COGNITIVE DIAGNOSTIC MODELS: METHODS FOR PRACTICAL APPLICATIONS

EDITED BY: Tao Xin, Chun Wang, Ping Chen and Liu Yanlou
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

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 978-2-88976-134-0
DOI 10.3389/978-2-88976-134-0

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

COGNITIVE DIAGNOSTIC MODELS: METHODS FOR PRACTICAL APPLICATIONS

Topic Editors:
Tao Xin, Beijing Normal University, China
Chun Wang, University of Washington, United States
Ping Chen, Beijing Normal University, China
Liu Yanlou, Qufu Normal University, China

Citation: Xin, T., Wang, C., Chen, P., Yanlou, L., eds. (2022). Cognitive Diagnostic Models: Methods for Practical Applications. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-134-0

Table of Contents

05 ***Editorial: Cognitive Diagnostic Models: Methods for Practical Applications***
Tao Xin, Chun Wang, Ping Chen and Yanlou Liu

08 ***The Development of an Instrument for Longitudinal Learning Diagnosis of Rational Number Operations Based on Parallel Tests***
Fang Tang and Peida Zhan

18 ***Estimating CDMs Using the Slice-Within-Gibbs Sampler***
Xin Xu, Jimmy de la Torre, Jiwei Zhang, Jinxin Guo and Ningzhong Shi

33 ***On the Sequential Hierarchical Cognitive Diagnostic Model***
Xue Zhang and Juntao Wang

46 ***Cognitive Diagnosis Modeling Incorporating Item-Level Missing Data Mechanism***
Na Shan and Xiaofei Wang

57 ***The Impact of Item Calibration Error on Variable-Length Cognitive Diagnostic Computerized Adaptive Testing***
Xiaojian Sun, Yanlou Liu, Tao Xin and Naiqing Song

66 ***A Semi-supervised Learning-Based Diagnostic Classification Method Using Artificial Neural Networks***
Kang Xue and Laine P. Bradshaw

78 ***Sample Size Requirements for Applying Diagnostic Classification Models***
Sedat Sen and Allan S. Cohen

94 ***Test Assembly for Cognitive Diagnosis Using Mixed-Integer Linear Programming***
Wenyi Wang, Juanjuan Zheng, Lihong Song, Yukun Tu and Peng Gao

108 ***Determining the Number of Attributes in Cognitive Diagnosis Modeling***
Pablo Nájera, Francisco José Abad and Miguel A. Sorrel

126 ***The Q-Matrix Anchored Mixture Rasch Model***
Ming-Chi Tseng and Wen-Chung Wang

135 ***Remedial Teaching and Learning From a Cognitive Diagnostic Model Perspective: Taking the Data Distribution Characteristics as an Example***
He Ren, Ningning Xu, Yuxiang Lin, Shumei Zhang and Tao Yang

146 ***An Optimal Choice of Cognitive Diagnostic Model for Second Language Listening Comprehension Test***
Yanyun Dong, Xiaomei Ma, Chuang Wang and Xuliang Gao

158 ***New Item Selection Method Accommodating Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing: Maximum Deviation and Maximum Limitation Global Discrimination Indexes***
Junjie Li, Lihua Ma, Pingfei Zeng and Chunhua Kang

169 ***An Improved Parameter-Estimating Method in Bayesian Networks Applied for Cognitive Diagnosis Assessment***
Ling Ling Wang, Tao Xin and Liu Yanlou

180 ***Model Selection for Cogitative Diagnostic Analysis of the Reading Comprehension Test***
Hui Liu and Yufang Bian

193 *An Adaptive Design for Item Parameter Online Estimation and Q-Matrix Online Calibration in CD-CAT*
Wenyi Wang, Yukun Tu, Lihong Song, Juanjuan Zheng and Teng Wang

212 *A Simplified Method for Predicting Pattern Match Ratio*
Xiaojuan Tang, Huiqiong Duan, Shuliang Ding and Mengmeng Mao

223 *Validating a Reading Assessment Within the Cognitive Diagnostic Assessment Framework: Q-Matrix Construction and Model Comparisons for Different Primary Grades*
Yan Li, Miaomiao Zhen and Jia Liu

236 *Two Test Assembly Methods With Two Statistical Targets*
Zheng Huijing, Li Junjie, Zeng Pingfei and Kang Chunhua

249 *A Multi-level Remedial Teaching Design Based on Cognitive Diagnostic Assessment: Taking the Electromagnetic Induction as an Example*
Rui Huang, Zengze Liu, Defu Zi, Qinmei Huang and Sudong Pan

frontiers | Frontiers in Psychology

EDITORIAL
published: 19 April 2022
doi: 10.3389/fpsyg.2022.895399

Editorial: Cognitive Diagnostic Models: Methods for Practical Applications

Tao Xin[1]*, *Chun Wang*[2], *Ping Chen*[1] and *Yanlou Liu*[3]

[1] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China, [2] College of Education, University of Washington, Seattle, WA, United States, [3] China Academy of Big Data for Education, Qufu Normal University, Qufu, China

Keywords: cognitive diagnostic model, diagnostic feedback information, computerized adaptive test, item-level model comparison, model-data fit evaluation

Editorial on the Research Topic

Cognitive Diagnostic Models: Methods for Practical Applications

Cognitive diagnostic models (CDMs) or diagnostic classification models are a relatively newer psychometric framework for collecting, analyzing, and reporting diagnostic data. They have received increasing attention in many disciplines, such as educational, psychological, and psychiatric measurement. Specifically, CDMs aim to provide discrete multivariate fine-grained diagnostic feedback information about examinees' strengths and weaknesses for developing targeted instruction and personalized support. In total, the 20 articles in this Research Topic focus on the methods and applications that contribute to the broad CDM literature by providing sustainable solutions to problems and issues emerged in practice.

Firstly, eight articles focus on providing new methods or insights that will guide the uses of CDMs in different applications for practitioners. To provide practitioners with a comprehensive guide of sample size on item recovery and classification accuracy, a thorough simulation study was conducted by Sen and Cohen to show the effects of sample size, test length, number of attributes and base rate of mastery on item parameter recovery and classification accuracy of the C-RUM, DINA, DINO, and LCDMREDUCED. Determination of the number of attributes is a fundamental issue in CDM applications. Nájera et al. evaluated the performance of a variety of dimensionality detection methods from the factor analysis literature, such as parallel analysis, minimum average partial, very simple structure, DETECT, empirical Kaiser criterion, exploratory graph analysis, and a machine learning factor forest model in discovering the number of attributes. They found that the parallel analysis with Pearson correlations and mean eigenvalue criterion, factor forest model, and model comparison with AIC are suitable in assessing the dimensionality of CDMs. A Q-matrix anchored mixture Rasch model was developed by Tseng and Wang for constructing a common scale between latent classes regardless of the ability distribution. To demonstrate the practical utility of the model, a real dataset from the Certificate of Proficiency in English was analyzed with the QAMRM, LCDM, and GDM, they find the proposed model has better model fