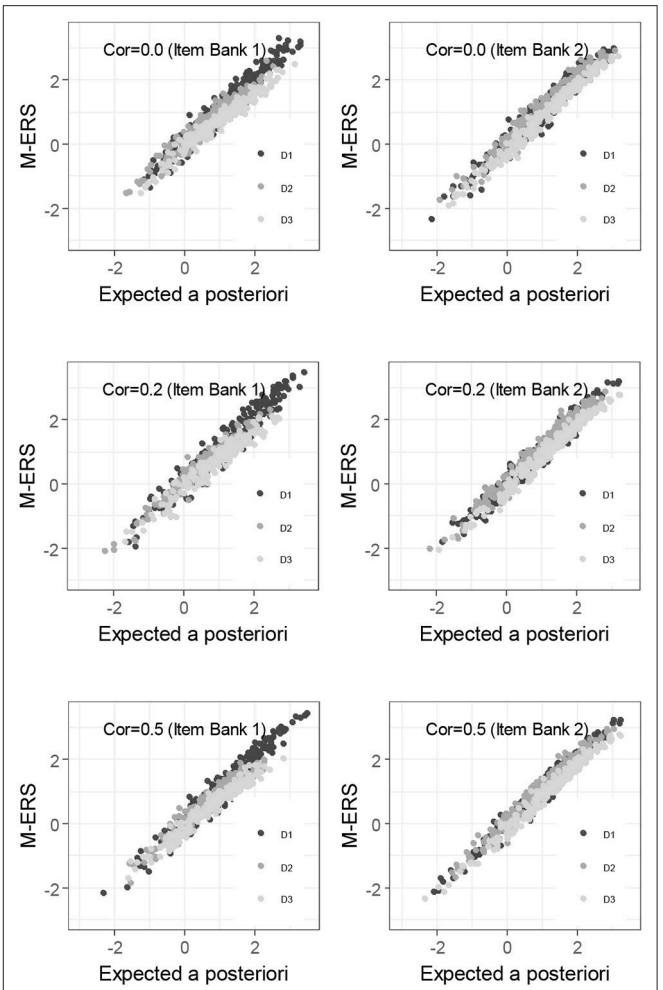


FIGURE 3 | Relations between EAP and M-ERS estimates.

STUDY 3: M-ERS FOR SPEED-ACCURACY TRADEOFF

There have been an increasing number of studies (e.g., Tuerlinckx and De Boeck, 2005; van der Linden, 2007; De Boeck et al., 2017) that account for response time as well as response accuracy in order to model the ability parameter. Of several statistical and psychological approaches to the response time modeling (van der Linden, 2009), one of the promising methods is to model response time and accuracy from the measurement perspective by two-step procedures; specifically, (a) setting up a scoring rule and (b) fitting a proper statistical model that conforms scores of the type. Klinkenberg et al. (2011) showed that the ERS method outperforms a standard CAT method (specifically, Eggen and Verschoor, 2006) when the speed-accuracy trade-off scores [so called high speed high stake (HSHS)] and the corresponding model were used. We do not know studies that model multidimensional ability trajectories based on both response time and accuracy within ERS. Therefore, we aim to explore the incorporation of the HSHS scoring rule in the proposed M-ERS method.

According to the HSHS scoring rule, the observed scores can be calculated as follows: $S_{ij} = (2Y_{ij} - 1)(d - T_{ij})$, where Y_{ij} is an accuracy for the learner i's response to the item j, d is a time limit, and T_{ij} is an time spent for the learner i until answering the item j. In this expression the residual time i.e., $(d - T_{ij})$ can compensate or penalize for the learner, corresponding to the learner's accuracy to the item. In particular, for a correct response (i.e., $Y_{ij} = 1$), the learner will gain the residual time as a score. Similarly, for an incorrect response (i.e., $Y_{ij} = 0$), the score will be reduced by the same amount. In current study, the maximum time given each item ($= d$) is restricted to be 1, so the residual time simply reflects the proportion of time left. Such a scoring scheme is especially useful to control for the case that the learner guesses instantaneously guess the given item (a quick incorrect answer). The expectation of the trade-off score, $E(S_{ij})$, for an item that is based on three abilities can naturally be extended from a unidimensional version in Maris and Van der Maas (2012). That is,

$$E\left(S_{ij}\right) = \frac{\exp\left(2(\sum_{m=1}^{3}\theta_{im} - \beta_j)\right) + 1}{\exp\left(2(\sum_{m=1}^{3}\theta_{im} - \beta_j)\right) - 1} - \frac{1}{\left(\sum_{m=1}^{3}\theta_{im} - \beta_j\right)}, \tag{11}$$

where m =1, 2, 3 (abilities). Specifically, the $E\left(S_{ij}\right)$ in Equation (10) provides the expected HSHS score for learner i's to solve the item j.

We conduct a simulation study to compare this approach with the M-ERS method based on the accuracy data only. The

TABLE 3 | Comparing IRT-based ERS algorithms for correctness vs. trade-off score (correctness and speediness combined).

		$\rho = 0.0$		$\rho = 0.2$		$\rho = 0.5$	
		|Bias|	**MSE**	**|Bias|**	**MSE**	**|Bias|**	**MSE**
ITEM BANK 1							
$\hat{\theta}_1$	Accuracy	0.009	0.300	0.008	0.286	0.009	0.289
	Speed-accuracy	0.008	0.245	0.008	0.232	0.007	0.213
$\hat{\theta}_2$	Accuracy	0.010	0.364	0.009	0.363	0.010	0.382
	Speed-accuracy	0.009	0.306	0.009	0.298	0.009	0.307
$\hat{\theta}_3$	Accuracy	0.010	0.400	0.010	0.395	0.010	0.392
	Speed-accuracy	0.009	0.337	0.009	0.336	0.009	0.332
ITEM BANK 2							
$\hat{\theta}_1$	Accuracy	0.008	0.291	0.008	0.294	0.008	0.290
	Speed-accuracy	0.008	0.245	0.008	0.251	0.008	0.252
$\hat{\theta}_2$	Accuracy	0.008	0.299	0.008	0.290	0.008	0.291
	Speed-accuracy	0.008	0.260	0.008	0.245	0.008	0.252
$\hat{\theta}_3$	Accuracy	0.009	0.337	0.009	0.353	0.009	0.315
	Speed-accuracy	0.008	0.299	0.009	0.310	0.008	0.278

|Bias|, absolute value of the bias averaged over learners and items; MSE, a mean squared error averaged over learners and items.

entire data-generating process follows what is described in the section "Study Design." The response time data is generated for each learner who solves each item, using a formula for expected response time from Maris and Van der Maas (2012), where the time limit for each item is consistently set at 1-min. Like in studies 1 and 2, the simulation conditions are combinations of patterns of dimensionality (Item banks 1 and 2) and the true correlations among ability parameters ($\rho_{mm'} = 0.0$, 0.2, and 0.5).

Table 3 comprises two tables that compare the performance of two M-ERS method when only accuracy data are used ("Accuracy") or both speed and accuracy data ("Speed-Accuracy") are used. Overall results suggest that M-ERS for speed-accuracy data shows smaller MSE, regardless of any simulation conditions.

REAL DATA ANALYSIS

Description

For illustrative purposes, we used a dataset collected from a web-based learning platform, "Number Sense" (Linsen et al., 2016) developed by KU Leuven, Belgium. It was designed as an item-based e-learning environment for 6- to 8-year-old children and includes approximate number discrimination tasks, symbolic comparison tasks, and symbolic and non-symbolic number line estimation tasks. In particular, current data were collected between Fall 2017 and Spring 2018, during one school year. It includes data from 299 students' responses to 330 items in total. Among the items, 168 of them are designed to measure (a) comparison ability and the remaining 162 items are designed to measure (b) number line estimation ability. There were no items that require both. All responses to the items are scored for accuracy i.e., the binary scale (correct/incorrect). Current log data do not include response times.

For the purpose of obtaining item parameters, data from 200 out of 299 students were used to fit a MIRT formula i.e., $P\left(Y_{ij}=1\right) = \frac{\exp(\theta_{j1}+\theta_{j2}-\beta_j)}{1+\exp(\theta_{j1}+\theta_{j2}-\beta_j)}$, where θ_{i1} and θ_{i2} reflect abilities in relation to the comparison and number line estimation, respectively. The remaining 99 students were used to illustrate the Elo algorithm. For the estimation procedure with this training set, the MCMC algorithm is implemented with R 3.3.3 (R Core Team, 2013). More specifically, JAGS (Plummer, 2015) was implemented by an R package "R2jags" (Su and Yajima, 2015) that provides wrapper functions for the Bayesian analysis program. For each analysis with the JAGS, four chains were run, and each ran for 10,000 iterations. We used a thinning parameter of four and used the first half as burn-in. (Gelman and Rubin's, 1995) statistics are used for a convergence diagnostic. Results of the Bayesian inference show that the posterior predictive mean for the correlation between the comparison and number line estimation abilities (i.e., θ_{i1} and θ_{i2}) was approximately $\hat{\rho} = 0.13$. The posterior predictive means for item discriminations and difficulties (i.e., $\hat{\alpha}_{mj}$ and $\hat{\beta}_j$) were used as known item parameters, and where therefore not updated within the Elo algorithm.

Results

Figure 4 shows the resulting ability trajectory of two randomly chosen students by fitting a standard ERS and the M-ERS. As in the simulation result, the figure presents the impact of assuming unidimensionality or multidimensionality of the ability parameter. It is noticeable that the ability estimates obtained by a standard ERS are in general greater than the two-dimensional ability estimates from the M-ERS. Based on the similarity of the Number Sense data to the data we generated in the simulation and the results we found there, we can suspect that the unidimensional ERS ability estimates for the Number Sense items are biased (upward), and the M-ERS has removed the bias. That implies that ignoring the multidimensional data structure may cause considerable bias in ability trajectory estimation in the learning environment, and therefore in a suboptimal adaptivity of the learning environment. After a longer sequence of items in the session, however, it is shown that the gaps among four approaches tend to be negligible.

CONCLUSION

In this paper, we have proposed an MIRT-based ERS method to address a dynamic estimation of the learner's progress in an adaptive practice environment where the learning items exhibit a multidimensional ability criteria. The model combines the idea of using a compensatory MIRT model to predict the learner performance with a fast and heuristic algorithm for tracking his or her irregular trend of ability parameters through the ERS.

First, we have shown that there occurs a considerable error in terms of updating the ability changes, when a unidimensional IRT is used in ERS when the truth is that there is a multidimensionality in a set of items. We have shown that the error in estimating the ability parameters can be alleviated with the compensatory IRT in ERS. Second, we have shown that at the initial step of learning, the error of ability estimates are bigger

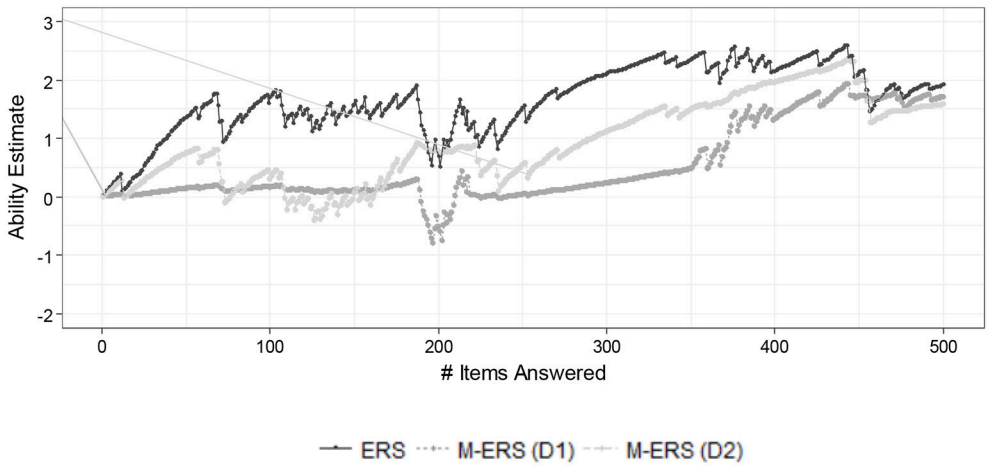


FIGURE 4 | Example of ability estimates for a student by standard ERS and M-ERS.

where each individual item involves more than one dimensions, as compared to the case where the item purely involves one of the multiple dimensions. However, we have found that the error has been noticeably reduced as more items are rendered. Third, we have extended the M-ERS method for the trade-off scores between response time and accuracy. Results show that bias and MSE of ability estimates are smaller when the HSHS (i.e., the speed-accuracy trade-off scoring that gives penalty for guessing) was incorporated in M-ERS than using the accuracy data only.

We believe that our approach offers the possibility to improve adaptivity when applied in an adaptive environment. In our simulation study, the items were chosen randomly across measurement occasions for each student, but it is also possible to administer items that optimize the item selection criteria. For example, the item can be chosen such that its difficulty level is as close as possible to the learner's current ability (e.g., 50% chance of answering correctly). In the ERS based on the IRT formula, such an item selection strategy can be flexibly adjusted to avoid too easy (e.g., 90% chance of answering correctly) or too hard (e.g., 25% chance of answering correctly) items to individual learners. That means that the ERS can provide a flexible item sequencing tool for adaptivity in which a series of items are updated in real time based on their ability or knowledge levels (Wauters et al., 2010).

Another idea that may arise when considering to deal with ability estimation in ERS is to handle the cold-start problem i.e., the system does not know a new learner's ability level in the beginning of learning stage, when the new learner comes into the e-learning system for the first time. The cold start problem may also occur when a learner leave the e-learning system for a while and return (i.e., between-session period) because external effects could lead to the ability level change. Finally, the current simulation study shows a few limitations. For instance, the true ability was assumed to be constant over time, although it tends to evolve in learning environments. Including a time trend can add additional challenges, such as the determination of a step size that is large enough to keep track of the evolving ability but not too large in order to avoid very instable ability estimates.

Nevertheless, we believe the results of current study provide valuable information about how to efficiently follow up estimate multidimensional ability changes in the e-learning environments in order to alleviate concerns about the ERS and catalyze the usefulness of the e-learning system in educational settings.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

JP developed the core ideas of the manuscript, the method, the design of the simulation study, as well as the real data analysis. WV supervised the entire work and gave feedback. FC and HvdM gave feedback on the results and the writing and contributed to the final manuscript.

FUNDING

This research includes a methodological approach and a real data example from the LEarning analytics for AdaPtiveSupport (LEAPS) project, funded by imec (Kapeldreef 75, B-3001, Leuven, Belgium) and the Agentschap Innoveren & Ondernemen. The LEAPS project aimed to develop a self-learning analytical system to enable adaptive learning. This support system can be integrated into educational games and in supporting software for difficult readers and forprofessional communication. Partners from a broad field of expertise work together within the consortium. Examples include educational and cognitive scientists, software developers, statisticians, experts in human-computer interaction and educational publishers. This extensive interdisciplinary collaboration makes it possible to create a much-needed and commercial solution.

REFERENCES

Adams, R., Wilson, M., and Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Appl. Psychol. Meas.* 21, 1–23. doi: 10.1177/0146621697211001

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educ. Psychol. Meas.* 74, 795–808. doi: 10.1177/0013164414527450

Bradley, R. A., and Terry, M. E. (1952). Rank Analysis Of Incomplete Block Designs: I. The method of paired comparisons. *Biometrika* 39:324. doi: 10.1093/biomet/39.3-4.324

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L. J., and Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: the case of arithmetic. *Cognition* 149, 40–55. doi: 10.1016/j.cognition.2016.01.004

Brinkhuis, M. J., Bakker, M., and Maris, G. (2015). Filtering data for detecting differential development. *J. Educ. Measure.* 52, 319–338. doi: 10.1111/jedm.12078

Chen, Y., Li, X., Liu, J., and Ying, Z. (2018). Recommendation system for adaptive learning. *Appl. Psychol. Meas.* 42, 24–41. doi: 10.1177/0146621617697959

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H. L., and Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PLoS ONE* 11:e0155149. doi: 10.1371/journal.pone.0155149

Corbett, A. T., and Anderson, J. R. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model. UserAdap. Inter.* 4, 253–278. doi: 10.1007/BF01099821

De Boeck, P., Chen, H., and Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses, *Br. J. Math. Stat. Psychol.* 70, 225–237. doi: 10.1111/bmsp.12094

Doebler, P., Alavash, M., and Giessing, C. (2015). Adaptive experiments with a multivariate Elo-type algorithm. *Behav. Res. Methods* 47, 384–394. doi: 10.3758/s13428-014-0478-7

Eggen, T. J. H. M., and Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Appl. Psychol. Meas.* 30, 379–393. doi: 10.1177/0146621606288890

Elo, A. E. (1978). *The Rating of Chess Players, Past and Present*. London: B.T. Batsford, Ltd.

Ferrini-Mundy, J., and Schmidt, W. H. (2005). International comparative studies in mathematics education: opportunities for collaboration and challenges for researchers. J. *Res. Math. Educ.* 36, 164–175. doi: 10.1007/978-3-319-42414-9_1

Gelman, A., and Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. Discussion of "Bayesian model selection in social research." *Soc. Method.* 25, 165–173.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired 1031 comparison experiments. *Appl. Stat.* 48, 377–394.

Hartig, J., and Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *J. Psychol.* 216, 89–101. doi: 10.1027/0044-3409.216.2.89

Hofman, A., Jansen, B., de Mooij, S., Stevenson, C., and van der Maas, H. (2018). A solution to the measurement problem in the idiographic approach using computer adaptive practicing. *J. Intel.* 6:14. doi: 10.3390/jintelligence6010014

Kiraly, F. J., and Qian, Z. (2017). *Modelling Competitive Sports: Bradley–Terry–Elo Models for Supervised and Online Learning of Paired Competition Outcomes*. arXiv 1701.08055.

Klinkenberg, S., Straatemeier, M., and van der Maas, H. L. J. (2011). Computer adaptive practice on Maths ability using a new item response model for on the fly ability and difficulty estimation. *Comput. Educ.* 57, 1813–1824. doi: 10.1016/j.compedu.2011.02.003

Linsen, S., Torbeyns, J., Verschaffel, L., Reynvoet, B., and De Smedt, B. (2016). The association between symbolic and nonsymbolic numerical magnitude processing and mental versus algorithmic subtraction in adults. *Acta Psychol.* 165, 34–42. doi: 10.1016/j.actpsy.2016.01.008

Maris, G., and Van der Maas, H. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika* 77, 615–633. doi: 10.1007/s11336-012-9288-y

NiŽnan, J., Pálanek, R., and Rihák, J. (2015). "Student models for prior knowledge estimation," in *Proceedings of the 8th International Conference on Educational Data Mining*. (Madrid: International Educational Data Mining Society), 109–116.

Pálanek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Comput. Educ.* 98, 169–179. doi: 10.1016/j.compedu.2016.03.017

Papousek, J., Pelanek, R., and Stanislav, V. (2014). "Adaptive practice of facts in domains with varied prior knowledge," in *Proceedings of the International Conference on Educational Data Mining (EDM)*, 6–13.

Park, J. Y., Joo, S. H., Cornillie, F., Van der Maas, H. L. J., and Van den Noortgate, W. (2018). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behav. Res. Method.* doi: 10.3758/s13428-018-1166-9. [Epub ahead of print].

Plummer, M. (2015). *Just Another Gibbs Sampler (JAGS)*. Retrieved from http://mcmc--jags.sourceforge.net

R Core Team (2013). *R: A Language and Environment for Statistical 1069 Computing (version 3.3.3)*. Vienna: R Foundation for 1070 Statistical Computing. Retrieved from: www.R-project.org

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Appl. Psychol. Measure.* 9, 401–441.

Robbins, H., and Monro, S. (1951). A Stochastic Approximation Method. *Ann. Math. Stat.* 22:400.

Savi, A. O., van der Maas, H. L., and Maris, G. K. (2015). Navigating massive open online courses. *Science* 347, 958–958. doi: 10.1126/science.347.6225.958

Su, Y. S., and Yajima, M. (2015). *R2jags: Using R to Run 'JAGS'. R Package Version* 0.5-7.

Tang, X., Chen, Y., Li, X., Liu, J., and Ying, Z. (2018). A reinforcement learning approach to personalized learning recommendation systems. *Br. J. Math. Stat. Psychol.* 72, 108–135. doi: 10.1111/bmsp.12144

Tuerlinckx, F., and De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika* 70, 629–650. doi: 10.1007/s11336-000-0810-3

van der Linden, W. J. (2000). "Constrained adaptive testing with shadow tests," in, *Computerized adaptive testing: Theory and practice.* eds W. J. van der Linden and C. A. W. Glas (Boston, MA: Kluwer), 27–52. doi: 10.1007/0-306-47531-6_2

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72:287. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *J. Educ. Measure.* 46, 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

van Groen, M. M., Eggen, T. J. H. M., and Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classifying respondents into multiple levels. *Appl. Psychol. Meas.* 38, 187-200. doi: 10.1177/0146621613509723

Wauters, K., Desmet, P., and Van Den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: possibilities and challenges. *J. Comput. Assist. Learn.* 26, 549–562. doi: 10.1111/j.1365-2729.2010.00368.x

Wauters, K., Desmet, P., and Van Den Noortgate, W. (2012). Item difficulty estimation: an auspicious collaboration between data and judgment. *Comput. Educ.* 58, 1183–1193. doi: 10.1016/j.compedu.2011.11.020

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Park, Cornillie, van der Maas and Van Den Noortgate. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 March 2019
doi: 10.3389/fpsyg.2019.00486

Statistical Analysis of Complex Problem-Solving Process Data: An Event History Analysis Approach

Yunxiao Chen[1]*, *Xiaoou Li*[2], *Jingchen Liu*[3] and *Zhiliang Ying*[3]

[1] Department of Statistics, London School of Economics and Political Science, London, United Kingdom, [2] School of Statistics, University of Minnesota, Minneapolis, MN, United States, [3] Department of Statistics, Columbia University, New York, NY, United States

Complex problem-solving (CPS) ability has been recognized as a central 21st century skill. Individuals' processes of solving crucial complex problems may contain substantial information about their CPS ability. In this paper, we consider the prediction of duration and final outcome (i.e., success/failure) of solving a complex problem during task completion process, by making use of process data recorded in computer log files. Solving this problem may help answer questions like "how much information about an individual's CPS ability is contained in the process data?," "what CPS patterns will yield a higher chance of success?," and "what CPS patterns predict the remaining time for task completion?" We propose an event history analysis model for this prediction problem. The trained prediction model may provide us a better understanding of individuals' problem-solving patterns, which may eventually lead to a good design of automated interventions (e.g., providing hints) for the training of CPS ability. A real data example from the 2012 Programme for International Student Assessment (PISA) is provided for illustration.

Keywords: process data, complex problem solving, PISA data, response time, event history analysis

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Frank Goldhammer,
German Institute for International Educational Research (LG), Germany
Jean-Paul Fox,
University of Twente, Netherlands

***Correspondence:**
Yunxiao Chen
y.chen186@lse.ac.uk

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *31 August 2018*
Accepted: *19 February 2019*
Published: *18 March 2019*

Citation:
Chen Y, Li X, Liu J and Ying Z (2019) Statistical Analysis of Complex Problem-Solving Process Data: An Event History Analysis Approach. Front. Psychol. 10:486. doi: 10.3389/fpsyg.2019.00486

1. INTRODUCTION

Complex problem-solving (CPS) ability has been recognized as a central 21st century skill of high importance for several outcomes including academic achievement (Wüstenberg et al., 2012) and workplace performance (Danner et al., 2011). It encompasses a set of higher-order thinking skills that require strategic planning, carrying out multi-step sequences of actions, reacting to a dynamically changing system, testing hypotheses, and, if necessary, adaptively coming up with new hypotheses. Thus, there is almost no doubt that an individual's problem-solving process data contain substantial amount of information about his/her CPS ability and thus are worth analyzing. Meaningful information extracted from CPS process data may lead to better understanding, measurement, and even training of individuals' CPS ability.

Problem-solving process data typically have a more complex structure than that of panel data which are traditionally more commonly encountered in statistics. Specifically, individuals may take different strategies toward solving the same problem. Even for individuals who take the same strategy, their actions and time-stamps of the actions may be very different. Due to such heterogeneity and complexity, classical regression and multivariate data analysis methods cannot be straightforwardly applied to CPS process data.

Possibly due to the lack of suitable analytic tools, research on CPS process data is limited. Among the existing works, none took a prediction perspective. Specifically, Greiff et al. (2015) presented a

case study, showcasing the strong association between a specific strategic behavior (identified by expert knowledge) in a CPS task from the 2012 Programme for International Student Assessment (PISA) and performance both in this specific task and in the overall PISA problem-solving score. He and von Davier (2015, 2016) proposed an N-gram method from natural language processing for analyzing problem-solving items in technology-rich environments, focusing on identifying feature sequences that are important to task completion. Vista et al. (2017) developed methods for the visualization and exploratory analysis of students' behavioral pathways, aiming to detect action sequences that are potentially relevant for establishing particular paths as meaningful markers of complex behaviors. Halpin and De Boeck (2013) and Halpin et al. (2017) adopted a Hawkes process approach to analyzing collaborative problem-solving items, focusing on the psychological measurement of collaboration. Xu et al. (2018) proposed a latent class model that analyzes CPS patterns by classifying individuals into latent classes based on their problem-solving processes.

In this paper, we propose to analyze CPS process data from a prediction perspective. As suggested in Yarkoni and Westfall (2017), an increased focus on prediction can ultimately lead us to greater understanding of human behavior. Specifically, we consider the simultaneous prediction of the duration and the final outcome (i.e., success/failure) of solving a complex problem based on CPS process data. Instead of a single prediction, we hope to predict at any time during the problem-solving process. Such a data-driven prediction model may bring us insights about individuals' CPS behavioral patterns. First, features that contribute most to the prediction may correspond to important strategic behaviors that are key to succeeding in a task. In this sense, the proposed method can be used as an exploratory data analysis tool for extracting important features from process data. Second, the prediction accuracy may also serve as a measure of the strength of the signal contained in process data that reflects one's CPS ability, which reflects the reliability of CPS tasks from a prediction perspective. Third, for low stake assessments, the predicted chance of success may be used to give partial credits when scoring task takers. Fourth, speed is another important dimension of complex problem solving that is closely associated with the final outcome of task completion (MacKay, 1982). The prediction of the duration throughout the problem-solving process may provide us insights on the relationship between the CPS behavioral patterns and the CPS speed. Finally, the prediction model also enables us to design suitable interventions during their problem-solving processes. For example, a hint may be provided when a student is predicted having a high chance to fail after sufficient efforts.

More precisely, we model the conditional distribution of duration time and final outcome given the event history up to any time point. This model can be viewed as a special event history analysis model, a general statistical framework for analyzing the expected duration of time until one or more events happen (see e.g., Allison, 2014). The proposed model can be regarded as an extension to the classical regression approach. The major difference is that the current model is specified over a continuous-time domain. It consists of a family of conditional models indexed by time, while the classical regression approach does not deal with continuous-time information. As a result, the proposed model supports prediction at any time during one's problem-solving process, while the classical regression approach does not. The proposed model is also related to, but substantially different from response time models (e.g., van der Linden, 2007) which have received much attention in psychometrics in recent years. Specifically, response time models model the joint distribution of response time and responses to test items, while the proposed model focuses on the conditional distribution of CPS duration and final outcome given the event history.

Although the proposed method learns regression-type models from data, it is worth emphasizing that we do not try to make statistical inference, such as testing whether a specific regression coefficient is significantly different from zero. Rather, the selection and interpretation of the model are mainly justified from a prediction perspective. This is because statistical inference tends to draw strong conclusions based on strong assumptions on the data generation mechanism. Due to the complexity of CPS process data, a statistical model may be severely misspecified, making valid statistical inference a big challenge. On the other hand, the prediction framework requires less assumptions and thus is more suitable for exploratory analysis. More precisely, the prediction framework admits the discrepancy between the underlying complex data generation mechanism and the prediction model (Yarkoni and Westfall, 2017). A prediction model aims at achieving a balance between the bias due to this discrepancy and the variance due to a limited sample size. As a price, findings from the predictive framework are preliminary and only suggest hypotheses for future confirmatory studies.

The rest of the paper is organized as follows. In Section 2, we describe the structure of complex problem-solving process data and then motivate our research questions, using a CPS item from PISA 2012 as an example. In Section 3, we formulate the research questions under a statistical framework, propose a model, and then provide details of estimation and prediction. The introduced model is illustrated through an application to an example item from PISA 2012 in Section 4. We discuss limitations and future directions in Section 5.

2. COMPLEX PROBLEM-SOLVING PROCESS DATA

2.1. A Motivating Example

We use a specific CPS item, *CLIMATE CONTROL* (CC)[1], to demonstrate the data structure and to motivate our research questions. It is part of a CPS unit in PISA 2012 that was designed under the "MicroDYN" framework (Greiff et al., 2012; Wüstenberg et al., 2012), a framework for the development of small dynamic systems of causal relationships for assessing CPS.

In this item, students are instructed to manipulate the panel (i.e., to move the top, central, and bottom control sliders; left side of **Figure 1A**) and to answer how the input variables (control

[1]The item can be found on the OECD website (http://www.oecd.org/pisa/test-2012/testquestions/question3/)

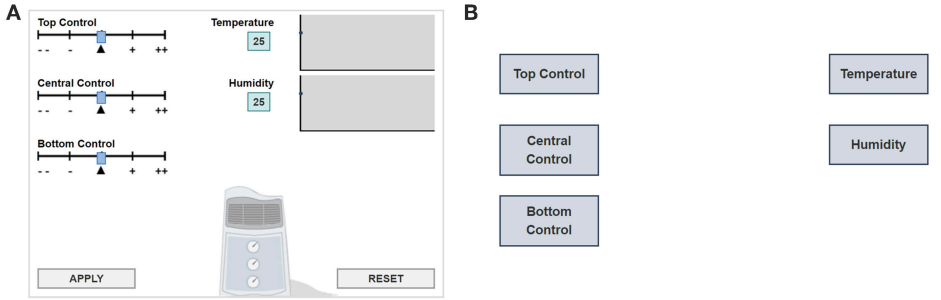


FIGURE 1 | (A) Simulation environment of CC item. **(B)** Answer diagram of CC item.

sliders) are related to the output variables (temperature and humidity). Specifically, the initial position of each control slider is indicated by a triangle "▲." The students can change the top, central and bottom controls on the left of **Figure 1** by using the sliders. By clicking "APPLY," they will see the corresponding changes in temperature and humidity. After exploration, the students are asked to draw lines in a diagram (**Figure 1B**) to answer what each slider controls. The item is considered correctly answered if the diagram is correctly completed. The problem-solving process for this item is that the students must experiment to determine which controls have an impact on temperature and which on humidity, and then represent the causal relations by drawing arrows between the three inputs (top, central, and bottom control sliders) and the two outputs (temperature and humidity).

PISA 2012 collected students' problem-solving process data in computer log files, in the form of a sequence of time-stamped events. We illustrate the structure of data in **Table 1** and **Figure 2**, where **Table 1** tabulates a sequence of time-stamped events from a student and **Figure 2** visualizes the corresponding event time points on a time line. According to the data, 14 events were recorded between time 0 (start) and 61.5 s (success). The first event happened at 29.5 s that was clicking "APPLY" after the top, central, and bottom controls were set at 2, 0, and 0, respectively. A sequence of actions followed the first event and finally at 58, 59.1, and 59.6 s, a final answer was correctly given using the diagram. It is worth clarifying that this log file does not collect all the interactions between a student and the simulated system. That is, the status of the control sliders is only recorded in the log file, when the "APPLY" button is clicked.

The process data for solving a CPS item typically have two components, knowledge acquisition and knowledge application, respectively. This CC item mainly focuses the former, which includes learning the causal relationships between the inputs and the outputs and representing such relationships by drawing the diagram. Since data on representing the causal relationship is relatively straightforward, in the rest of the paper, we focus on the process data related to knowledge acquisition and only refer a student's problem-solving process to his/her process of exploring the air conditioner, excluding the actions involving the answer diagram.

TABLE 1 | An example of computer log file data from CC item in PISA 2012.

Time	Event
0	Start.
29.5	Set top, central, and bottom controls at 2, 0, and 0, respectively, and click APPLY.
32.4	Set top, central, and bottom controls at 0, 0, and 0, respectively, and click APPLY.
35.2	Click RESET.
36.2	Set all three controls at 0, and click APPLY.
⋮	⋮
58	Connecting "top control" with "temperature."
59.1	Connecting "central control" with "humidity."
59.6	Connecting "bottom control" with "humidity."
61.5	Success.

Intuitively, students' problem-solving processes contain information about their complex problem-solving ability, whether in the context of the CC item or in a more general sense of dealing with complex tasks in practice. However, it remains a challenge to extract meaningful information from their process data, due to the complex data structure. In particular, the occurrences of events are heterogeneous (i.e., different people can have very different event histories) and unstructured (i.e., there is little restriction on the order and time of the occurrences). Different students tend to have different problem-solving trajectories, with different actions taken at different time points. Consequently, time series models, which are standard statistical tools for analyzing dynamic systems, are not suitable here.

2.2. Research Questions

We focus on two specific research questions. Consider an individual solving a complex problem. Given that the individual has spent t units of time and has not yet completed the task, we would like to ask the following two questions based on the information at time t: How much additional time does the individual need? And will the individual succeed or fail upon the time of task completion?

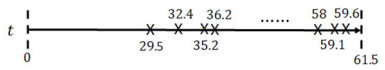


FIGURE 2 | Visualization of the structure of process data from CC item in PISA 2012.

Suppose we index the individual by i and let T_i be the total time of task completion and Y_i be the final outcome. Moreover, we denote $\mathbf{H}_i(t) = (h_{i1}(t), ..., h_{ip}(t))^\top$ as a p-vector function of time t, summarizing the event history of individual i from the beginning of task to time t. Each component of $\mathbf{H}_i(t)$ is a feature constructed from the event history up to time t. Taking the above CC item as an example, components of $\mathbf{H}_i(t)$ may be, the number of actions a student has taken, whether all three control sliders have been explored, the frequency of using the reset button, etc., up to time t. We refer to $\mathbf{H}_i(t)$ as the event history process of individual i. The dimension p may be high, depending on the complexity of the log file.

With the above notation, the two questions become to simultaneously predict T_i and Y_i based on $\mathbf{H}_i(t)$. Throughout this paper, we focus on the analysis of data from a single CPS item. Extensions of the current framework to multiple-item analysis are discussed in Section 5.

3. PROPOSED METHOD

3.1. A Regression Model

We now propose a regression model to answer the two questions raised in Section 2.2. We specify the marginal conditional models of Y_i and T_i given $\mathbf{H}_i(t)$ and $T_i > t$, respectively. Specifically, we assume

$$P(Y_i = 1|\mathbf{H}_i(t), T_i > t) = \Phi(b_{11}h_{i1}(t) + \cdots + b_{1p}h_{ip}(t)), \quad (1)$$

$$E(\log(T_i - t)|\mathbf{H}_i(t), T_i > t) = b_{21}h_{i1}(t) + \cdots + b_{2p}h_{ip}(t), \quad (2)$$

and

$$Var(\log(T_i - t)|\mathbf{H}_i(t), T_i > t) = \sigma^2, \quad (3)$$

where Φ is the cumulative distribution function of a standard normal distribution. That is, Y_i is assumed to marginally follow a probit regression model. In addition, only the conditional mean and variance are assumed for $\log(T_i - t)$. Our model parameters include the regression coefficients $B = (b_{jk})_{2\times p}$ and conditional variance σ^2. Based on the above model specification, a pseudo-likelihood function will be devived in Section 3.3 for parameter estimation.

Although only marginal models are specified, we point out that the model specifications (1) through (3) impose quite strong assumptions. As a result, the model may not most closely approximate the data-generating process and thus a bias is likely to exist. On the other hand, however, it is a working model that leads to reasonable prediction and can be used as a benchmark model for this prediction problem in future investigations.

We further remark that the conditional variance of $\log(T_i - t)$ is time-invariant under the current specification, which can be further relaxed to be time-dependent. In addition, the regression model for response time is closely related to the log-normal model for response time analysis in psychometrics (e.g., van der Linden, 2007). The major difference is that the proposed model is not a measurement model disentangling item and person effects on T_i and Y_i.

3.2. Prediction

Under the model in Section 3.1, given the event history, we predict the final outcome based on the success probability $\Phi(b_{11}h_{i1}(t) + \cdots + b_{1p}h_{ip}(t))$. In addition, based on the conditional mean of $\log(T_i - t)$, we predict the total time at time t by $t + \exp(b_{21}h_{i1}(t) + \cdots + b_{2p}h_{ip}(t))$. Given estimates of B from training data, we can predict the problem-solving duration and final outcome at any t for an individual in the testing sample, throughout his/her entire problem-solving process.

3.3. Parameter Estimation

It remains to estimate the model parameters based on a training dataset. Let our data be (τ_i, y_i) and $\{\mathbf{H}_i(t) : t \geq 0\}$, $i = 1, ..., N$, where τ_i and y_i are realizations of T_i and Y_i, and $\{\mathbf{H}_i(t) : t \geq 0\}$ is the entire event history.

We develop estimating equations based on a pseudo likelihood function. Specifically, the conditional distribution of Y_i given $\mathbf{H}_i(t)$ and $T_i > t$ can be written as

$$f_1(y|\mathbf{H}_i(t), \tau > t; \mathbf{b}_1) = \Phi(\mathbf{b}_1^\top \mathbf{H}_i(t))^y (1 - \Phi(\mathbf{b}_1^\top \mathbf{H}_i(t)))^{1-y},$$

where $\mathbf{b}_1 = (b_{11}, ..., b_{1p})^\top$. In addition, using the log-normal model as a working model for $T_i - t$, the corresponding conditional distribution of T_i can be written as

$$f_2(\tau|\mathbf{H}_i(t), \tau > t; \mathbf{b}_2, \sigma) = \frac{1}{(\tau - t)\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(\tau - t) - (\mathbf{b}_2^\top \mathbf{H}_i(t)))^2}{2\sigma^2}\right), \quad (4)$$

where $\mathbf{b}_2 = (b_{21}, ..., b_{2p})^\top$. The pseudo-likelihood is then written as

$$L(B, \sigma) = \prod_{i=1}^{N}\prod_{j=1}^{J} \left(f_1(y_i|\mathbf{H}_i(t_j), \tau_i > t_j; \mathbf{b}_1) f_2(\tau_i|\mathbf{H}_i(t_j), \tau_i > t_j; \mathbf{b}_2, \sigma)\right)^{1_{\{\tau_i > t_j\}}}, \quad (5)$$

where t_1, ..., t_J are J pre-specified grid points that spread out over the entire time spectrum. The choice of the grid points will be discussed in the sequel. By specifying the pseudo-likelihood based on the sequence of time points, the prediction at different time is taken into accounting in the estimation. We estimate the model parameters by maximizing the pseudo-likelihood function $L(B, \sigma)$.

In fact, (5) can be factorized into

$$L(B, \sigma) = L_1(\mathbf{b}_1)L_2(\mathbf{b}_2, \sigma),$$

where

$$L_1(\mathbf{b}_1) = \prod_{i=1}^{N}\prod_{j=1}^{J}(f_1(y_i|\mathbf{H}_i(t_j), \tau_i > t_j; \mathbf{b}_1))^{1_{\{\tau_i > t_j\}}}, \quad (6)$$

and

$$L_2(\mathbf{b}_2, \sigma) = \prod_{i=1}^{N}\prod_{j=1}^{J}(f_2(\tau_i|\mathbf{H}_i(t_j), \tau_i > t_j; \mathbf{b}_2, \sigma))^{1_{\{\tau_i > t_j\}}}. \quad (7)$$

Therefore, $\mathbf{b}_1$ is estimated by maximizing $L_1(\mathbf{b}_1)$, which takes the form of a likelihood function for probit regression. Similarly, $\mathbf{b}_2$ and σ are estimated by maximizing $L_2(\mathbf{b}_2, \sigma)$, which is equivalent to solving the following estimation equations,

$$\sum_{i=1}^{N}\sum_{j=1}^{J} 1_{\{\tau_i > t_j\}}\big(\log(\tau_i - t_j) - \mathbf{b}_2^{\top}\mathbf{H}_i(t_j)\big)h_{ik}(t_j) = 0, k = 1, ..., p, \quad (8)$$

and

$$\sum_{i=1}^{N}\sum_{j=1}^{J} 1_{\{\tau_i > t_j\}}\big(\sigma^2 - (\log(\tau_i - t_j) - \mathbf{b}_2^{\top}\mathbf{H}_i(t_j))^2\big) = 0. \quad (9)$$

The estimating equations (8) and (9) can also be derived directly based on the conditional mean and variance specification of $\log(T_i - t)$. Solving these equations is equivalent to solving a linear regression problem, and thus is computationally easy.

3.4. Some Remarks

We provide a few remarks. First, choosing suitable features into $\mathbf{H}_i(t)$ is important. The inclusion of suitable features not only improves the prediction accuracy, but also facilitates the exploratory analysis and interpretation of how behavioral patterns affect CPS result. If substantive knowledge about a CPS task is available from cognition theory, one may choose features that indicate different strategies toward solving the task. Otherwise, a data-driven approach may be taken. That is, one may select a model from a candidate list based on certain cross-validation criteria, where, if possible, all reasonable features should be consider as candidates. Even when a set of features has been suggested by cognition theory, one can still take the data-driven approach to find additional features, which may lead to new findings.

Second, one possible extension of the proposed model is to allow the regression coefficients to be a function of time t, whereas they are independent of time under the current model. In that case, the regression coefficients become functions of time, $b_{jk}(t)$. The current model can be regarded as a special case of this more general model. In particular, if $b_{jk}(t)$ has high variation along time in the best predictive model, then simply applying the current model may yield a high bias. Specifically, in the current estimation procedure, a larger grid point tends to have a smaller sample size and thus contributes less to the pseudo-likelihood function. As a result, a larger bias may occur in the prediction at a larger time point. However, the estimation of the time-dependent coefficient is non-trivial. In particular, constraints should be imposed on the functional form of $b_{jk}(t)$ to ensure a certain level of smoothness over time. As a result, $b_{jk}(t)$ can be accurately estimated using information from a finite number of time points. Otherwise, without any smoothness assumptions, to predict at any time during one's problem-solving process, there are an infinite number of parameters to estimate. Moreover, when a regression coefficient is time-dependent, its interpretation becomes more difficult, especially if the sign changes over time.

Third, we remark on the selection of grid points in the estimation procedure. Our model is specified in a continuous time domain that supports prediction at any time point in a continuum during an individual's problem-solving process. The use of discretized grid points is a way to approximate the continuous-time system, so that estimation equations can be written down. In practice, we suggest to place the grid points based on the quantiles of the empirical distribution of duration based on the training set. See the analysis in Section 4 for an illustration. The number of grid points may be further selected by cross validation. We also point out that prediction can be made at any time point on the continuum, not limited to the grid points for parameter estimation.

4. AN EXAMPLE FROM PISA 2012

4.1. Background

In what follows, we illustrate the proposed method via an application to the above CC item[2]. This item was also analyzed in Greiff et al. (2015) and Xu et al. (2018). The dataset was cleaned from the entire released dataset of PISA 2012. It contains 16,872 15-year-old students' problem-solving processes, where the students were from 42 countries and economies. Among these students, 54.5% answered correctly. On average, each student took 129.9 s and 17 actions solving the problem. Histograms of the students' problem-solving duration and number of actions are presented in **Figure 3**.

4.2. Analyses

The entire dataset was randomly split into training and testing sets, where the training set contains data from 13,498 students and the testing set contains data from 3,374 students. A predictive model was built solely based on the training set and then its performance was evaluated based on the testing set. We used $J = 9$ grid points for the parameter estimation, with t_1 through t_9 specified to be 64, 81, 94, 106, 118, 132, 149, 170, and 208 s, respectively, which are the 10% through 90% quantiles of the empirical distribution of duration. As discussed earlier, the number of grid points and their locations may be further engineered by cross validation.

4.2.1. Model Selection

We first build a model based on the training data, using a data-driven stepwise forward selection procedure. In each step,

[2]The log file data and code book for the CC item can be found online: http://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm.

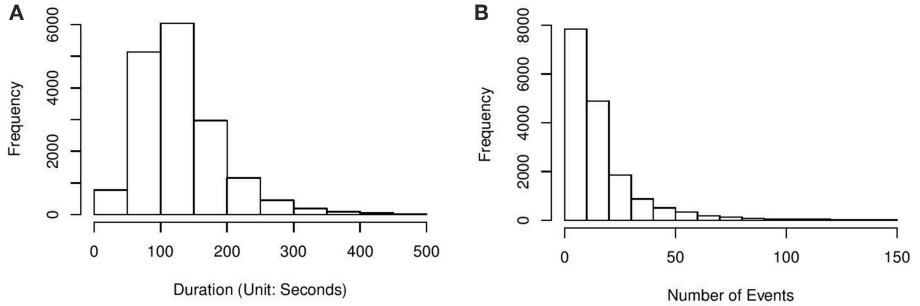


FIGURE 3 | (A) Histogram of problem-solving duration of the CC item. **(B)** Histogram of the number of actions for solving the CC item.

we add one feature into $\mathbf{H}_i(t)$ that leads to maximum increase in a cross-validated log-pseudo-likelihood, which is calculated based on a five-fold cross validation. We stop adding features into $\mathbf{H}_i(t)$ when the cross-validated log-pseudo-likelihood stops increasing. The order in which the features are added may serve as a measure of their contribution to predicting the CPS duration and final outcome.

The candidate features being considered for model selection are listed in **Table 2**. These candidate features were chosen to reflect students' CPS behavioral patterns from different aspects. In what follows, we discuss some of them. For example, the feature $I_i(t)$ indicates whether or not all three control sliders have been explored by simple actions (i.e., moving one control slider at a time) up to time t. That is, $I_i(t) = 1$ means that the vary-one-thing-at-a-time (VOTAT) strategy (Greiff et al., 2015) has been taken. According to the design of the CC item, the VOTAT strategy is expected to be a strong predictor of task success. In addition, the feature $N_i(t)/t$ records a student's average number of actions per unit time. It may serve as a measure of the student's speed of taking actions. In experimental psychology, response time or equivalently speed has been a central source for inferences about the organization and structure of cognitive processes (e.g., Luce, 1986), and in educational psychology, joint analysis of speed and accuracy of item response has also received much attention in recent years (e.g., van der Linden, 2007; Klein Entink et al., 2009). However, little is known about the role of speed in CPS tasks. The current analysis may provide some initial result on the relation between a student's speed and his/her CPS performance. Moreover, the features defined by the repeating of previously taken actions may reflect students' need of verifying the derived hypothesis on the relation based on the previous action or may be related to students' attention if the same actions are repeated many times. We also include $1, t, t^2$, and t^3 in $\mathbf{H}_i(t)$ as the initial set of features to capture the time effect. For simplicity, country information is not taken into account in the current analysis.

Our results on model selection are summarized in **Figure 4** and **Table 3**. The pseudo-likelihood stopped increasing after 11 steps, resulting a final model with 15 components in $\mathbf{H}_i(t)$. As we can see from **Figure 4**, the increase in the cross-validated log-pseudo-likelihood is mainly contributed by the inclusion of features in the first six steps, after which the increment is quite marginal. As we can see, the first, second, and sixth features entering into the model are all related to taking simple actions, a strategy known to be important to this task (e.g., Greiff et al., 2015). In particular, the first feature being selected is $I_i(t)$, which confirms the strong effect of the VOTAT strategy. In addition, the third and fourth features are both based on $N_i(t)$, the number of actions taken before time t. Roughly, the feature $1_{\{N_i(t)>0\}}$ reflects the initial planning behavior (Eichmann et al., 2019). Thus, this feature tends to measure students' speed of reading the instruction of the item. As discussed earlier, the feature $N_i(t)/t$ measures students' speed of taking actions. Finally, the fifth feature is related to the use of the RESET button.

TABLE 2 | The list of candidate features to be incorporated into the model.

	Feature	Explanation
1.	$N_i(t)$	Number of actions taken up to time t.
2.	$N_i(t)/t$	Frequency of actions up to time t.
3.	$1_{\{N_i(t)>0\}}$	Indicator of whether an action has been taken before time t.
4.	$S_i(t)$	Number of simple actions (i.e., moving one control slider at a time) taken up to time t.
5.	$S_i(t)/t$	Frequency of simple actions up to time t.
6.	$1_{\{S_i(t)>0\}}$	Indicator of whether a simple action has been taken before time t.
7.	$I_i(t)$	An indicator function, $I_i(t) = 1$ if all three control sliders have been explored via simple actions up to time t and $I_i(t) = 0$, otherwise.
8.	$R_i(t)$	Number of RESET used up to time t.
9.	$R_i(t)/t$	Frequency of RESET up to time t.
10.	$1_{\{R_i(t)>0\}}$	Indicator of whether RESET has been used before time t.
11.	$RP_i(t)$	Number of times that previously taken actions (excluding RESET)are repeated.
12.	$RP_i(t)/t$	Frequency of repeating previously taken actions (excluding RESET).
13.	$1_{\{RP_i(t)>0\}}$	Indicator of repeating previously taken actions (excluding RESET).

4.2.2. Prediction Performance on Testing Set

We now look at the prediction performance of the above model on the testing set. The prediction performance was evaluated

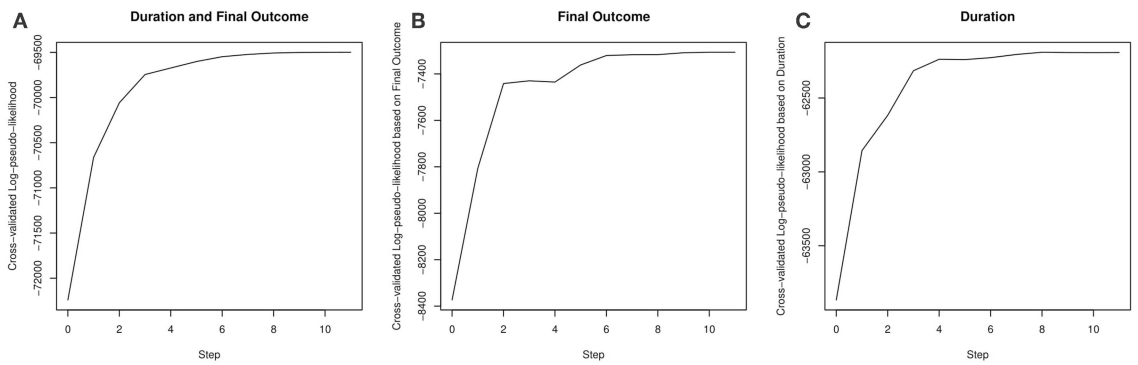


FIGURE 4 | The increase in the cross-validated log-pseudo-likelihood based on a stepwise forward selection procedure. **(A–C)** plot the cross-validated log-pseudo-likelihood, corresponding to $L(B, \sigma)$, $L_1(\mathbf{b}_1)$, $L_2(\mathbf{b}_2, \sigma)$, respectively.

TABLE 3 | Results on model selection based on a stepwise forward selection procedure.

Step	Var.add	Lik	Lik.out	Lik.dur
0.	$1, t, t^2, t^3$	–72241.7	–63867.9	–8373.7
1.	$I_i(t)$	–70663.0	–62856.1	–7806.9
2.	$1_{\{S_i(t)>0\}}$	–70058.3	–62617.0	–7441.4
3.	$1_{\{N_i(t)>0\}}$	–69744.9	–62315.2	–7429.7
4.	$N_i(t)/t$	–69672.7	–62237.6	–7435.1
5.	$1_{\{R_i(t)>0\}}$	–69601.3	–62239.9	–7361.4
6.	$S_i(t)/t$	-69547.6	–62226.8	–7320.8
7.	$RP_i(t)/t$	–69522.5	–62205.1	–7317.4
8.	$1_{\{RP_i(t)>0\}}$	–69507.0	-62190.0	–7317.0
9.	$R_i(t)$	–69500.8	–62191.9	–7308.9
10.	$N_i(t)$	–69499.4	–62192.6	–7306.8
11.	$RP_i(t)$	–69498.5	–62191.8	–7306.7

The columns "Lik," "Lik.out," and "Lik.dur" give the value of the cross-validated log-pseudo-likelihood, corresponding to $L(B, \sigma)$, $L_1(\mathbf{b}_1)$, $L_2(\mathbf{b}_2, \sigma)$, respectively.

at a larger set of time points from 19 to 281 s. Instead of reporting based on the pseudo-likelihood function, we adopted two measures that are more straightforward. Specifically, we measured the prediction of final outcome by the Area Under the Curve (AUC) of the predicted Receiver Operating Characteristic (ROC) curve. The value of AUC is between 0 and 1. A larger AUC value indicates better prediction of the binary final outcome, with AUC $= 1$ indicating perfect prediction. In addition, at each time point t, we measured the prediction of duration based on the root mean squared error (RMSE), defined as

$$\sqrt{\frac{\sum_{i=N+1}^{N+n} 1_{\{\tau_i > t\}} (\tau_i - \hat{\tau}_i(t))^2}{\sum_{i=N+1}^{N+n} 1_{\{\tau_i > t\}}}},$$

where τ_i, $i = N + 1, ..., N + n$, denotes the duration of students in the testing set, and $\hat{\tau}_i(t)$ denotes the prediction based on information up to time t according to the trained model.

Results are presented in **Figure 5**, where the testing AUC and RMSE for the final outcome and duration are presented. In particular, results based on the model selected by cross validation ($p = 15$) and the initial model ($p = 4$, containing the initial covariates 1, t, t^2, and t^3) are compared. First, based on the selected model, the AUC is never above 0.8 and the RMSE is between 53 and 64 s, indicating a low signal-to-noise ratio. Second, the students' event history does improve the prediction of final outcome and duration upon the initial model. Specifically, since the initial model does not take into account the event history, it predicts the students with duration longer than t to have the same success probability. Consequently, the test AUC is 0.5 at each value of t, which is always worse than the performance of the selected model. Moreover, the selected model always outperforms the initial model in terms of the prediction of duration. Third, the AUC for the prediction of the final outcome is low when t is small. It keeps increasing as time goes on and fluctuates around 0.72 after about 120 s.

4.2.3. Interpretation of Parameter Estimates

To gain more insights into how the event history affects the final outcome and duration, we further look at the results of parameter estimation. We focus on a model whose event history $\mathbf{H}_i(t)$ includes the initial features and the top six features selected by cross validation. This model has similar prediction accuracy as the selected model according to the cross-validation result in **Figure 4**, but contains less features in the event history and thus is easier to interpret. Moreover, the parameter estimates under this model are close to those under the cross-validation selected model, and the signs of the regression coefficients remain the same.

The estimated regression coefficients are presented in **Table 4**. First, the first selected feature $I_i(t)$, which indicates whether all three control sliders have been explored via simple actions, has a positive regression coefficient on final outcome and a negative coefficient on duration. It means that, controlling the rest of the parameters, a student who has taken the VOTAT strategy tends to be more likely to give a correct answer and to complete in a

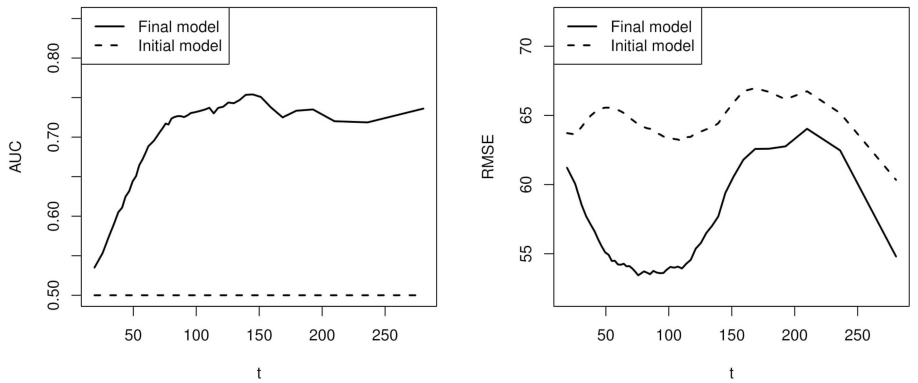


FIGURE 5 | A comparison of prediction accuracy between the model selected by cross validation and a baseline model without using individual specific event history.

TABLE 4 | Estimated regression coefficients for a model for which the event history process contains the initial features based on polynomials of t and the top six features selected by cross validation.

	Feature	$\hat{b}_1$	$\hat{b}_2$
1.	1	3.1×10^{-1}	4.8
2.	t	-5.9×10^{-3}	-2.7×10^{-3}
3.	t^2	3.1×10^{-6}	-4.5×10^{-7}
4.	t^3	1.7×10^{-8}	3.5×10^{-8}
5.	$I_i(t)$	5.2×10^{-1}	-8.4×10^{-1}
6.	$1_{\{S_i(t)>0\}}$	6.8×10^{-1}	-2.1×10^{-1}
7.	$1_{\{N_i(t)>0\}}$	-3.1×10^{-1}	-6.6×10^{-1}
8.	$N_i(t)/t$	-1.1	-1.4
9.	$1_{\{R_i(t)>0\}}$	3.7×10^{-1}	3.8×10^{-2}
10.	$S_i(t)/t$	3.0	7.9×10^{-1}

shorter period of time. This confirms the strong effect of VOTAT strategy in solving the current task.

Second, besides $I_i(t)$, there are two features related to taking simple actions, $1_{\{S_i(t)>0\}}$ and $S_i(t)/t$, which are the indicator of taking at least one simple action and the frequency of taking simple actions. Both features have positive regression coefficients on the final outcome, implying larger values of both features lead to a higher success rate. In addition, $1_{\{S_i(t)>0\}}$ has a negative coefficient on duration and $S_i(t)/t$ has a positive one. Under this estimated model, the overall simple action effect on duration is $\hat{b}_{25} I_i(t) + \hat{b}_{26} 1_{\{S_i(t)>0\}} + \hat{b}_{2,10} S_i(t)/t$, which is negative for most students. It implies that, overall, taking simple actions leads to a shorter predicted duration. However, once all three types of simple actions have been taken, a higher frequency of taking simple actions leads to a weaker but sill negative simple action effect on the duration.

Third, as discussed earlier, $1_{\{N_i(t)>0\}}$ tends to measure the student's speed of reading the instruction of the task and $N_i(t)/t$ can be regarded as a measure of students' speed of taking actions. According to the estimated regression coefficients, the data suggest that a student who reads and acts faster tends to complete the task in a shorter period of time with a lower accuracy. Similar results have been seen in the literature of response time analysis in educational psychology (e.g., Klein Entink et al., 2009; Fox and Marianti, 2016; Zhan et al., 2018), where speed of item response was found to negatively correlated with accuracy. In particular, Zhan et al. (2018) found a moderate negative correlation between students' general mathematics ability and speed under a psychometric model for PISA 2012 computer-based mathematics data.

Finally, $1_{\{R_i(t)>0\}}$, the use of the RESET button, has positive regression coefficients on both final outcome and duration. It implies that the use of RESET button leads to a higher predicted success probability and a longer duration time, given the other features controlled. The connection between the use of the RESET button and the underlying cognitive process of complex problem solving, if it exists, still remains to be investigated.

5. DISCUSSIONS

5.1. Summary

As an early step toward understanding individuals' complex problem-solving processes, we proposed an event history analysis method for the prediction of the duration and the final outcome of solving a complex problem based on process data. This approach is able to predict at any time t during an individual's problem-solving process, which may be useful in dynamic assessment/learning systems (e.g., in a game-based assessment system). An illustrative example is provided that is based on a CPS item from PISA 2012.

5.2. Inference, Prediction, and Interpretability

As articulated previously, this paper focuses on a prediction problem, rather than a statistical inference problem. Comparing with a prediction framework, statistical inference tends to draw stronger conclusions under stronger assumptions on the data generation mechanism. Unfortunately, due to the complexity of CPS process data, such assumptions are not only hardly satisfied, but also difficult to verify. On the other hand, a prediction framework requires less assumptions and thus is more

suitable for exploratory analysis. As a price, the findings from the predictive framework are preliminary and can only be used to generate hypotheses for future studies.

It may be useful to provide uncertainty measures for the prediction performance and for the parameter estimates, where the former indicates the replicability of the prediction performance and the later reflects the stability of the prediction model. In particular, patterns from a prediction model with low replicability and low stability should not be overly interpreted. Such uncertainty measures may be obtained from cross validation and bootstrapping (see Chapter 7, Friedman et al., 2001).

It is also worth distinguishing prediction methods based on a simple model like the one proposed above and those based on black-box machine learning algorithms (e.g., random forest). Decisions based on black-box algorithms can be very difficult to understood by human and thus do not provide us insights about the data, even though they may have a high prediction accuracy. On the other hand, a simple model can be regarded as a data dimension reduction tool that extracts interpretable information from data, which may facilitate our understanding of complex problem solving.

5.3. Extending the Current Model

The proposed model can be extended along multiple directions. First, as discussed earlier, we may extend the model by allowing the regression coefficients b_{jk} to be time-dependent. In that case, nonparametric estimation methods (e.g., splines) need to be developed for parameter estimation. In fact, the idea of time-varying coefficients has been intensively investigated in the event history analysis literature (e.g., Fan et al., 1997). This extension will be useful if the effects of the features in $\mathbf{H}_i(t)$ change substantially over time.

Second, when the dimension p of $\mathbf{H}_i(t)$ is high, better interpretability and higher prediction power may be achieved by using Lasso-type sparse estimators (see e.g., Chapter 3 Friedman et al., 2001). These estimators perform simultaneous feature selection and regularization in order to enhance the prediction accuracy and interpretability.

Finally, outliers are likely to occur in the data due to the abnormal behavioral patterns of a small proportion of people. A better treatment of outliers will lead to better prediction performance. Thus, a more robust objective function will be developed for parameter estimation, by borrowing ideas from the literature of robust statistics (see e.g., Huber and Ronchetti, 2009).

5.4. Multiple-Task Analysis

The current analysis focuses on analyzing data from a single task. To study individuals' CPS ability, it may be of more interest to analyze multiple CPS tasks simultaneously and to investigate how an individual's process data from one or multiple tasks predict his/her performance on the other tasks. Generally speaking, one's CPS ability may be better measured by the information in the process data that is generalizable across a representative set of CPS tasks than only his/her final outcomes on these tasks. In this sense, this cross-task prediction problem is closely related to the measurement of CPS ability. This problem is also worth future investigation.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This research was funded by NAEd/Spencer postdoctoral fellowship, NSF grant DMS-1712657, NSF grant SES-1826540, NSF grant IIS-1633360, and NIH grant R01GM047845.

REFERENCES

Allison, P. D. (2014). *Event history analysis: Regression for longitudinal event data*. London: Sage.

Danner, D., Hagemann, D., Schankin, A., Hager, M., and Funke, J. (2011). Beyond IQ: a latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence* 39, 323–334. doi: 10.1016/j.intell.2011.06.004

Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., and Naumann, J. (2019). The role of planning in complex problem solving. *Comput. Educ.* 128, 1–12. doi: 10.1016/j.compedu.2018.08.004

Fan, J., Gijbels, I., and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Anna. Statist.* 25, 1661–1690. doi: 10.1214/aos/1031594736

Fox, J.-P., and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivar. Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. New York, NY: Springer.

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., and Funke, J. (2012). Dynamic problem solving: a new assessment perspective. *Appl. Psychol. Measur.* 36, 189–213. doi: 10.1177/0146621612439620

Halpin, P. F., and De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika* 78, 793–814. doi: 10.1007/s11336-013-9329-1

Halpin, P. F., von Davier, A. A., Hao, J., and Liu, L. (2017). Measuring student engagement during collaboration. *J. Educ. Measur.* 54, 70–84. doi: 10.1111/jedm.12133

He, Q., and von Davier, M. (2015). "Identifying feature sequences from process data in problem-solving items with N-grams," in *Quantitative Psychology Research*, eds L. van der Ark, D. Bolt, W. Wang, J. Douglas, and M. Wiberg, (New York, NY: Springer), 173–190.

He, Q., and von Davier, M. (2016). "Analyzing process data from problem-solving items with n-grams: insights from a computer-based large-scale assessment," in *Handbook of Research on Technology Tools for Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf (Hershey, PA: IGI Global), 750–777.

Huber, P. J., and Ronchetti, E. (2009). *Robust Statistics*. Hoboken, NJ: John Wiley & Sons.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., and Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses

and response times. *Psychol. Methods* 14, 54–75. doi: 10.1037/a0014877

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press .

MacKay, D. G. (1982). The problems of flexibility, fluency, and speed–accuracy trade-off in skilled behavior. *Psychol. Rev.* 89, 483–506. doi: 10.1037/0033-295X.89.5.483

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

Vista, A., Care, E., and Awwal, N. (2017). Visualising and examining sequential actions as behavioural paths that can be interpreted as markers of complex behaviours. *Comput. Hum. Behav.* 76, 656–671. doi: 10.1016/j.chb.2017.01.027

Wüstenberg, S., Greiff, S., and Funke, J. (2012). Complex problem solving–More than reasoning? *Intelligence* 40, 1–14. doi: 10.1016/j.intell.2011.11.003

Xu, H., Fang, G., Chen, Y., Liu, J., and Ying, Z. (2018). Latent class analysis of recurrent events in problem-solving items. *Appl. Psychol. Measur.* 42, 478–498. doi: 10.1177/0146621617748325

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Zhan, P., Jiao, H., and Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *Br. J. Math. Statist. Psychol.* 71, 262–286. doi: 10.1111/bmsp.12114

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Chen, Li, Liu and Ying. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 May 2019
doi: 10.3389/fpsyg.2019.01122

Termination Rules for Variable-Length CD-CAT From the Information Theory Perspective

Lei Guo [1,2,3] *and Chanjin Zheng* [4,5]*

[1] Faculty of Psychology, Southwest University, Chongqing, China, [2] Southwest University Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality, Chongqing, China, [3] Chongqing Collaborative Innovation Center for Brain Science, Chongqing, China, [4] Department of Educational Psychology, Faculty of Education, East China Normal University, Shanghai, China, [5] Words Up Your Way, Beijing, China

Cognitive diagnostic computerized adaptive testing (CD-CAT) aims to take full advantage of both cognitive diagnosis (CD) and CAT. Cognitive diagnostic models (CDMs) attempt to classify students into several attribute profiles so as to evaluate their strengths and weaknesses while the CAT system selects items from the item pool to realize that goal as efficiently as possible. Most of the current research focuses on developing the item selection strategies and uses a fixed-length termination rule in CAT. Nevertheless, a variable-length termination rule is more appropriate than the fixed-length rule in order to bring out the full potential of CD-CAT. The current study discussed the inherent issue of instability over different numbers of attributes with the previous termination rules (the Tatsuoka rule and the two-criterion rule), proposed three termination rules from the information theory perspective, and revealed the connection between the previous methods and one of the information-based termination rules that will be discussed, further demonstrating the instability issue. Two simulation studies were implemented to evaluate the performance of these methods. Simulation results indicated that the SHE rule demonstrated strong stability across different numbers of attributes and different CDMs and should be recommended for application.

Keywords: computerized adaptive testing, cognitive diagnostic model, information theory, Shannon entropy, Kullback–Leibler distance, variable-length CD-CAT

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Mark D. Reckase,
Michigan State University, United States
Chester Chun Seng Kam,
University of Macau, China

***Correspondence:**
Chanjin Zheng
russelzheng@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *04 August 2018*
Accepted: *29 April 2019*
Published: *29 May 2019*

Citation:
Guo L and Zheng C (2019) Termination Rules for Variable-Length CD-CAT From the Information Theory Perspective. Front. Psychol. 10:1122. doi: 10.3389/fpsyg.2019.01122

INTRODUCTION

The goal of cognitive diagnosis is to obtain the students' status of mastering specific attributes measured by items in psychological and educational assessment. In recent decades, various cognitive diagnosis models (CDMs) have been developed to evaluate the attribute profiles or latent classes for each student, which designates whether each of the measured attributes or skills has been mastered (Tatsuoka, 1983; Mislevy et al., 2000; Junker and Sijtsma, 2001; Rupp et al., 2010).

One main application of CDM that has been published by many researches is in combination with computerized adaptive testing (CAT), which can be termed as cognitive diagnostic computerized adaptive testing (CD-CAT; Cheng, 2009; Huebner, 2010). The major benefit of CAT is that a tailored test can be generated for each individual *via* selecting items from the item pool according to their responses to previous items. Generally speaking, CAT will get the same precision of ability estimation as a traditional paper and pencil test by using fewer items. In other words, CAT can provide a high-efficient estimate for latent trait of interest (Weiss and Kingsbury, 1984). Thus, it is obvious that CD-CAT may have a performance comparable to Item Response Theory (IRT)-CAT.

To date, numerous studies have been done to examine the property of CD-CAT (Cheng, 2009; Wang et al., 2011, 2012; Wang, 2013; Kaplan and de la Torre, 2015; Zheng and Wang, 2017). However, most previous studies focused on proposing item selection strategies and almost used the fixed-length rule to stop the CD-CAT. It is possible to implement a needlessly long test to some students and an undesirably short test to others when the fixed-length termination rule is adopted. Consequently, it often leads to different measurement precision for different students. In practice, one may prefer that every student has nearly the same degree of estimate precision, which is a major strength of CAT over non-adaptive testing (Weiss and Kingsbury, 1984). The termination rule issue in CD-CAT has begun to attract some attention from researchers. Tatsuoka (2002) suggested that the CD-CAT stops when the examinee's posterior probability of a given attribute profile exceeded 0.80 (hereafter denoted as the Tatsuoka rule). Hsu et al. (2013) proposed a two-criterion termination rule by adding another criterion to the Tatsuoka rule. Cheng (2008) mentioned the possibility of proposing termination rules from the information theory perspective, but no theoretical explanation or empirical study was provided. The current study demonstrates the derivation of three termination rules from the information theory perspective and evaluates the termination rules using simulation studies.

In the following, first, the previous methods (i.e., the Tatsuoka rule and the two-criterion rule) for variable-length CD-CAT are summarized and their inherent issue of instability over different numbers of attributes will be discussed. Second, we introduce three information-based termination rules for CD-CAT. The connection between the previous methods and one of the information-based termination rules is shown, which further demonstrates the instability issue. Third, following this, two simulation studies are conducted to assess the performance of the new termination rules over different numbers of attributes and CDMs with regard to the instability issue. Finally, some important issues in variable-length termination rules will be discussed.

THE PREVIOUS RULES FOR VARIABLE-LENGTH CD-CAT AND THEIR ISSUES

To our knowledge, two termination rules for CD-CAT have been proposed, namely, the Tatsuoka rule and the two-criterion rule, respectively. Tatsuoka (2002) suggested that a CD-CAT stops when the examinee's posterior probability of a given attribute profile exceeded 0.80, i.e., the posterior probability of one latent class (PPLS) is bigger than 0.80. The principle is that the more peaked the posterior probability distribution is, the more dependable the classification is (Huebner, 2010). Inspired by the Tatsuoka rule, Hsu et al. (2013) recommended to add another criterion for the second largest PPLS. Thus, the modified termination rule for variable-length CD-CAT using the following two criteria were proposed:

Criterion 1: CD-CAT will be stopped when the largest PPLS is not smaller than a predetermined value (e.g., 0.70).

Criterion 2: CD-CAT will be stopped when the largest PPLS is not smaller than a predetermined value (e.g., 0.70) and the second largest PPLS is not larger than a predetermined value (e.g., 0.10).

The key of the two-criterion rule is to determine the threshold for the second largest PPSL. The following formula can be used to determine the lower bound and upper bound for the second largest PPSL.

$$P_{2nd} = \frac{1 - P_{1st}}{2^K - 1} + \frac{(2^K - 2) \times (1 - P_{1st}) \times d}{2^K - 1}, 0 \leq d \leq 1 \quad (1)$$

where $P_{1\text{st}}$ and $P_{2\text{nd}}$ are the prespecified largest and second largest PPSL, K represents the number of attributes, and d is the weighted value for $P_{2\text{nd}}$. Based on the simulation results, Hsu et al. (2013) offered two suggestions:

1. One can set the value of $P_{1\text{st}}$ as high as 0.90 or 0.95 if the high-stakes tests are implemented. Thus, only Criterion 1 will be needed and Criterion 2 is not necessary.
2. One can set the value of $P_{1\text{st}}$ at 0.70 or lower, and the d value can be set between 0.25 and 0.50, or simply set $P_{2\text{nd}} = 0.10$ if the low-stakes tests are implemented.

The Tatsuoka rule is intuitive and simple, but with an increase in the number of attributes, which leads to the exponential increase of the number of attribute profile, it discards more and more information contained in the other attribute profiles since it only cares about the one with the largest probability mass. It is a sensible conjecture that there is an unstable issue with the Tatsuoka rule, namely, the realized accuracy of the attribute profile estimate might not be consistent across different numbers of attributes under the same model by implementing the Tatsuoka rule. Hsu et al. (2013) recognized this fact and attempted to solve this issue by setting a lower and upper bound for the second largest PPSL. One of the factors that influence the determination of the second largest PPSL is the number of attributes. But the fine-tuning of the second largest PPSL is of *ad hoc* nature, which makes the implementation difficult. The practical recommendation for d or $P_{2\text{nd}}$ taking a value of 0.1 can ameliorate this problem, but it was made based on the simulation study for only the case of six attributes and it may bring the instability issue again.

The current study proposes some new termination rules from the information perspective and evaluates their performance for different numbers of attributes under two major CDMs. Statistically speaking, the development of termination rules for CD-CAT aims to identify some statistical tools that describe certain characteristics of the posterior distribution of cognitive profiles. Both Tatsuoka and Hsu and his associates used the point(s) in the distribution with the largest concentration of the probability mass and discarded the remaining, and thus their methods can be labeled as a partial information approach. It is also worth pointing out that Tatsuoka did not carry out any empirical simulation study on the termination rule he proposed and Hsu et al. (2013) did not explore the performance of the two-criterion rule for different numbers of attributes, although it is an important factor in Equation (1). Another more powerful tool that describes a distribution is information indexes, which

can capture the characteristics of the whole distribution and thus can be considered as a full information approach. The major advantage of the new methods is that they incorporate the information of several attributes easily and they are expected to provide a simple consistent termination rule without demanding delicate fine-tuning as the two-criterion rule requires.

INFORMATION THEORY FOR CDM AND INFORMATION-BASED TERMINATION RULES FOR CD-CAT

Information Theory for CDM

A brief introduction to information theory, which is heavily borrowed from Cover and Thomas (2012) and Chang et al. (2016), is given below. Since the models involved in cognitive diagnosis are discrete, only the discrete version of various information indexes is presented where possible.

Information was first introduced by Fisher (1925). An important development in the information theory, introduced by Shannon (1948), was that of entropy. Shannon entropy is used to describe the uncertainty in the distribution of a random variable. Specifically, its value becomes maximum when distribution is uniform and minimum when distribution is a single point mass. In cognitive diagnosis, we need to classify an examinee into a certain attribute profile, so the posterior distributions were expected to be a point mass. This means the smaller the Shannon entropy value is, the more accurate the classification is. Let $\boldsymbol{\alpha}_c = (\alpha_{c1}, \alpha_{c2}, ..., \alpha_{cK})$ $(c = 1, ..., 2^K)$ be the attribute profile of examinees, and there were 2^K attribute profiles totally. Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{2^K})$ represent the posterior probability vector, and the element π_c is corresponding probability for $\boldsymbol{\alpha}_c$. Note that $\pi_c > 0$ $(c = 1, \ldots, 2^K)$, and $\sum_{c=1}^{2^K} \pi_c = 1$. The Shannon entropy of $\boldsymbol{\pi}$ is expressed as follows:

$$H(\boldsymbol{\pi}) = \sum\nolimits_{c=1}^{2^K} \pi_c \log(1/\pi_c) \tag{2}$$

The notion of entropy was extended to relative entropy by Kullback and Leibler (1951) and thus it was also denoted as the Kullback–Leibler (KL) distance. The relative entropy $KL(p||q)$ measures the divergence between distributions p and q. Cover and Thomas (2012) gave the original expression for the KL distance, i.e., $KL(p||q) = E_p\left[\log \frac{p(x)}{q(x)}\right]$. KL distance is non-negative and equals zero if distributions p and q are identical and becomes large as the distributions diverge. In cognitive diagnosis, the KL distance between Y_{ij} conditioning on estimated attribute profile $f(Y_{ij}|\widehat{\boldsymbol{\alpha}})$ and the conditional distribution of Y_{ij} given another attribute profile $\boldsymbol{\alpha}_c$, i.e., $f(Y_{ij}|\boldsymbol{\alpha}_c)$, is expressed as follows:

$$KL_j(\widehat{\boldsymbol{\alpha}}||\boldsymbol{\alpha}_c) = \sum_{y=0}^{1} \log\left(\frac{P(Y_{ij} = y|\widehat{\boldsymbol{\alpha}})}{P(Y_{ij} = y|\boldsymbol{\alpha}_c)}\right) P(Y_{ij} = y|\widehat{\boldsymbol{\alpha}}) \tag{3}$$

where Y_{ij} is the response of 1 (correct) and 0 (incorrect) to item j for examinee i. The larger the KL index value is, the more accurate the classification is.

The distinct difference between Shannon entropy and KL distance is that Shannon entropy uses some absolute values to describe one distribution while the KL distance tries to capture the distance between two distributions; thus, we can develop information-based terminations rules from this absolute-vs.-relative perspective.

INFORMATION-BASED TERMINATION RULES FOR CD-CAT

Some work has been done to develop item selection algorithms for CD-CAT from the information perspective (Xu et al., 2003; Cheng, 2009). The derivation of the termination rules from the information perspective is straightforward and can be obtained by simply replacing the random variable by the posterior distribution of the attribute profiles. The information-based termination rules suggest that a test can stop when:

a) The Shannon entropy of the posterior distribution becomes reasonably small (denoted as the SHE rule):

$$H(g_t) < \varepsilon \tag{4}$$

where g_t is the corresponding posterior distribution when an examinee answers t items. $\varepsilon \in R^+$ is a very small positive number. The SHE rule is equivalent to verify that the uncertainty of the posterior distribution has been reduced to a prescribed absolute level (obviously, this falls into the category of the absolute approach). Most of the posterior mass density is more concentrated and a few points (attribute profiles in CDM) occupy majority of the probability in posterior distribution. Because we hope one attribute profile will take up most of the probability, the CD-CAT test stops when Equation (4) is satisfied.

b) The KL distance (relative entropy) between two adjacent posterior distributions becomes small enough (denoted as the KL-distance rule):

$$D(g_t||g_{t-1}) < \varepsilon \tag{5}$$

where g_{t-1} is the posterior distribution of attribute profiles after $(t-1)$ items have been administered. The rationale for the KL-distance rule is that if the posterior distribution change between responding t items and $(t-1)$ items is negligible, the final attribute profile will be confirmed. Thus, the CD-CAT test stops when Equation (5) is satisfied.

c) The change of Shannon entropy for the adjacent posterior distributions becomes reasonably small (denoted as the SHE-difference rule):

$$\left|H(g_t) - H(g_{t-1})\right| < \varepsilon \tag{6}$$

The SHE-difference rule and the KL-distance rule follow a similar line of thinking and both of them fall into the category of the relative approach. Both of them involve the comparison of the two adjacent posterior distributions with the test stopping when the difference between the posterior estimate and the immediate previous one is small enough to

reach a predetermined level; i.e., the posterior estimate for the true attribute profile cannot be significantly improved given the current estimate and item selection method.

In summary, the above three information-based rules introduced in this section can fall into two categories: an absolute approach and a relative approach. The SHE rule is an absolute approach while the other two are relative approaches.

THE CONNECTION AND DIFFERENCE BETWEEN THE PREVIOUS RULES AND THE SHE RULE

The Tatsuoka rule and the two-criterion rule can be re-expressed as the SHE rule. This reformulation can further demonstrate the issues with the previous methods discussed above. For the Tatsuoka rule, P_{1st} is required to be larger than 0.8. This is equivalent to the following:

1) The addend for P_{1st} in the SHE rule is required to be smaller than

$$0.8 * \log_e(\frac{1}{0.8}) = 0.179 \tag{7}$$

2) The remaining probabilities excluding P_{1st} satisfies the assumption that

$$\sum_{i=1}^{2^K-1} P_i = 1 - 0.8 = 0.2 \tag{8}$$

In other words, if the preset value for P_{1st} has been set at 0.8, the remaining 2^K-1 attribute profiles share the rest of probability. In the worst case, the 2^K-1 attribute profiles share 0.2 equally, which signifies that they are equally probable. Thus, the Shannon entropy value of this probability distribution equals $0.8*\log_e(\frac{1}{0.8})+(\frac{0.2}{2^K-1}*\log_e(\frac{2^K-1}{0.2}))*(2^K-1)$. In the best scenario, the second largest PPSL (P_{2nd}) takes all the probability mass 0.2, i.e., the remaining probabilities are all 0s; the Shannon entropy value of this probability distribution equals $0.8 * \log_e(\frac{1}{0.8}) + 0.2 * \log_e(\frac{1}{0.2}) = 0.5$.

Two important observations can be made. First, a certain termination criterion value of the Tatsuoka rule corresponds to an interval of the SHE rule, and the range only depends on the number of attributes. **Table 1** shows the various ranges and the lower (upper) bound in the SHE rule under different numbers of attributes when P_{1st} is set at 0.7 or 0.8 in the Tatsuoka rule.

As shown in **Table 1**, the lower bound is always a constant when P_{1st} is set at a fixed value. However, the upper bound and the range of Shannon entropy rely on the number of attributes. Specifically, with the increase of the attribute number, the upper bound and range become larger. Consequently, for one particular Tatsuoka rule criterion, the larger the range is, the more possible values the classification accuracy can take.

Second, there is considerable overlap for the interval for two neighboring Tatsuoka rule values. For example, the lower and upper bounds are 0.611 and 2.273, respectively, when $P_{1st} = 0.7$ ($K = 8$), and the values become 0.5 and 1.609 when $P_{1st} = 0.8$ ($K = 8$). The size of overlaps is 0.998 (= 1.609–0.611). The overlap implies that the finalized classification accuracy might be similar or reversed (a higher classification accuracy rate for a lower criterion) for two different Tatsuoka rule criteria, which is undesirable.

TABLE 1 | Correspondence between the Tatsuoka and the SHE rule under different numbers of attributes.

P_{1st}	Number of attributes	Shannon entropy		
		Lower bound	Upper bound	Range
0.7	4	0.611	1.423	0.812
	5	0.611	1.641	1.030
	6	0.611	1.854	1.243
	7	0.611	2.064	1.453
	8	0.611	2.273	1.662
0.8	4	0.500	1.042	0.542
	5	0.500	1.187	0.687
	6	0.500	1.329	0.829
	7	0.500	1.469	0.969
	8	0.500	1.609	1.109

It is clear that the Tatsuoka rule is not as refined as the SHE rule. The final realized classification accuracy of one particular criterion from the Tatsuoka rule may vary depending on how many attributes there are. This correspondence between the two methods further reveals the root cause of the instability issue with the Tatsuoka rule.

A similar reformulation can be done for the two-criterion rule. The two-criterion rule with $P_{1st} = 0.7$ and $P_{2nd} = 0.1$, in terms of the SHE rule, is equivalent to the following:

1) The addend for the P_{1st} in the SHE rule is smaller than

$$0.7 * \log_e(\frac{1}{0.7}) = 0.250 \tag{9}$$

2) The addend for the P_{2nd} and other addends in the SHE rule are smaller than

$$0.1 * \log_e(\frac{1}{0.1}) = 0.230 \tag{10}$$

3) The remaining probabilities excluding P_{1st} and P_{2nd} satisfy the assumption that.

$$\sum_{i=1}^{2^K-2} P_i = 1 - 0.7 - 0.1 = 0.2 \tag{11}$$

Following the same line of reasoning, correspondence between the two-criterion rule and the SHE rule under different numbers of attributes can also be derived. **Table 2** shows the various ranges and the lower (upper) bounds in the SHE rule under different numbers of attributes when P_{1st} is set at 0.7 or 0.8 and P_{2nd} is fixed at 0.1 in the two-criterion rule.

TABLE 2 | Correspondence between the two-criterion rule and the SHE rule under different numbers of attributes.

P_{1st}	P_{2nd}	Number of attributes	Shannon entropy		
			Lower bound	Upper bound	Range
0.7	0.1	4	0.802	1.343	0.541
		5	0.802	1.489	0.687
		6	0.802	1.630	0.828
		7	0.802	1.771	0.969
		8	0.802	1.910	1.108
0.8	0.1	4	0.639	0.910	0.271
		5	0.639	0.982	0.343
		6	0.639	1.053	0.414
		7	0.639	1.123	0.484
		8	0.639	1.193	0.554

TABLE 3 | The taxonomy for the termination rules.

	Partial information	Full information
Absolute approach	Tatsuoka rule Two-criterion rule	SHE rule
Relative approach	–	SHE-difference rule KL-distance rule

Similar observations can be made, although there is some reduction in the size of the corresponding SHE interval for the two-criterion rule and their overlap.

In summary, all the termination rules can be summarized in a new taxonomical framework as in **Table 3**. It provides a basis for better understanding and discussion of the advantages and disadvantages of all methods, and the following two simulation studies will be designed to evaluate the absolute-vs.-relative and partial-vs.-full information comparison, respectively.

SIMULATION STUDIES

The DINA and Fusion Model

Two commonly used CDMs are the fusion model (Hartz, 2002) and the Deterministic Input; Noisy And gate (DINA) model (Junker and Sijtsma, 2001). An essential component underlying CDMs is the **Q**-matrix (Tatsuoka, 1983). Assume a test contains J items and K attributes, the **Q**-matrix is usually defined as a $J \times K$ matrix. The element that is related to the kth attribute for the jth item can be written as q_{jk}. $q_{jk} = 1$ if item j measures the attribute k, and $q_{jk} = 0$ otherwise.

The DINA model assumes that only when the examinee has mastered all attributes required by the item can he respond correctly. In fact, two possible behaviors, namely, "slip" and "guess," may occur when examinees respond to the items. Slip represents that the examinee gives an incorrect response to the item even though (s)he has mastered all the required attributes of this item, and guess indicates that the examinee gives a correct response to the item even though (s)he has not mastered all the required attributes of this item. With these characteristics, the correct response probability to the jth item for the ith examinee is

$$P(Y_{ij} = 1|\boldsymbol{\alpha_i}) = (1 - s_j)^{\eta_{ij}} g_j{}^{1-\eta_{ij}} \quad (12)$$

where $\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK})$ is the attribute profile of examinee i. $\alpha_{ik} = 1$ if ith examinee possesses attribute k, and $\alpha_{ik} = 0$ otherwise. s_j and g_j are the slip parameter and guess parameter, respectively. $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}$ is a latent variable that represents the examinee i's ideal response to item j. Note that if examinee i has mastered all the required attributes of item j, $\eta_{ij} = 1$; otherwise, $\eta_{ij} = 0$.

To introduce the fusion model, two types of item parameters are needed to be defined first: a) the parameter π_j^* denotes the probability of correct response to item j if examinees have mastered all measured attributes, and b) the parameter r_{jk}^* denotes the penalty for not having mastered attribute k of item j. Thus, the correct response probability in the fusion model arrives as

$$P\left(Y_{ij} = 1|\boldsymbol{\alpha_i}, \pi_j{}^*, r_{jk}{}^*, c_j\right) = \pi_j{}^* \prod_{k=1}^{K} r_{jk}{}^{*(1-\alpha_{ik})q_{jk}} P_{c_j}(\theta_i) \quad (13)$$

where $P_{c_j}(\theta_i)$ is the Rasch model in which the item difficulty parameter is c_j and θ_i is the ability parameter for examinee i to explain the additional contribution except those specified attributes in the **Q**-matrix. Usually, $P_{c_j}(\theta_i)$ is set at 1 (Henson and Douglas, 2005; McGlohen and Chang, 2008; Wang et al., 2011). With this constraint, the fusion model becomes the Reduced Reparameterized Unified Model [R-RUM; (Hartz, 2002)]. This practice is adopted in this study.

Item Selection Method

Xu et al. (2003) introduce the KL distance into CD-CAT and use the KL index as an item selection strategy. In order for KL distance to be able to indicate item j's global discrimination power between $f(Y_{ij}|\widehat{\boldsymbol{\alpha}})$ and all possible attribute profiles, the KL index was proposed to describe the sum of KL distance between $f(Y_{ij}|\widehat{\boldsymbol{\alpha}})$ and all $f(Y_{ij}|\boldsymbol{\alpha}_c)$s:

$$KL_j(\widehat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^K} \left[\sum_{y=0}^{1} \log \left(\frac{P(Y_{ij} = y|\widehat{\boldsymbol{\alpha}})}{P(Y_{ij} = y|\boldsymbol{\alpha}_c)} \right) P(Y_{ij} = y|\widehat{\boldsymbol{\alpha}}) \right] \quad (14)$$

where K is the number of attributes, and there will be 2^K possible attribute profiles.

The item with the maximum KL value, given the attribute profile of $\widehat{\boldsymbol{\alpha}}$, will be administered from the item pool. Furthermore, to feature the different importance of different attribute profiles, the supplement in Equation (14) is weighted by the posterior probability, and this modification can be called PWKL information (Cheng, 2009). The selection criterion in PWKL information is expressed as follows:

$$PWKL_j(\widehat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^K} \left\{ \sum_{y=0}^{1} \left[\log \left(\frac{P(Y_{ij} = y|\widehat{\boldsymbol{\alpha}})}{P(Y_{ij} = y|\boldsymbol{\alpha}_c)} \right) P(Y_{ij} = y|\widehat{\boldsymbol{\alpha}}) \right] g(\boldsymbol{\alpha}_c|y_{t-1}) \right\} \quad (15)$$

where $g(\boldsymbol{\alpha}_c|y_{t-1}) = p(\boldsymbol{\alpha}_c)\prod_{j=1}^{t-1} P(Y_{ij}=1|\boldsymbol{\alpha}_c)^{y_{ij}}[1-P(Y_{ij}=1|\boldsymbol{\alpha}_c)]^{1-y_{ij}}$ is the posterior probability of $\boldsymbol{\alpha}_c$, $p(\boldsymbol{\alpha}_c)$ is the prior probability, and y_{t-1} is the response vector for examinee i on previous $t-1$ items.

Study 1: Absolute vs. Relative Approach

Design

The item pool consisted of 300 items and no maximum test length was imposed in order to investigate the performance of all methods without any constraints. Each attribute was set to be measured by 40% of all items and make sure that each item at least measured one attribute. For the DINA model, the s_j and g_j parameters were both generated from $U(0.05, 0.25)$. For the fusion model, the item parameters ${\pi^*}_j$ and ${r_{jk}}^*$ were generated from $U(0.75, 0.95)$ and $U(0.2, 0.95)$, respectively (Henson and Douglas, 2005). A total of 2,000 examinees were generated assuming that every examinee has 50% probability of mastering each attribute. That is, there were 64 equally distributed attribute profiles in the population if a test measured six attributes.

The major goal of this simulation was to evaluate the stability of absolute and relative approaches across different numbers of attributes and different CDMs. Three factors were manipulated in this study. First, there were two models used in the study: the DINA model and the fusion model. Second, the number of attributes varied from 4 to 8. Finally, three information-based termination rules were investigated. The Tatsuoka rule and the two-criterion rule as partial information absolute methods were also included as baselines. For the SHE-difference rule and KL-distance rule, there were five levels for ε : 0.01, 0.05, 0.1, 0.15, and 0.2. Levels for ε were set at 0.3, 0.6, 0.9, 1.2, 1.5, and 1.8 for the SHE rule. The termination criterion for the Tatsuoka rule P_{1st} was set as either 0.5, 0.6, 0.7, 0.8, or 0.9, while for the two-criterion method, the criterion for P_{2nd} to be set as 0.1 was added as well. Thus, there were $(5+5+5+5+6)\times 5\times 2 = 260$ conditions.

The major dependent variables were the same as in Hsu et al. (2013). Say: (a) classification accuracy of attribute profiles, pattern correct classification rate (PCCR), calculated as the percentage of examinees whose attribute profiles were estimated correctly. For the interpretation of the result, we care more about

TABLE 4 | Classification accuracy for attribute profile and test length using the Tatsuoka rule.

#Attribute	P1st	DINA					FM				
		M	SD	Max	Min	PCCR	M	SD	Max	Min	PCCR
4	0.9	7.1	2.1	20	3	0.941	12.7	5.2	42	6	0.926
	0.8	5.5	1.6	18	2	0.869	9.9	4.2	33	5	0.869
	0.7	4.5	1.1	14	2	0.780	7.3	2.7	23	4	0.779
	0.6	4.3	0.9	11	2	0.739	6.4	2.4	31	4	0.725
	0.5	4.1	0.7	5	2	0.752	5.7	2.0	15	3	0.665
5	0.9	9.3	2.8	27	4	0.934	17.6	7.3	95	7	0.922
	0.8	7.6	2.2	21	3	0.863	13.9	5.2	41	5	0.853
	0.7	6.8	2.0	17	3	0.799	11.1	4.1	41	5	0.788
	0.6	5.5	1.1	14	3	0.737	9.4	3.6	28	4	0.714
	0.5	5.2	1.0	14	3	0.698	7.5	2.6	28	3	0.664
6	0.9	12.9	12.3	300	5	0.935	23.1	15.3	300	9	0.923
	0.8	10.6	10.3	300	4	0.856	18.0	7.5	70	7	0.855
	0.7	8.7	2.7	24	4	0.773	15.5	6.5	59	6	0.777
	0.6	7.2	1.8	25	4	0.732	11.4	4.2	38	5	0.705
	0.5	6.4	1.0	15	4	0.679	11.4	4.5	37	4	0.657
7	0.9	19.4	15.0	300	6	0.925	31.9	25.0	300	9	0.931
	0.8	14.2	12.8	300	6	0.842	24.5	11.8	117	8	0.853
	0.7	9.9	2.2	21	6	0.757	18.4	5.4	57	10	0.765
	0.6	10.6	3.9	44	5	0.723	16.9	6.9	73	6	0.703
	0.5	8.0	1.7	21	4	0.672	14.6	6.0	56	6	0.649
8	0.9	33.8	17.3	300	6	0.926	45.3	40.3	300	10	0.926
	0.8	24.1	14.9	300	6	0.845	33.3	25.1	300	9	0.847
	0.7	13.9	4.9	43	5	0.767	25.7	13.7	131	7	0.764
	0.6	11.5	3.8	36	5	0.708	22.2	10.6	115	7	0.712
	0.5	11.3	3.9	34	5	0.672	20.4	10.2	100	6	0.632

#Attribute, the number of attributes; P1st, the largest PPLS.

TABLE 5 | Classification accuracy for attribute profile and test length using the two-criterion rule.

#Attribute	P1st	P2nd	DINA					FM				
			M	SD	Max	Min	PCCR	M	SD	Max	Min	PCCR
4	0.9	0.1	7.1	2.1	20	3	0.941	12.7	5.2	42	6	0.926
	0.8	0.1	5.5	1.7	18	2	0.867	9.9	4.2	48	5	0.878
	0.7	0.1	4.8	1.3	14	3	0.793	7.7	3.0	24	3	0.777
	0.6	0.1	4.2	0.7	7	2	0.761	6.7	2.7	26	4	0.760
	0.5	0.1	4.2	0.8	8	1	0.796	7.3	2.8	24	4	0.773
5	0.9	0.1	9.3	2.8	27	4	0.934	17.6	7.3	95	7	0.922
	0.8	0.1	7.6	2.2	21	3	0.871	13.8	5.4	51	5	0.842
	0.7	0.1	6.8	1.9	20	3	0.792	11.3	4.6	44	5	0.780
	0.6	0.1	6.0	1.6	17	3	0.765	9.1	3.7	35	4	0.721
	0.5	0.1	6.1	1.8	22	3	0.776	9.4	3.6	27	4	0.732
6	0.9	0.1	12.9	13.5	300	5	0.941	23.1	15.3	300	9	0.913
	0.8	0.1	10.5	7.3	300	4	0.832	18.3	8.0	108	8	0.856
	0.7	0.1	8.5	2.5	28	3	0.787	14.2	8.3	47	6	0.774
	0.6	0.1	7.7	2.3	22	3	0.759	13.6	5.9	96	5	0.735
	0.5	0.1	7.4	2.1	26	4	0.744	11.0	4.2	42	5	0.697
7	0.9	0.1	19.4	15.0	300	6	0.912	31.9	25.0	300	9	0.929
	0.8	0.1	14.2	12.8	300	5	0.850	25.5	17.8	300	8	0.837
	0.7	0.1	10.0	7.5	26	6	0.792	18.5	14.6	50	10	0.753
	0.6	0.1	10.3	3.7	37	5	0.730	17.2	9.5	105	6	0.709
	0.5	0.1	8.0	1.9	28	4	0.675	16.1	7.8	78	6	0.699
8	0.9	0.1	33.8	17.3	300	6	0.917	45.3	40.3	300	10	0.938
	0.8	0.1	25.6	14.9	300	6	0.824	34.0	29.8	300	10	0.853
	0.7	0.1	12.4	3.6	35	6	0.787	24.7	12.8	164	8	0.750
	0.6	0.1	12.4	4.1	38	6	0.706	23.6	12.5	177	7	0.732
	0.5	0.1	10.8	3.5	32	5	0.657	20.4	10.9	114	7	0.656

P1st, the largest PPLS; P2nd, the second largest PPLS.

the stability of PCCR for one particular termination criterion for different numbers of attributes and CDMS than PCCR itself; and (b) the test length at the end of the CD-CAT.

Results

Tables 4–**8** show the PCCR values and the concerned statistics, such as mean (M), standard deviation (SD), maximum (Max), and minimum (Min) of the test length at the end of the CD-CAT across all examinees. The results are summarized as the following.

In terms of the performance of the absolute and relative approaches across different numbers of attributes, the methods from the absolute approach maintained better stability than those from the relative approach across different numbers of attributes. **Table 6** showed that the classification accuracy for different numbers of attributes was approximately the same for both models except when the number of attributes was small (namely, four or five attributes). **Figure 1** is the visual representation of this result for the DINA model. **Tables 4**, **5** indicate that the Tatsuoka rule and the two-criterion rule presented a very similar trend in terms of stability of the classification accuracy across different numbers of attributes. In contrast, those in the relative approach were severely influenced by the number of attributes. **Tables 7**, **8** indicate that each termination criterion from either the SHE-difference rule or the KL-distance rule produced differential classification accuracy for different numbers of attributes under both models. For example, as shown in **Figure 2**, under the DINA model, the most conservative termination criterion, 0.01, from the KL-distance rule produced 5% difference in PCCR for four to eight attributes (from 0.998 to 0.945) while the most liberal termination criterion, 0.20, yielded even a more prominent gap of 20% for four to eight attributes (from 0.864 to 0.664). Similar results can be readily identified in the fusion model.

In the aspect of cross-model stability, the differential performance between the absolute and relative approach was even more striking. Take the SHE rule and the KL-distance rule as an example. **Figures 3**, **4** show the classification accuracy for the SHE rule and KL-distance rule under the DINA and fusion models for eight attributes. The SHE rule produced similar classification accuracy for both models under all the different termination criteria while the KL-distance rule yielded drastically different classification accuracy for the two models.

TABLE 6 | Classification accuracy for attribute profile and test length using the SHE rule.

#Attribute	ε	DINA					FM				
		M	SD	Max	Min	PCCR	M	SD	Max	Min	PCCR
4	0.3	8.4	2.7	21	4	0.977	14.8	5.7	63	7	0.974
	0.6	6.4	1.8	16	3	0.899	11.0	4.1	30	5	0.901
	0.9	5.7	1.7	19	3	0.856	9.8	3.4	38	5	0.848
	1.2	4.4	0.9	12	3	0.790	6.8	2.4	19	4	0.782
	1.5	3.9	0.9	10	2	0.734	6.5	2.4	21	3	0.722
	1.8	2.9	0.6	5	2	0.379	4.7	1.6	15	3	0.550
5	0.3	10.6	3.2	28	5	0.976	18.9	7.3	69	8	0.963
	0.6	8.8	2.8	25	4	0.900	15.3	6.4	57	6	0.911
	0.9	7.5	2.4	25	3	0.858	13.8	5.2	44	6	0.842
	1.2	7.1	1.9	21	4	0.793	11.0	4.3	36	5	0.782
	1.5	6.5	1.7	17	3	0.743	9.5	3.6	32	4	0.723
	1.8	4.6	0.8	9	3	0.546	7.8	2.6	26	4	0.644
6	0.3	12.8	7.4	300	6	0.973	17.7	15.7	300	10	0.959
	0.6	11.4	3.3	26	5	0.915	23.1	10.9	136	8	0.908
	0.9	10.5	3.1	25	5	0.859	18.8	7.5	74	6	0.849
	1.2	8.6	2.6	30	4	0.792	15.2	5.7	60	6	0.775
	1.5	8.0	2.6	24	3	0.743	13.3	5.0	43	6	0.727
	1.8	6.8	1.9	20	3	0.671	13.5	5.7	62	5	0.685
7	0.3	15.7	3.6	37	8	0.971	30.0	13.8	300	16	0.961
	0.6	14.1	3.4	38	7	0.901	24.2	11.5	300	13	0.900
	0.9	11.8	2.7	28	7	0.857	21.9	6.2	51	11	0.856
	1.2	10.0	2.0	25	6	0.789	19.7	6.1	61	10	0.774
	1.5	9.0	1.9	25	6	0.743	16.4	4.8	47	9	0.727
	1.8	9.7	3.2	33	4	0.680	16.5	6.6	55	6	0.683
8	0.3	20.0	6.7	58	7	0.974	46.4	24.6	194	11	0.967
	0.6	16.9	5.7	54	6	0.907	40.7	24.3	205	9	0.915
	0.9	15.9	6.1	51	6	0.858	32.4	18.5	231	10	0.860
	1.2	14.7	5.2	50	6	0.792	29.5	17.8	240	8	0.776
	1.5	13.3	5.0	52	5	0.742	25.3	11.8	120	8	0.730
	1.8	12.6	4.8	50	6	0.682	22.6	10.6	122	7	0.690

ε is the symbol in Equations (4)–(6).

In summary, the absolute approach, two previous methods, and the SHE rule did a much better job than the relative approach in terms of stability across differing numbers of attributes and different CDMs.

Simulation study 1 also provided some preliminary result for the partial-vs.-full information comparison. Within the absolute approach, the full information approach (the SHE rule) was slightly more consistent with respect to the classification accuracy than the partial information approaches (the Tatsuoka rule and the two-criterion rule). More interestingly, there are reversed classification accuracies for both the Tatsuoka rule and the two-criterion rule. For example, **Table 4** shows that for the DINA model with four attributes, the classification rate for the termination criterion 0.6 is 0.739, which is smaller than 0.752, the one for the criterion 0.5. The two-criterion rule suffered from this problem for both models with four and five attributes as shown in **Table 5**.

To further reveal the differential performance between the partial-vs.-full information approaches, Study 2 attempted to explore this issue under a more realistic application setting with a larger number of attributes.

Study 2: Full vs. Partial Information

Design

Study 2 aimed to investigate the performance of the absolute full information approach (the SHE rule) and the absolute partial information approach (the two-criterion rule) for a large number of attributes. As shown from the results of the study 1, classification accuracies were certainly high when termination criteria were set at stringent levels; we do not expect too much

TABLE 7 | Classification accuracy for attribute profile and test length using the SHE-difference rule.

#Attribute	ε	DINA					FM				
		M	SD	Max	Min	PCCR	M	SD	Max	Min	PCCR
4	0.01	14.0	3.9	39	7	0.998	24.4	8.2	74	6	0.972
	0.05	10.7	3.1	34	5	0.987	15.8	5.7	39	4	0.896
	0.10	9.8	2.6	26	4	0.967	9.6	4.6	30	3	0.749
	0.15	9.4	2.8	22	2	0.946	6.3	3.3	25	3	0.630
	0.20	7.9	2.2	21	2	0.927	6.5	2.6	17	3	0.672
5	0.01	17.8	5.0	45	5	0.993	28.2	10.7	106	4	0.942
	0.05	12.7	3.6	29	5	0.975	16.3	6.7	49	4	0.794
	0.10	11.6	3.2	29	5	0.926	12.1	4.9	37	2	0.712
	0.15	10.5	3.3	25	3	0.877	6.3	3.8	24	2	0.536
	0.20	8.4	2.1	19	3	0.857	5.6	3.3	19	2	0.503
6	0.01	20.5	6.0	56	7	0.985	31.7	13.3	113	4	0.903
	0.05	14.1	4.7	36	4	0.906	18.1	8.0	58	2	0.746
	0.10	13.4	4.6	32	2	0.864	11.2	5.6	36	2	0.599
	0.15	12.3	3.5	28	4	0.859	6.0	3.0	24	2	0.433
	0.20	10.1	3.5	25	2	0.766	5.3	3.7	27	2	0.353
7	0.01	23.8	5.2	60	9	0.992	40.1	13.4	103	7	0.914
	0.05	17.7	4.2	42	7	0.956	22.1	8.9	54	6	0.744
	0.10	16.1	3.6	38	8	0.912	16.2	6.3	39	5	0.599
	0.15	14.8	3.3	36	7	0.903	8.5	4.9	28	4	0.403
	0.20	12.6	3.3	25	4	0.826	3.9	2.2	21	2	0.211
8	0.01	26.9	9.1	72	5	0.931	37.8	20.2	158	2	0.793
	0.05	16.9	6.5	45	3	0.787	17.2	9.8	76	2	0.539
	0.10	11.1	5.9	36	2	0.542	10.3	5.9	43	2	0.404
	0.15	9.7	4.8	31	3	0.485	5.4	2.8	26	2	0.218
	0.20	8.2	3.9	24	2	0.433	4.1	2.0	21	2	0.229

difference among these methods which echoes the first practical recommendation for P_{1st} and P_{2nd} from Hsu et al. (2013). In order to better investigate the performance of full and partial information approaches with respect to stability, more liberal termination criteria should be adopted. The termination criteria for the SHE rule were changed to 1.6, 1.8, and 2.0. The three termination criteria for the two-criterion rule were 0.6, 0.7, and 0.8. In addition, from a practical perspective, the number of attributes can be as many as 14 (McGlohen and Chang, 2008; Jang, 2009; Roman, 2009), and the classification accuracy of the attribute profile is not necessarily as high as 0.8 or even 0.9 since formative assessment is usually low-stakes. Hence, the number of attributes was set to be either 8, 10, or 12. Thus, the performance of the termination rules under these conditions carries a practical implication. Since there are more attributes, a larger item pool is needed for the simulation. The item pool used in study 2 was generated in the same way as in study 1 except that it consisted of 1,000 items instead of 300 items. Due to the large number of attributes, it might take a lot of items for some examinees to finish the test, so the maximum number of items an examinee can take in a CD-CAT test was set to 100, which was 10% of the total number of items.

The basic setup for study 2 was similar to that for study 1. There were three factors in this simulation study: CDMs, number of attributes, and termination rules. The major dependent variables were the same as in study 1. The ratio of examinees who reached the maximum test length as a confounding variable was also reported.

Results

Tables 9, **10** summarize the results for the simulation. The 12-attribute condition showed that the proportions of examinees reaching the maximum test length in the two-criterion rule were higher than those in the SHE rule under both models in the corresponding conditions. Beyond that, most results indicated that the proportions of examinees attaining the maximum test length were small under a variety of conditions in the study, so this confounding variable was well-controlled. The effect of the proportion of examinees using the maximum length stopping rule will be discussed in detail in the Discussion section. The eight-attribute condition can be considered as a partial replication study of study 1 since the only difference is the bank size, which increased from 300 to 1,000. The results for this

TABLE 8 | Classification accuracy for attribute profile and test length using the KL-distance rule.

#Attribute	ε	DINA					FM				
		M	SD	Max	Min	PCCR	M	SD	Max	Min	PCCR
4	0.01	10.3	2.8	26	5	0.998	15.6	5.4	50	8	0.978
	0.05	8.7	2.6	28	4	0.962	11.1	4.4	37	5	0.890
	0.10	6.7	2.0	20	4	0.913	8.8	3.2	28	4	0.790
	0.15	6.2	2.0	23	4	0.886	4.5	2.2	15	2	0.528
	0.20	6.1	1.7	17	4	0.864	3.9	1.4	10	2	0.516
5	0.01	13.3	3.9	36	6	0.988	23.1	8.4	65	8	0.966
	0.05	10.5	3.3	30	4	0.939	14.4	5.0	46	6	0.896
	0.10	8.7	2.7	27	4	0.884	10.3	3.6	29	3	0.738
	0.15	7.6	2.2	24	3	0.843	7.2	2.1	17	3	0.642
	0.20	7.5	2.1	19	4	0.809	4.2	1.5	12	2	0.435
6	0.01	15.8	4.5	56	6	0.985	26.5	10.0	92	10	0.953
	0.05	12.5	3.4	34	3	0.934	17.9	6.1	42	7	0.851
	0.10	9.6	2.6	25	3	0.830	11.9	4.1	36	3	0.678
	0.15	9.3	2.4	25	3	0.800	5.6	3.2	19	2	0.401
	0.20	8.5	2.0	24	2	0.768	5.6	1.8	14	2	0.386
7	0.01	17.2	3.9	36	10	0.976	29.7	8.2	86	16	0.931
	0.05	14.0	3.1	41	8	0.905	18.4	5.3	60	6	0.740
	0.10	11.3	2.7	29	7	0.809	9.4	3.4	27	5	0.466
	0.15	10.0	2.3	25	6	0.772	7.0	2.3	19	4	0.408
	0.20	9.8	2.2	23	6	0.743	3.4	2.1	14	2	0.170
8	0.01	21.4	6.1	51	10	0.945	39.5	13.0	92	13	0.912
	0.05	16.0	4.5	45	8	0.887	22.0	6.6	55	5	0.710
	0.10	12.6	3.6	33	4	0.747	14.4	4.1	40	2	0.574
	0.15	12.1	3.2	32	3	0.688	6.8	3.3	22	2	0.289
	0.20	11.1	2.5	25	2	0.664	3.8	2.4	13	2	0.152

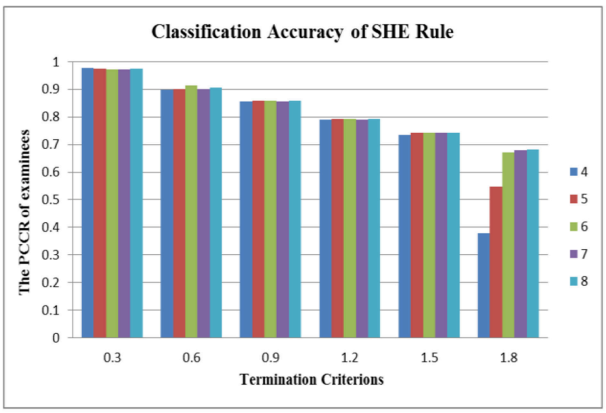


FIGURE 1 | Stability of the SHE rule across different numbers of attributes in the DINA model.

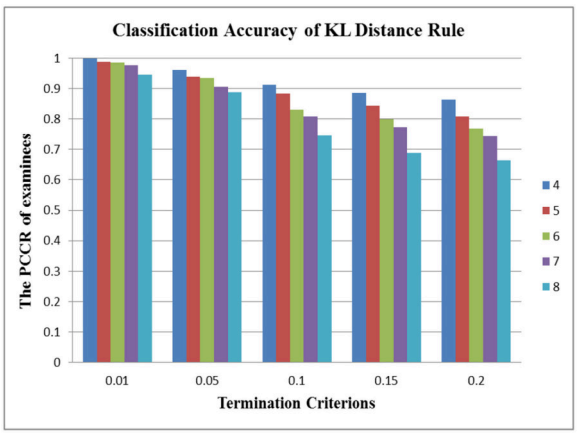


FIGURE 2 | Stability of the KL-distance rule across different numbers of attributes in the DINA model.

condition were very similar to those from study 1 and thus the possible confounding bank effect was also eliminated from study 2.

This simulation produced similar results for the two rules under the large number of attributes to study 1. The SHE rule demonstrated strong stability across both the number of

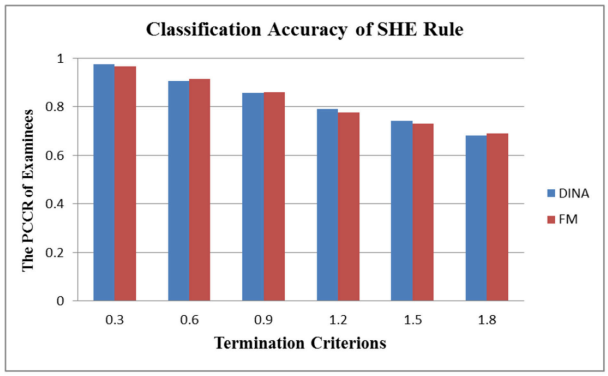


FIGURE 3 | Stability of the SHE rule across different models for eight attributes.

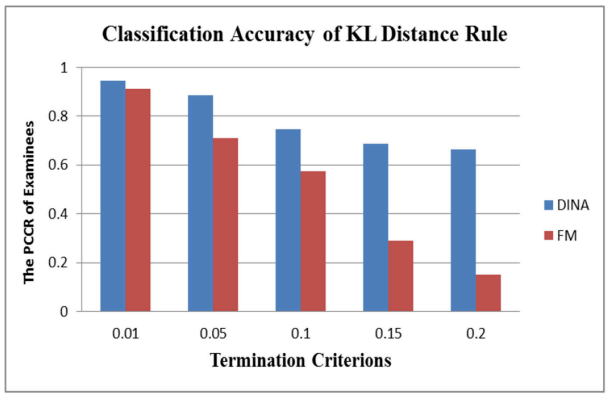


FIGURE 4 | Stability of the KL-distance rule across different models for eight attributes.

attributes and the CDMs while the two-criterion rule had some irregularity for some conditions. The classification accuracy for the three numbers of attributes was almost equal to 0.73, 0.70, and 0.65, respectively, for three termination criteria (1.6, 1.8, and 2.0). However, some termination criteria from the two-criterion rule yielded different classification accuracies for different numbers of attributes. For example, for the termination criterion $P_{1st} = 0.6$, the classification accuracy under the DINA model was 0.747, 0.706, and 0.677, respectively, for three different numbers of attributes (8, 10, and 12). Similar results can be easily identified for the fusion model.

In terms of cross-model constancy, the SHE rule also presented strong stability of classification accuracy. The two-criterion rule improved, but there were also inconsistencies of classification accuracy between the DINA model and the fusion model. The biggest difference was equal to 0.061 (= 0.738–0.677), which appeared on the condition of $P_{1st} = 0.6$ and 12 attributes.

DISCUSSION

Cognitive diagnostic assessment (CDA) informs an examiner about the attribute mastery pattern of every student so that designing effective remedial interventions in formative instruction can be administered (Leighton and Gierl, 2007a; Cui et al., 2012). CD-CAT as the computerized adaptive version of the CDA needs a flexible termination rule that can stop the test at an appropriate level to achieve that goal. This study provided a theoretical derivation of information-based termination rules proposed by Cheng (2008) and demonstrated the instability issue with previous methods from the information theory perspective. Two multi-factor simulation studies were conducted to evaluate the new three termination rules.

Some important observations can be made. The first point worth noting is that not all the full information methods outperform the previous methods, and the absolute full information method, the SHE rule, is the best with regard to the cross-attribute and cross-model stability. From the two simulation studies, we identified the termination criteria for the termination rule ranging from 0.3 to 1.8, which could produce a smooth decreasing trend of the estimate accuracy from about 0.97–0.6. The classification accuracy was not affected by the number of attributes (if it is more than five) or by the models. This implies that the SHE rule is a very flexible and effective method to stop the variable-length CD-CAT.

Then, there are some common problems shared by the Tatsuoka rule and the two-criterion rule. First, they are affected by the number of attributes, although their between-model performances are decent. Some careful consideration must be given with regard to the number of attributes for the item pool. Second, if some liberal criteria are used, such as P_{1st} is 0.6, 0.7, or 0.8 for large numbers of attributes, the problem of instability across different numbers of attributes is exacerbated. This reflects the inherent problem with the partial information rules. In CDA, the number of attribute patterns increases exponentially with the number of attributes. For a large number of attributes, the partial information rules do not have an effective control and thus there is a wide range of classification accuracy for differing numbers of attributes, although they, as members of the absolute approach, can guarantee a lower bound of the classification accuracy as the SHE rule does.

Lastly, the use of the maximum test length rule, in combination with the variable-length termination rule, and the proportion of examinees using this rule are important in the variable-length CD-CAT application. As noted above, the number of attribute profiles increases exponentially with that of attributes. When the number of attributes is large, the number of attribute profiles is so huge that it will take a lot of items—in some instances, even the entire item pool—for some examinees to satisfy the requirement prescribed by the termination rule. Thus, it is necessary to set the maximum test length even if a variable-length termination rule is adopted, and this treatment is often imposed in real CAT programs. It is also necessary to monitor the proportion of examinees hitting the maximum test length. If that proportion is high, then there might be some problems that merit further investigation, such as the criterion for the variable-length termination rule being too conservative, or there not being

TABLE 9 | Summary statistics for the SHE rule.

#Attribute	ε	DINA						FM					
		M	SD	Max	Min	%(ML)	PCCR(P)	M	SD	Max	Min	%(ML)	PCCR(P)
8	1.6	12.4	3.2	32	5	0	0.732	18.6	6.2	53	7	0	0.737
	1.8	11.1	2.8	32	4	0	0.682	17.0	6.1	59	7	0	0.708
	2.0	10.2	3.1	30	4	0	0.640	16.4	5.9	52	6	0	0.654
10	1.6	16.9	7.6	100	7	0.3	0.733	28.2	11.2	100	10	0.05	0.739
	1.8	15.3	6.1	100	6	0.15	0.692	27.7	11.5	95	9	0	0.710
	2.0	15.3	6.3	100	6	0.15	0.642	26.5	10.0	88	8	0	0.652
12	1.6	27.2	20.2	100	8	6	0.733	42.3	18.1	100	13	2	0.733
	1.8	26.2	20.9	100	8	6.4	0.691	37.3	16.7	100	12	1	0.709
	2.0	23.6	17.6	100	9	4	0.645	39.6	17.6	100	11	1	0.653

%(ML) = %(Max Length), the ratio of examinees attaining the maximum test length; PCCR(P), the pattern correct classification rate for examinees who finished the CD-CAT using termination rules.

TABLE 10 | Summary statistics for the two-criterion rule.

#Attribute	P_{1st}	P_{2nd}	DINA						FM					
			M	SD	Max	Min	%(ML)	PCCR(P)	M	SD	Max	Min	%(ML)	PCCR(P)
8	0.8	0.1	13.0	3.8	34	6	0	0.829	23.5	8.0	67	9	0	0.843
	0.7	0.1	11.9	3.5	33	6	0	0.801	20.3	7.1	64	7	0	0.787
	0.6	0.1	10.3	2.6	25	4	0	0.747	17.6	6.3	54	7	0	0.709
10	0.8	0.1	21.3	12.7	100	8	2	0.849	33.7	13.3	100	12	0.2	0.843
	0.7	0.1	18.6	12.4	100	8	1.8	0.768	30.1	12.8	100	10	0.1	0.779
	0.6	0.1	17.8	13.4	100	7	2.2	0.706	27.3	11.9	91	10	0	0.683
12	0.8	0.1	31.0	23.6	100	11	9.7	0.858	48.3	20.6	100	13	4.9	0.867
	0.7	0.1	30.7	25.5	100	11	11.2	0.750	43.5	19.1	100	12	3.2	0.784
	0.6	0.1	26.3	23.1	100	9	8.4	0.677	40.3	18.2	100	11	1.8	0.738

enough high-quality items in the pool. One possible solution to this issue is to make use of the attribute hierarchical structure (Leighton and Gierl, 2007b) to cut down the number of possible attribute profiles and then construct an informative prior for the distribution of the attribute profile.

Several issues require further investigation. In real testing situations, different CDMs, different item selection algorithms, item exposure control methods, content and attribute balancing, and item pool quality are all possible elements that could affect the performance of all the rules; more simulation studies are needed to investigate these situations. In the current study, we only used the DINA and fusion model as examples and the result with regard to the cross-model stability should be interpreted with caution. We carefully chose the two models that are the two ends of the spectrum of the existing CDMs and a similar conclusion regarding the SHE rule is expected for other models, but further study in this aspect is still warranted. Although two simulation studies were conducted, some real-life data studies are also necessary to investigate the performance of these termination rules in real situations. In addition, as one anonymous reviewer pointed out, the simulated examinees would answer an average of only 26% of the items correctly using the procedure described in the first simulation study based on the DINA model with eight attributes. There could be some reasons for this, such as the quality of item pool, the Q-matrix of test, distribution of attribute profiles in the population, and CDMs. An interesting study in the future is to investigate how the generation procedures of examinees' attribute profiles affect the classification accuracy and responses.

AUTHOR CONTRIBUTIONS

LG proposed the idea of this article and wrote all the simulation study code. CZ is the corresponding author who mainly organized and wrote the article.

FUNDING

This research was funded by the Ministry of education of humanities and social science research youth fund (15YJC190003), the Fundamental Research Funds for the Central Universities (SWU1809106), and the Peak Discipline Construction Project of Education at East China Normal University.

REFERENCES

Chang, H. H., Wang, C., and Ying, Z. (2016). "Information theory and its application to testing," in *Handbook of Item Response Theory, Models, Statistical Tools, and Applications*, eds W. J. van der Linden and R. K. Hambleton (Boca Raton, FL: Chapman & Hall/CRC Press), 105–118.

Cheng, Y. (2008). *Computerized Adaptive Testing: New Developments and Applications*. Unpublished Doctoral Dissertation. Champaign, IL: University of Illinois at Urbana-Champaign.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

Cover, T. M., and Thomas, J. A. (2012). *Elements of Information Theory*. Hoboken, NJ: John Wiley & Sons.

Cui, Y., Gierl, M. J., and Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Edu. Meas.* 49, 19–38. doi: 10.1111/j.1745-3984.2011.00158.x

Fisher, R. A. (1925). "Theory of statistical estimation," in *Paper Presented at the Mathematical Proceedings of the Cambridge Philosophical Society* (Cambridge).

Hartz, S. M. C. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*. Unpublished Doctoral Dissertation. Champaign, IL: University of Illinois at Urbana-Champaign.

Henson, R., and Douglas, J. (2005). Test construction for cognitive diagnosis. *Appl. Psychol. Meas.* 29, 262–277. doi: 10.1177/0146621604272623

Hsu, C. L., Wang, W. H., and Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Appl. Psychol. Meas.* 37, 563–582. doi: 10.1177/0146621613488642

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Pract. Assess. Res. Eval.* 15, 1–7. Available online at: https://www.learntechlib.org/p/75591/

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for Fusion Model application to LanguEdge assessment. *Lang. Test.* 26, 31–73. doi: 10.1177/0265532208097336

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kaplan, M., and de la Torre, J. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Appl. Psychol. Meas.* 39, 167–188. doi: 10.1177/0146621614554650

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86. doi: 10.1214/aoms/1177729694

Leighton, J. P., and Gierl, M. J. (2007a). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Edu. Meas. Issues Pract.* 26, 3–16. doi: 10.1111/j.1745-3992.2007.00090.x

Leighton, J. P., and Gierl, M. J. (2007b). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

McGlohen, M., and Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behav. Res. Methods* 40, 808–821. doi: 10.3758/BRM.40.3.808

Mislevy, R. J., Almond, R. G., Yan, D., and Steinberg, L. S. (2000). *Bayes Nets in Educational Assessment: Where do the Numbers Come from*. CRESST/Educational Testing Service: CSE Technical Reports 518.

Roman, S. (2009). *Fitting Cognitive Diagnostic Assessment to the Concept Assessment Tool for Statics (CATS)*. PhD Unpublished Doctoral Dissertation. Lafayette, IN: Purdue University.

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: The Guilford Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27:379. doi: 10.1002/j.1538-7305.1948.tb01338.x

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *J. R. Stat. Soc. Ser. C* 51, 337–350. doi: 10.1111/1467-9876.00272

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Edu. Meas.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educ Psychol. Meas.* 73, 1017–1035. doi: 10.1177/0013164413498256

Wang, C., Chang, H. H., and Douglas, J. (2012). Combining CAT with cognitive diagnosis: a weighted item selection approach. *Behav. Res. Methods* 44, 95–109. doi: 10.3758/s13428-011-0143-3

Wang, C., Chang, H. H., and Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *J. Edu. Meas.* 48, 255–273. doi: 10.1111/j.1745-3984.2011.00145.x

Weiss, D. J., and Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *J. Edu. Meas.* 21, 361–375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Xu, X., Chang, H., and Douglas, J. (2003). "A simulation study to compare CAT strategies for cognitive diagnosis," in *Paper Presented at the the Annual Meeting of National Council on Measurement in Education* (Chicago, IL).

Zheng, C., and Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Appl. Psychol. Meas.* 41, 561–576. doi: 10.1177/0146621617707509

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Guo and Zheng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

METHODS
published: 06 November 2018
doi: 10.3389/fpsyg.2018.02142

Integrating Differential Evolution Optimization to Cognitive Diagnostic Model Estimation

Zhehan Jiang[1]* *and Wenchao Ma*[2]

[1] Department of University Libraries, University of Alabama, Tuscaloosa, AL, United States, [2] Department of Educational Studies, University of Alabama, Tuscaloosa, AL, United States

A log-linear cognitive diagnostic model (LCDM) is estimated via a global optimization approach- differential evolution optimization (DEoptim), which can be used when the traditional expectation maximization (EM) fails. The application of the DEoptim to LCDM estimation is introduced, explicated, and evaluated via a Monte Carlo simulation study in this article. The aim of this study is to fill the gap between the field of psychometric modeling and modern machine learning estimation techniques and provide an alternative solution in the model estimation.

OPEN ACCESS

Keywords: differential evolution optimization, cognitive diagnostic model, LCDM, estimation, EM algorithm

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Chun Wang,
University of Minnesota Twin Cities, United States
Peida Zhan,
Zhejiang Normal University, China

***Correspondence:**
Zhehan Jiang
zjiang17@ua.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *15 May 2018*
Accepted: *18 October 2018*
Published: *06 November 2018*

Citation:
Jiang Z and Ma W (2018) Integrating Differential Evolution Optimization to Cognitive Diagnostic Model Estimation. Front. Psychol. 9:2142. doi: 10.3389/fpsyg.2018.02142

Assessments have been widely used in education as a part of a summative program for many purposes, such as evaluating whether students have reached the desired proficiency level and determining whether students should be given a scholarship. However, in the past decades, stakeholders have shown a strong interest in the information of students' strengths and weaknesses of their knowledge and skills. This has led to fruitful exploration in the field of psychometrics of how to extract diagnostic information to enhance classroom instruction and learning. Cognitive diagnostic models (CDMs) are a set of psychometric models developed to identify whether a student masters a set of fine-grained skills, such as addition, subtraction, multiplication, and division in math ability assessments. For example, question "2+4–1" measures addition and subtraction, and "4 × 2/3" measures multiplication and division. Although it seems straightforward to conclude that a student may not master addition or subtraction if s\he fails 2+4–1, it is indeed much more complicated in practice in that students may answer a question correctly by guessing or fail a question due to carelessness. As a result, formal psychometric models such as CDMs should be employed for data analysis to make sure the inferences are valid. In addition to educational testing, CDMs are useful in psychological measurement. For example, the literature indicates that neuro-vegetative symptoms are a general construct that contains three attributes: depression (DEP), fatigue (FAT), and sleeplessness (SLE; Rabinowitz et al., 2011). Using CDMs allows researchers/practitioners to investigate the attributes of a given patient. Among the item data types, a binary scale is the most common one that has been adopted in many surveys and measures.

Prior to the data analysis using CDMs, whether a skill is required for answering a question needs to be determined by content experts and/or cognitive psychologists and specified in a binary matrix (Q-matrix; Tatsuoka, 1983) as illustrated in **Table 1** such that theory-granted structure can be applied to the measurement of interest. Rows of the Q-matrix represent questions and columns represent skills. Element 1 indicates that the skill is measured by the question and 0 indicates that the sill is not measured.

TABLE 1 | A Q-matrix sample.

	Skill 1	Skill 2	Skill 3
Question 1	1	0	0
Question 2	1	0	1
Question 3	0	1	1
Question 4	1	1	0

Recent advances in modeling development have produced several general CDMs, such as the Log-linear CDM (LCDM; Henson et al., 2009) and, equivalently, the generalized Deterministic Input; Noisy "And" gate model (G-DINA; de la Torre, 2011). The LCDM provides great flexibility such as (1) subsuming most latent variables, (2) enabling both additive and non-additive relationships between skills and questions simultaneously, and (3) syncing with other psychometric models. Rupp et al. (2010, p. 163) proved that LCDM can be constrained to core CDMs such as Deterministic Input; Noisy "And" gate (DINA; Junker and Sijtsma, 2001) model, Noisy Input; Deterministic "And" gate (NIDA; Junker and Sijtsma, 2001) model, and the Reduced Reparameterized Unified Model (RRUM; Hartz, 2002), and Deterministic Input; Noisy "Or" gate (DINO, Templin and Henson, 2006) model.

The LCDM is essentially a restricted latent class model (Day, 1969; Wolfe, 1970; Titterington et al., 1985), and mathematically, it can be defined as:

$$P\left(\boldsymbol{Y}_p = \boldsymbol{y_p}\right) = \sum_{c=1}^{C}\left(v_c \prod_{i=1}^{I} \pi_{ci}^{y_{pi}} \left(1-\pi_{ci}\right)^{1-y_{pi}}\right), \quad (1)$$

where $\boldsymbol{y}_p = (y_{p1}, y_{p2}, \ldots, y_{pI})$ is the binary response vector of person p on a test comprised of I items, and element y_{pi} is the response on item i. v_c is the probability of membership in latent class c, and π_{ci} is the probability of correct response to item i by person p from latent class c. The log-likelihood of observing item responses of N persons can be expressed as

$$L = \sum_{p=1}^{N} \log\left\{\sum_{c=1}^{C}\left(v_c \prod_{i=1}^{I} \pi_{ci}^{y_{pi}} \left(1-\pi_{ci}\right)^{1-y_{pi}}\right)\right\}. \quad (2)$$

Further, Equation 2 can also be converted to:

$$L = \sum_{p=1}^{N} \log\left\{\sum_{c=1}^{C}\left(\exp\left(\log(v_c) + \log(\prod_{i=1}^{I} \pi_{ci}^{y_{pi}} \left(1-\pi_{ci}\right)^{1-y_{pi}})\right)\right)\right\}, \quad (3)$$

where $\log(\prod_{i=1}^{I} \pi_{ci}^{y_{pi}} \left(1-\pi_{ci}\right)^{1-y_{pi}})$ can be replaced by $\sum_{i=1}^{I} \log(\pi_{ci}^{y_{pi}} \left(1-\pi_{ci}\right)^{1-y_{pi}})$ due to the mathematical property of *log* operation.

Assume the number of attributes is A, the mastery profile of the attributes for a random person is denoted by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_A)$, where element α_a is either 1 or 0. In total, there are 2^A possible attribute profiles and correspondingly 2^A latent classes. For example, when A=4, a person with attribute profile $\boldsymbol{\alpha}$= (1, 1, 1, 0) has mastered the first three attributes except the last one. As illustrated in **Table 1**, a Q-matrix of size I^*A is necessary for a LCDM, where the (i, a) element q_{ia} is 1 when item i measures attribute a and 0 otherwise. The conditional probability of person p with attribute profile $\boldsymbol{\alpha}_c$ answering item i correctly can be written by

$$\pi_{ci} = P\left(y_{pi}=1 \middle| \boldsymbol{\alpha}_c\right) = \frac{\exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)\right)}{1+\exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)\right)}, \quad (4)$$

where $\boldsymbol{q}_i$ is the set of Q-matrix entries for item i, $\lambda_{i,0}$ is the intercept parameter, where $\boldsymbol{\lambda}_i$ represents a vector of length 2^A-1 that contains main effect and interaction effect parameters of item i, and $\boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)$ is a vector of the same length with linear combinations of the $\boldsymbol{\alpha}_c$ and $\boldsymbol{q}_i$. Particularly, $\boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)$ can be expanded to:

$$\boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right) = \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ca} q_{ia} + \sum_{a=1}^{A-1} \sum_{a'>a}^{A} \lambda_{i,2,(a,a')} \alpha_{ca} \alpha_{ca'} q_{ia} q_{ia'} + \ldots, \quad (5)$$

where $\lambda_{i,1,(a)}$ and $\lambda_{i,2,(a,a')}$ are the main effect for attribute α_a and the two-way interaction effect for α_a and $\alpha_{a'}$. Since elements of $\boldsymbol{\alpha}_c$ and $\boldsymbol{q}_i$ are binary, $\boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)$ contains binary elements, which indicate effects that are estimates of interest. For an item measuring n attributes, n-way interaction effects should be specified in $\boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)$. **Table 2** provides a sample of a measure with three attributes: the first item that measures one attribute only (i.e., α_1) has two estimates, where the third item which is associated with all given attributes contains eight estimates.

LCDM ESTIMATION

Traditionally, estimating LCDMs refers to the expectation maximization (EM) algorithm (Bock and Aitkin, 1981) that maximizes the marginal likelihood; this is the most commonly-seen algorithm in the CDM literature. In addition to the EM algorithm, Markov chain Monte Carlo (MCMC) techniques can be, theoretically, used to estimate the LCDM, but to date its application remains upon simpler CDMs such as the DINA model (da Silva et al., 2017; Jiang and Carter, 2018a). This study focuses on the EM algorithm due to its practicality and popularity. The EM algorithm is an intertwined updating mechanism consisting of E- and M-steps. With the provisional item parameter and probability of membership estimates from iteration t-1 (i.e., $\boldsymbol{\lambda}$s and $\boldsymbol{v}$s), the posterior class probability for person p can be obtained in the E-step by

$$H\left(C=c \middle| \boldsymbol{Y_p}=\boldsymbol{y_p}\right) = \frac{v_c \prod_{i=1}^{I} \pi_{pi}^{y_{pi}} \left(1-\pi_{pi}\right)^{1-y_{pi}}}{\sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{pi}^{y_{pi}} \left(1-\pi_{pi}\right)^{1-y_{pi}}} \quad (6)$$

Based on Equation (6), the expected number of persons in latent class c and the expected number of persons in latent class c who

TABLE 2 | A 3-item sample of expressions of a log-linear cognitive diagnostic model.

Item	α_1	α_2	α_3	Expanded $\lambda_{i,0}+\lambda_i^T h(\alpha_c, q_i)$ expression	Shortened expression
1	1	0	0	$\lambda_{1,0}+\lambda_{1,1}(1)+\lambda_{1,2}(0)+\lambda_{1,3}(0)+\lambda_{1,12}(1\times0)+\lambda_{1,13}(1\times0)+\lambda_{1,23}(0\times0)+\lambda_{1,123}(1\times0\times0)$	$\lambda_{1,0}+\lambda_{1,1}(1)$
2	0	1	1	$\lambda_{2,0}+\lambda_{2,1}(0)+\lambda_{2,2}(1)+\lambda_{2,3}(1)+\lambda_{2,12}(0\times1)+\lambda_{2,13}(0\times1)+\lambda_{2,23}(1\times1)+\lambda_{2,123}(0\times1\times1)$	$\lambda_{2,0}+\lambda_{2,2}(1)+\lambda_{2,3}(1)+\lambda_{2,23}(1)$
3	1	1	1	$\lambda_{3,0}+\lambda_{3,1}(1)+\lambda_{3,2}(1)+\lambda_{3,3}(1)+\lambda_{3,12}(1\times1)+\lambda_{3,13}(1\times1)+\lambda_{3,23}(1\times1)+\lambda_{3,123}(1\times1\times1)$	$\lambda_{3,0}+\lambda_{3,1}(1)+\lambda_{3,2}(1)+\lambda_{3,3}(1)+\lambda_{3,12}(1)+\lambda_{3,13}(1)+\lambda_{3,23}(1)+\lambda_{3,123}(1)$

answer item i correctly can be obtained by:

$$n_c = \sum_{p=1}^{N} H\left(C=c \,\middle|\, \boldsymbol{Y_p}=\boldsymbol{y_p}\right), \text{ and}$$

$$r_{ci} = \sum_{p=1}^{N} y_{pi} H\left(C=c \,\middle|\, \boldsymbol{Y_p}=\boldsymbol{y_p}\right),$$

respectively. In the M-step, the following function is maximized with respect to item parameters $\boldsymbol{\lambda}$:

$$\ell = \sum_{i=1}^{I}\sum_{c=1}^{2^A}\left[r_{ci}\log\pi_{ci} + (n_c - r_{ci})\log(1-\pi_{ci})\right],$$

and the probability of membership is updated by

$$\nu_c = \frac{\sum_{p=1}^{N} H\left(C=c \,\middle|\, \boldsymbol{Y_p}=\boldsymbol{y_p}\right)}{N}.$$

Maximizing objective function ℓ usually requires Newton or Fisher scoring methods, where first- and second-order derivatives i.e., $\frac{\partial L}{\partial \lambda}\cdot\left(\frac{\partial^2 L}{\partial \lambda^2}\right)^{-1}$ where the first component is a vector and the second component is a matrix) of the objective function are needed. If $\frac{\partial^2 L}{\partial \lambda^2}$ becomes 0, the iteration process will stop and therefore fail to converge.

As a restricted latent class models, LCDM estimation faces the risk of local maxima (Jin et al., 2016). Theoretically, to obtain valid and accurate estimates, the model estimation should converge at a global maximum of the likelihood function, however, the mixture component of a mixture model is likely to trap the aforementioned EM updates to local maxima. In addition, label switching can occur and therefore lead to a misinterpretation of an estimation. For instance, a person mastering all attributes of interest can be mistakenly labeled as one with zero-mastery. Basing on the traditional EM approach, Rupp et al. (2010) add constraints to the parameter estimates (e.g., ensuring main effects are non-negative); this constraint approach substantially reduces the risks of local maxima and label switching (Lao and Templin, 2016). Using *Mplus* (Muthén and Muthén, 2013), a commercial software designed for latent variable modeling that by default deploys the traditional EM approach, Templin and Hoffman (2013) outline the procedures to specify syntax with parameter constraints for the LCDM estimation. Note that in the LCDM estimation, the EM approach in *Mplus* is turned into an accelerated version, meaning its updating steps are replaced with Quasi-Newton and Fisher scoring, this, however, still falls under the family of the traditional EM algorithm. Although Templin and Hoffman's *Mplus* practice has been implemented in many published works and is proved to be efficient (see Bradshaw and Templin, 2014; Li et al., 2016; Ravand, 2016 for example), it is still not avoiding the convergence failure issue: Templin and Bradshaw (2014) conduct a simulation study with vast conditions each of which was replicated 500 times, where the result shows the numbers of converged replications range from 330 to 447. To avoid the convergence issue while maintaining the properties of the EM approach, we introduce a machine-learning technique named Differential Evolution to estimate LCDMs.

DIFFERENTIAL EVOLUTION

Global optimization under machine-learning umbrella has gained tremendous attention from researchers, mathematicians as well as professionals in the field of engineering, finance, and scientific areas (Mohamed et al., 2012). Many applications of this kind impose complex optimization problems such traditional estimation techniques based upon derivatives become cumbersome or even impossible. To avoid the mathematical deriving procedures yet provide reliable solutions to complex models, Differential Evolution (DE) is invented (Storn and Price, 1997), developed, and applied to practice in different fields (e.g., Paterlini and Krink, 2006; Das et al., 2008; Rocca et al., 2011). Inspired by Darwinian evolution that entails the idea of mutation, crossover, and selection, DE is an enhanced version of derivative-free evolutionary algorithms and has been recognized as a simple yet efficient optimization approach in solving a variety of benchmark problems. The complete DE algorithm cycle can be found in **Figure 1**; in particular, the algorithm starts by sampling D candidate solutions to the problem of interest, where each candidate solution can be either a scalar or a vector (if there are more than one estimate). The mutation procedure takes place by performing simple arithmetic operations (i.e., addition, subtraction, and multiplication) among the existing solutions (namely parent solutions). The resultant mutation outcomes are then crossed over with the parent solutions to produce new candidate (offspring) solutions. Finally, in a one-to-one selection process of each pair of offspring and parent vectors, candidate solutions that fit the model better are passes into the next evolutionary cycle. This cycle iterates until the estimation

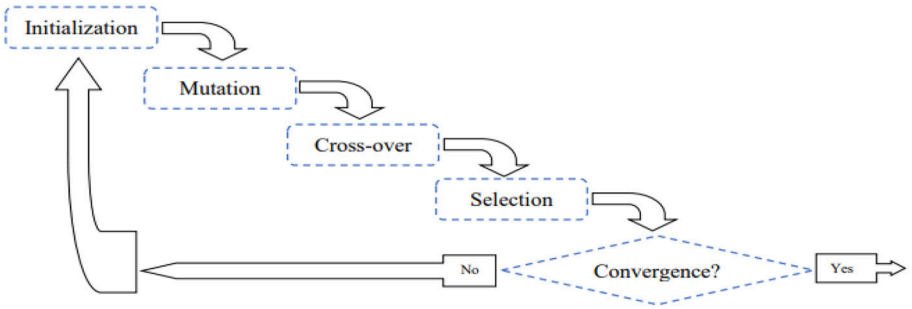


FIGURE 1 | The iterative cycle of differential evolution algorithm.

converge. Mathematical and algorithmic details can be found in the following paragraph.

Let R be the number of estimates, $\boldsymbol{\lambda}_{LO}$ and $\boldsymbol{\lambda}_{HI}$ be the lower and the upper limits (vectors) of the parameters of the estimates, and $G(\cdot)$ be the objective function. Initial candidate solutions $\boldsymbol{\lambda}_d = (\lambda_{d1}, \lambda_{d2}, \ldots, \lambda_{dR})$ for $d = 1, 2, \ldots, D$ can be generated by (1) randomly drawing samples from certain distribution(s) or (2) specifying values with educated guesses, where D is the number of candidate solutions. Mutation procedure can be achieved via different strategies: (1) DE /rand/ 1, (2) DE/current-to-rest/1, and (3) DE/best/1. In particular, for a given set of candidate solutions $\boldsymbol{\lambda}_d$ for $d = 1, 2, \ldots, D$, the mutation outcomes $\boldsymbol{m}_d$ can be calculated as:

$$
\begin{aligned}
DE/rand/1 \qquad & \boldsymbol{m}_d = \boldsymbol{\lambda}_{\delta_o} + F_d{}^*(\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2}) \\
DE/current-to-rest/1 \qquad & \boldsymbol{m}_d = \boldsymbol{\lambda}_{\delta_o} + F_d{}^*(\boldsymbol{\lambda}_{best} - \boldsymbol{\lambda}_{\delta_2}) \\
& \qquad\quad + F_d{}^*(\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2}) \\
DE/best/1 \qquad & \boldsymbol{m}_d = \boldsymbol{\lambda}_{best} + F_d{}^*(\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2})
\end{aligned}
$$

Where δ_o, δ_1, and δ_2 are distinct integers uniformly sampled from 1 to D, $\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2}$ is the difference vector that would be used to mutate two selected parent candidates (e.g., DE/rand/1), $\boldsymbol{\lambda}_{best}$ is the best candidate solution at the current iteration, and finally F_d is the mutation scaling factor that is randomly drawn from a uniform distribution on the interval (0, 1). Some $\boldsymbol{m}_d$ may be produced beyond the constraints set by $\boldsymbol{\lambda}_{lo}$ and $\boldsymbol{\lambda}_{hi}$; some effective solutions to the violation include (1) re-generating a candidate solution until it is valid and (2) setting penalty to the objective function. If an element r in a candidate solution encounter the boundary issue, a quick fix by setting the violating elements to be the middle between boundaries and the that of its parent solution. That is, $m_{dr} = \frac{\lambda_{LOr} + \lambda_{dr}}{2}$ for $m_{dr} < \boldsymbol{\lambda}_{LOr}$ and $m_{dr} = \frac{\lambda_{HIr} + \lambda_{dr}}{2}$ for $m_{dr} > \boldsymbol{\lambda}_{HIr}$. After obtaining $\boldsymbol{m}_d$ from the mutation procedure, a "binomial" crossover operation forms the offspring candidate solutions: let CR be a crossover probability that controls the fraction of the elements that are copied from the parent candidate solution and u_{dr} be a candidate solution, if a random number z_r sampled from a uniform distribution (0, 1) is smaller than CR, the element r of the offspring of u_{dr} is m_{dr}, and λ_{dr} otherwise. The default CR is usually set to 0.5 for a balanced stochastic move. Finally, if $G(\boldsymbol{u}_d)$ is better than $G(\boldsymbol{\lambda}_d)$, $\boldsymbol{u}_d$ would replace $\boldsymbol{\lambda}_d$ to serve as a parent solution for the next iteration. The DE algorithm can be tailored to a parallel computing platform; technically each candidate solution can be calculated in an independent computational unit such that queuing time can be shortened. That said, instead of sequentially updating the candidate solutions, a parallel DE algorithm can perform simultaneous updates.

To illustrate how the DE algorithm functions, an example of a simple regression estimation is provided here. Let independent variable $\boldsymbol{x} = [22, 14, 15, 12, 10, 26, 11, 28]$ and dependent variable $\boldsymbol{y} = [44, 29, 30, 27, 24, 51, 25, 56]$ resulting in $\hat{\boldsymbol{\beta}} = [4.98, 1.78]$ with the ordinary least squares (OLS) estimator. Using the DE algorithm in this case sets the objective function $G(\lambda)$ to $-1^* \sum (\boldsymbol{y} - \hat{\boldsymbol{y}})^2$, which ideally should be maximized to -7.19 according to the OLS result. To keep the demonstration simple, let the number of candidate solutions $D = 3$ and initial values for $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, and $\boldsymbol{\lambda}_3$ were arbitrarily set to [2, 1], [−3, 5], and [1, 2]. At the initial iteration, the best solution was [1, 2] as $G(\boldsymbol{\lambda_3}) = -24$ where $G(\boldsymbol{\lambda_1})$ and $G(\boldsymbol{\lambda_2})$ were −2400 and −21672. Therefore, $\boldsymbol{\lambda}_{best}$ at this stage became $\boldsymbol{\lambda_3}$. With certain random draws for a given mutation calculation (e.g., DE/best/1), $\boldsymbol{m}_1$, $\boldsymbol{m}_2$, and $\boldsymbol{m}_3$ happened to be [3.5, 0.8], [−1, 3.5], and [2, 1.8]. Let $CR = 0.5$, if a random generation produced $z_1 = 0.7$ and $z_2 = 0.4$ for example, the first offspring $\boldsymbol{u_1}$ became [3.5, 1] by taking elements from $\boldsymbol{\lambda_1}$ and $\boldsymbol{m_1}$. This resulted in $G(\boldsymbol{u_1}) = -2022$, which is larger than $G(\boldsymbol{\lambda_1})$, and therefore the new $\boldsymbol{\lambda_1}$ would be replaced by $\boldsymbol{u_1}$. On the other hand, if $\boldsymbol{u_3}$ became [1, 1.8] which produced $G(\boldsymbol{u_1}) = -116.8$, then the $\boldsymbol{\lambda_3}$ remained still. This process continues until $G(\boldsymbol{\lambda})$ converges to −7.19.

In this paper, we integrated the DE into the EM algorithm to estimate LCDMs[1] To make the proposed approach easy to follow, we name it EM-DEoptim algorithm from here. Especially, the method for updating item parameters within the M-step is replaced by the DE algorithm, while the rest of the EM procedures remain identical. To be concrete, the objective function that the EM-DEoptim maximizes is Equation 3, given v_c for each latent class is known. As the DE is a

[1]The snippet code can be found https://alabama.box.com/s/cbuxetk19b1pk1invi1gnbqxd8hw5fij.

stochastic and global optimization technique, the EM-DEoptim is expected to encounter fewer occurrences of the local maxima problem than the traditional EM algorithm (Celeux et al., 1996). In addition, as addressed above, the EM-DEoptim is based upon derivative-free framework such that it can be easily fitted to arbitrarily customized LCDMs without re-deriving the gradient functions nor re-approximating information matrix. For example, if constraining the main effects of Item i and Item i' to be equal while still allowing others to be estimated freely is needed, the EM-DEoptim algorithm can handle the situation by simply assigning the same labels to the constrained parts in the likelihood function expression, where the traditional EM algorithm needs altering the derivatives. This advantage can effectively prevent the aforementioned un-differentiable situations. Last but not least, the computational speed of the EM-DEoptim algorithm, although not outperform the traditional EM algorithm in a singular operation environment, can be substantially improved via parallel computing facilities that are naturally suited to modern machine-learning-based techniques.

SIMULATION STUDY

We conducted a simulation study to demonstrate the utility of the EM-DEoptim algorithm. Specifically, the study involved two investigations: the number of times that the traditional EM algorithm fails and the comparison between the EM-DEoptim algorithm and the traditional EM algorithm in terms of the parameter recovery. In the simulation study, the numbers of attributes A were set to 3, 4, 5. The Q-matrix was randomly generated: when there were 3 attributes ($A = 3$), a balanced Q-matrix in which each item measures either one or two attributes was utilized; similarly, at the condition of 4 and 5 attributes, each item measures two to three attributes. The number of items I was set to 30 and the number of persons N was set to 300. The attributes were generated via two steps: continuous values were initially generated from a multinormal distribution MV ($\mathbf{0}$, Σ) of which the diagonal elements of Σ were constrained to 1 and the off-diagonal values (i.e., correlations between attributes) were randomly drew from a uniform distribution ranging from 0.7 to 0.9, and these continuous values were further converted onto the binary scale by comparing the values with zero (i.e., 1 if the value is larger than zero and 0 otherwise). Finally, the item parameters were specified to two level: high-quality group that sets main effects = 2, intercepts = −1.5, and interaction effects = 0.5, and low-quality group that makes main effects = 0.2, intercepts = −0.5, and interaction effects = 0.1.

The traditional EM algorithm was realized via the package *CDM* (George et al., 2016; alternatively, one can choose the package *GDINA* by Ma and de la Torre, 2018), where the EM-DEoptim algorithm was executed in *R* (R Core Team, 2018). The stop criterion in *CDM* was set to 1,000 iterations or the change of likelihood value <0.001, where the EM-DEoptim algorithm was forced to stop if the iteration number reaches to 1,000 or the likelihood value remains identical for 10 iterations. In this study, the DE configurations were set to default (Ardia et al., 2011): DE/current-to-rest/1 with $F_d = 0.8$, $CR = 0.5$, and 500 candidate solutions, where ± 20 is used to constrain the parameter estimates. The machine used was Dell Precision 3520 with 16GB RAM and a 2.90 GHz i7-7820 4-core Intel processor. The study was replicated for 200 times.

The dependent variables in this part of the study are (1) the number of convergence failure of the EM algorithm, (2) relative bias (RBIAS) and root mean squared error (RMSE), and (3) the attribute classification accuracy measured by each attribute and each profile. Overall, there was only two failed convergence failures when the item quality was high, where the low-quality item parameters led to seven failures: two cases in the situation of $A = 4$ and five cases when $A = 5$. On the other hand, EM-DEoptim had no unexpected terminations during the iterations. **Table 3** shows the attribute classification accuracy rates. Both algorithms produced very similar results, where some patterns can be discovered: (1) the more attributes the estimation face, the less accurate the attribute estimates are yielded, (2) the higher the item parameter quality is, the more accurate the attribute estimates are produced, and (3) the profile accuracy is more sensitive to the item parameter quality.

Similar to the attribute estimates, the item parameter recovery presented similar pattern for both algorithms as listed in **Table 4**. The biases and MSEs were higher when (1) the number of attributes was larger and (2) the item parameter quality is higher. In addition, main effect estimates were more accurate and efficient than both interaction and intercept effects. This finding is not uncommon in complex psychometric models (Jiang et al., 2016). When the item parameter quality is low, and/or the number of attribute is large (e.g., 5), the EM-DEoptim performed better than the traditional EM algorithm. An important reason is that the boundary constraints imposed by the EM-DEoptim algorithm can limit the estimates into a certain range. Although not a main focus of the studies, the computing speed showed a substantial difference: the average time (in seconds) for 3-, 4-, and 5 attributes were 4.45, 22.55, and 78.64 for the traditional EM algorithm, while the EM-DEoptim took 61.22, 354.18, and 1228.76.

REAL DATA APPLICATION

The dataset used in this session is an assessment of a health profession administered to 3491 test takers (Jiang and Raymond, 2018). The number of items is 200 each of which measures

TABLE 3 | attribute accuracy rate of the simulation study.

		EM		DE-EMoptim	
A	Quality	Attribute	Profile	Attribute	Profile
3	High	0.848	0.634	0.847	0.642
4	High	0.816	0.505	0.821	0.489
5	High	0.768	0.336	0.755	0.342
3	Low	0.516	0.104	0.517	0.104
4	Low	0.509	0.039	0.513	0.044
5	Low	0.504	0.017	0.468	0.013

TABLE 4 | Item parameter estimates of the simulation study.

RBIAS		EM			EM-DEoptim		
A	Quality	Main	Intercept	Interaction	Main	Intercept	Interaction
3	High	−0.433	−0.270	−1.896	−0.463	−0.280	−1.196
4	High	−1.744	−0.948	−5.036	−1.740	−0.938	−4.106
5	High	−1.906	−0.957	−4.975	−1.906	−0.957	−2.675
3	Low	−5.389	−8.906	−2.160	−2.389	−4.406	−1.960
4	Low	−9.526	−9.950	−3.696	−8.626	−7.950	−4.026
5	Low	−11.99	−6.419	−10.027	−9.495	−5.419	−7.027
RMSE		**EM**			**EM-Deoptim**		
3	High	6.346	2.124	8.349	6.890	1.924	4.336
4	High	14.576	5.035	15.679	15.079	6.422	13.853
5	High	21.643	6.230	26.327	18.223	7.360	18.707
3	Low	15.423	11.630	18.478	12.863	12.112	18.481
4	Low	22.512	15.165	26.064	19.299	16.865	17.446
5	Low	25.410	15.109	15.319	16.506	14.409	14.319

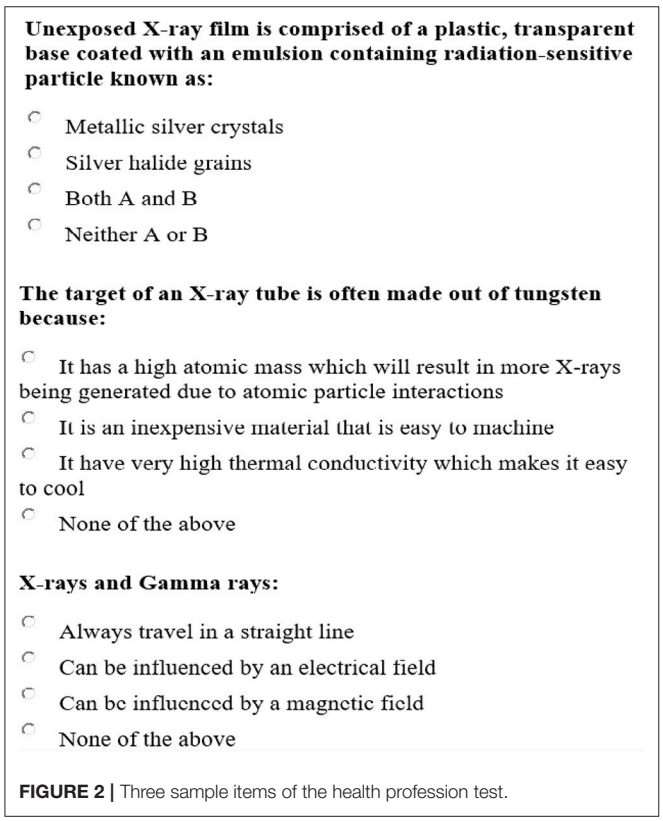


FIGURE 2 | Three sample items of the health profession test.

one attribute. Therefore, there are five attributes in total: the knowledge of radiation biology (Items #1-45), the knowledge of equipment operation (Items #46-67), the image acquisition and evaluation capacity (Items #68-112), the knowledge of imaging procedures (Items #113-162), and ethics (Items #163-200). Three samples of the items can be found in **Figure 2**.

Two common model fit indices are reported as: (1): mean of absolute deviations in observed and expected correlations (MADcor; DiBello et al., 2007) is 0.041 and standardized mean square root of squared residuals (SRMSR; Maydeu-Olivares, 2013; Maydeu-Olivares and Joe, 2014) is 0.05. Overall, the model has an adequate fit. Note that more model fit indices such as χ^2-like statistics (Orlando and Thissen, 2000) are recommended. This paper focuses on the estimation. More model fit details can be found in Hu et al. (2016) and Sorrel et al. (2017).

Rounding the number of digits to three after the decimal point, one can see that 16 classes are nearly empty and therefore are labeled as "others" in **Figure 3** (see Jiang and Carter, 2018b for more visual aids). Nearly 40% of the test takers master all five attributes. According to Templin and Bradshaw (2014), many empty classes indicate potential hierarchies of attribute structure, however, the parameter estimates can be relatively robust even the non-hierarchical modeling is adopted here. **Figure 4** shows the distributions of the parameter estimates grouped by parameter types and attribute identifications. Attribute #3 had the highest means of both intercepts and main effects: 2.65 and 1.79. The means of intercepts and main effects of Attribute #5 were −1.05 and 0.20.

To compare the estimates with other estimation approaches, we also implemented a Bayesian technique-Hamiltonian Monte Carlo-to the analysis by adopting uninformative priors for both item parameters and the class membership probability: the mean and standard deviation for item parameters were 0 and 20, while the Dirichelet prior parameters were all set to 1 (see Jiang and Carter, 2018a for details). The correlations of item parameter estimates were relatively high: 0.77, 0.84, and 0.69 for intercept, main effect, and interaction effects. On the other hand, the attribute agreement was lower than that of the item parameter estimates: the average ratio for all attributes was 0.67, where the value dropped to 0.39 when it comes to the match of the class membership classification. This makes sense as the Dirichelet

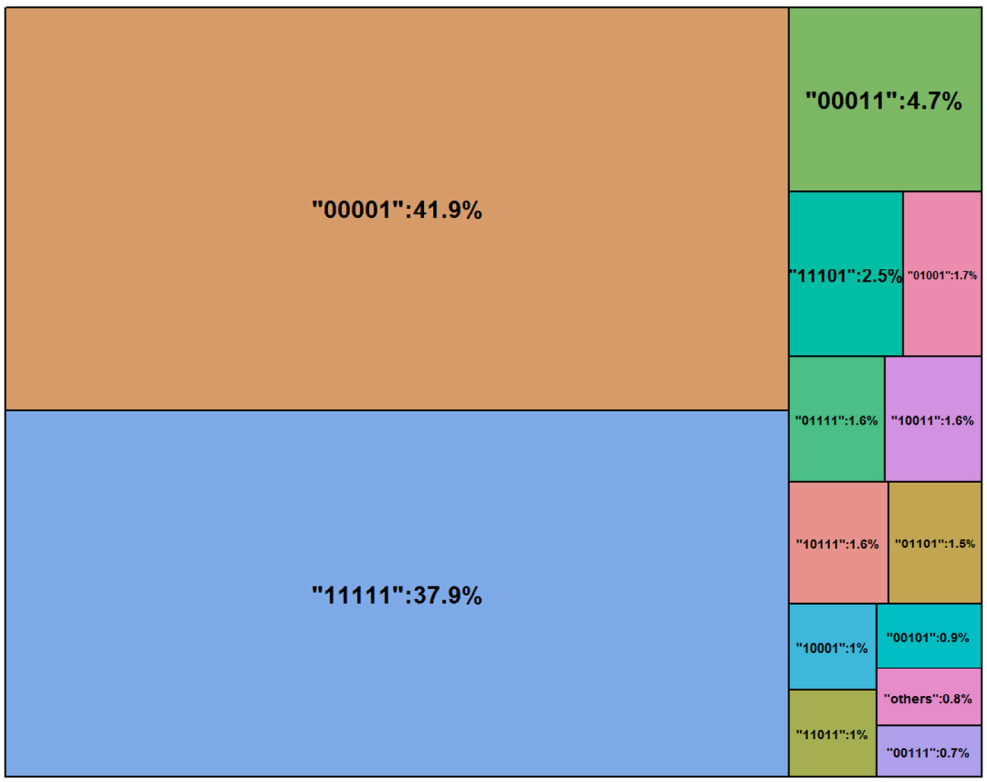


FIGURE 3 | Estimated class probabilities via LCDM.

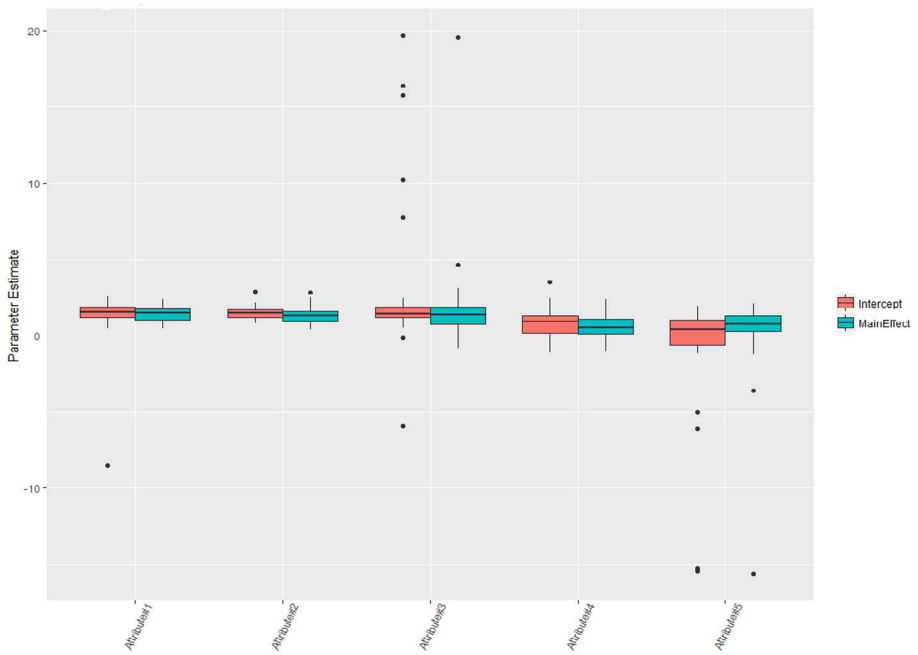


FIGURE 4 | Item parameter estimates of the health profession test.

prior had forced the assignment on each latent class and therefore the result tended to be more different from those that were fully determined by the EM algorithm.

DISCUSSION AND CONCLUSION

The purpose of this paper is to propose a machine-learning based algorithm for the estimation of LCDMs. In particular, the proposed estimator is a combination of the EM framework and the DEoptim algorithm, which has been popular in neural networks and business analytics fields. The performance of the proposed algorithm is evaluated through a simulation study of which the results indicate that it is an appropriate option to handle LCDM estimation task. This paper, however, does not suggest that the proposed algorithm should replace the EM algorithm in practice; at the situations where the EM algorithm fails to produce estimates due to the unsuccessful derivative updates, the EM-DEoptim algorithm can be an alternative.

The proposed EM-DEoptim algorithm and the traditional EM algorithm implemented in M*plus* produced virtually identical parameter estimates, and the former seems less frequently to fail. The average computational time for M*plus* estimation with the multiple-core option is 15 min. The difference is caused by the features of the algorithms: the EM algorithm based upon Quasi-Newton and Fisher scoring updates estimates with directional steps (i.e., the iteration always leads to better solutions), while the DEoptim part is truly stochastic such that the updating procedures may be wasted. Even though the DEoptim mechanism is fundamentally less directional than Quasi-Newton and Fisher scoring, The EM-DEoptim algorithm perform cannot very similar to the EM algorithm. Theoretically, the EM-DEoptim algorithm can be many times faster than what it is now if the entire function is constructed in *C++* or *Fortran*; currently only the DEoptim is implemented in *C++* through the package *RcppDE*, where the entire algorithm is written in base *R* software scripting language. Research has shown that using compiler package with *R* often takes less than half of time executing the same function than that of without packages (e.g., Aruoba and Fernández-Villaverde, 2015). In addition, given the DEoptim algorithm is composed of basic calculation, performing the proposed algorithm in a vectorization approach and therefore with graphics processing units (GPUs) is expected to accelerate the estimations.

AUTHOR CONTRIBUTIONS

ZJ proposes the idea about integrating differential evolution optimization into the EM framework in the LCDM estimation and deploys the functional algorithm in simulation studies; WM produces both literature review and technical detail supports to this manuscript.

FUNDING

ZJ was sponsored by the Research Grants Committee from Research and Economic Development, the University of Alabama (No. RG14790).

REFERENCES

Ardia, D., Boudt, K., Carl, P., Mullen, K. M., and Peterson, B. G. (2011). Differential evolution with DEoptim. *R J.* 3, 27–34.

Aruoba, S. B., and Fernández-Villaverde, J. (2015). A comparison of programming languages in macroeconomics. *J. Econ. Dyn. Control* 58, 265–273. doi: 10.1016/j.jedc.2015.05.009

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Bradshaw, L., and Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika* 79, 403–425. doi: 10.1007/s11336-013-9350-4

Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Stat. Comput. Simul.* 55, 287–314. doi: 10.1080/00949659608811772

da Silva, M. A., de Oliveira, E. S., Davier, A. A., and Bazán, J. L. (2017). Estimating the DINA model parameters using the No-U-Turn Sampler. *Biom. J.* 60, 352–368. doi: 10.1002/bimj.201600225

Das, S., Abraham, A., and Konar, A. (2008). Automatic clustering using an improved differential evolution algorithm. *IEEE Trans. Syst. Man Cybernet. A Syst. Hum.* 38, 218–237. doi: 10.1109/TSMCA.2007.909595

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* 56, 463–474. doi: 10.1093/biomet/56.3.463

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

DiBello, L. V., Roussos, L. A., and Stout, W. F. (2007). "Review of cognitively diagnostic assessment and a summary of psychometric models," in *Handbook of Statistics*, Vol. 26, eds C. R. Rao and S. Sinharay (Amsterdam: Elsevier), 979–1030.

George, A. C., Robitzsch, A., Kiefer, T., Gro,ß, J., and Ünlü, A. (2016). The R package CDM for cognitive diagnostic models. *J. Stat. Softw.* 74, 1–24. doi: 10.18637/jss.v074.i02

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality.* Ph.D. Doctoral dissertation, ProQuest Information & Learning.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnostic models using log-linear models with latent variables. *Psychometrika* 74:191. doi: 10.1007/s11336-008-9089-5

Hu, J., Miller, M. D., Huggins-Manley, A. C., and Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnostic models. *Int. J. Test.* 16, 119–141. doi: 10.1080/15305058.2015.1133627

Jiang, S., Wang, C., and Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Front. Psychol.* 7:109. doi: 10.3389/fpsyg.2016.00109

Jiang, Z., and Carter, R. (2018a). Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnostic model via Stan. *Behav. Res. Methods*, 1–12. doi: 10.3758/s13428-018-1069-9

Jiang, Z., and Carter, R. (2018b). Visualizing library data interactively: two demonstrations using R language. *Library Hi Tech. News.* 35, 14–17. doi: 10.1108/LHTN-01-2018-0003

Jiang, Z., and Raymond, M. (2018). The use of multivariate generalizability theory to evaluate the quality of subscores. *Appl. Psychol. Meas.* doi: 10.1177/0146621618758698. [Epub ahead of print].

Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., and Jordan, M. I. (2016). "Local maxima in the likelihood of gaussian mixture models: structural results and algorithmic consequences," in *Advances in Neural Information Processing*

Systems (Barcelona), 4116–4124. Available online at: https://papers.nips.cc/book/advances-in-neural-information-processing-systems-29-2016

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Lao, H., and Templin, J. (2016). "Estimation of diagnostic classification models without constraints: issues with class label switching, in *Paper Presented at Annual Meeting of the National Council on Measurement in Education* (Washington, DC: District of Columbia).

Li, H., Hunter, V., and Lei, P. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Lang. Test.* 33, 391–409. doi: 10.1177/0265532215590848

Ma, W., and de la Torre, J. (2018). *GDINA: The Generalized DINA Model Framework [Computer Software Version 2.1]*. Available online at: https://CRAN.R-project.org/package=GDINA

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measurement* 11, 71–137. doi: 10.1080/15366367.2013.831680

Maydeu-Olivares, A., and Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivar. Behav. Res.* 49, 305–328. doi: 10.1080/00273171.2014.911075

Mohamed, A. W., Sabry, H. Z., and Khorshid, M. (2012). An alternative differential evolution algorithm for global optimization. *J. Adv. Res.* 3, 149–165. doi: 10.1016/j.jare.2011.06.004

Muthén, L. K., and Muthén, B. O. (2013). *Mplus User's Guide (Version 6.1)[Computer Software and Manual]*. Los Angeles, CA: Muthén & Muthén.

Orlando, M., and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Appl. Psychol. Meas.* 24, 50–64. doi: 10.1177/01466216000241003

Paterlini, S., and Krink, T. (2006). Differential evolution and particle swarm optimisation in partitional clustering. *Comput. Stat. Data Anal.* 50, 1220–1247. doi: 10.1016/j.csda.2004.12.004

Rabinowitz, A. R., Fisher, A. J., and Arnett, P. A. (2011). Neurovegetative symptoms in patients with multiple sclerosis: fatigue, not depression. *J. Int. Neuropsychol. Soc.* 17, 46–55. doi: 10.1017/S1355617710001141

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available online at: http://www.Rproject.org/

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assess.* 34, 782–799. doi: 10.1177/0734282915623053

Rocca, P., Oliveri, G., and Massa, A. (2011). Differential evolution as applied to electromagnetics. *IEEE Antennas Propagat. Mag.* 53, 38–49. doi: 10.1109/MAP.2011.5773566

Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., and Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnostic modeling. *Appl. Psychol. Meas.* 41, 614–631. doi: 10.1177/0146621617707510

Storn, R., and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optimiz.* 11, 341–359. doi: 10.1023/A:1008202821328

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Measur.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin, J., and Bradshaw, L. (2014). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika* 79, 317–339. doi: 10.1007/s11336-013-9362-0

Templin, J., and Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educ. Measur. Issues Prac.* 32, 37–50. doi: 10.1111/emip.12010

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnostic models.*Psychol. Methods* 11:287. doi: 10.1037/1082-989X.11.3.287

Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley.

Wolfe, J. H. (1970). Profile clustering by multivariate mixture analysis. *Multivariate Behav. Res.* 5, 329–350. doi: 10.1207/s15327906mbr0503_6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Jiang and Ma. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 April 2018
doi: 10.3389/fpsyg.2018.00607

Joint Testlet Cognitive Diagnosis Modeling for Paired Local Item Dependence in Response Times and Response Accuracy

Peida Zhan[1]*, *Manqian Liao*[2] and *Yufang Bian*[1]

[1] Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China, [2] Measurement, Statistics and Evaluation, Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, United States

In joint models for item response times (RTs) and response accuracy (RA), local item dependence is composed of local RA dependence and local RT dependence. The two components are usually caused by the same common stimulus and emerge as pairs. Thus, the violation of local item independence in the joint models is called paired local item dependence. To address the issue of paired local item dependence while applying the joint cognitive diagnosis models (CDMs), this study proposed a joint testlet cognitive diagnosis modeling approach. The proposed approach is an extension of Zhan et al. (2017) and it incorporates two types of random testlet effect parameters (one for RA and the other for RTs) to account for paired local item dependence. The model parameters were estimated using the full Bayesian Markov chain Monte Carlo (MCMC) method. The 2015 PISA computer-based mathematics data were analyzed to demonstrate the application of the proposed model. Further, a brief simulation study was conducted to demonstrate the acceptable parameter recovery and the consequence of ignoring paired local item dependence.

Keywords: cognitive diagnosis models, response time models, response times, local item dependence, testlet, DINA model, PISA

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Yong Luo,
National Center for Assessment in Higher Education (Qiyas), Saudi Arabia
Chun Wang,
University of Minnesota, United States

***Correspondence:**
Peida Zhan
pdzhan@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *06 February 2018*
Accepted: *10 April 2018*
Published: *25 April 2018*

Citation:
Zhan P, Liao M and Bian Y (2018) Joint Testlet Cognitive Diagnosis Modeling for Paired Local Item Dependence in Response Times and Response Accuracy. Front. Psychol. 9:607. doi: 10.3389/fpsyg.2018.00607

INTRODUCTION

Nowadays, it becomes a common practice to collect response time (RT) data as the computer-based tests are applied to large-scale assessments. RT represents the amount of time a respondent spends on an item. It serves as an additional source of information about the working speed of a respondent as well as the time intensity of an item. In the past few decades, a number of studies have been done to model the RTs. Before the year of 2007, the RT modeling studies such as Thissen (1983), Verhelst et al. (1997), and Wang and Hanson (2005) were motivated by the speed-accuracy trade-off (Luce, 1986). However, this trade-off only reflected a within-person relationship between speed and accuracy (van der Linden, 2009) where, given a fixed set of items, a respondent's speed is dependent on his or her accuracy. Therefore, the relationship between speed and accuracy should be modeled at a higher level. To this end, van der Linden (2007) proposed a hierarchical modeling framework to explain the higher-level relationship between speed and accuracy. In this framework, RTs and RA were separately modeled at the first level whereas two correlational structures were modeled at the second level. The correlational structures accounted for either the dependence between person latent speed and latent ability parameters

and that between item accuracy-related and item time-related parameters. A comparison study suggested that the hierarchical modeling framework yielded more reasonable outcomes in both real and simulated data than other RT modeling approaches (Suh, 2010). The hierarchical modeling framework was generalized to integrate different measurement models due to its flexible nature (e.g., Klein Entink et al., 2009a,b; Wang et al., 2013; Meng et al., 2015; Molenaar et al., 2015; Wang and Xu, 2015; Fox and Marianti, 2016). However, almost all the previous studies in RT modeling were based on unidimensional item response theory (IRT) models but none used multidimensional measurement models.

Multidimensional tests and cognitive diagnostic assessments become more and more prevalent given the increasing demand for diagnostic test feedback containing refined information. In general, cognitive diagnostic assessments aim at evaluating respondent's mastery status (e.g., mastery or non-mastery) of latent skills or attributes. This information can be provided to teachers or clinicians so that they can determine the remedial instructions or targeted interventions accordingly. Although numerous cognitive diagnosis models (CDMs) have been developed (for review, see Rupp et al., 2010) based on various cognitive and psychological assumptions, almost all of them only utilized information on RA. Recently, Zhan et al. (2017) proposed a joint cognitive diagnosis modeling approach to simultaneously model RTs and RA. In the study of Zhan et al. (2017), the deterministic-inputs, noisy "and" (DINA) model (Macready and Dayton, 1977; Haertel, 1989; Junker and Sijtsma, 2001) and the lognormal RT model (van der Linden, 2006) were used as the measurement models for RA and RTs, respectively. A higher-order latent structure (de la Torre and Douglas, 2004) was introduced to account for the relationship between latent attributes and a continuous higher-order latent ability. Furthermore, a bivariate normal distribution was used to model the relationship between the higher-order latent ability and the latent speed. A similar approach was proposed by Minchen (2017). Unlike Minchen's approach, Zhan et al. (2017)'s approach explicitly modeled the correlation between different item parameters (i.e., within-item characteristic dependency; Fox, 2010; Zhan et al., manuscript submitted for publication) by assuming that they followed a multivariate normal distribution.

A key assumption in the joint models of RA and RTs is local item dependence. Specifically, the observed RA responses are conditionally independent of each other given an individual score in latent ability or a specific latent attribute mastery status, which is denoted as *local RA independence*; in the meanwhile, all the RTs are conditionally independent of each other given the an individual score in latent speed, which is denoted as *local RT independence*. In other words, in the joint models, local item independence is composed of local RA independence and local RT independence, which is known as *paired local item independence*. However, the assumption of local item independence is often violated in educational tests, resulting in local item dependence. One of the most common scenarios that lead to local item dependence is the presence of testlet, where several items are based on a common context (Wainer and Kiely, 1987).

A testlet is defined as a cluster of items that share a common stimulus. The local item dependence resulted from a testlet is called testlet effect. Testlet has been widely adopted in educational tests. For example, in a reading comprehension test, a testlet is formed when a bundle of items are based on the same reading passage. The testlet design makes the assessment process more efficient (DeMars, 2012). While responding to the items within the same testlet, the students only need to process the scenario once and the context information can be applied to all the items in the testlet. However, the testlet design makes it more difficult to measure student's reading ability as the student's performance may be affected by their knowledge or interest in the reading passage content besides their reading ability (Yen, 1993). Thus, item responses within the same testlet may be locally dependent on each other.

Testlet response theory modeling (Wang and Wilson, 2005; Wainer et al., 2007) is one of the most popular approaches to handle testlet effect or local item dependency. As a bi-factor multidimensional IRT model (DeMars, 2006; Li et al., 2006), the testlet response theory model assumes that all the item responses are accounted for by a common factor of latent ability, while the responses within a testlet are further explained by a random testlet effect factor. It has been demonstrated that the presence of testlet effect affects model parameter estimates, equating process, and test reliability estimates (e.g., Sireci et al., 1991; Bradlow et al., 1999; Wang and Wilson, 2005; Wainer et al., 2007; Jiao et al., 2012, 2013; Zhan et al., 2014; Jiao and Zhang, 2015; Tao and Cao, 2016). However, all the studies above only addressed the local RA dependence but none accounted for the local RT dependence.

As aforementioned, the paired local item independence is composed of local RA independence and local RT independence. Given that the item clusters which cause local RA dependence would also result in local RT dependence, and local RA dependence and local RT dependence should emerge in pairs. Thus, the violation of paired local item independence is called *paired local item dependence*. In other words, local RA dependence and its corresponding local RT dependence are caused by the same stimulus but are reflected in different forms (i.e., RA and RTs). To address the paired local item dependence in the IRT framework, Im (2017) proposed a hierarchical testlet model, in which local RA dependence was handled by a testlet response theory model whereas local RT dependence was handled by a lognormal RT testlet model.

In cognitive diagnosis, however, only a few studies focused on accounting for local RA dependence (e.g., Hansen, 2013; Zhan et al., 2015; Hansen et al., 2016), and, to our knowledge, none examined local RT dependence. As aforementioned, the joint CDMs assume paired local item independence. Thus, the purpose of this study is to extend the joint cognitive diagnosis modeling approach (Zhan et al., 2017) in order to address the potential paired local item dependence in RTs and RA. The rest of the paper starts with a review of the testlet-DINA model (Zhan et al., 2015) and the lognormal RT testlet model (Im, 2017). Then the proposed joint testlet-DINA model is introduced. It is followed by a real data analysis using the Program for International Student Assessment (PISA) 2015 computer-based mathematics data, which serves to demonstrate the application of the proposed

model. Finally, a brief simulation study is presented used to demonstrate the model parameter recovery and the consequence of ignoring paired local item dependence.

JOINT TESTLET COGNITIVE DIAGNOSIS MODELING

The Testlet-DINA Model

To account for the local RA dependence in cognitive diagnosis, Hansen (2013) and Hansen et al. (2016) proposed a higher-order, hierarchical CDM which can be viewed as a combination of the two-tier item factor model (Cai, 2010) and the log-linear CDM (Henson et al., 2009). Like the two-tier item factor model, Hansen's model could only account for local RA dependence which was resulted from a single source. Zhan et al. (2015) proposed two within-item multidimensional testlet effect CDMs which was able to account for local RA dependence that was resulted from multiple sources simultaneously (Rijmen, 2011; Zhan et al., 2014). The two models included a compensatory model which allowed attributes to compensate each other and a non-compensatory model which assumed that respondents need to master all the required attributes in order to have a high correct response probability. For simplicity, the testlet-DINA model in this study only refers to the non-compensatory model, which is written as

$$\text{logit}(P(Y_{ni}=1)) = \beta_i + \delta_i \prod_{k=1}^{K} \alpha_{nk}^{q_{ik}} + \sum_{m=1}^{M} u_{im}\gamma_{nm}, \quad (1)$$

where Y_{ni} denotes the dichotomous response of person n to item i; $\alpha_n = (\alpha_{n1}, \ldots, \alpha_{nK})'$denotes person n's attribute pattern, K is the number of required attributes; β_i and δ_i are the intercept and interaction parameters for item i, respectively; The Q-matrix (Tatsuoka, 1983) is an I-by-K confirmatory matrix with element q_{ik} indicating whether the attribute k is required to correctly answer the item i (i.e., $q_{ik} = 1$ if the attribute is required, and 0 otherwise); $\gamma_{nm} \sim N(0, \sigma^2_{\gamma_m})$is the RA testlet effect of the mth testlet, which represents the interaction effect between person n and items within testlet m on RA. Usually, the value of $\sigma^2_{\gamma_m}$indicates the magnitude of testlet effect (Wang and Wilson, 2005; Wainer et al., 2007). A large variance is associated with a large testlet effect. All the γ_{nm}s are assumed to be independent with each others; Let M be the total number of testlets in the test, the U-matrix (Zhan et al., 2014) is an I-by-M confirmatory matrix with element u_{im} indicating whether item i belongs to testlet m (i.e., $u_{im} = 1$ if item i belongs to testlet m, and 0 otherwise).

Obviously, when all elements in the U-matrix equal to 0 (means no tesltet in the test) or all $\sigma^2_{\gamma_m} = 0$ (means no testlet effect), the testlet-DINA model reduces to the reparameterized DINA model (DeCarlo, 2011; von Davier, 2014).

The Lognormal RT Testlet Model

To account for the local RT dependence, Im (2017) proposed the lognormal RT testlet model. The lognormal RT testlet model is an extension of the regular lognormal RT model (van der Linden, 2006) by introducing a random testlet effect parameter, but it can also be taken as a special case of the multidimensional lognormal RT model (Zhan et al., manuscript submitted for publication). Let T_{ni} be the observed RT of person n to item i, the lognormal RT testlet model can be expressed as

$$\begin{aligned} T_{ni} &\sim f(t_{ni}; \tau_n, \lambda_{nm}, \omega_i, \xi_i) \\ &= \frac{\omega_i}{t_{ni}\sqrt{2\pi}} \exp(-\frac{1}{2}(\omega_i(\log t_{ni} - (\xi_i - \tau_n - \lambda_{nm})))^2), \end{aligned} \quad (2)$$

where $\log t_{ni}$ be the logarithm of RT, which is used to transform the positively skewed distribution of RT to a more symmetric shape; τ_n be the latent speed of person n; ξ_i be the time-intensity of item i; ω_i be the discriminating power of item i, which can be treated as a time-kurtosis parameter; $\lambda_{nm} \sim N(0, \sigma^2_{\lambda_m})$be the mth RT testlet effect parameter to address local RT dependence, which represents the interaction between person n and items within testlet m in RT. The larger the variance, the larger the testlet effect is. All λ_{nm}s are assumed to be independent of each other.

Equation (2) can be extended to account for potential within-item multidimensional testlet effect

$$\begin{aligned} T_{ni} &\sim f(t_{ni}; \tau_n, \lambda_n, \omega_i, \xi_i) \\ &= \frac{\omega_i}{t_{ni}\sqrt{2\pi}} \exp(-\frac{1}{2}(\omega_i(\log t_{ni} - (\xi_i - \tau_n - \sum_{m=1}^{M} u_{im}\lambda_{nm})))^2), \end{aligned} \quad (3)$$

where all the parameters have been defined above. Equation (3) is regarded as the within-item multidimensional testlet effect lognormal RT model, which can be seen as a special case of the multidimensional lognormal RT model (Zhan et al., manuscript submitted for publication). For simplicity, Equation (3) can be equivalently expressed as

$$\log T_{ni} \sim N(\xi_i - \tau_n - \sum_{m=1}^{M} u_{im}\lambda_{nm}, \omega_i^{-2}). \quad (4)$$

When there is only one source of local RT dependence, the within-item multidimensional testlet effect lognormal RT model reduces to the lognormal RT testlet model (Im, 2017). Further, when all the elements in the U-matrix equal to 0 or $\sigma^2_{\lambda_m} = 0$for all testlets, the within-item multidimensional testlet effect lognormal RT model reduces to the regular lognormal RT model (van der Linden, 2006).

The Joint Testlet-DINA Model

The joint testlet-DINA model is specified as follows: Y_{ni} and $\log T_{ni}$ are separately modeled at the first level following the convention of joint cognitive diagnosis modeling approach and the hierarchical testlet model; a higher-order latent structural model is used to account for the relationship between binary latent attributes and a continuous higher-order latent ability; further, at the higher level, three variance-covariance structures are imposed to model the dependencies among person parameters, item parameters, and testlet effect parameters. A graphical representation of the joint testlet-DINA model is given in **Figure 1**.

First, the testlet-DINA model (Equation 1) and the within-item multidimensional testlet effect lognormal RT model

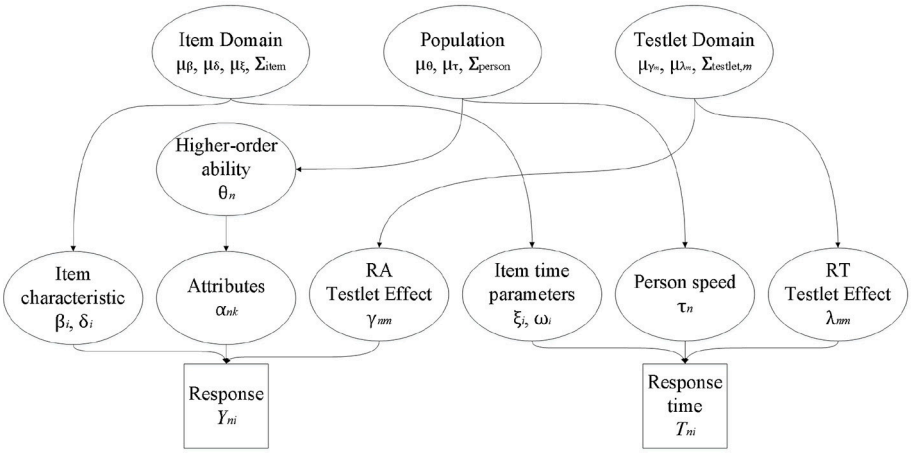


FIGURE 1 | A graphical representation of the joint testlet-DINA model.

(Equation 4) are used as the measurement models for RA and RTs, respectively.

Then, the higher-order latent structural model is used to link the correlated attributes, which is given by

$$\text{logit}(P(\alpha_{nk} = 1)) = \nu_k \theta_n - \kappa_k, \tag{5}$$

where $P(\alpha_{nk} = 1)$ is the probability of mastery of attribute k by person n; θ_n is a higher-order (general) ability of person n, which is assumed to follow a standard normal distribution for identification purpose; and ν_k and κ_k are the slope and difficulty parameters for attribute k.

Further, item parameters are assumed to follow a trivariate normal distribution

$$\boldsymbol{\Psi}_i = \begin{pmatrix} \beta_i \\ \delta_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\xi \end{pmatrix}, \Sigma_{item} \right). \tag{6}$$

Additionally, since the residual error variance, ω_i^{-2}, is assumed to be independently distributed (Zhan et al., 2017), it is not included in $\boldsymbol{\Psi}_i$.

Likewise, person parameters are assumed to follow a bivariate normal distribution

$$\Theta_n = \begin{pmatrix} \theta_n \\ \tau_n \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \Sigma_{person} \right). \tag{7}$$

In addition, testlet effect parameters in testlet m are assumed to follow a bivariate normal distribution

$$\Gamma_{nm} = \begin{pmatrix} \gamma_{nm} \\ \lambda_{nm} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_{testlet,m} \right). \tag{8}$$

If there are M testlets, there will be M bivariate normal distributions. In addition, it should be noted that, in the proposed model, the u_{im} in RT model (Equation 3) has the same value as the u_{im} in RA model (Equation 1) because of the paired local item dependence. In summary, Equations (1, 4–8), together, constitute the joint testlet-DINA model. Constraints are set for identification purpose (i.e., $\mu_\theta = 0, \sigma_\theta^2 = 1; \mu_\tau = 0$). The first two constraints are consistent with those set in the higher-order latent trait model while the third removes the tradeoff between ξ_i and τ_n from a lognormal model. After addressing the paired local item dependence, four conditional independence assumptions are made: the α_{nk} are conditionally independent given $\boldsymbol{\theta}_n$; the Y_{ni} are conditionally independent given $\boldsymbol{\alpha}_n$ and γ_{nm}; the $\log T_{ni}$ are conditionally independent given τ_n and λ_{nm}; and Y_{ni} and $\log T_{ni}$ for a particular item i are conditionally independent given person parameters and testlet effect.

Bayesian Parameter Estimation

Parameters in the joint testlet-DINA model can be estimated using the full Bayesian approach with the Markov chain Monte Carlo (MCMC) method. In this study, free software JAGS (Version 4.3.0; Plummer, 2015) was used to estimate the parameters. JAGS uses a default option of the Gibbs sampler (Gelfand and Smith, 1990). Sample code were presented in Appendix. A tutorial of using JAGS for Bayesian CDM estimation can be found in Zhan (2017).

To begin with, under the assumption of local independence, Y_{ni}, $\log T_{ni}$ and α_{nk} are independently distributed, which is written as

$$\begin{aligned} Y_{ni} &\sim \text{Bernoulli}(P(Y_{ni} = 1)), \\ \log T_{ni} &\sim N(\xi_i - \tau_n - \textstyle\sum_{m=1}^{M} u_{im}\lambda_{nm}, \omega_i^{-2}), \\ \alpha_{nk} &\sim \text{Bernoulli}(P(\alpha_{nk} = 1)). \end{aligned}$$

The priors of item parameters are assumed to be a trivariate normal distribution, written as

$$\begin{pmatrix} \beta_i \\ \delta_i \\ \xi_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\xi \end{pmatrix}, \Sigma_{item} \right), \omega_i^{-2} \sim \text{InvGamma}(1, 1). \tag{9}$$

Further, the hyper priors are specified as

$$\begin{aligned}\mu_\beta &\sim N(-2.197, 2),\\ \mu_\delta &\sim N(4.394, 2)I(\mu_\delta > 0),\\ \mu_\xi &\sim N(3, 2),\\ \boldsymbol{\Sigma}_{\text{item}} &\sim \text{InvWishart}(\mathbf{R}_{\text{item}}, 3),\end{aligned}$$

where $\mathbf{R}_{\text{item}}$ is a tridimensional identity matrix.

The priors of person parameters are set as

$$\begin{pmatrix}\theta_n\\ \tau_n\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \boldsymbol{\Sigma}_{\text{person}}\right).$$

As suggested by Zhan et al. (2017), the Cholesky decomposition of the $\boldsymbol{\Sigma}_{\text{person}}$ is used

$$\boldsymbol{\Sigma}_{\text{person}} = \boldsymbol{\Delta}_{\text{person}}\boldsymbol{\Delta}'_{\text{person}}$$

where

$$\boldsymbol{\Delta}_{\text{person}} = \begin{pmatrix}1 & 0\\ \varphi & \psi\end{pmatrix}$$

is a low triangular matrix with positive entries on the diagonal and unrestricted entries below the diagonal; $\boldsymbol{\Delta}'_{\text{person}}$ is the conjugate transpose of $\boldsymbol{\Delta}_{\text{person}}$. The priors of the elements in $\boldsymbol{\Delta}_{\text{person}}$ are specified as $\varphi \sim N(0, 1)$,$\psi \sim \text{Gamma}(1, 1)$.

Then, the priors of the higher-order structural parameters are specified as

$$\kappa_k \sim N(0, 4), \nu_k \sim N(0, 4)I(\nu_k > 0).$$

In addition, the priors of testlet effect parameters in testlet m are specified as

$$\begin{pmatrix}\gamma_{nm}\\ \lambda_{nm}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \boldsymbol{\Sigma}_{\text{testlet,m}}\right),$$

with the hyper priors of $\boldsymbol{\Sigma}_{\text{testlet,m}} \sim \text{InvWishart}(\mathbf{R}_{\text{testlet,m}}, 2)$, where $\mathbf{R}_{\text{testlet},m}$ is a two-dimensional identity matrix for testlet m.

Finally, the posterior mean and the posterior mode are used as the estimates for the continuous parameters (e.g., β_i, δ_i, θ_n, and τ_n) and categorical parameters (e.g., α_{nk}), respectively.

REAL DATA ANALYSIS

Data

In this study, the PISA 2015 computer-based mathematics data were used. 17 computer-scored dichotomous items from M1 and M2 testing clusters were selected and used in the analysis. The complete-case method was implemented to handle the missing data. That is, only the respondents without missing values in any of the 17 items were used. As a result, the dataset used for analysis contained the dichotomous response data and continuous RT data for 8,606 respondents from 58 countries/economies. The natural logarithm of RTs (i.e., log RTs) were used for modeling. According to the PISA 2015 mathematics assessment framework (OECD, 2016), 11 attributes were assessed, including change and relationships (α1), space and shape (α2), quantity (α3), uncertainty and data (α4), personal (α5), occupational (α6), societal (α7), scientific (α8), formulating situations mathematically (α9), employing mathematical concepts, facts, procedures and reasoning (α10), and interpreting, and applying and evaluating mathematical outcomes (α11). The first four attributes are associated with the mathematical content knowledge that is targeted for use in the items. The next four attributes are associated with the mathematical context that is needed to place additional demands on the problem-solver (Watson and Callingham, 2003; OECD, 2016). The last three attributes are associated with the mathematical processes that connect the context of the mathematics problem with problem-solving (OECD, 2016). In addition, the 17 items contained four testlets, namely, population pyramids (m1), diving (m2), cash withdrawal (m3), and chair lift (m4). Only one source of local item dependence was considered in this study (i.e., an item only belongs to one testlet). The Q-matrix and the U-matrix are presented in **Table 1**.

Analysis

In addition to the joint testlet-DINA model, the joint responses and times DINA (denoted as the JRT-DINA) model (Zhan et al., 2017) was also used to fit the data for comparison purpose. The JRT-DINA model can be seen as a special case of the joint testlet-DINA model where all random testlet effect parameters are set to be zero. For both models, two Markov chains with random starting points were used and 10,000 iterations were run for each chain. The first 5,000 iterations in each chain were discarded as burn-in. In order to save space in memory[1], the thinning interval was set to be five. As a result, 2,000 iterations were retained for model parameter inferences. The potential scale reduction factor (PSRF; Brooks and Gelman, 1998) was computed to assess the convergence of each parameter. PSRF values lower than 1.1 or 1.2 were used as convergence criteria in previous studies (Brooks and Gelman, 1998; de la Torre and Douglas, 2004). In this study, the PSRFs were generally lower than 1.05, indicating good convergence in the specific setting.

The AIC (Akaike, 1974), BIC (Schwarz, 1978), and DIC (Spiegelhalter et al., 2002) were computed for model comparison. Posterior predictive model checking (PPMC; Gelman et al., 2014) was used to evaluate model-data fit. Posterior predictive probability (PPP) values near 0.5 indicate that there are no systematic differences between the observed and predicted values, suggesting an adequate model-data fit. As the research in the absolute model-fit statistics for joint models was limited, this study followed Zhan et al. (2017) to evaluate the model fit of the RA and RT models separately. The sum of the squared Pearson residuals for person n and item i (Yan et al., 2003) was used as a discrepancy measure to evaluate the overall fit of the RA model,

[1] All calculations were conducted on a laptop with 32GB of memory. Insufficient space of memory was caused by no thinning.

TABLE 1 | Q- and U-matrix for PISA 2015 computer-based mathematics items.

Items	Q-matrix											U-matrix			
	α1	α2	α3	α4	α5	α6	α7	α8	α9	α10	α11	m1	m2	m3	m4
CM033Q01		1			1						1				
CM474Q01			1		1					1					
CM155Q01	1							1		1		1			
CM155Q04	1							1			1	1			
CM411Q01			1				1			1			1		
CM411Q02				1			1				1		1		
CM803Q01				1		1			1						
CM442Q02			1				1				1				
CM034Q01		1				1			1						
CM305Q01		1					1			1					
CM496Q01			1				1		1					1	
CM496Q02			1				1			1				1	
CM423Q01				1	1						1				
CM603Q01			1					1		1					
CM571Q01	1							1			1				
CM564Q01			1				1		1						1
CM564Q02				1			1		1						1

Blank means "0."

TABLE 2 | Models fit for PISA 2015 computer-based mathematics.

Model	−2LL	AIC	BIC	DIC	NP	*ppp*_RA	*ppp*_RT
Joint testlet-DINA	**387466**	**387648**	**388291**	**525481**	91	0.486	0.547
JRT-DINA	414438	414596	415154	530742	79	0.521	0.539

−2LL, −2 log-likelihood; AIC, Akaike's information criterion; BIC, Bayesian information criterion; NP, number of parameters; ppp, posterior predictive p-value; RA, item response accuracy; RT, item response time.

which is written as

$$D(Y_{ni};\boldsymbol{\alpha}_n,\beta_i,\delta_i)=\sum_{n=1}^{N}\sum_{i=1}^{I}\left(\frac{Y_{ni}-P(Y_{ni}=1)}{\sqrt{P(Y_{ni}=1)(1-P(Y_{ni}=1))}}\right)^2,$$

where $P(Y_{ni}=1)$ has the same definition as that in Equation (1). On the other hand, the sum of the standardized error function of $\log T_{ni}$ for person n and item i (Marianti et al., 2014; Fox and Marianti, 2017) was used as a discrepancy measure to evaluate the overall fit of the RT model, which is given by

$$D(\log T_{ni};\xi_i,\tau_n,\omega_i)=\sum_{n=1}^{N}\sum_{i=1}^{I}\left(\omega_i(\log T_{ni}-(\xi_i-\tau_n-\sum\nolimits_{m=1}^{M}u_{im}\lambda_{nm}))\right)^2.$$

TABLE 3 | Item mean vector and variance and covariance matrix estimates for PISA 2015 computer-based mathematics items.

	$\boldsymbol{\mu}_{\text{item}}$	Σ_{item}	β	δ	ξ
μ_β	−1.232 (0.278)	β	1.436 (0.558)	−0.645	−0.700
μ_δ	2.394 (0.231)	δ	−0.749 (0.384)	0.938 (0.377)	0.450
μ_ξ	4.197 (0.113)	ξ	−0.408 (0.198)	0.212 (0.146)	0.236 (0.092)

Covariance in lower triangular matrix and correlation coefficient in upper triangular matrix, respectively, in $\boldsymbol{\Sigma}_{item}$; standard error (standard deviation of the posterior distribution) is in parentheses; β, item intercept; δ, item interaction; ξ, item time-intensity.

Results

The joint testlet-DINA model was favored based on the AIC, BIC, and DIC, as is shown in **Table 2**. In addition, the likelihood deviances (i.e., −2 log likelihood or −2LL) of these two models were 387,466 and 414,438, respectively (Δ −2LL = 26,972, df = 12, $p < 0.001$). Therefore, the joint testlet-DINA model fitted the data significantly better than the JRT-DINA model, indicating that paired local item dependence existed among items within testlets. In the joint testlet-DINA model, the PPP values of the RA model and the RT model were 0.486 and 0.547, respectively, which indicated an adequate model-data fit. Thus, only the results pertaining to the joint testlet-DINA model are discussed next (the difference between two models see Figures S1, S2 in Appendix).

Table 3 presents the estimated item mean vector and the estimated item variance-covariance matrix. $\rho_{\beta\delta}$ was estimated to be −0.645, which means that higher item intercept parameters were associated with lower item interaction parameters. $\rho_{\beta\xi}$ and $\rho_{\delta\xi}$ were estimated to be −0.700 and 0.450, respectively, indicating that items with higher intercept parameters tended to have lower time-intensity parameters; by contrast, items with higher interaction parameters tended to be have higher time-intensity parameters. Further, **Figure 2** presents the estimated item parameters. All the β_i estimates were negative except the 1st and the 13th items, which means that the guessing probabilities (i.e., $\frac{\exp(\beta_i)}{1+\exp(\beta_i)}$) of these two items were higher than 0.5.

Table 4 presents the estimated person variance and covariance matrix. $\rho_{\theta\tau}$ was estimated to be −0.196, which means that a low negative correlation was observed between the higher-order ability and the latent speed parameters. The negative correlation was consistent with the results in Zhan et al. (2017). One reasonable explanation is that low-ability respondents lack motivation in taking the low-stakes test (Wise and Kong, 2005). Thus, the low-ability respondents may have shorter RTs and a greater number of incorrect responses than the high-ability respondents. In addition, the variance of latent speed was quite small (i.e., 0.073), which means the variability in latent speed among all respondents was small.

Table 5 presents the four estimated testlet effect variance-covariance matrices. As aforementioned, a larger variance of testlet effect parameters indicates a larger testlet effect. The variances of the four RA testlet effect parameters were estimated to be 0.438, 0.260, 2.800, and 0.414, respectively. Compared to the variance of the latent trait (i.e., 1.00), the RA testlet effects ranged from small to large[2]. By contrast, the variances of the four RT testlet effect parameters were estimated to be 0.110, 0.083, 0.226, and 0.212, respectively. Although the RT testlet effects were small in terms of the absolute values, their ratios to the variance of latent speed (i.e., 0.073) were

TABLE 4 | Person variance and covariance matrix estimates for PISA 2015 computer-based mathematics items.

Σ_{person}	θ	τ
θ	1	−0.196
τ	−0.053 (0.004)	0.073 (0.001)

Covariance in lower triangular matrix and correlation coefficient in upper triangular matrix, respectively; standard error (standard deviation of the posterior distribution) is in parentheses.

[2]According to previous studies (e.g., Wainer and Wang, 2000; Wang and Wilson, 2005), the value of 0.25, 0.5, and 1.0 is corresponding to small, moderate, and large testlet effect, respectively.

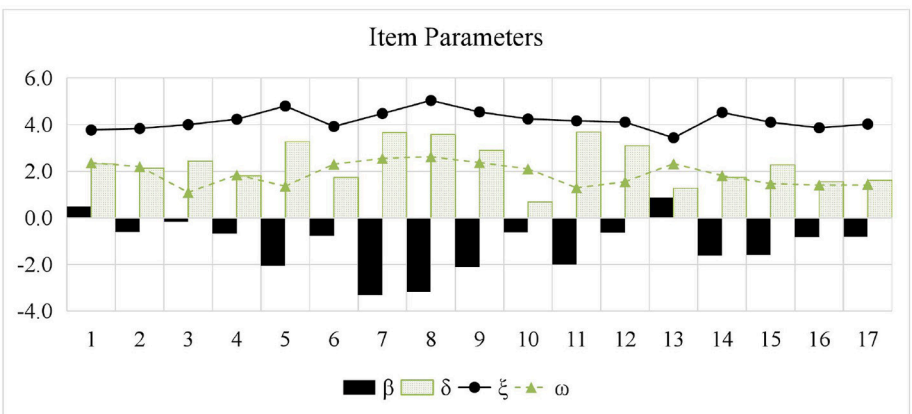


FIGURE 2 | Item parameter estimates for PISA 2015 computer-based mathematics items. β, item intercept; δ, item interaction; ξ, item time-intensity; ω, item time-kurtosis.

TABLE 5 | Testlet effect variance and covariance matrix estimates for PISA 2015 computer-based mathematics items.

	m1: population pyramids		m2: diving		m3: cash withdrawal		m4: chair lift	
$\Sigma_{testlet}$	γ	λ	γ	λ	γ	λ	γ	λ
γ	0.438 (0.072)	−0.268	0.260 (0.070)	−0.065	2.800 (0.220)	0.022	0.414 (0.067)	−0.187
λ	−0.059 (0.012)	0.110 (0.007)	−0.010 (0.010)	0.083 (0.005)	0.018 (0.022)	0.226 (0.008)	−0.056 (0.013)	0.212 (0.008)

Covariance in lower triangular matrix and correlation coefficient in upper triangular matrix, respectively; standard error (standard deviation of the posterior distribution) is in parentheses.

around 1.507, 1.137, 3.096, and 2.904, respectively, indicating that the RT testlet effects were large in this dataset. In addition, low correlation was observed between each pair of RA testlet effect and RT testlet effect, indicating that these two types of testlet effects were separable. This is an unexpected result. A moderate or a high correlation was expected since, theoretically speaking, local RA dependence and local RT dependence should be caused by the same stimulus. More practical evidence needs to be accumulated from future studies to explain the results.

Figure 3 presents the posterior mixing proportions of the 20 most frequent attribute patterns out of the 2,048 possible attribute patterns. Only 73 patterns were observed in the estimated attribute profiles. Attribute pattern (11111111111) was the most prevalent with a percentage of 40.19%; the second most prevalent pattern was (10100100000) with a percentage of 23.41%.

A BRIEF SIMULATION STUDY

Design and Data Generation

A brief simulation study was conducted to examine the parameter recovery of the proposed model and the consequence of ignoring the potential paired local item dependence in analysis. The simulated dataset contained 1,000 respondents and 30 items measuring five attributes. The Q-matrix is presented in **Figure 4**. The last 20 items were evenly divided into 4 testlets. Specifically, testlet 1 consisted of items 11 $\sim$ 15, testlet 2 consisted of items 16 $\sim$ 20, testlet 3 consisted of items 21 $\sim$ 25, and testlet 4 consisted of items 26 $\sim$ 30. For simplicity, the four pairs of RA and RT testlet effects were generated from a same bivariate normal distribution,

$$\begin{pmatrix}\gamma_{nm}\\ \lambda_{nm}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.50 & \\ -0.25 & 0.50\end{pmatrix}\right),$$

where $\rho_{\gamma\lambda} = -0.5$. Typically, setting the testlet effect as 0.5 indicates a moderate testlet effect (Wang and Wilson, 2005; Wainer et al., 2007). In addition, each item was assumed to belong to only one testlet. Item parameters were generated from a trivariate normal distribution,

$$\begin{pmatrix}\beta_i\\ \delta_i\\ \xi_i\end{pmatrix} \sim N\left(\begin{pmatrix}-2.197\\ 4.394\\ 4.000\end{pmatrix}, \begin{pmatrix}1.00 & & \\ -0.80 & 1.00 & \\ -0.25 & 0.15 & 0.25\end{pmatrix}\right),$$

where $\rho_{\beta\delta} = -0.8$, $\rho_{\beta\xi} = -0.5$, and $\rho_{\delta\xi} = 0.3$, which were set according to the estimates from the real data analysis (Zhan et al., 2017); ω_i were generated from $N(2, 0.25)$. Person parameters

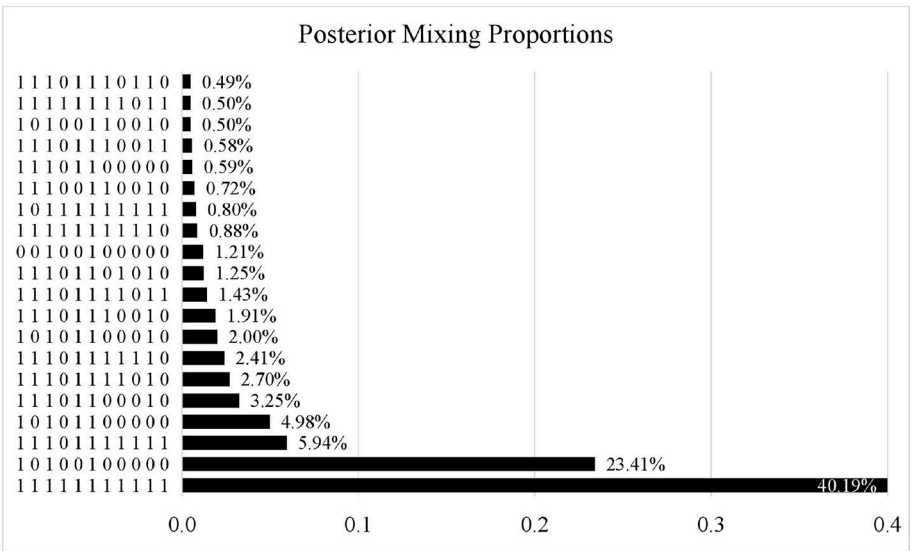


FIGURE 3 | Posterior mixing proportions for PISA 2015 computer-based mathematics items. only the 20 most frequent attribute patterns are displayed.

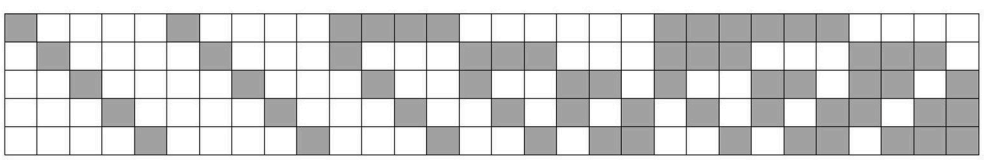


FIGURE 4 | *K*-by-*I* Q' matrix for simulation study. blank means "0," gray means "1".

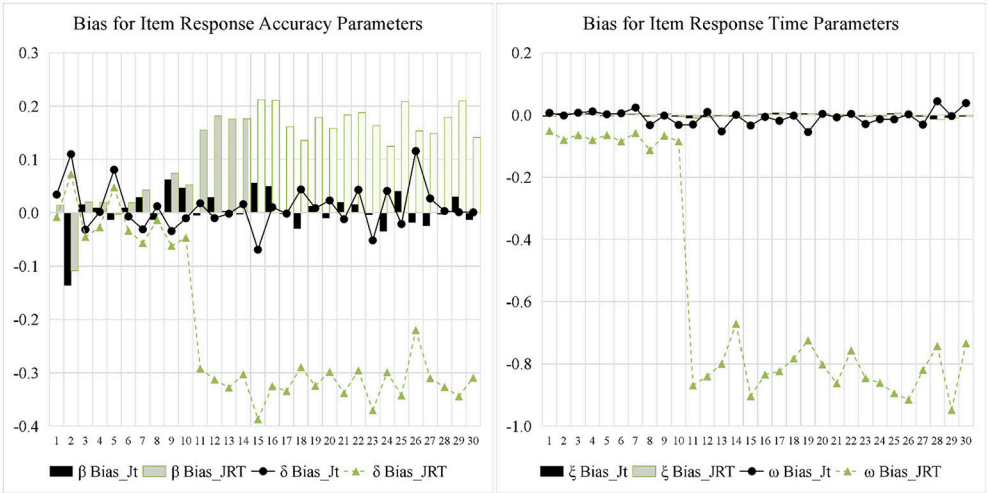


FIGURE 5 | Bias for item parameter in simulation study. Jt, joint testlet-DINA model; JRT, JRT-DINA model; β, item intercept; δ, item interaction; ξ, item time-intensity; ω, item time-kurtosis.

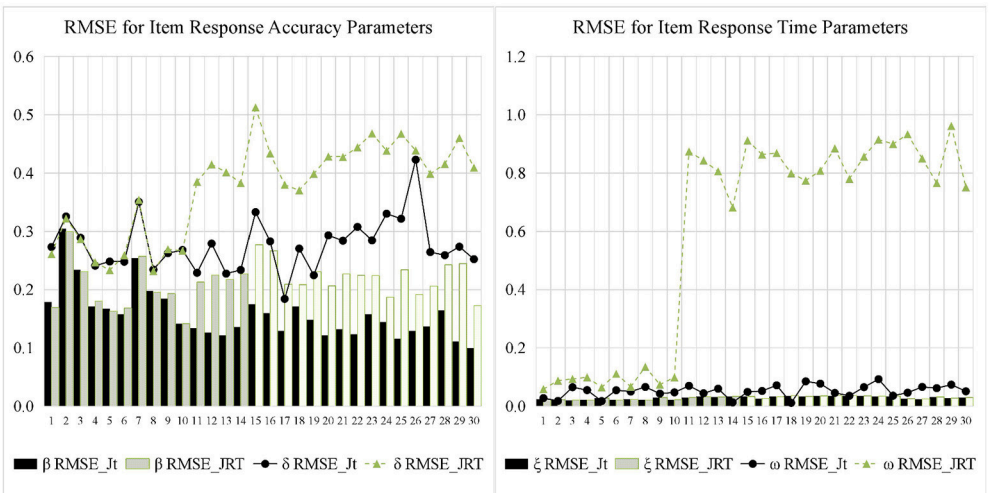


FIGURE 6 | Root mean square error (RMSE) for item parameter in simulation study. Jt, joint testlet-DINA model; JRT, JRT-DINA model; β, item intercept; δ, item interaction; ξ, item time-intensity; ω, item time-kurtosis.

were generated from a bivariate normal distribution,

$$\begin{pmatrix}\theta_n\\ \tau_n\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.00 & \\ -0.25 & 0.25\end{pmatrix}\right),$$

where $\rho_{\theta\tau} = -0.5$. For higher-order structural parameters, $\nu_k = 1.5$ for all the attributes and $\boldsymbol{\kappa}_k = (-1.0, -0.5, 0.0, 0.5, 1.0)$, indicating moderate correlations among attributes. The mastery status of each person on each attribute was generated from a Bernoulli distribution with the parameter, $P(\alpha_{nk} = 1)$ which was computed based on Equation (5).

Analysis

Thirty replications were implemented. Both the joint testlet-DINA model and the JRT-DINA model were fit to the simulated data. In each replication, the number of chains, burn-in iterations, and post-burn-in iterations were consistent with those in the real data analysis. Convergence was well achieved (see Figure S3 in Appendix). The bias and root mean square error (RMSE) were used to evaluate parameter recovery, which were calculated as $bias(\hat{\upsilon}) = \sum_{r=1}^{R} \frac{\hat{\upsilon}_r - \upsilon}{R}$ and $RMSE(\hat{\upsilon}) = \sqrt{\sum_{r=1}^{R} \frac{(\hat{\upsilon}_r - \upsilon)^2}{R}}$, where $\hat{\upsilon}$ and υ are the estimated and true value of model parameters, respectively; R is the number of replications. In addition, the correlation between the true and estimated value of model parameters was computed. In terms of the classification accuracy, the attribute correct classification rate (ACCR) and pattern correct classification rate (PCCR) were computed as ACCR $= \frac{\sum_{r=1}^{R}\sum_{n=1}^{N} W_{nk}}{R \times N}$ and PCCR $=$

$\frac{\sum_{r=1}^{R}\sum_{n=1}^{N}\prod_{k=1}^{K}W_{nk}}{R\times N}$, where $W_{nk}=1$ if $\alpha_{nk}=\hat{\alpha}_{nk}$, and $W_{nk}=0$ otherwise.

Results

In all the 30 replications, the joint tesltet-DINA model was favored by AIC, BIC and DIC, which indicates that the three fit indices can select the best-fitting model correctly.

Figures 5, **6** display the recovery of the item parameters for the two models. According to the results of the last 20 items with testlet structure, the performance of the JRT-DINA model was significantly affected by the paired local item dependence. Specifically, ignoring paired local item dependence in analysis would result in overestimation of item intercept parameters, underestimation of item interaction parameters, and underestimation of item time-kurtosis parameters. However, it had little effect on the recovery of item time-intensity parameters. In addition, most of the 10 items without testlet structure had smaller absolute bias in parameter estimates from the joint testlet-DINA model than from the JRT-DINA model; the RMSE of the parameter estimates from the joint testlet-DINA model was equal to or smaller than those from the JRT-DINA model. **Table 6** further summarizes the item parameter recovery by presenting the mean absolute bias, the mean RMSE, and the correlation between estimated and true values of all the items. Again, it can be seen that ignoring the paired local item dependence mainly affected the recovery of item time-kurtosis parameters. In addition, the item RT parameters were recovered better than the item RA parameters in joint models.

Figures 7, **8** display the recovery of the person parameters for the two models. The two models performed similarly on recovering the higher-order ability parameter. In terms of the latent speed parameters, the bias was similar for the two models, but the RMSE from the JRT-DINA model was significantly larger than that from the joint testlet-DINA model. The results indicate that ignoring the paired local item dependence in analysis would result in large variability in latent speed parameters but had little effect on the recovery of higher-order ability parameters. **Table 7** further summarizes the recovery of person parameters. The two models mainly differed in the mean RMSE of latent speed across person. In addition, the recovery of latent speed parameters was better than that of the higher-order ability parameters.

Table 8 presents the recovery of individual attributes and attribute patterns. The joint testlet-DINA model was higher than the JRT-DINA model in both ACCR and PCCR, which indicates that ignoring the paired local item dependence would slightly reduce attribute and pattern correct classification rates (PCCRs).

Table 9 presents the recovery of item, person and testlet variance-covariance matrices. First, in terms of the item variance-covariance matrix, the bias was similar for the two models, but

TABLE 6 | Summary of the item parameter recovery in simulation study.

Index	Item parameter	Joint testlet-DINA	JRT-DINA
MA_Bias	β	0.025	0.127
	δ	0.029	0.225
	ξ	0.004	0.004
	ω	**0.017**	**0.572**
M_RMSE	β	0.158	0.214
	δ	0.277	0.374
	ξ	0.029	0.029
	ω	**0.052**	**0.591**
Correlation	β	0.986	0.982
	δ	0.958	0.946
	ξ	0.999	0.999
	ω	**0.973**	**0.123**

MA_Bias, mean absolute value of bias across all items; M_RMSE, mean value of root mean square error across all items; Correlation, correlation between estimated and true values of all items; β, item intercept; δ, item interaction; ξ, item time-intensity; ω, item time-kurtosis; JRT-DINA, joint responses and times DINA model.

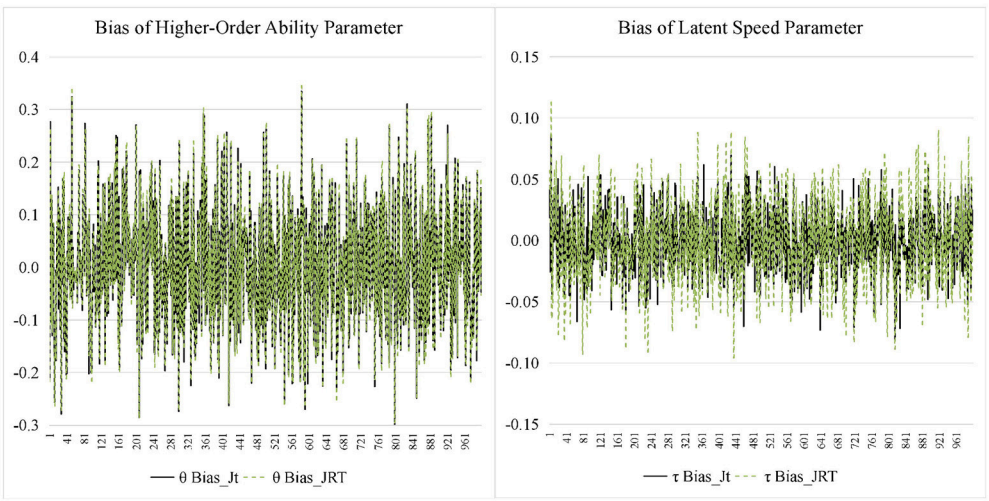


FIGURE 7 | Bias for person parameter in simulation study. Jt, joint testlet-DINA model; JRT, JRT-DINA model; θ, higher-order latent ability; τ, latent speed.

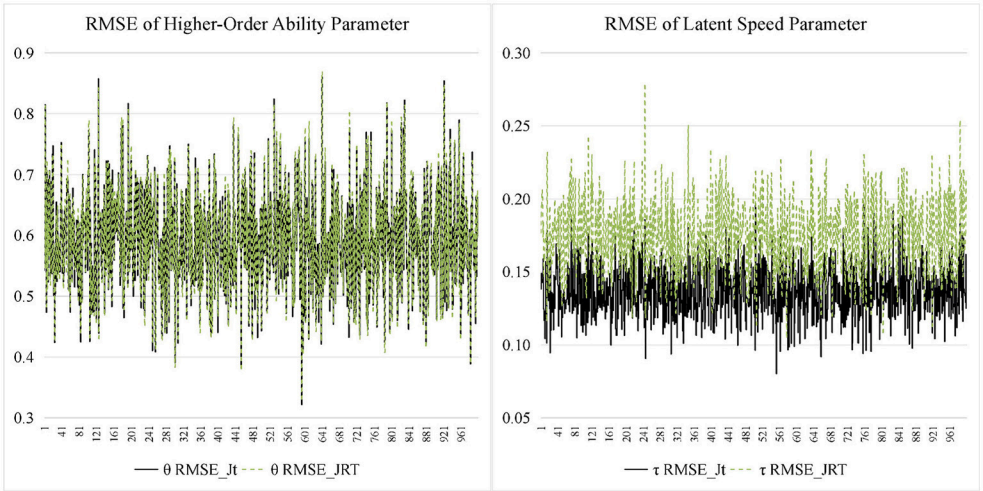


FIGURE 8 | Root mean square error (RMSE) for person parameter in simulation study. Jt, joint testlet-DINA model; JRT, JRT-DINA model; θ, higher-order latent ability; τ, latent speed.

TABLE 7 | Summary of the person parameter recovery in simulation Study.

Index	Item parameter	Joint testlet-DINA	JRT-DINA
MA_Bias	θ	0.088	0.088
	τ	0.020	0.026
M_RMSE	θ	0.593	0.595
	τ	**0.137**	**0.175**
Cor	θ	0.803	0.801
	τ	0.961	0.939

MA_Bias, mean absolute value of bias across all persons; M_RMSE, mean value of root mean square error across all persons; Correlation, correlation between estimated and true values of all persons; θ, higher-order latent ability; τ, latent speed.

TABLE 8 | Attribute and pattern correct classification rate in simulation study.

Analysis model	ACCR					PCCR
	α1	α2	α3	α4	α5	
Joint testlet-DINA	0.974	0.961	0.968	0.973	0.980	0.872
JRT-DINA	0.974	0.961	0.967	0.973	0.979	0.870

ACCR, attribute correct classification rate; PCCR, pattern correct classification rate.

the RMSE from the joint testlet-DINA model was larger than that from the JRT-DINA model. Second, the latent speed variance was recovered better in the joint testlet-DINA model than in the JRT-DINA model. Third, all the four testlet variance-covariance matrices were well recovered. The recovery of the RT testlet effect variance parameters was better than that of the RA testlet effect variance parameters.

Table 10 presents the recovery of item mean vector components and higher-order structural parameters. The item mean vector component estimates from the joint testlet-DINA model had smaller absolute bias and RMSE than those from the JRT-DINA model. The two models performed similarly on recovering the higher-order structural parameters. The results indicate that ignoring the paired local item dependence in analysis would result in less precise item mean vector component estimates, but had little effect on the higher-order structural parameter recovery.

Overall, the model parameters of the joint testlet-DINA model were well recovered by using the proposed MCMC estimation algorithm. Additionally, ignoring the paired local item dependence in analysis would result in biased model parameter estimates and lower correct classification rates. Specifically, it would result in overestimation of item intercept parameters, underestimation of item interaction parameters, and underestimation of item time-kurtosis parameters. It would lead to less precise estimates of latent speed parameters and item mean vector components. It would also reduce attribute and PCCRs. However, it had little effect on the recovery of item time-intensity parameters, the higher-order ability parameters, or the higher-order structural parameters.

CONCLUSION AND DISCUSSION

To address the paired local item dependence in RT and RA when applying the joint CDMs, this study proposed a joint testlet cognitive diagnosis modeling approach. As an extension of the joint cognitive diagnosis modeling approach (Zhan et al., 2017), the proposed approach modeled the relationship between each pair of RA testlet effect and RT testlet effect using correlational structure. Specifically, the testlet-DINA model and the within-item multidimensional testlet effects lognormal RT model were adopted as the RA model and RT model, respectively. The model parameters were estimated using the full Bayesian MCMC method. The 2015 PISA computer-based mathematics data were

TABLE 9 | Recovery of Variance and Covariance Matrices in Simulation Study.

Parameter			Joint testlet-DINA		JRT-DINA	
			Bias	RMSE	Bias	RMSE
$\boldsymbol{\Sigma}_{\text{item}}$	Variance of intercept	σ^2_{β}	0.043	0.230	−0.032	0.208
	Covariance of intercept and interaction	$\sigma_{\beta\delta}$	−0.001	0.211	0.057	0.202
	Covariance of intercept and time-intensity	$\sigma_{\beta\xi}$	−0.008	0.096	0.004	0.089
	Variance of interaction	σ^2_{δ}	0.035	0.253	−0.033	0.242
	Covariance of interaction and time-intensity	$\sigma_{\delta\xi}$	0.007	0.099	−0.004	0.089
	Variance of time-intensity	σ^2_{ξ}	0.062	0.088	0.062	0.088
$\boldsymbol{\Sigma}_{\text{person}}$	Covariance of ability and speed	$\sigma_{\theta\tau}$	0.004	0.020	0.003	0.021
	Variance of speed	σ^2_{τ}	−0.001	0.010	0.019	0.022
$\boldsymbol{\Sigma}_{\text{testlet},1}$	Variance of 1st RA testlet effect	$\sigma^2_{\gamma 1}$	−0.010	0.098		
	Covariance of 1st pair of testlet effects	$\sigma_{\gamma 1\lambda.1}$	0.008	0.048		
	Variance of 1st RT testlet effect	$\sigma^2_{\lambda.1}$	0.002	0.032		
$\boldsymbol{\Sigma}_{\text{testlet},2}$	Variance of 2nd RA testlet effect	$\sigma^2_{\gamma 2}$	0.013	0.104		
	Covariance of 2ndt pair of testlet effects	$\sigma_{\gamma 2\lambda.2}$	0.000	0.038		
	Variance of 2nd RT testlet effect	$\sigma^2_{\lambda.2}$	0.009	0.028		
$\boldsymbol{\Sigma}_{\text{testlet},3}$	Variance of 3rd RA testlet effect	$\sigma^2_{\gamma 3}$	0.005	0.108		
	Covariance of 3rd pair of testlet effects	$\sigma_{\gamma 3\lambda.3}$	0.009	0.034		
	Variance of 3rd RT testlet effect	$\sigma^2_{\lambda.3}$	0.006	0.025		
$\boldsymbol{\Sigma}_{\text{testlet},4}$	Variance of 4th RA testlet effect	$\sigma^2_{\gamma 4}$	0.014	0.107		
	Covariance of 4th pair of testlet effects	$\sigma_{\gamma 4\lambda.4}$	−0.003	0.041		
	Variance of 4th RT testlet effect	$\sigma^2_{\lambda.4}$	0.008	0.028		

RMSE, root mean square error.

TABLE 10 | Recovery of item mean vector and higher-order structural parameters.

Parameter			Joint testlet-DINA		JRT-DINA	
			Bias	RMSE	Bias	RMSE
$\boldsymbol{\mu}_{\text{item}}$	Mean intercept	μ_{β}	−0.001	0.178	0.110	0.204
	Mean interaction	μ_{δ}	−0.006	0.214	−0.229	0.311
	Mean time-intensity	μ_{ξ}	0.012	0.094	0.013	0.095
$\boldsymbol{\kappa}$	Difficulty of attribute 1	κ_1	0.008	0.114	0.008	0.119
	Difficulty of attribute 2	κ_2	0.002	0.108	0.002	0.107
	Difficulty of attribute 3	κ_3	−0.010	0.111	−0.010	0.111
	Difficulty of attribute 4	κ_4	0.048	0.121	0.048	0.119
	Difficulty of attribute 5	κ_5	0.007	0.103	0.002	0.099
$\boldsymbol{\nu}$	Slope of attribute 1	ν_1	−0.006	0.151	−0.006	0.150
	Slope of attribute 2	ν_2	0.049	0.191	0.052	0.198
	Slope of attribute 3	ν_3	−0.007	0.190	−0.008	0.189
	Slope of attribute 4	ν_4	0.105	0.230	0.106	0.227
	Slope of attribute 5	ν_5	−0.056	0.170	−0.060	0.168

RMSE, root mean square error.

analyzed to demonstrate the application of the proposed model. The real data analysis results are summarized as follows: (a) a negative correlation was observed between the higher-order ability and latent speed; (b) a negative correlation was observed between the item intercept parameters and the item time-intensity parameters; (c) a positive correlation was observed between the item interaction parameters and the item time-intensity parameters; (d) the magnitude of RA testlet effects varied from small to large whereas the magnitude of RT testlet effects was large; and (e) low correlation coefficients between the RA and RT testlet effects were found. Overall, most results in this real data analysis were consistent with those in Zhan et al. (2017) that used PISA 2012 computer-based mathematics data. Further, a simulation study was conducted to examine model parameter recovery of the proposed model and the consequence of ignoring testlet effects. The results indicated that the model parameters of the proposed model can be well recovered. Additionally, ignoring the paired local item dependence in analysis would result in biased model parameter estimates and low individual correct classification rates.

Despite the promising results, further research is needed. First, only a DINA-based testlet model and a lognormal RT-based testlet model were used for illustration in this study. In the future study, other CDMs (e.g., von Davier, 2008; Henson et al., 2009; de la Torre, 2011) and RT models (e.g., Klein Entink et al., 2009b; Wang et al., 2013) can be used as the measurement models of RA and RTs. Second, in this study, the proposed model was evaluated using a brief simulation where only a limited number of factors were manipulated. More factors (e.g., test length, number of attributes, magnitude of testlet effects, etc.) and replications are recommended in future studies. Third, the

model-data fit of RA and RT models was evaluated separately because of the lack of model-data fit indices for the joint models. In the future studies, absolutely model-fit indices designed for joint models can be explored and further be applied to evaluate the current modeling approach. Fourth, in educational and psychological measurements, latent speed can be defined as the ratio of the amount of labor spent on the items with respect to time (van der Linden, 2011). Due to the multidimensional nature of labors, latent speed may also be a multidimensional concept, each dimension of which corresponds to a specific type of labor. The latent speed was treated as a unidimensional latent trait in this study although the RT testlet effect can be regarded as a specific factor that is relevant to the working speed. Recently, Zhan et al., Manuscript submitted for publication proposed a multidimensional lognormal RT model to account for the potential multidimensionality of latent speed. One possible extension of the current joint modeling approach is to account for the multidimensional latent speed. Fifth, as noted by one of the anonymous reviewers, if there are many testlets, there will be many bivariate covariance matrices to be estimated, leading to large computational burden. Further exploration is needed to deal with this challenging issue. Sixth, in this study, respondents were assumed to be from the same population group, but, in reality, they may be from different groups (e.g., male and female). Multiple group joint modeling (e.g., Jiao et al., 2017) and mixture modeling (e.g., von Davier, 2008) can be incorporated into the current modeling approach in the future. Seventh, in practice, students are nested within classrooms, and classrooms are further nested within schools. Thus, multilevel modeling (e.g., Fox and Glas, 2001; Jiao et al., 2012; Jiao and Zhang, 2015) extension can also be a future direction. Finally, the generalizability of the results from this study is limited given that only data from a low-stakes test were analyzed. More empirical studies based on data from other tests, especially high-stakes tests, are needed.

AUTHOR CONTRIBUTIONS

PZ: Design of the study, data analysis, paper writing, and revision; ML: Data cleaning, interpretation of data for the work, and paper revision; YB: Preliminary idea construction, paper revision, and proofreading.

FUNDING

The first author was sponsored by the Graduate Student Self-manage Project from Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University (No. BJSP-2016A1-15001).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00607/full#supplementary-material

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *Autom. Control IEEE Trans.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika* 64, 153–168. doi: 10.1007/BF02294533

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comp. Graph. Stat.* 7, 434–455.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika* 75, 581–612. doi: 10.1007/s11336-010-9178-0

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Appl. Psychol. Measur.* 35, 8–26. doi: 10.1177/0146621610377081

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *J. Educ. Measur.* 43, 145–168. doi: 10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2012). Confirming testlet effects. *Appl. Psychol. Measur.* 36, 104–121. doi: 10.1177/0146621612437403

Fox, J. P., and Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66, 271–288. doi: 10.1007/BF02294839

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer.

Fox, J.-P., and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivar. Behav. Res.* 51, 540–553. doi: 10.1080/00273171.2016.1171128

Fox, J.-P., and Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *J. Educ. Measur.* 54, 243–262. doi: 10.1111/jedm.12143

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85, 398–409. doi: 10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Measur.* 26, 301–321. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hansen, M. (2013). *Hierarchical Item Response Models for Cognitive Diagnosis*. Doctoral dissertation, University of California, LA.

Hansen, M., Cai, L., Monroe, S., and Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *Br. J. Math. Stat. Psychol.* 69, 225–252. doi: 10.1111/bmsp.12074

Henson, R., Templin, J., and Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Im, K. S. (2017). *The Hierarchical Testlet Response Time Model: Bayesian Analysis of a Testlet Model for Item Responses and Response Times*. Doctoral dissertation, University of Kansas, Lawrence, KS.

Jiao, H., and Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *Br. J. Math. Stat. Psychol.* 68, 65–83. doi: 10.1111/bmsp.12035

Jiao, H., Kamata, A., Wang, S., and Jin, Y. (2012). A multilevel testlet model for dual local dependence. *J. Educ. Measur.* 49, 82–100. doi: 10.1111/j.1745-3984.2011.00161.x

Jiao, H., Wang, S., and He, W. (2013). Estimation methods for one-parameter testlet models. *J. Educ. Measur.* 50, 186–203. doi: 10.1111/jedm.12010

Jiao, H., Zhan, P., and Liao, M. (2017). "A joint multigroup testlet model for responses and response times accounting for differential item and time functioning," in *Invited Talk at the Fifth Conference on the Statistical Methods in Psychometrics in Columbia University* (New York, NY).

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Measur.* 25, 258–272. doi: 10.1177/01466210122032064

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika* 74, 21–48. doi: 10.1007/s11336-008-9075-y

Klein Entink, R. H., van der Linden, W. J., and Fox, J.-P. (2009b). A Box-Cox normal model for response times. *Br. J. Math. Stat. Psychol.* 62, 621–640. doi: 10.1348/000711008X374126

Li, Y., Bolt, D. M., and Fu, J. (2006). A comparison of alternative models for testlets. *Appl. Psychol. Measur.* 30, 3–21. doi: 10.1177/0146621605275414

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. New York, NY: Oxford University Press.

Macready, G. B., and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *J. Educ. Behav. Stat.* 2, 99–120. doi: 10.3102/10769986002002099

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., and Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *J. Educ. Behav. Stat.* 39, 426–451. doi: 10.3102/1076998614559412

Meng, X.-B., Tao, J., and Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *J. Educ. Measur.* 52, 1–27. doi: 10.1111/jedm.12060

Minchen, N. (2017). *Continuous Response in Cognitive Diagnosis Models: Response Time Modeling, Computerized Adaptive Testing, and Q-matrix Validation.* Doctoral dissertation, Rutgers, The State University of New Jersey.

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *Br. J. Math. Stat. Psychol.* 68, 197–219. doi: 10.1111/bmsp.12042

OECD (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Paris: PISA, OECD Publishing.

Plummer, M. (2015). *JAGS Version 4.0.0 User Manual.* Lyon. Available online at: http://sourceforge.net/projects/mcmc-jags/

Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI Monogr. Ser.* 4, 59–74.

Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sireci, S. G., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *J. Educ. Meas.* 28, 237–247. doi: 10.1111/j.1745-3984.1991.tb00356.x

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B.* 64, 583–616. doi: 10.1111/1467-9868.00353

Suh, H. (2010). *A Study of Bayesian Estimation and Comparison of Response Time Models in Item Response Theory.* Doctoral dissertation, University of Kansas, Lawrence, KS.

Tao, W., and Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Appl. Meas. Educ.* 29, 108–121. doi: 10.1080/08957347.2016.1138956

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Measur.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Thissen, D. (1983). "Timed testing: an approach using item response theory," in *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing,* ed D. J. Weiss (New York, NY: Academic Press), 179–203.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *J. Educ. Measur.* 46, 247–272. doi: 10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J. (2011). Test design and speededness. *J. Educ. Measur.* 48, 44–60. doi: 10.1111/j.1745-3984.2010.00130.x

Verhelst, N. D., Verstralen, H. H. F. M., and Jansen, M. G. (1997). "A logistic model for time-limit tests," in *Handbook of Modern Item Response Theory,* eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 169–185.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* 61, 287–307. doi: 10.1348/000711007X193957

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: two variants of a model equivalency. *Br. J. Math. Stat. Psychol.* 67, 49–71. doi: 10.1111/bmsp.12003

Wainer, H., Bradlow, E. T., and Wang, X. H. (2007). *Testlet Response Theory and Its Applications*. New York, NY: Cambridge University Press.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *J. Educ. Measur.* 24, 185–201. doi: 10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., and Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *J. Educ. Meas.* 37, 203–220. doi: 10.1002/j.2333-8504.2001.tb01851.x

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Chang, H., and Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

Wang, T., and Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Appl. Psychol. Measur.* 29, 323–339. doi: 10.1177/0146621605275984

Wang, W.-C., and Wilson, M. (2005). The Rasch testlet model. *Appl. Psychol. Measur.* 29, 126–149. doi: 10.1177/0146621604271053

Watson, J. M., and Callingham, R. (2003). Statistical literacy: a complex hierarchical construct. *Stat. Educ. Res. J.* 2, 3–46.

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Measur. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Yan, D., Mislevy, R. J., and Almond, R. G. (2003). *Design and Analysis in a Cognitive Assessment (ETS Research Report Series, RR-03-32).* Princeton, NJ: ETS.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *J. Educ. Measur.* 30, 187–213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Zhan, P. (2017). Using JAGS for Bayesian cognitive diagnosis models: a tutorial. *arXiv preprint arXiv:1708.02632*. Available online at: https://arxiv.org/abs/1708.02632

Zhan, P., Jiao, H., and Liao, D. (2017). Cognitive diagnosis modeling incorporating item response times. *Br. J. Math. Stat. Psychol.* doi: 10.1111/bmsp.12114

Zhan, P., Li, X., Wang, W.-C., Bian, Y., and Wang, L. (2015). The multidimensional testlet-effect cognitive diagnostic models. *Acta Psychol. Sin.* 47, 689–701. doi: 10.3724/SP.J.1041.2015.00689

Zhan, P., Wang, W.-C., Wang, L., and Li, X. (2014). The multidimensional testlet-effect Rasch model. *Acta Psychol. Sin.* 46, 1208–1222. doi: 10.3724/SP.J.1041.2014.01208

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Zhan, Liao and Bian. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 May 2019
doi: 10.3389/fpsyg.2019.01011

Characterizing Interactive Communications in Computer-Supported Collaborative Problem-Solving Tasks: A Conditional Transition Profile Approach

Jiangang Hao and Robert J. Mislevy*

Educational Testing Service, Princeton, NJ, United States

Communication in a collaborative problem-solving activity plays a pivotal role in the success of the collaboration in both academia and the workplace. Computer-supported collaboration makes it possible to collect large-scale communication data to investigate the process at a finer granularity. In this paper, we introduce a conditional transition profile (CTP) to characterize aspects of each team member's communication. Based on the data from a large-scale empirical study, we found that participants in the same team tend to show similar CTP compared to participants from different teams. We also found that team members who showed more "negotiation" after the partner "shared" information tended to show more improvement after the collaboration while those who continued sharing ideas while their partners were negotiating tended to improve less.

Keywords: collaborative problem solving, communication, transition matrix, stochastic process, assessment

OPEN ACCESS

Edited by:
Jason C. Immekus,
University of Louisville, United States

Reviewed by:
Mark Billinghurst,
University of South Australia, Australia
Bernard Veldkamp,
University of Twente, Netherlands

****Correspondence:***
Jiangang Hao
jhao@ets.org

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *11 September 2018*
Accepted: *16 April 2019*
Published: *08 May 2019*

Citation:
Hao J and Mislevy RJ (2019) Characterizing Interactive Communications in Computer-Supported Collaborative Problem-Solving Tasks: A Conditional Transition Profile Approach. Front. Psychol. 10:1011. doi: 10.3389/fpsyg.2019.01011

1. INTRODUCTION

Technology advancement allows computer-supported collaboration to be widely adopted in both academia and the workplace. Compared to face-to-face collaboration, online collaboration significantly reduces the effort and cost of organizing joint work, making it ideal for a wide range of collaborative activities (Stahl et al., 2006). The communication data in computer-supported collaboration contain rich information regarding the collaboration process. Understanding the communication process will help to identify pathways to more successful collaboration outcomes. Such knowledge can further inform the development of real-time facilitation or intervention mechanisms to scaffold the collaboration.

The analysis of communication data (or discourse analysis as it is often called in the computer-supported collaborative learning (CSCL) community) usually starts with the coding or labeling of each turn (or several turns that constitute large speech units) of communications based on a framework (rubrics) being developed to address specific research questions. For example, a number of coding frameworks have been developed to analyze different aspects of the communications among team members, such as the coding framework for collaborative problem solving (CPS) skills (Liu et al., 2015), for the interactive patterns in collaboration (Andrews et al., 2017), for cohesion and language (Graesser et al., 2004; Dowell et al., 2016), and for dialog acts (Allen and Core, 1997). Based on human-coded discourse, natural language processing (NLP) techniques can

be employed to automate the annotation to an accuracy level that is close to human coding (Rosé et al., 2008; Rus et al., 2015; Flor et al., 2016; Hao et al., 2017a).

The codings of discourses are numerical representations of the communication data and can be used as input variables for developing higher level feature representations of the communication process, or for developing statistical models of the process. Given that the communication data and codings often involve multiple interacting team members, it is of interest to develop feature variables that characterize both team performance and individual performance. Traditional discourse analysis usually uses the frequency of different codings (e.g., Dowell et al., 2016) or sequence of codings (e.g., Hao et al., 2016) as the high-level representations of the communication. However, such representations fail to capture the information of how a specific member responds to different types of utterances from others throughout the communication process. To address this issue, in this paper, we introduce a conditional transition profile (CTP) approach to form representations of each team member's responses to different types of utterances (based on a given coding framework) from other members. In collaborative work, what one member says is important, but how a member responds to the others' utterances may contain more information about the member's skills in collaboration. The CTP approach provides a quantitative measure of how a team member responds to other team members. To illustrate the effectiveness of the method, we apply the CTP to data collected through a large-scale online collaborative task from the ETS collaborative science assessment prototype (ECSAP) project and show an example of how the team members' CTPs were related to their performance improvements after the collaboration.

2. CONDITIONAL TRANSITION PROFILE

Suppose we have a coding framework that has k different categories, the t-th turn of the communication can be characterized by a k dimensional state vector $\mathbf{X}_t$, with elements either 0 or 1, indicating whether a given category is assigned to this turn of discourse[1]. For coding frameworks that require mutually exclusive codings, the state vector will have only one element as 1 and all others as 0. The states in a communication process can be considered from both the team level and the individual level. At each level, the most straightforward measure is the cumulative counts of the different states. A CPS profile based on the counts of states at the team level has been introduced to characterize the overall collaboration process of the team (Hao et al., 2016). In this CPS profile, we considered the counts of different states (unigram) and consecutive state pairs (bigram), though the approach can be extended to include the counts of n sequential states (n-gram). It has been shown that different CPS profiles are related to different collaboration outcomes of the team (Hao et al., 2016).

In the current paper, we further generalize the CPS profile from characterizing the whole team process to characterizing each team member's communication process. The most straightforward way to generalize the CPS profile is the direct counts of different states from each team member instead of all the team members. However, in a communication, what one member (target team member) says depends heavily on the other members' preceding discourses. As such, counting the states of a target team member by conditioning on other partners' preceding discourse states should encode more information about the individual's communicative moves in context than merely counting all the states together. As such, we introduce a conditional transition profile for each team member as follows.

TABLE 1 | Conditional transition profile of the communication.

	State 1	**State 2**	**State 3**	**· · ·**
State 1	N_{11}	N_{12}	N_{13}	· · ·
State 2	N_{21}	N_{22}	N_{23}	· · ·
State 3	N_{31}	N_{32}	N_{33}	· · ·
· · ·	· · ·	· · ·	· · ·	· · ·

The columns correspond to the states of the discourse from the targeted team member and the rows correspond to the states of preceding discourses from the other team members.

For a sequence of coded discourses[2], we can represent the states of communication in **Table 1**, where the column name indicates the states of the discourse from the targeted team member and the rows indicate the states of the discourse from the most immediate preceding discourse category from other team members. The numbers in the cells are the counts of the occurrences of the states specified by the corresponding row and column names. It is worth noting that we consider only the most immediate turns of discourses and ignore longer range dependency, though the extension to longer range dependency is straightforward. The reason for doing this is that the majority of short online conversations do not display long range dependency (some empirical evidence of this can be found in Hao et al., 2017a). The elements of a CTP are defined as follows,

$$C(D_i|\bar{D}_j) = N_{ij} \tag{1}$$

where D_i denotes the state (coding category) i of the discourse from the targeted team member and $\bar{D}_j$, denotes the state j of the immediately preceding discourse from other team members. Here i runs for the columns and j runs for the rows. N_{ij} is the count of occurrences of the state in the corresponding cell. Note that this matrix is very similar to the (weighted) adjacency matrix widely used in graph theory, except that the latter is traceless (Biggs, 1993).

In many practical applications, the relative ratios of the categories are often considered important. A representation of the ratios can be obtained by normalizing each cell of the table by the sum of its row.

$$T(D_i|\bar{D}_j) = N_{ij} / \left(\sum_i N_{ij} \right) \tag{2}$$

[1]In practice, the categories or states are assigned either by human coders or automated coding algorithms.

[2]**Table 2** shows an empirical example of coded discourses.

We call this the normalized CTP. In practice, as some elements could be zero due to a small sample size, so smoothing techniques, such as Laplace smoothing (Schütze et al., 2008), can be used to estimate the elements of the normalized CTP as follows,

$$T(D_i|\bar{D}_j) = \frac{N_{ij} + \alpha}{\sum_i N_{ij} + \alpha k} \tag{3}$$

where $\alpha > 0$ is a smoothing parameter. We call the $C(D_i|\bar{D}_j)$ as conditional transition profile and $T(D_i|\bar{D}_j)$ as normalized conditional transition profile. Generally speaking, the $C(D_i|\bar{D}_j)$ contains more information than $T(D_i|\bar{D}_j)$ as the latter can be derived from the former but not the other way around. $T(D_i|\bar{D}_j)$ characterizes the probability of the transition among states and could be more generalizable than $C(D_i|\bar{D}_j)$ under some circumstances. A reliable estimate of the elements in $T(D_i|\bar{D}_j)$ requires that the number of the occurrences in each cell should be large enough, which suggests that one may want to use the $C(D_i|\bar{D}_j)$ instead of $T(D_i|\bar{D}_j)$ if the count numbers are low. In the above definition of the CTP, we consider the counts by conditioning only the most immediately preceding turn by others. One can extend this to higher order association for situations where long-range dependency prevails in the communication.

It is worth noting that the normalized CTP resembles the stochastic matrix (also known as Markov matrix) if the underlying communication process is a discrete time Markov process that meets the following condition (Van Kampen, 1992; Grimmett and Stirzaker, 2001).

$$\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1}, \cdots, \mathbf{X}_1) = \mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1}) \tag{4}$$

where t denotes the t^{th} step of the process. A transition matrix (or stochastic matrix) $\mathbf{P}$ with elements

$$\mathrm{P}_{ij}(t) = \mathrm{P}(\mathbf{X}_t = \mathbf{x}_j|\mathbf{X}_{t-1} = \mathbf{x}_i) \tag{5}$$

will characterize the transition structure of the Markov process. If a Markov process is stationary (homogeneous), e.g., the following equation holds for all t, i, and j:

$$\mathrm{P}(\mathbf{X}_t = \mathbf{x}_j|\mathbf{X}_{t-1} = \mathbf{x}_i) = \mathrm{P}(\mathbf{X}_1 = \mathbf{x}_j|\mathbf{X}_0 = \mathbf{x}_i) \tag{6}$$

and we can readily predict the probability of different states for the $(t+1)^{th}$ turn based on the preceding turn and the initial turn through the following equation,

$$\mathbf{X}_t = \mathbf{X}_{t-1}\mathbf{P} = \mathbf{X}_0\mathbf{P}^t \tag{7}$$

One notable difference between the normalized CTP and the stochastic matrix of Markov process is that the former is not defined on a closed set of states as one team member's states are dependent on other team members' states instead of her own. As such, the (normalized) CTP introduced above is more a way to numerically represent an aspect of the coded communication process for each team member rather than claiming the mathematical properties associated with the stochastic matrix of a Markov process, though some methods based on the stochastic matrix may still be borrowed to analyze the normalized CTP.

In the next section, we will show how the CPT approach can be used to characterize empirical communication data.

3. EMPIRICAL STUDY

3.1. Task and Data

We carried out the ECSAP project to explore the assessment of communications in large-scale online CPS activities. The goal is to investigate what CPS skills can be detected in the communications and how these skills are related to collaboration outcomes. The details of ECSAP are beyond the scope of this paper, and we refer the readers to Hao et al. (2017b) for a description of the study. The core part of the ECSAP is a simulation-based task that allows two human participants to collaborate through a chat window to complete a set of questions and tasks about volcano science (Hao et al., 2015). **Figure 1** shows two screenshots of the simulation-based collaborative task. In the simulation task, the participants were shown some tutorials about the factors related to volcano eruption. Then, they were asked to answer about fifteen questions, during which they need to carry out some small experiments, such as deploying seismometers around a virtual volcano to collect data, to assist them in answering the questions. The first seven questions are selected responses which allow us to impose a set of structured system prompts to maximize the information elicitation. For each of the seven questions, the system prompts each team member to respond individually at first and then prompts the team members to collaborate with each other to discuss their answers via a chat window. After the collaboration, each member is given a chance to revise her initial answer. By checking the difference in the scores on the initial and revised answers, we can calculate each person's gain/loss from the collaboration. The remaining eight questions require manipulation of the tools in the simulation, which makes it more difficult to impose the initial-discuss-revise procedure. They are not addressed in the current analysis. In addition to this simulation-based collaborative task, we also administered a general science knowledge test (Rundgren et al., 2012) to each participant to measure her content-relevant knowledge.

We collected data through a crowdsourcing data collection platform, Amazon Mechanical Turk (Kittur et al., 2008). We recruited 1,000 participants located in the United States with at least one year of college education and randomly assigned them into 500 dyads to complete the simulation-based collaborative task. Seventy-eight percent of the participants were White, 7% were Black or African American, 5% were Asian, 5% were Hispanic or Latino, and 5% were multiracial. Half of the participants are males and half are females, and the age ranges from 25 to 54. Most of the participants have prior experience of online communication, though not necessarily collaborative problem solving. After removing the teams that did not complete the task successfully, we were left with 474 dyads. In each team's response, there are about 80 turns of chat in total and

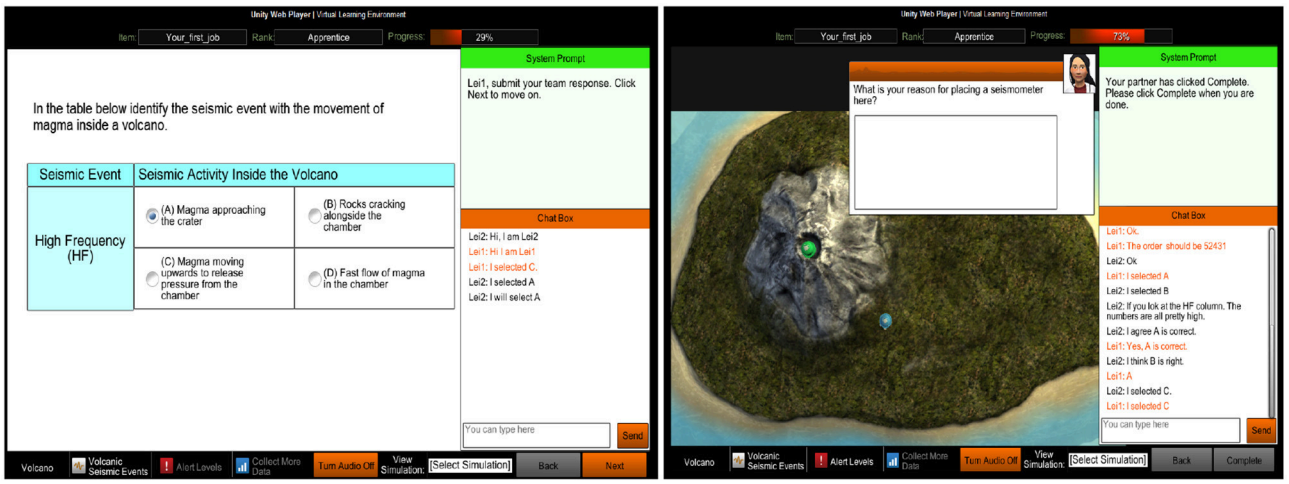


FIGURE 1 | Two screenshots of the simulation-based collaborative task used in the ECSAP.

about 30 turns around the first seven questions. We noticed that many teams did not precisely follow the initial-collaborate-revise procedure we set forth and started some non-prompted discussions when they were asked to answer alone. In our analysis, we consider only the teams that have no more than two non-prompted discussions. After this cut, we were left with 237 out of the 474 dyads. The analyses in this paper are based on this subset unless otherwise stated.

The data from each collaborative session include both the responses to the questions in the simulation and the text-chat communication between the team members around each question. The responses to the questions were scored based on the rubrics shown in Zapata-Rivera et al. (2014). We developed a framework for coding the communication data in CPS (Liu et al., 2015) based on CSCL literature and the assessment frameworks from PISA 2015 (Organization for Economic Co-operation and Development, 2013) and ATC21S (Griffin et al., 2012). This framework considers four skills, namely, sharing ideas, negotiating ideas, regulating problem-solving and maintaining communication, which have been identified to be highly relevant to the CPS activity we are targeting. Each turn of the chat communications was coded into one of the four categories of skills based on our CPS framework. **Table 2** shows some example chats and states. Two human raters were trained on the CPS framework, and they double-coded a subset of the discourse data (15% of the data). The unit of coding is each turn of a conversation or each conversational utterance. The inter-rater agreement in terms of unweighted kappa is 0.67.

TABLE 2 | Example of a part of annotated chat data from one teams.

Topic	Member	Chat	State
IntroduceYourselves	A	Hi	Maintaining
IntroduceYourselves	B	Hi, I'm Jennifer	Maintaining
Question1A	A	chose b, cause its rocks cracking that cause the high frequency events	Sharing
Question1A	B	yes, same here	Negotiating
Question1B	A	d sound right to you?	Regulating
Question1B	B	I couldn't remember, I thought it was C	Regulating
Question1B	A	you are right	Negotiating
QuestionsP2	B	A and B?	Regulating
QuestionsP2	A	yes, that's what i got	Negotiating
QuestionsP3	A	52431?	Regulating
QuestionsP3	B	I was only sure about 5 and 1 being first and last	Sharing
QuestionsP3	B	4 is probably second to last	Sharing
ExampleSeisQuestion1	B	A?	Regulating
ExampleSeisQuestion1	A	picked a	Sharing
ExampleSeisQuestion2	A	thoughts?	Regulating
ExampleSeisQuestion2	B	b?	Regulating
ExampleSeisQuestion2	A	same	Negotiating
ExampleSeisQuestion3	A	obviously c	Sharing
ExampleSeisQuestion3	B	c	Sharing
...	...	...	...

The topic column indicates the specific question around which the conversations happened. The member column indicates which member in the dyadic team produced the discourse.

3.2. Methods

Given that there are about 30 turns of conversations in each team and there are four different coding categories, the expected count in each cell of the four by four matrix is relatively low—about two. Therefore, we choose to use the CTP instead of the normalized version in this paper. The central research question we want to address is the usefulness of the CTP representation of each participant's communication process. As one aspect of this question, we investigated whether such a representation of the communication process is related to the participant's gain or loss as measured based on their total score changes between the initial and revised responses. The hypothesis is that if the CTP is an effective method for characterizing the collaboration process, it

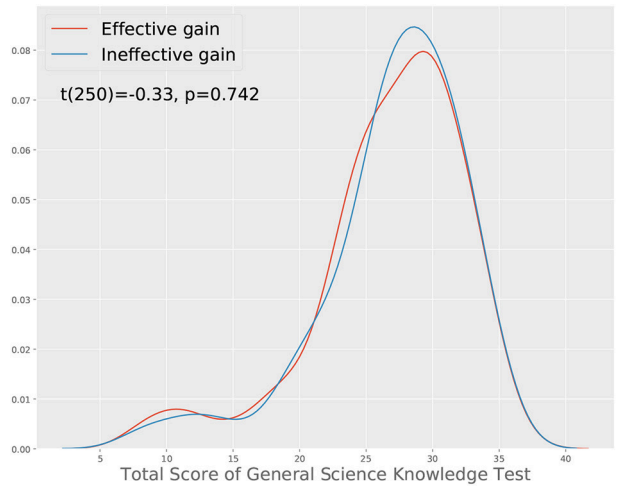


FIGURE 2 | Comparison of the total scores from participants who gain effectively and ineffectively from the collaboration. A *t*-test shows that the two groups have similar contents-relevant science knowledge.

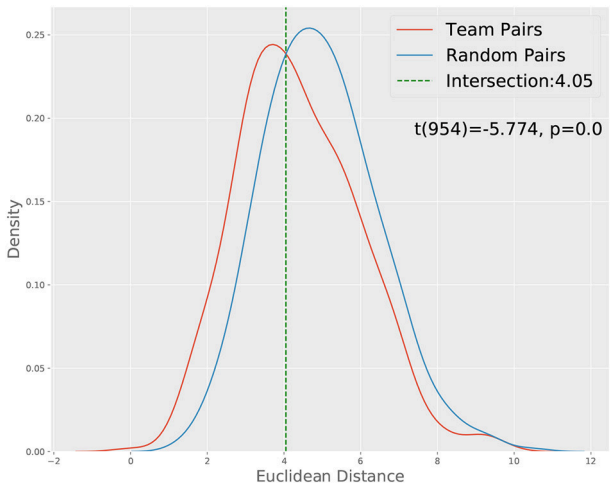


FIGURE 3 | Distance distribution of team pairs and random pairs. A *t*-test show that the two distributions' means are significantly different.

should have implications for the collaboration outcomes. We try the following two approaches to gain some in-depth knowledge of the relationship between a team member's communication process and her outcome from the collaboration.

In the first approach, we started with the total score changes and examine how the CTPs are different in different groups. Specifically, we divide the participants into two groups, labeled effective gain and ineffective gain. Each participant in the effective gain group has a positive total score change while each in the ineffective gain group has a negative or zero total score change. One may notice that such a grouping may systematically penalize people with higher content-relevant knowledge, as they have a higher chance to have a correct initial response to a given item, so it is not possible to further improve. To ensure that we are considering people with comparable content-relevant knowledge, we removed the participants who correctly answered more than five of the seven questions in their initial response. After controlling on this, we have 151 and 101 participants in the effective gain and ineffective gain groups respectively. We verified that they have comparable content-relevant knowledge by comparing their performance in the general science knowledge test, as shown in **Figure 2**. The findings from this approach may be useful in informing the teaching or training of what features of the communication process lead to more effective collaboration outcomes.

In the second approach, we started with the communication process by clustering the participants based on their CTPs, then examined the total score changes in each of the clusters. To perform the cluster analysis, we flattened each CTP into a 16-dimensional vector by appending rows one after another, then calculated Euclidean distances based on the vectors between pairs of participants as a similarity measure of their communication processes. Based on this similarity measure, we first perform a hierarchical clustering analysis using Ward linkage (Ward, 1963) to cluster the participants and then examine the difference of the outcomes in terms of the total score change in different clusters. The findings from this approach can help to uncover similar patterns from the communication process that are associated with similar or different collaboration outcomes, which may also lead to meaningful feedback for a better teaching or training strategies for improving collaboration.

Both approaches may thus lead to actionable procedures in practice to diagnose issues in a computer-supported collaboration and provide feedback to better scaffold the collaboration. For example, after an online collaboration, if we found students who tend to respond to partners in a particular way often show poor collaboration outcomes, we can design coaching or training program to help them to change their ways of communication to ways that are more likely to lead to successful collaboration. The consistency of the findings from the two approaches will substantiate the efficacy of the CTP method for characterizing the communication process in a collaborative activity; whether these characterizations support effective feedback is beyond the scope of the present article.

4. RESULTS

Before we present the results corresponding to the two approaches described above, we would like first to check whether CTPs between team members are more similar compared to those between random pairs of participants. Given the interdependent nature of dyadic communication, we might expect the CTPs between the team members to be more correlated than those between random pairs of participants, which can serve as a check of the plausibility of the CTP approach. We carried out such an analysis based on the full dataset, i.e., without taking out those teams with more than three non-prompted conversations and show the results in **Figure 3**, where we compare the Euclidean distance between the CTPs from team members and random pairs. The result confirms

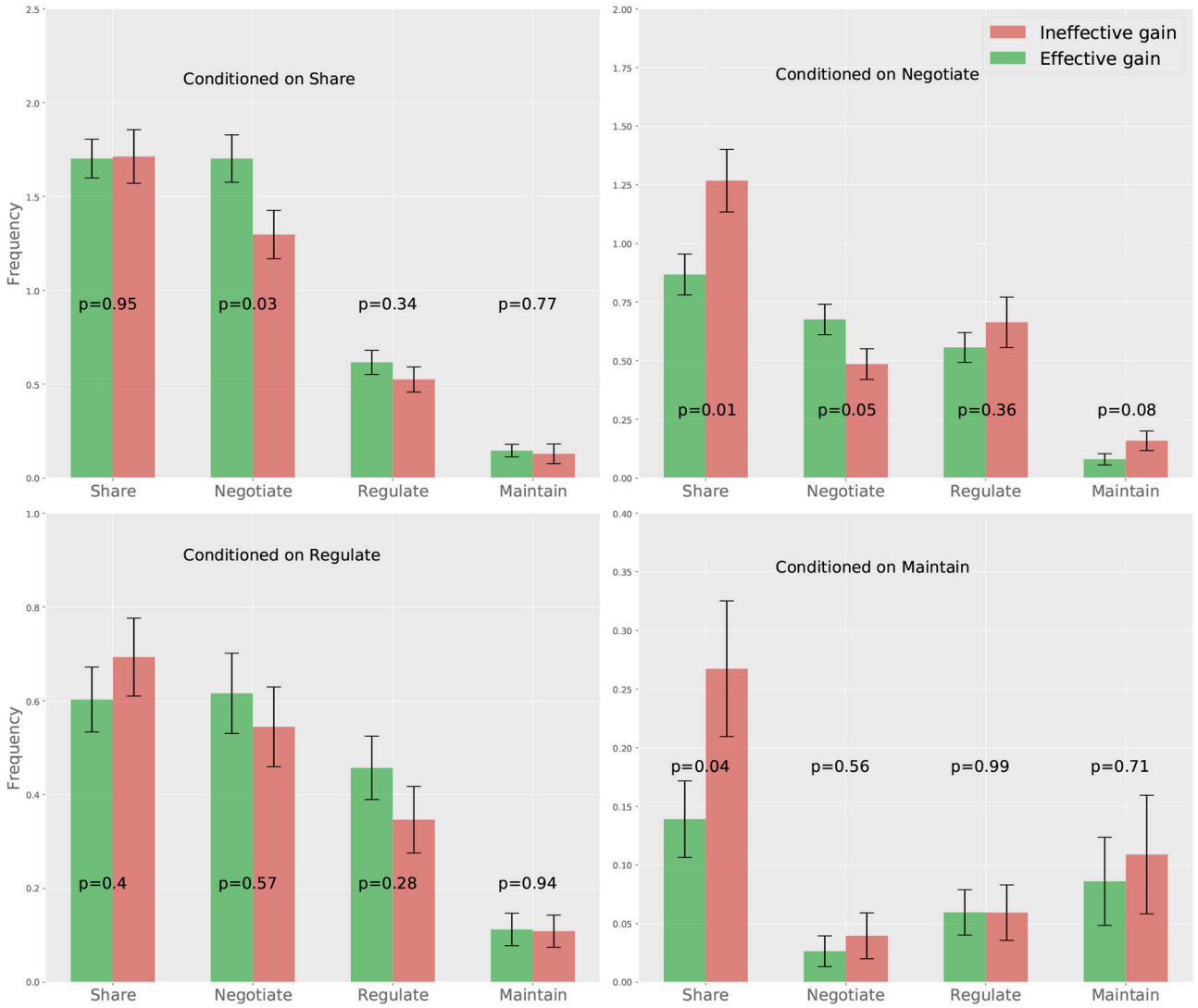


FIGURE 4 | Mean and standard error of the CTPs correspond to the effective and ineffective gain groups. The p-values of pairwise *t*-tests for different CTP components are also presented.

our hypothesis of the interdependence of the communication between team members, which also lends support to the effectiveness of the CTP approach for characterizing the team member's communication process.

The results from our first approach is shown in **Figure 4**, where we compare each element of the CTPs corresponding to the effective and ineffective gain groups via independent *t*-tests (2-tailed)[3]. The results show that the effective gain group has significantly more "negotiate" following the partner's "share" and "negotiate", while the ineffective gain group shows significantly more "share" following the partner's "negotiate" and "maintain." This findings suggests that a person is more likely to demonstrate improved performance if she shows more "negotiate" following her partner's "share" and "negotiate." However, a person is less likely to get an improved response if she shows more "share" upon her partner's "negotiate" and "maintain." This suggests the fact that negotiation is essential for gaining more from a collaboration, while excessively sharing information will contribute negatively, which is consistent with our earlier findings at the team level (Hao et al., 2016).

For the second approach, we show the dendrogram of the hierarchical clustering analysis in **Figure 5**. By examining the distance among the clusters at different levels, we noted that cutting the inter-cluster separations by the elbow point of the inter-cluster distances leads to four clusters. Each cluster is colored differently in **Figure 5** and the number of members in

[3]Note that multiple comparison happens in this case. As the Bonferroni correction is well-known to be too stringent for discovery-oriented studies, we adopted the False Discover Rate (FDR Benjamini and Hochberg, 1995) approach by setting the level of FDR to 0.2, which means we tolerate 20% of the discoveries to be false. At this FDR level, the adjusted p-value for significance is still 0.05 (which is a coincidence).

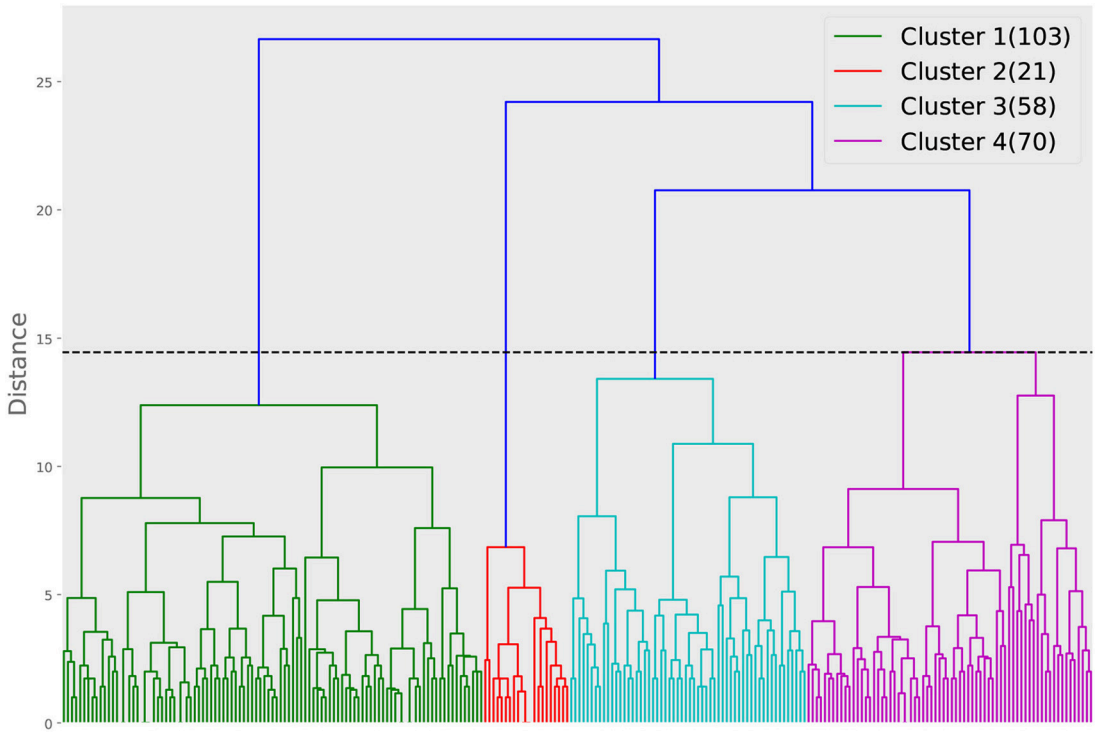


FIGURE 5 | Dendrogram of the hierarchical clustering based on the Euclidean distance calculated from the CTPs. The horizontal dashed line is the distance cut corresponding to the elbow point of the inter-cluster distances. The numbers in the bracket in the legend show how many participants are in each of the clusters.

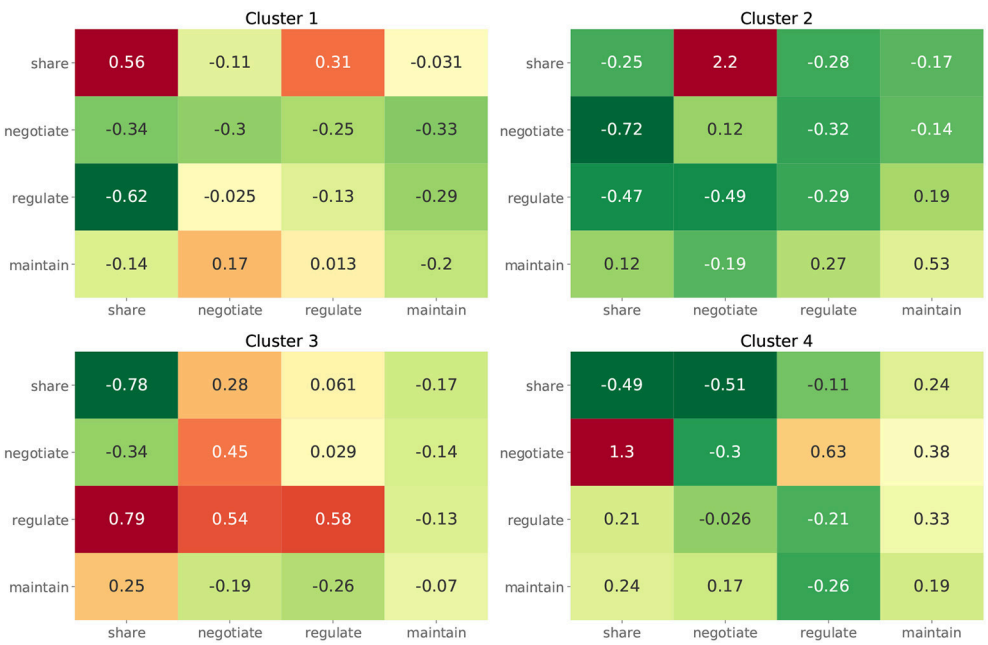


FIGURE 6 | The effect size in terms of Cohen's d between the CPTs of participants from each cluster and from all participants.

each cluster is shown in the legend. To gain more insight into the differences among the four clusters, we compare their CTPs against the CTP of the overall participants by looking at the effect size in terms of Cohen's d. A positive value implies the people in that cluster show more conditional actions corresponding to that cell than the overall population, while a negative value

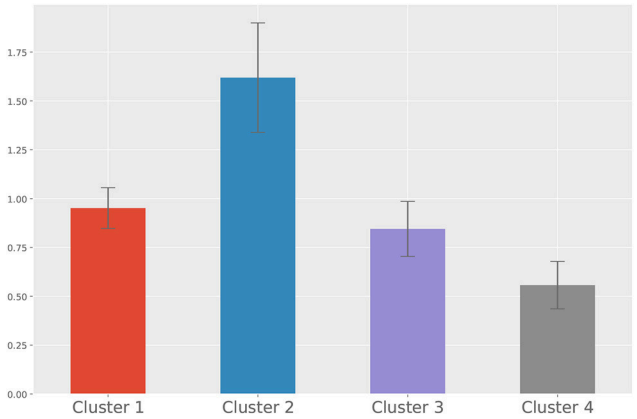


FIGURE 7 | The means and standard errors of the total score changes from each cluster.

implies the other way around. The results are shown in **Figure 6**. A general guideline (Sawilowsky, 2009) for interpreting the effect size is that a Cohen's d equal and greater than 0.8 is considered large effect. Then, in each panel of **Figure 6**, readers can identify how the corresponding cluster is different from the overall participants. Such a plot can give readers a general sense of the major difference between the clusters. **Figure 7** further shows the total score changes in each cluster. The participants in cluster 2 show significantly more positive gain compared to people in other clusters. Connecting back to **Figure 6**, one can immediately identify the main feature of the cluster 3, e.g., participants show more "negotiate" actions when partners "share" information, which is consistent with the results from the first approach.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced a CTP approach to characterize individual team member's communication process in computer-supported collaborations. Based on a large-scale empirical study and using two different approaches starting from the collaboration outcome and the communication process respectively, we show the CTP approach can effectively characterize aspects of one's communication process.

The purpose of the current study was to demonstrate the use of the CTP matrix rather than examine collaboration patterns in a controlled experiment. However, the results of applying CTP to the empirical study suggest that RM's one might try to negotiate while his/her team partner is sharing and negotiating ideas with him/her if he/she wants to gain more from the collaboration. Just sharing ideas seems less likely to help you gain more from collaboration, and even lead to worse outcomes if you do so while your partner is negotiating with you. This finding is consistent with our previous findings at the team level (Hao et al., 2016) and findings in the CSCL literature (Scardamalia and Bereiter, 1994; Stahl, 2006). Moreover, such findings can be incorporated into the teaching of collaborative problem solving skills, and can also be included into real-time feedback mechanisms for scaffolding collaboration.

Despite the effectiveness of CTP, the approach has several known limitations. The first is that it does not capture timing information that could contain useful information concerning, for example, the participation and engagement of the team members regarding their communication and collaboration. Timing is often strongly dependent on the specific task design, however, and its relationship with the other aspects of a collaboration can vary significantly from task to task. As such, a time-dependent version of the CTP with proper inclusion of timing data may provide a better characterization of the process in a given task situation but at the cost of reduced generalizability.

The second is that the CTP does not address possible random errors of the states, such as those introduced during the coding process. A future line of work that may help to improve along this direction may be the introduction of hidden states and emission probabilities to connect the hidden states to the observed states to accommodate the random errors, as Hidden Markov Models (Baum and Petrie, 1966).

The third is that the CTP may become very sparse if there are many coding categories and multiple participants. The average count of each element in the CTP scales down as $1/(nk^2)$ with n as the number of team members and k as the number of coding categories. Users need to make sensible decisions regarding whether to use this method if the communication sequence is very short. A future line of work to address this limitation could consider latent variable modeling, such as factor analysis, though which one can identify a small set of factors to deal with the sparsity.

Finally, the communication process data used in this paper is relatively short, only about thirty turns on average when considering the first seven questions. Though some statistically significant effects have been detected at the subgroup level (thanks to a large number of participants), it does not allow us to reveal more details of each team member's process. In ongoing work, we have collected new data using a task hosted on the ETS Platform for Collaborative Assessment and Learning (Hao et al., 2017c). The new task elicits over 120 turns of communication in each team. We will report the findings based on the new data set in future work.

ETHICS STATEMENT

The data collection is approved through ETS's IRB.

AUTHOR CONTRIBUTIONS

JH contributed to task development, data collection and analysis, research idea and method development, and presentation. RM contributed to research idea and method development, presentation and interpretation.

FUNDING

The work is funded through the Research Allocation Funding at Educational Testing Service.

REFERENCES

Allen, J., and Core, M. (1997). *Draft of Damsl: Dialog Act Markup in Several Layers*. Available online at: http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/

Andrews, J. J., Kerr, D., Mislevy, R. J., Davier, A., Hao, J., and Liu, L. (2017). Modeling collaborative interaction patterns in a simulation-based task. *J. Educ. Measure.* 54, 54–69. doi: 10.1111/jedm.12132

Baum, L. E., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.* 37, 1554–1563. doi: 10.1214/aoms/1177699147

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Biggs, N. (1993). *Algebraic Graph Theory, 2nd edn*. Cambridge: Cambridge University Press; Cambridge Mathematical Library.

Dowell, N. M., Graesser, A. C., and Cai, Z. (2016). Language and discourse analysis with coh-metrix: applications from educational material to learning environments at scale. *J. Learn. Anal.* 3, 72–95. doi: 10.18608/jla.2016.33.5

Flor, M., Yoon, S.-Y., Hao, J., Liu, L., and von Davier, A. (2016). "Automated classification of collaborative problem solving interactions in simulated science tasks," in *Proceedings of 11th Workshop on Innovative Use of NLP for Building Educational Applications* (San Diego, CA). doi: 10.18653/v1/W16-0504

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behav. Res. Methods Instrum. Comput.* 36, 193–202. doi: 10.3758/BF03195564

Griffin, P., McGaw, B., and Care, E. (2012). *Assessment and Teaching of 21st Century Skills*. Springer. doi: 10.1007/978-94-007-2324-5

Grimmett, G., and Stirzaker, D. (2001). *Probability and Random Processes*. New York, NY: Oxford University Press.

Hao, J., Chen, L., Flor, M., Liu, L., and von Davier, A. A. (2017a). Cps-rater: automated sequential annotation for conversations in collaborative problem-solving activities. *ETS Res. Report Ser.* 2017, 1–9. doi: 10.1002/ets2.12184

Hao, J., Liu, L., von Davier, A., and Kyllonen, P. (2015). *Assessing Collaborative Problem Solving With Simulation Based Tasks.* Gothenburg: International Society of the Learning Sciences, Inc.[ISLS].

Hao, J., Liu, L., von Davier, A., Kyllonen, P., and Kitchen, C. (2016). "Collaborative problem solving skills versus collaboration outcomes: Findings from statistical analysis and data mining," in *EDM*, eds T. Barnes, M. Chi, and M. Feng (Raleigh, NC: International Conference on Educational Data Mining), 382–387.

Hao, J., Liu, L., von Davier, A. A., and Kyllonen, P. C. (2017b). "Initial steps towards a standardized assessment for collaborative problem solving (cps): practical challenges and strategies," in *Innovative Assessment of Collaboration*, eds A. A. von Davier, M. Zhu, and P. C. Kyllonen (Springer), 135–156. doi: 10.1007/978-3-319-33261-1-9

Hao, J., Liu, L., von Davier, A. A., Lederer, N., Zapata-Rivera, D., Jakl, P., et al. (2017c). Epcal: Ets platform for collaborative assessment and learning. *ETS Res. Report Ser.* 2017, 1–14. doi: 10.1002/ets2.12181

Kittur, A., Chi, E. H., and Suh, B. (2008). "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence: ACM), 453–456. doi: 10.1145/1357054.1357127

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., and Zapata-Rivera, D. (2015). "A tough nut to crack: Measuring collaborative problem solving," in *Handbook of Research on Technology Tools for Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf (Hershey, PA: IGI Global), 344. doi: 10.4018/978-1-4666-9441-5.ch013

Organization for Economic Co-operation and Development (2013). *Pisa 2015 Draft Collaborative Problem Solving Assessment Framework*. Paris: OECD Publishing.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., et al. (2008). Analyzing collaborative learning processes automatically: exploiting the advances of computational linguistics in computer-supported collaborative learning. *Int. J. Comput Supp. Collabor. Learn.* 3, 237–271. doi: 10.1007/s11412-007-9034-0

Rundgren, C.-J., Rundgren, S.-N. C., Tseng, Y.-H., Lin, P.-L., and Chang, C.-Y. (2012). Are you slim? developing an instrument for civic scientific literacy measurement (slim) based on media coverage. *Public Understand. Sci.* 21, 759–773. doi: 10.1177/0963662510377562

Rus, V., Niraula, N. B., Maharjan, N., and Banjade, R. (2015). "Automated labelling of dialogue modes in tutorial dialogues," in *FLAIRS Conference* (Hollywood, FL), 205–210.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *J. Modern Appl. Statist. Methods* 8, 467–474. doi: 10.22237/jmasm/1257035100

Scardamalia, M., and Bereiter, C. (1994). Computer support for knowledge-building communities. *J Learn. Sci.* 3, 265–283. doi: 10.1207/s15327809jls0303-3

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to Information Retrieval*, Vol 39. Cambridge, UK: Cambridge University Press.

Stahl, G. (2006). *Group Cognition: Computer Support for Building Collaborative Knowledge (Acting With Technology)*. The MIT Press.

Stahl, G., Koschmann, T., and Suthers, D. (2006). Computer-supported collaborative learning: an historical perspective. *Cambridge Handbook Learn Sci.* 2006, 409–426. doi: 10.1017/CBO9781511816833.025

Van Kampen, N. G. (1992). *Stochastic Processes in Physics and Chemistry*, Vol 1. Amsterdam: Elsevier.

Ward J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.* 58, 236–244. doi: 10.1080/01621459.1963.10500845

Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., and Katz, I. R. (2014). "Assessing science inquiry skills using trialogues," in *Intelligent Tutoring Systems*, eds S. Trausan-Matu, K. E. Boyer, M. E. Crosby, and K. Panourgia (Cham: Springer), 625–626.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Hao and Mislevy. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 February 2019
doi: 10.3389/fpsyg.2019.00369

Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data

*Jianlin Yuan[1], Yue Xiao[2] and Hongyun Liu[2,3]**

[1] Educational Science Research Institute, Hunan University, Changsha, Hunan, China, [2] Faculty of Psychology, Beijing Normal University, Beijing, China, [3] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China

As one of the important 21st-century skills, collaborative problem solving (CPS) has aroused widespread concern in assessment. To measure this skill, two initiative approaches have been created: the human-to-human and human-to-agent modes. Between them, the human-to-human interaction is much closer to the real-world situation and its process stream data can reveal more details about the cognitive processes. The challenge for fully tapping into the information obtained from this mode is how to extract and model indicators from the data. However, the existing approaches have their limitations. In the present study, we proposed a new paradigm for extracting indicators and modeling the dyad data in the human-to-human mode. Specifically, both individual and group indicators were extracted from the data stream as evidence for demonstrating CPS skills. Afterward, a within-item multidimensional Rasch model was used to fit the dyad data. To validate the paradigm, we developed five online tasks following the asymmetric mechanism, one for practice and four for formal testing. Four hundred thirty-four Chinese students participated in the assessment and the online platform recorded their crucial actions with time stamps. The generated process stream data was handled with the proposed paradigm. Results showed that the model fitted well. The indicator parameter estimates and fitting indexes were acceptable, and students were well differentiated. In general, the new paradigm of extracting indicators and modeling the dyad data is feasible and valid in the human-to-human assessment of CPS. Finally, the limitations of the current study and further research directions are discussed.

Keywords: collaborative problem solving, process stream data, indicator extracting, dyad data, multidimensional model

OPEN ACCESS

Edited by:
Qiwei He,
Educational Testing Service,
United States

Reviewed by:
Zhi Wang,
Columbia University, United States
Patrick Charles Kyllonen,
Educational Testing Service,
United States

***Correspondence:**
Hongyun Liu
hyliu@bnu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *01 September 2018*
Accepted: *06 February 2019*
Published: *26 February 2019*

Citation:
Yuan J, Xiao Y and Liu H (2019) Assessment of Collaborative Problem Solving Based on Process Stream Data: A New Paradigm for Extracting Indicators and Modeling Dyad Data. Front. Psychol. 10:369. doi: 10.3389/fpsyg.2019.00369

INTRODUCTION

In the field of education, some essential abilities named Key Competencies (Rychen and Salganik, 2003) or 21st Century Skills (Partnership for 21st Century Skills, 2009; Griffin et al., 2012) have been identified. Students must master these skills if they want to live a successful life in the future. Collaborative problem solving is one of the important 21st century skills. Since computers have

substituted for workers to complete many explicitly rule-based tasks (Autor et al., 2003), non-routine problem-solving abilities and complex communication and social skills are becoming increasingly valuable in the labor market (National Research Council, 2011). This set of special skills can be generalized as the construct of Collaborative Problem Solving (Care and Griffin, 2017).

The importance of CPS has spurred researchers in the educational area to assess and teach the skill. However, effectively measuring CPS challenges the current assessment area (Wilson et al., 2012; Graesser et al., 2017, 2018). Because of the complexity of CPS, the traditional testing approaches, such as the paper-pencil test, are inappropriate for it. Therefore, two initiative approaches have been created and applied to the assessment of CPS (Scoular et al., 2017), which are the human-to-human mode and the human-to-agent mode. The human-to-human mode was created by the Assessment and Teaching of 21st Century Skills (ATC21S) project for measuring CPS (Griffin and Care, 2014). It requires two students to collaborate and communicate with each other to solve problems and achieve a common goal. A computer-based testing system has been developed to undisturbedly record students' operation actions, such as chatting, clicking buttons, and dragging objectives, and to generate process stream data (also called log file data; Adams et al., 2015). ATC21S also puts forward a conceptual framework of CPS (Hesse et al., 2015), which includes social and cognitive components. The social component refers to the collaboration part of CPS and the cognitive component refers to the problem solving part. Within the social dimension, there are three strands that are participation, perspective taking, and social regulation. The cognitive dimension includes two strands, task regulation and learning and knowledge building. Each strand contains several elements or subskills, and a total of 18 elements are identified in the framework. Indicators mapped to the elements are extracted from the log file data, and then are used to estimate individual ability (Adams et al., 2015). The Programme for International Student Achievement (PISA) employed the human-to-agent mode for the CPS assessment in 2015 (OECD, 2017a). A computer-based testing system for it has been developed, where computer agents are designed to interact with test-takers. The agents can generate chat messages and perform actions, and test-takers need to make responses (Graesser et al., 2017; OECD, 2017b). These responses, like answers of traditional multiple-choice items, can be directly used to estimate individual CPS ability.

There are many discussions about which is the better way to assess CPS between the two approaches. ATC21S takes the view that the human-to-human interaction is more likely to yield a valid measure of collaboration while the human-to-agent interaction does not conform with the real-world situation (Griffin et al., 2015). Graesser et al. (2017) indicate that the human-to-agent mode provides consistency and control over the social collaboration and that thus it is more suitable for the large-scale assessment. Studies have also shown that each approach involves limitations and have suggested further research to find comprehensive conclusions (Rosen and Foltz, 2014; Scoular et al., 2017). However, from the perspective of data collection, process stream data generated by the human-to-human mode is a record of the whole process of students' actions in computer-based assessment. Based on the data, researchers can reproduce the process of how students collaborate and solve problems, which provides insight into students' cognitive processes and problem solving strategies. In addition, technological advance promotes researchers in assessment area to focus on the process of solving problems or completing tasks, not just the test results. For example, numerous studies of problem solving assessment took a procedural perspective with the assistance of some technology-based assessment systems (PIAAC Expert Group in Problem Solving in Technology-Rich Environments, 2009; Zoanetti, 2009; Greiff et al., 2013; OECD, 2013). These systems could collect the process data and record problem-solving results simultaneously. Thus, the assessment can reveal more about students' thinking process. By comparison, responses of multiple-choice items in the human-to-agent mode can only provide limited information. Therefore, we choose the human-to-human mode in the current study.

However, process stream data cannot be directly used to estimate individual ability. The theory of Evidence-centered Design (ECD) indicates that measurement evidence must be identified from these complicated data before latent constructs are inferred (Mislevy et al., 2003). In the context of educational assessment, existing methods for identifying measurement evidence from process data can be classified into two types. One type is derived from the field of machine learning and data mining, such as Clustering and Classification (Herborn et al., 2017; Tóth et al., 2017), Natural Language Processing and Text Mining (He and von Davier, 2016; He et al., 2017), Graphic Network models (Vista et al., 2016; Zhu et al., 2016), and Bayesian Networks (Zoanetti, 2010; Almond et al., 2015). These data-driven approaches aggregate process data to detect specific behaviors or behavioral patterns that are related to problem-solving outcomes as measurement evidence. Another type of methods can be seen as the theory-driven behavior coding, which means that specific behaviors or behavioral patterns in process data are coded as indicators to demonstrate corresponding skills. This approach was adopted in the CPS assessment of ATC21S. ATC21S defined two categories of indicators: direct and inferred indicators (Adams et al., 2015). Direct indicators can be identified clearly, such as a particular action performed by a student. Inferred indicators are related to sequential actions that represent specific behavioral patterns (Adams et al., 2015). The presence or absence of particular actions or behavioral patterns is the direct evidence that can be used to infer students' abilities. If a corresponding action or behavioral pattern exists in process stream data, the indicator is scored as 1. Otherwise, it is scored as 0. From the perspective of measurement, indicators play the role of traditional items for estimating individual ability.

The theory-driven behavior coding seems effective to obtain measurement evidence from process data, but there exists a problem, that is, how to extract indicators for the dyad members in the human-to-human assessment mode. The ATC21S project adopted the asymmetric mechanism as the basic principle for task design (Care et al., 2015), which is

also called jigsaw (Aronson, 2002) or hidden-profiles (Sohrab et al., 2015) in other research. The asymmetric design means that different information and resources are assigned to the two students in the same group so as to facilitate collaborative activities between them. As a result, they will perform different actions during the process of completing tasks, such as different operations, chat messages, and work products, and will generate their unique process stream data. ATC21S only extracted the same indicators for the two students. This means that the unique information contained in each student's process stream data is ignored, while this information can demonstrate individual skills. Therefore, a comprehensive strategy must be considered to address the complexity of indicator extracting.

Another important problem related to the human-to-human mode is the non-independence between the dyad partners (Griffin et al., 2015). In the ATC21S project, two unacquainted individuals are assigned to work on a common task together. Because of the asymmetric design, they need to exchange information, share resources, negotiate and manage possible conflicts, and cooperate with each other. Each individual member cannot progress through the tasks without his/her partner's assistance. This kind of dependence is called the dyad relationship (Alexandrowicz, 2015). Therefore, a concerned issue is whether the dyad dependence would affect individual scores (Griffin et al., 2015). In the measurement, the dyad relationship violates the local independence assumption of the measurement model. The ATC21S project used the unidimensional Rasch model and the multidimensional Rasch model in calibration (Griffin et al., 2015), and neglected the dyad dependence. However, group assessment has caught the attention of researchers in the measurement field. New approaches and models have been proposed for effective measurement within group settings (von Davier, 2017). Methodologies, such as weighted analysis and multilevel models, were suggested to allow group dependence (Wilson et al., 2012). Wilson et al. (2017) utilized item response models with and without random group effect to model dyad data. Results indicated that the model with the group effect fit better (Wilson et al., 2017). Andrews et al. (2017) used the Andersen/Rasch (A/R) multivariate IRT model to explore the propensities of dyads who followed certain interaction patterns. Alexandrowicz (2015) proposed a multidimensional IRT model to analyze dyad data in social science, in which each individual member had their unique indicators. Researchers have also proposed several innovative statistical models, such as stochastic point process and Hawkes process, to analyze the dyadic interaction (Halpin and De Boeck, 2013; von Davier and Halpin, 2013; Halpin et al., 2017). Olsen et al. (2017) extended the additive factors model to account for the effect of collaboration in the cooperative learning setting. Besides, computational psychometrics that incorporates techniques from educational data mining and machine learning has been introduced into the measurement of CPS (von Davier, 2017). For example, Polyak et al. (2017) used Bayes' rule and clustering analysis in real-time analysis and post-game analysis, respectively. However, there is no definite conclusion on how to model the dyad data.

The Present Study

We agree with the view that the human-to-human interaction is more likely to reveal the complexity and authenticity of collaboration in the real world. Therefore, following the approach of ATC21S, this study employed the human-to-human mode in the assessment of CPS. Students were grouped in pairs to complete the same tasks. The asymmetric mechanism was adopted for task design. Particular actions or behavioral patterns were identified as observable indicators for inferring individual ability. Distinct from the ATC21S approach, we considered a new paradigm for extracting indicators and modeling the dyad data. The main work involved in this study can be classified into three parts.

(1) Following the asymmetric mechanism, we developed five tasks and integrated them into an online testing platform. Process stream data were generated by the platform when the test was going on.
(2) Because of the asymmetry of tasks, we hold that there are unique performances of each member in the dyad for demonstrating their individual skills. Therefore, we extracted individual indicators for each dyad member based on his/her unique process stream data. At the same time, we also identified group indicators that reflected the dyad's contribution and wisdom.
(3) Based on the special design of indicators, we utilized a multidimensional IRT model to fit the dyad data, in which each dyad member was attached with their individual indicators and group indicators.

DESIGN AND DATA

Conceptual Framework of CPS

The CPS framework proposed by ATC21S was adopted in this study, while its detailed description can be seen in Hesse et al. (2015). A total of 18 elements were identified. ATC21S has given a detailed illustration of each element, including its implication and different performance levels (Hesse et al., 2015). The specification provides full insight into the complex skills. More importantly, it serves as the criterion for identifying indicators in this study.

Task Design and Development

We developed five tasks in the present study. To complete each task, two students needed to compose a group. These tasks were designed following the asymmetric mechanism. The two students would obtain different information and resources so they have to cooperate with each other. The current assessment was planned for 15-year-old students, and the problem scenarios of all tasks were related to students' daily life. To illustrate the task design, one of these five tasks, named Exploring Air Conditioner, is presented in **Appendix**. This task was adapted from the task of Climate Control released by PISA2012 (OECD, 2012), which was applied to the assessment of individual problem solving in a computer-based interactive environment. We adapted it for the context of CPS assessment.

To capture students' actions, we predefined a series of events for each task, which can be classified into two types: common and unique events. The common events refer to universal events that would happen in all collaborative assessment tasks, such as the start and the end of a task, chat messages. The unique events occur in specific tasks due to the nature of the behaviors and interactions elicited in these tasks (Adams et al., 2015). **Table 1** presents examples of event specifications for the task of Exploring Air Conditioner. Each event is defined from four aspects, including the event name, the student who might trigger it, the record format, and the explanation for how to capture it. The event specification plays an important role in the computer-based interactive assessment. Firstly, the events represent the key actions and system variables. These actions provide insight into the cognitive process of performing the task. Secondly, the event specification provides a uniform format for recording students' behaviors, which is beneficial to explain the process stream data.

Based on the design of problem scenarios and event specifications, the mainstream techniques of J2EE and MySQL database were adopted for implementing the five tasks. Besides, an online testing platform of multi-user architecture was developed for delivery of all tasks, providing convenience for user login, task navigation, and system administration. The development of tasks and the testing platform followed an iterative process of software development. With the mature platform, students' actions with time stamps could be undisturbedly recorded into the MySQL database as the test progressed, thus the process stream data could be generated.

Data Collection

Procedures

Before the test, we established a set of technical standards for the computer device and internet access to choose schools with perfect Information and Communication Technology (ICT) infrastructure. Since most students and teachers are unfamiliar with the web-based human-to-human assessment of CPS, a special procedure of test administration was considered in the present study. The whole testing process took 70 min, which was divided into two stages. The practice stage was about 10 min, during which examiners needed to illustrate to students what was the human-to-human assessment of CPS. Meanwhile, one task was used as an exercise to help students understand rules. After the practice, the other four tasks were used as assessment tasks in the formal test stage, and 60 min were assigned. Students were demanded to follow the test rules just like what they did in a traditional test, except that they needed to collaborate with their partners via the chat box. Examiners only provided technological assistance during the period. Student's data generated in the four assessment tasks would be used for indicator extracting and subsequent data analysis.

Participants

Four hundred thirty-four students with an average age of approximately 15 years old participated in the assessment, including 294 students from urban schools and 140 students from rural schools in China. All students possess basic ICT skills, such as typing words, sending email, and browsing websites. Since the present study does not focus on the problem of team composition, all the students were randomly grouped in pairs and each student was assigned to a role (A or B) in the group. During the test, students would act as the same role and two members in the dyad group were anonymous to each other.

Ethics Statement

Before we conducted the test, the study was reviewed and approved by the research committee in Beijing Normal University, as well as by the committee in local government. The school teachers, students, and students' parents had clear understanding about this project and how the data were collected. All the students were required to take the written informed consent form to their parents and ask their parents to sign it if they agreed with it.

Process Stream Data

As mentioned above, we predefined a series of events for each task, which represent specific actions and system variables. When the test was in progress, students' actions with time stamps would be fully recorded into a database and then process stream data would be generated. **Figure 1** presents a part of the process stream data from the task of Exploring Air Conditioner, which is exported from MySQL database. The process stream data is constituted by all the events generated by dyad members from the start to the end of tasks, including students' actions, chat messages and status changes of system variables. Each event was recorded as a single row and tagged with the corresponding student identifier, the task identifier, the event content, the role of the actor in the dyad, and the time of the event.

TABLE 1 | Examples of events defined in the task of Exploring Air Conditioner.

Event type	Event name	Role	Record format	Explanation of capturing an action
Common events	Task start	A, B	task start	Record the start of a task
	Task end	A, B	task end	Record the end of a task
	Chat	A, B	free-form chat messages	Record the content of chat messages
Unique events	Control A	A	controlA: *status*	Record the action of changing the position of the slider in the control A
	Control B	A	controlB: *status*	Record the action of changing the position of the slider in the control B
	Control C	B	controlC: *status*	Record the action of changing the position of the slider in the control C
	Control D	B	controlD: *status*	Record the action of changing the position of the slider in the control D
	Apply	A, B	apply: *A* or *B*	Record the action of clicking the button of Apply

student_name	task_name	events_name	log_content	log_time	stuent_role
张莺瑶	空调的探索	task_start	task_start	8:25:59	B
刘若伊	空调的探索	task_start	task_start	8:25:59	A
张莺瑶	空调的探索	chat	哈哈哈哈你好	8:25:59	B
刘若伊	空调的探索	controlA	controlA:0	8:26:00	A
张莺瑶	空调的探索	controlC	controlC:0	8:26:00	B
刘若伊	空调的探索	controlB	controlB:0	8:26:00	A
张莺瑶	空调的探索	controlD	controlD:static	8:26:00	B
刘若伊	空调的探索	chat	你好	8:26:07	A
刘若伊	空调的探索	controlA	controlA:0	8:26:37	A
张莺瑶	空调的探索	chat	我这里有旋钮cd	8:26:38	B
张莺瑶	空调的探索	controlC	controlC:0	8:26:44	B
张莺瑶	空调的探索	controlC	controlC:0	8:26:44	B
刘若伊	空调的探索	chat	我有ab	8:26:45	A
刘若伊	空调的探索	controlA	controlA:+	8:26:48	A
张莺瑶	空调的探索	controlD	controlD:UD_lower	8:26:48	B
张莺瑶	空调的探索	controlD	controlD:UD_high	8:26:50	B
刘若伊	空调的探索	controlA	controlA:0	8:26:51	A
刘若伊	空调的探索	controlA	controlA:-	8:26:52	A
张莺瑶	空调的探索	controlD	controlD:static	8:26:53	B
刘若伊	空调的探索	controlA	controlA:--	8:26:57	A
刘若伊	空调的探索	controlA	controlA:-	8:26:59	A
刘若伊	空调的探索	controlA	controlA:0	8:27:00	A
刘若伊	空调的探索	controlA	controlA:+	8:27:01	A
刘若伊	空调的探索	controlA	controlA:++	8:27:02	A
刘若伊	空调的探索	controlA	controlA:+	8:27:03	A
刘若伊	空调的探索	controlA	controlA:0	8:27:04	A
张莺瑶	空调的探索	controlC	controlC:+	8:27:06	B

FIGURE 1 | A part of process stream data from Exploring Air Conditioner.

Data Processing

Data processing included two steps. First, indicators that serve as measurement evidence were identified and extracted from process stream data. This procedure is an analogy to item scoring in traditional tests. Second, to estimate individual ability precisely, we used a multidimensional Rasch model to fit the dyad data. The quality of indicators and the test was also evaluated in this stage.

INDICATOR EXTRACTING

Rationale for Indicator Extracting

From the perspective of measurement, it is hard to directly judge the skill level of each student based on the process stream data. According to the theory of ECD (Mislevy et al., 2003), measurement evidence must be identified from process stream data for inferring latent ability. Since the abstract construct of CPS has been deconstructed into concrete elements or subskills, it is easier to find direct evidence for demonstrating these subskills or elements than the whole construct. To build up the reasoning chain from process stream data to assessment inference, a theoretical rationale has been commonly taken in many process-oriented assessments, which is that "students' skills can be demonstrated through behaviors which are captured in the form of processes" (Vista et al., 2016). In other words, the observable features of performance data can be used to differentiate test-takers in high and low ability levels (Zoanetti and Griffin, 2017). If the rules of behavior coding that link the process data and inference are established, specific actions or sequential actions in process stream data can be coded into rule-based indicators for assessment (Zoanetti, 2010; Adams et al., 2015; Vista et al., 2016; Zoanetti and Griffin, 2017). This procedure is called indicator extracting in the current study.

In the present study, indicator extracting includes two steps. First, the theoretical specification of indicators was set up, which illustrates why each indicator can be identified and how to extract it. Second, all the indicators were evaluated by experts and the validated indicators were used to score process stream data. Thus, the scoring results of each student were obtained.

Indicator Specification

Based on single events or sequential actions in process stream data, we defined both direct and inferred indicators mapped to elements of the CPS framework. The direct indicator could be clearly identified from a single event, such as the success or failure of a task and a correct or false response to a question. However, the inferred indicator identified from a sequence of actions must be rigorously evaluated. **Table 2** outlines examples for illustrating the specifications of inferred indicators.

As can be seen from **Table 2**, the specification of each indicator includes five aspects. First, all indicators were named following a coding rule. Taking the indicator 'T1A01' as an example, 'T1' represents the first task, 'A' represents that it is identified for student A and is an individual indicator ('G' represents a group indicator), and '01' is a numerical code in the task. Then, the

TABLE 2 | Examples of indicator specifications.

Indicator name	Mapping element	Definition of the indicator	Algorithm	Output
T1A01	Action	The number of messages and actions generated by student A, reflecting his/her activeness in collaboration.	In the process stream data of student A, count all the events that he/she generated.	The count value.
T1G02	Interaction	The number of interactive chat blocks (A, B) between two students, reflecting their interaction. Consecutive chats without interrupted actions from the same student are counted as one. (e.g., A,B,A,B = 2 chat blocks; AA,B,A,B = 2 chat blocks)	Step 1: Find all sequences of consecutive chat messages without any interrupted actions in the process stream data of Student A and B. Step 2: Count the number of chat blocks in each chat sequence. Add one to the value of the indicator if one chat block is found.	The count value.

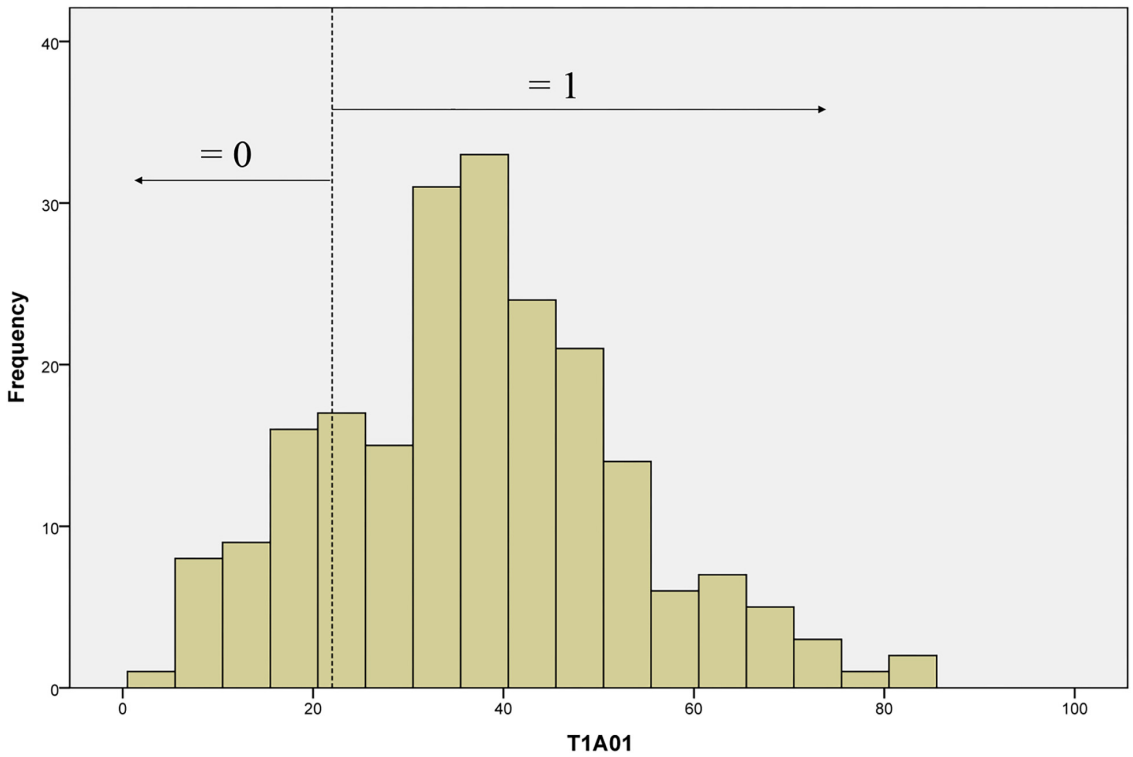


FIGURE 2 | The frequency distribution of T1A01.

mapping element shows what element of CPS this indicator is related to. The definition provides a theoretical description of why it can be identified. The algorithm elaborates the detailed process of how to extract it from process stream data, which is the basis for developing the scoring program. In the last column of the table, the type of the scoring result is simply described. There are two types of output: the count value and the dichotomous value.

A New Paradigm of Extracting Indicators

Distinct from ATC21S, we defined two types of indicators, group and individual indicators. The group indicators are used to illustrate the underlying skills of the two students as a dyad, reflecting the endeavor and contribution of the group. As the indicator T1G02 in **Table 2**, the interactive conversation cannot be completed by any individual member and it needs the two students' participation. Another typical group indicator is identified from task outcomes, that is, the success or failure of each task. The individual indicators are used to demonstrate the underlying skills of the dyad members. Owing to the asymmetric task design, the two members in a group would take different and unique actions or sequential actions, which are used to identify these indicators.

Indicator Validation and Scoring

We defined 8 group indicators and 44 individual indicators (23 for student A and 21 for student B) across the four assessment tasks. To reduce the errors of indicator specifications

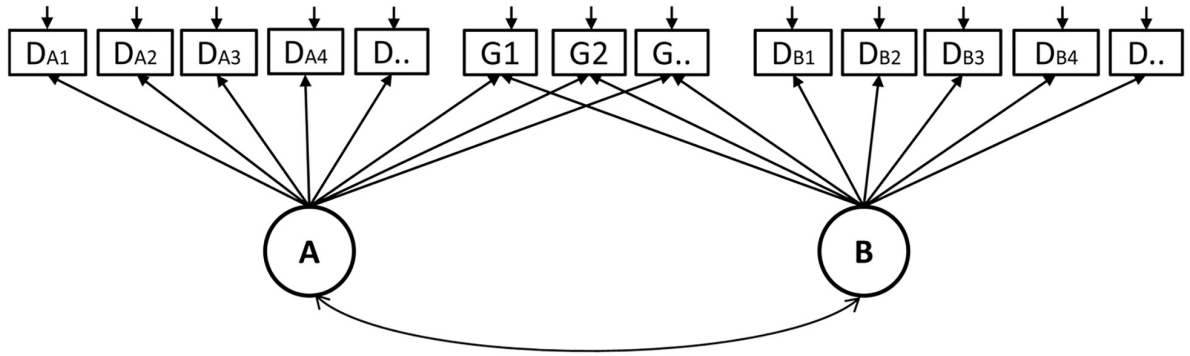


FIGURE 3 | A diagram of the within-item Rasch model for the dyad data.

caused by subjective judgment, indicators were validated by means of expert evaluation. A five-member panel constituted by domain and measurement experts were consulted to evaluate all indicator specifications. Materials, including problem scenario designs, event definitions, samples of process stream data, and all indicator specifications, were provided to them. Experts were demanded to evaluate whether the indicator specifications were reasonable and to give suggestions for modification. An iterative process including evaluation and modification of indicator specifications was used. The process was repeated until all experts agreed on the modified version of all indicators.

Because it is unpractical to score process stream data of all students by human rating, an automatic scoring program was developed based on R language, according to the final specifications of all indicators. We randomly selected 15 groups (30 students) from the sample and obtained their scores separately by the scoring program and a trained human rater. The Kappa consistency coefficient determining the validation of the automatic scoring was calculated for each dichotomously scored indicator. For a few indicators with low Kappa values, we modified their scoring algorithm until their consistency was acceptable. The final results of Kappa consistency for all indicators were shown in Section "Indicator Validation Results." We did not use the Kappa coefficient for indicators with count values, i.e., frequency-based indicators, since the coefficient was based on categorical data. Instead, the reliability of automatic scoring for these indicators were rigorously checked by the research team. The scoring results of each indicator, which were generated by the scoring program and the human rater, were compared based on the randomly selected data of 3 to 5 students. Once there were any differences, we modified the scoring algorithm until the automatic scoring results were the same as scores given by the human rater. After the validation, the process stream data of 434 participants were scored by the automatic scoring program.

Conversion of Frequency-Based Indicators

For model estimation, the count values of frequency-based indicators needed to be converted into discrete values. Since the unique nature of the scoring approach for process data, there is little existing literature that could be used as a guide for the conversion. ATC21S proposed several approaches (Adams et al., 2015), and two of them were adopted in the study. Specifically, we did the transformation by setting thresholds according to

TABLE 3 | Kappa consistency of indicators between the scoring program and the human rater.

Indicators	Kappa coefficient	Indicators	Kappa coefficient	Indicators	Kappa coefficient
T1A03	0.659	T4A02	0.605	T4B03	0.852
T1A07	0.595	T1B03	1.000	T4B01	0.474
T1A09	0.857	T1B07	0.842	T1G01	0.857
T2A01	0.857	T1B09	0.471	T2G01	1.000
T3A04	0.587	T2B01	1.000	T3G01	0.789
T3A06	0.706	T3B04	1.000	T4G01	1.000
T4A03	0.700	T3B06	0.400		

TABLE 4 | Model fit of the two-dimensional Rasch model.

Sample size	Final deviance	Separation reliability	Reliability of dimension 1	Reliability of dimension 2	Correlation of dimension 1 and 2
217	12869.646	0.981	0.886	0.891	0.561

the empirical frequency distribution or the meaning of count values. First, some indicators were converted by setting cut-off values according to their distributions from empirical data. For instance, the frequency distribution of T1A01 (the first indicator in **Table 2**), as shown in **Figure 2**, had a mean of 37.18 and a standard deviation of 15.74. This indicator was mapped to the element of Action in CPS framework and evaluated student activeness in the task. Obviously, a more active student would generate more behaviors and chats. Following the approach of ATC21S (Adams et al., 2015), the cut-off value was set at 22, to which the mean minus a standard deviation (21.44) was rounded up. Thus, students whose number of behaviors and chats less than 22 ($n < 22$) got a score of 0, while those with the number more than 22 ($n \geq 22$) got a score of 1. Second, some frequency-based indicators only contain limited values and each count value was easily interpretable. Thus, a particular value with special meaning could be set as the threshold to transform the indicator. Based on the two approaches, all frequency-based indicators were converted to dichotomous or polytomous variables. Then, all indicators could serve as evidence in the measurement model for inferring students' ability.

MODELING DYAD DATA

Model Definition

In the human-to-human assessment mode of CPS, two students in the same group establish a dyad relationship; hence we call the scoring results dyad data. As mentioned above, how to model the dyad data is a central concern in the assessment of CPS (Wilson et al., 2012; Griffin et al., 2015). Researchers have proposed a number of models to account for the non-independence between the dyad members, such as the multilevel IRT models (Wilson et al., 2017), Hawkes process (Halpin and De Boeck, 2013), and the multidimensional IRT models (Alexandrowicz, 2015). Since group and individual indicators were simultaneously extracted in this study, we employed a multidimensional IRT model to fit the dyad data. The multidimensional model is the extension of the unidimensional model when more than one latent trait is assumed to exist in a test. Some researchers have employed multidimensional IRT models to fit dyad data (Alexandrowicz, 2015). This enlightened us to apply the multidimensional model to the human-to-human assessment of CPS, where two members in a dyad are regarded as two different dimensions.

There are two types of multidimensional models: within-item and between-item multidimensional models (Adams et al., 1997). In this study, we chose the within-item multidimensional Rasch model for the dyad data. As depicted in **Figure 3**, student A and B are regarded as two dimensions, where the latent factor A and B, respectively represent the CPS ability of the role A and B. The indicator D_{A1}, D_{A2}, ..., attached to factor A, are individual indicators of student A. Similarly, D_{B1}, D_{B2}, ..., are individual indicators for student B. The indicator G1, G2, ..., are group indicators that are simultaneously attached to factor A and B. Specifically, the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; Adams et al., 1997) was adopted to fit the data and its formula is

$$P\left(X_{ik}=1;A,B,\xi|\theta\right)=\frac{\exp\left(b_{ik}\theta+\alpha_{ik}^{'}\xi\right)}{\sum_{k=1}^{K_i}\exp\left(b_{ik}\theta+\alpha_{ik}^{'}\xi\right)} \quad (1)$$

where θ is a vector representing the person's location in a multidimensional space and is equal to (θ_A, θ_B) in the current study. The notations of A, B, and ξ represent the design matrix, the scoring matrix, and the indicator parameter vector, respectively. $X_{ik} = 1$ represents a response in the kth category of indicator i. The design matrix A is expressed as

$$A=\left[\begin{array}{cccc:ccc}
0&0&0&0&0&0&0\\
1&0&0&0&0&0&0\\
0&0&0&0&0&0&0\\
0&1&0&0&0&0&0\\
0&1&1&0&0&0&0\\
\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\
0&0&0&0&0&0&0\\
0&0&0&0&0&1&0\\
0&0&0&0&0&1&1
\end{array}\right]. \quad (2)$$

where each row corresponds to a category of an indicator and each column represents an indicator parameter. For example, indicator 1 and 2 have two and three categories respectively, which correspond to the first to second row and the third to fifth row in the above matrix. The scoring matrix B specifies how the individual and group indicators were attached to dimension θ_A and θ_B, which is expressed as

$$B=\left[\begin{array}{cc}
0&0\\
1&0\\
0&0\\
1&0\\
2&0\\
\vdots&\vdots\\
0&0\\
0&1\\
0&0\\
0&1\\
0&2\\
\vdots&\vdots\\
0&0\\
1&1\\
0&0\\
1&1\\
2&2\\
\vdots&\vdots
\end{array}\right]. \quad (3)$$

TABLE 5 | Results of indicator parameter estimates and fit.

Indicator	Discrimination	Difficulty	Error	MNSQ	Confidence interval		T
T1A01	0.24	−1.554	0.180	1.01	0.79	1.21	0.1
T1A04	0.34	−0.309	0.083	1.09	0.87	1.13	1.3
T1A07	0.42	0.139	0.142	0.95	0.93	1.07	−1.5
T1A09	0.26	1.189	0.163	0.99	0.84	1.16	−0.1
T2A01	0.46	−1.860	0.196	0.93	0.74	1.26	−0.5
T2A03	0.28	1.516	0.176	0.97	0.80	1.20	−0.3
T2A01	0.28	−0.723	0.149	0.99	0.89	1.11	−0.2
T3A01	0.39	−1.943	0.201	0.94	0.73	1.27	−0.4
T3A04	0.35	0.935	0.154	0.94	0.87	1.13	−0.9
T3A06	0.22	0.835	0.151	1.01	0.88	1.12	0.1
T3A01	0.15	−1.243	0.120	1.09	0.84	1.16	1.1
T3A02	0.15	0.048	0.084	1.30	0.88	1.12	4.5
T3A03	0.41	1.027	0.090	0.95	0.79	1.21	−0.4
T4A01	0.30	−1.681	0.190	0.98	0.76	1.24	−0.1
T4A03	0.50	0.655	0.151	0.88	0.90	1.10	−2.3
T4A04	0.38	0.794	0.154	0.92	0.88	1.12	−1.3
T4A01	0.44	0.408	0.116	0.94	0.83	1.17	−0.7
T4A02	0.39	0.412	0.147	0.94	0.91	1.09	−1.3
T4A03	0.49	−0.179	0.121	0.89	0.82	1.18	−1.3
T1B01	0.33	−1.781	0.193	0.95	0.75	1.25	−0.4
T1B04	0.25	−0.067	0.079	1.18	0.88	1.12	2.9
T1B07	0.31	0.077	0.141	0.99	0.93	1.07	−0.3
T1B09	0.36	1.025	0.157	0.96	0.86	1.14	−0.6
T2B01	0.39	−1.566	0.179	0.95	0.78	1.22	−0.4
T2B03	0.31	1.038	0.157	0.97	0.86	1.14	−0.4
T2B01	0.32	−0.896	0.152	0.96	0.87	1.13	−0.6
T3B01	0.36	−2.009	0.206	0.95	0.71	1.29	−0.3
T3B04	0.29	1.172	0.161	0.96	0.85	1.15	−0.5
T3B06	0.30	0.636	0.146	1.00	0.90	1.10	0.0
T3B01	0.22	−0.708	0.101	1.12	0.85	1.15	1.5
T3B02	0.04	0.036	0.085	1.34	0.88	1.12	5.0
T3B03	0.42	1.113	0.087	0.97	0.79	1.21	−0.3
T4B01	0.43	−1.748	0.193	0.96	0.75	1.25	−0.3
T4B03	0.37	0.329	0.145	0.96	0.92	1.08	−1.0
T4B01	0.22	−0.147	0.144	0.87	0.93	1.07	−3.8
T4B03	0.04	0.679	0.092	0.88	0.84	1.16	−1.5
T1G01	0.42	−0.277	0.080	1.18	0.87	1.13	2.6
T1G02	0.43	−0.035	0.076	1.18	0.88	1.12	2.9
T2G01	0.37	−0.439	0.043	1.10	0.82	1.18	1.1
T2G02	0.22	0.322	0.083	0.90	0.87	1.13	−1.4
T3G01	0.04	0.015	0.064	1.05	0.82	1.18	0.6
T3G02	0.50	0.163	0.061	1.07	0.83	1.17	0.8
T4G01	0.51	−0.157	0.070	1.01	0.81	1.19	0.1
T4G02	0.47	0.055	0.067	1.12	0.82	1.18	1.2

In Conquest, discrimination is the product moment correlation between the case scores on the indicator and the corresponding case raw scores.

where each row corresponds to a category of an indicator and each column denotes a dimension. In the above matrix B, for example, the first five rows denote that the first indicator (scored as 0 or 1) and the second indicator (scored as 0, 1, or 2) are individual indicators scored on the first dimension (θ_A). The middle several rows correspond to those individual indicators scoring on the second dimension (θ_B). The last several rows indicate those indicators measuring both dimensions, i.e., group indicators.

Calibration

Indicator calibration was performed by ConQuest 3.0, which included two stages. At the first stage, all the indicators (44 individual indicators and 8 group indicators) were calibrated with the one-parameter multidimensional Rasch model. Since

the Rasch model only provides difficulty estimates, indicator discrimination was calculated by the traditional CTT (Classical Testing Theory) method in ConQuest. To evaluate the indicator quality, we used some important indicator indexes, such as discrimination, difficulty, and Infit mean square (Information Weighted Mean Squared residual goodness of fit statistic, often represented as MNSQ). In addition, researchers suggested special sequential actions in the process of problem solving were related to task performance (He and von Davier, 2016). This enlightened us to use the correlation between procedural indicator and the corresponding task outcome as a criterion for evaluating indicator quality. It was assumed that good procedural performance is always associated with a better outcome. After comprehensive consideration, the indicators, of which the MNSQ outside the range of 0.77 and 1.33, the discrimination and correlation below zero, were excluded from the subsequent analysis. In the second stage, the selected indicators were used to estimate individual ability. Model fit indexes, indicator parameter estimates, and the case distribution based on these indicators provided by ConQuest were used to evaluate test quality.

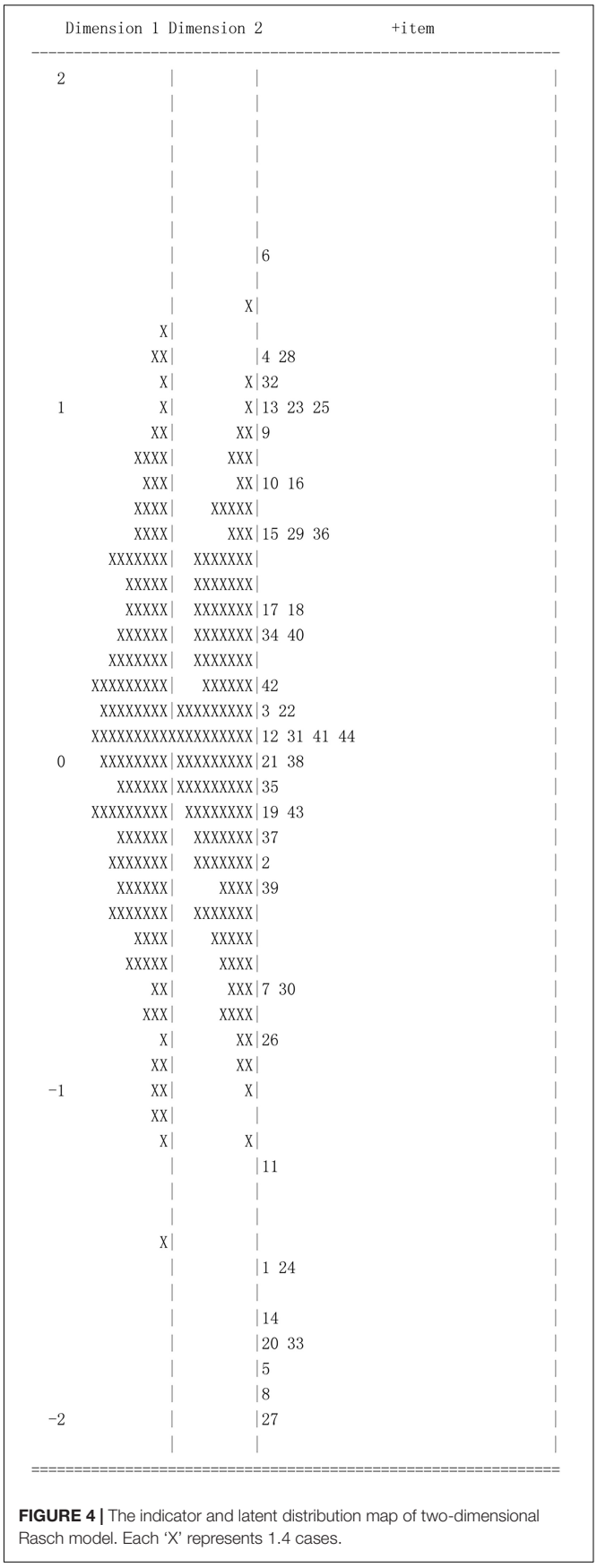


FIGURE 4 | The indicator and latent distribution map of two-dimensional Rasch model. Each 'X' represents 1.4 cases.

RESULTS

Calibration is an exploratory process when it is carried out in test development. For saving space, here we only present the results in the second stage of calibration, which provides the final evidence for the test quality. A total number of 36 individual indicators and 8 group indicators were calibrated in the second stage. The results of calibration and indicator validation are as follows.

Indicator Validation Results

The interrater reliability of twenty dichotomously scored indicators were validated by computing the Kappa consistency between the scoring program and the human rater. The results are shown in **Table 3**. According to the magnitude guideline, the consistency was excellent with a Kappa value over 0.75 and was fair to good with the value from 0.4 to 0.75 (Fleiss et al., 2013). As seen in **Table 3**, all indicators' Kappa value are over 0.4 and there are 12 indicators with excellent Kappa consistency, indicating the reliability of automatic scoring.

Model Fit

Model fit results are shown in **Table 4**. The sample size is the number of dyad groups, indicating a total number of 217 groups (434 students) participated in the assessment. Separation reliability describes how well the indicator parameters are separated (Wu et al., 2007), and the value of 0.981 indicates an excellent performance of test reliability. Dimension 1 and 2, respectively represent student A and B. Reliability of dimensions represents the degree of person separation. The value of 0.886 and 0.891 indicate that the test is sensitive enough to distinguish students at high and low ability levels. Wright and Masters (1982) showed that the indicator separation index and person separation index could be respectively used as an index of construct validity and criterion validity. Therefore, the results in the present study indicate the adequate validity of the test. The

dimension correlation is calculated by estimated scores of student A and B, and the value of 0.561 indicates that dyad members are dependent on each other to a certain extent.

Indicator Parameter Estimates and Fit

Indicator parameter estimates and fit indexes are presented in **Table 5**. The indicator difficulty estimates are within the range of −2.0 to 1.156 and have an average value of −0.107. Indicator discrimination, calculated by traditional CTT item analysis, falls within the range from 0.22 to 0.51 for most indicators. The MNSQ estimates and confidence interval are reported with *T*-value, and the accepted value of MNSQ ranges from 0.77 to 1.33 (Griffin et al., 2015). The MNSQ values of most indicators fall inside their confidence intervals and the absolute values of their corresponding T statistics are smaller than 2.0. As can be seen, the MNSQ of all indicators are reasonable and has an average value of 1.0, indicating good indicator fit.

Indicator and Latent Distribution

ConQuest can output indicator and case distribution, in which the indicator difficulty and the student ability are mapped to the same logit scale. **Figure 4** presents the distribution of indicator difficulty and student ability in the second stage of calibration. Dimension 1 and 2, respectively represent student A and B. Since the mean of latent ability is constrained as zero in ConQuest, students' abilities are concentrated in the zero point of logit scale and approximate a Gaussian distribution. On the right of the map, indicators are dispersedly distributed from easy to difficult. There are 8 indicators whose difficulty parameters are below the lowest level of ability, indicating they were very easy for all students.

Descriptive Analysis of Testing Results

Of the 434 participants, the minimum and maximum score respectively are −2.17 logits and 2.15 logits. Student ability estimates vary in the full range of 4.319 with a standard deviation of 0.68, indicating that students were well differentiated by the current assessment. **Table 6** presents the descriptive statistics of students' ability of successful group and failure group in each task. There are more students who successfully completed task 1 and 2 than those who failed, while the case is opposite for task 3 and 4. To some extent, this indicates the latter two tasks may be more difficult than the former two tasks. In addition, in all tasks, the mean ability of the students who successfully completed the task is higher than that of the unfinished students. It is consistent with common sense, indicating students' ability estimation is reliable.

DISCUSSION

The current study employed a human-to-human interaction approach initiated by the ATC21S project to measure the collaborative problem solving construct. Following the asymmetric mechanism, we designed and developed five tasks which two students need to partner with each other to work through. Moreover, we integrated the tasks into an online testing platform. There are several reasons impelling us to adopt the human-to-human interaction in the CPS assessment. One advantage is that it approximates to the situation in real life (Griffin et al., 2015) because it requires the real people to collaborate with each other and provides an open environment, such as a free-form chat box, for them to communicate. More importantly, the process stream data obtained provide informative insights into the process of collaboration and problem solving.

The task design is crucial in the present study, which includes the problem scenario design and the definition of events. The problem scenario design aims to elicit students' latent ability of CPS effectively. Therefore, we adopted the asymmetric mechanism for it, which required dyad members to pool their knowledge and resources to achieve a common goal. The event definition is about how to record students' actions in the process stream data. To solve it, we predefined a number of crucial events that represent key actions and system variables for each task. They are indispensable observations for understanding the process of performing tasks and provide a uniform format for recording the data stream. In addition, the technical architecture of tasks and the testing platform are important for developing a stable test system according to our experience, especially a well-constructed multi-user synchronization mechanism.

To tap the rich information from the process stream data, we need to identify indicators that could be mapped to the elements of the conceptual framework as measurement evidence. It has been found that particular sequential actions could be used as rule-based indicators for assessment (Zoanetti, 2010; Adams et al., 2015; Vista et al., 2016). Therefore, we identified specific actions or sequential actions as markers of complex problem solving process in the current study. However, distinct

TABLE 6 | Descriptive statistics of students' ability of successful and failure group in each task.

	N	Minimum	Maximum	Mean	Std. deviation
Failure group in task 1	156	−2.169	1.183	−0.237	0.626
Successful group in task 1	254	−2.029	2.149	0.184	0.647
Failure group in task 2	210	−2.169	1.521	−0.205	0.629
Successful group in task 2	222	−1.975	2.149	0.181	0.676
Failure group in task 3	352	−2.170	1.900	−0.139	0.636
Successful group in task 3	82	−0.784	2.149	0.572	0.549
Failure group in task 4	226	−2.169	1.267	−0.210	0.585
Successful group in task 4	148	−1.150	2.150	0.485	0.552

from the ATC21S approach, we defined two kinds of indicators, individual and group indicators, which reflect the underlying skills of individuals and groups, respectively. Owing to the asymmetry of resources, two members in a dyad would perform differently and generate unique process stream data, while their group performance would also be recorded. Therefore, we could investigate the CPS ability at both the individual and group level.

Another problem concerned by the present study is how to model the dyad data. ATC21S extracted the same indicators for dyad members and the dyad data was modeled by traditional methods. Hence, the local independence assumption of the measurement model was violated. We adopted the two-dimensional within-item Rash model to analyze the dyad data based on the new paradigm of indicator extracting, taking the dyad dependence into account. Results indicated that the model fit well and that indicator parameters and participants were separated well. All the indicator parameter estimates and indicator fit indexes were also reasonable and acceptable. Along with the logit scale, indicators were dispersedly distributed from easy to difficult. In general, the results of data analysis demonstrate that the new paradigm of extracting indicators and modeling the dyad data is a feasible method for CPS assessment.

LIMITATION

As a tentative practice of CPS assessment, the current study also has some limitations. First, most indicators identified in the study are based on the events of operation actions, while students' chat messages are not utilized effectively. Chatting is the only way for the two students to communicate in the human-to-human interaction. Thus, the messages contain abundant information that can be used as measurement evidence. However, extracting indicators from chats requires the technique of semantic analysis. We did not do that work due to our limitation of Chinese semantic analysis. Second, for some elements in the conceptual framework, such as audience awareness and transactive memory, there are no indicators that can be mapped. This is because it is unable to find corresponding sequential actions from process stream data. It is necessary to extract more indicators to ensure an effective measurement of CPS. Third, following the ATC21S' approach, we set up cut-off values for frequency-based indicators based on their distributions of empirical data. This choice of thresholds is tentative and further research is needed for setting more accurate values. Fourth, in the present study, we randomly assigned participants into dyad groups and did not consider group composition, because the current work focuses on how to extract indicators and model the dyad data. However, it is obvious that the group composition would affect the process and results of the collaboration. Further research can consider employing advanced techniques to extract more reliable indicators or exploring the strategies for student grouping.

AUTHOR CONTRIBUTIONS

JY contributed to task design, scoring, data analysis, manuscript writing, and revision. HL contributed to organizing the study and manuscript revision. YX contributed to data collection and manuscript revision.

FUNDING

This study was supported by the National Natural Science Foundation of China (31571152), Beijing Advanced Innovation Center for Future Education and Special Fund for Beijing Common Construction Project (019-105812), Beijing Advanced Innovation Center for Future Education, and National Education Examinations Authority (GJK2017015).

ACKNOWLEDGMENTS

We would like to thank the company of Iflytek for their technical support of developing test platform.

REFERENCES

Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., and Care, E. (2015). "Automatic coding procedures for collaborative problem solving," in *Assessment and Teaching of 21st Century Skills: Methods and Approach*, eds P. Griffin and E. Care (Dordrecht: Springer), 115–132.

Adams, R. J., Wilson, M., and Wang, W. C. (1997). The multidimensional random coefficients multinominal logit model. *Appl. Psychol. Meas.* 21, 1–23. doi: 10.1177/0146621697211001

Alexandrowicz, R. W. (2015). "Analyzing dyadic data with IRT models," in *Dependent Data in Social Sciences Research*, eds M. Stemmler, A. von Eye, and W. Wiedermann (Cham: Springer), 173–202. doi: 10.1007/978-3-319-20585-4_8

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., and Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. New York, NY: Springer. doi: 10.1007/978-1-4939-2125-6

Andrews, J. J., Kerr, D., Mislevy, R. J., von Davier, A., Hao, J., and Liu, L. (2017). Modeling collaborative interaction patterns in a simulation-based task. *J. Educ. Meas.* 54, 54–69. doi: 10.1111/jedm.12132

Aronson, E. (2002). "Building empathy, compassion, and achievement in the jigsaw classroom," in *Improving Academic Achievement: Impact of Psychological Factors on Education*, ed. J. Aronson (San Diego, CA: Academic Press), 209–225.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: an empirical exploration. *Q. J. Econ.* 118, 1279–1333. doi: 10.1162/003355303322552801

Care, E., and Griffin, P. (2017). "Assessment of collaborative problem-solving processes," in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 227–243.

Care, E., Griffin, P., Scoular, C., Awwal, N., and Zoanetti, N. (2015). "Collaborative problem solving tasks," in *Assessment and Teaching of 21st Century Skills: Methods and Approach*, eds P. Griffin and E. Care (Dordrecht: Springer), 85–104.

Fleiss, J. L., Levin, B., and Paik, M. C. (2013). *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley & Sons.

Graesser, A., Kuo, B. C., and Liao, C. H. (2017). Complex problem solving in assessments of collaborative problem solving. *J. Intell.* 5:10. doi: 10.3390/jintelligence5020010

Graesser, A. C., Foltz, P. W., Rosen, Y., Shaffer, D. W., Forsyth, C., and Germany, M. L. (2018). "Challenges of assessing collaborative problem solving," in *Assessment and Teaching of 21st Century Skills: Methods and Approach*,

eds P. Griffin and E. Care (Cham: Springer), 75–91. doi: 10.1007/978-3-319-65368-6_5

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., and Funke, J. (2013). Computer-based assessment of complex problem solving: concept, implementation, and application. *Educ. Technol. Res. Dev.* 61, 407–421. doi: 10.1007/s11423-013-9301-x

Griffin, P., and Care, E. (eds) (2014). *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Dordrecht: Springer.

Griffin, P., Care, E., and Harding, S. M. (2015). "Task characteristics and calibration," in *Assessment and Teaching of 21st Century Skills: Methods and Approach*, eds P. Griffin and E. Care (Dordrecht: Springer), 133–178.

Griffin, P., McGaw, B., and Care, E. (eds) (2012). *Assessment and Teaching of 21s Century Skills*. Dordrecht: Springer. doi: 10.1007/978-94-007-2324-5

Halpin, P. F., and De Boeck, P. (2013). Modelling dyadic interaction with hawkes processes. *Psychometrika* 78, 793–814. doi: 10.1007/s11336-013-9329-1

Halpin, P. F., von Davier, A. A., Hao, J., and Liu, L. (2017). Measuring student engagement during collaboration. *J. Educ. Meas.* 54, 70–84. doi: 10.1111/jedm.12133

He, Q., Veldkamp, B. P., Glas, C. A., and de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment* 24, 157–172. doi: 10.1177/1073191115602551

He, Q., and von Davier, M. (2016). "Analyzing process data from problem-solving items with n-grams: insights from a computer-based large-scale assessment," in *Handbook of Research on Technology Tools for Real-World Skill Development*, eds Y. Rosen, S. Ferrara, and M. Mosharraf (Hershey, PA: IGI Global), 750–777. doi: 10.4018/978-1-4666-9441-5.ch029

Herborn, K., Mustafiæ, M., and Greiff, S. (2017). Mapping an experiment-based assessment of collaborative behavior onto collaborative problem solving in PISA 2015: a cluster analysis approach for collaborator profiles. *J. Educ. Meas.* 54, 103–122. doi: 10.1111/jedm.12135

Hesse, F., Care, E., Buder, J., Sassenberg, K., and Griffin, P. (2015). "A framework for teachable collaborative problem solving skills," in *Assessment and Teaching of 21st Century Skills: Methods and Approach*, eds P. Griffin and E. Care (Dordrecht: Springer), 37–56.

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Rep. Ser.* 2003:i-29. doi: 10.1016/j.amepre.2015.10.013

National Research Council (2011). *Assessing 21st Century Skills: Summary of a Workshop*. Washington, DC: National Academies Press.

OECD (2012). *Explore PISA 2012 Mathematics, Problem Solving and Financial Literacy Test Questions*. Available at: http://www.oecd.org/pisa/test-2012/

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Available at: http://www.oecd-ilibrary.org/education/pisa-2012-assessment-and-analytical-framework_9789264190511-en

OECD (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. Available at: http://www.oecd.org/fr/publications/pisa-2015-assessment-and-analytical-framework-9789264255425-en.htm

OECD (2017b). *PISA 2015 Collaborative Problem-Solving Framework*. Available at: https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Olsen, J., Aleven, V., and Rummel, N. (2017). Statistically modeling individual students' learning over successive collaborative practice opportunities. *J. Educ. Meas.* 54, 123–138. doi: 10.1111/jedm.12137

Partnership for 21st Century Skills (2009). *P21 Framework Definitions*. Available at: http://www.p21.org /documents/P21_Framework_Definitions.pdf

PIAAC Expert Group in Problem Solving in Technology-Rich Environments (2009). *PIAAC Problem Solving in Technology-Rich Environments: A Conceptual Framework, OECD Education Working Papers, No. 36, OECD Publishing, Paris*. Available at: http://dx.doi.org/10.1787/220262483674 doi: 10.1787/220262483674

Polyak, S. T., von Davier, A. A., and Peterschmidt, K. (2017). Computational psychometrics for the measurement of collaborative problem solving skills. *Front. Psychol.* 8:2029. doi: 10.3389/fpsyg.2017.02029

Rosen, Y., and Foltz, P. W. (2014). Assessing collaborative problem solving through automated technologies. *Res. Pract. Technol. Enhanc. Learn.* 9, 389–410.

Rychen, D. S., and Salganik, L. H. (eds) (2003). *Key Competencies for a Successful Life and a Well-Functioning Society*. Gottingen: Hogrefe & Huber.

Scoular, C., Care, E., and Hesse, F. W. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *J. Educ. Meas.* 54, 12–35. doi: 10.1111/jedm.12130

Sohrab, S. G., Waller, M. J., and Kaplan, S. (2015). Exploring the hidden-profile paradigm: a literature review and analysis. *Small Group Res.* 46, 489–535. doi: 10.1177/1046496415599068

Tóth, K., Rölke, H., Goldhammer, F., and Barkow, I. (2017). "Educational process mining: new possibilities for understanding students' problem-solving skills," in *The Nature of Problem Solving. Using Research to Inspire 21st Century Learning*, eds B. Csapó and J. Funke (Paris: OECD), 193–209.

Vista, A., Awwal, N., and Care, E. (2016). Sequential actions as markers of behavioural and cognitive processes: extracting empirical pathways from data streams of complex tasks. *Comput. Educ.* 92, 15–36. doi: 10.1016/j.compedu.2015.10.009

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *J. Educ. Meas.* 54, 3–11. doi: 10.1111/jedm.12129

von Davier, A. A., and Halpin, P. F. (2013). Collaborative problem solving and the assessment of cognitive skills: psychometric considerations. *ETS Res. Rep. Ser.* 2013:i-36. doi: 10.1002/j.2333-8504.2013.tb02348.x

Wilson, M., Bejar, I., Scalise, K., Templin, J., Wiliam, D., and Irribarra, D. T. (2012). "Perspectives on methodological issues," in *Assessment and Teaching of 21st Century Skills*, eds P. Griffin, B. McGaw, and E. Care (Dordrecht: Springer), 67–141. doi: 10.1007/978-94-007-2324-5_3

Wilson, M., Gochyyev, P., and Scalise, K. (2017). Modeling data from collaborative assessments: learning in digital interactive social networks. *J. Educ. Meas.* 54, 85–102. doi: 10.1111/jedm.12134

Wright, B. D., and Masters, G. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R., and Haldane, S. A. (2007). *ACER Conquest Version 2.0: Generalized Item Response Modelling Mannual*. Camberwell: ACER Press.

Zhu, M., Shu, Z., and von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *J. Educ. Meas.* 53, 190–211. doi: 10.1111/jedm.12107

Zoanetti, N. (2010). Interactive computer based assessment tasks: how problem-solving process data can inform instruction. *Aust. J. Educ. Technol.* 26, 585–606. doi: 10.14742/ajet.1053

Zoanetti, N., and Griffin, P. (2017). Log-file data as indicators for problem-solving processes. 177–191.

Zoanetti, N. P. (2009). *Assessment of Student Problem-Solving Processes with Interactive Computer-Based Tasks*. Doctoral dissertation. The University of Melbourne, Melbourne.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Yuan, Xiao and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX 1: THE TASK DESCRIPTION OF EXPLORING AIR CONDITIONER

The interface of each task is in a unified form as shown in **Figure A1**. For each student, an instruction is presented at the top of the page to describe the problem scenario. A chat box for communication is in the right panel. Navigation buttons are placed at the bottom.

The task of Exploring Air Conditioner includes two pages. **Figure A1** shows the first pages seen by the two students, respectively in the task. There are four controls on an air conditioner, which correspond to the regulation of temperature, humidity, and swing. Two students are demanded to explore the function of each control together. On the first page, student A can operate control A and B, and student B can operate control C and D. The function panel showing the temperature, humidity, and swing levels is shared by two students, which means the function status is simultaneously affected by two students' operations. To complete the task, they have to exchange information, negotiate strategies for problem solving and coordinate their operations. But above all, they must follow the rule that "change one condition at a time." After figuring out each control's function, they can use the navigation button to jump to the next page where students need to submit their exploration results.

FIGURE A1 | Screenshots of first pages in Exploring Air Conditioner.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership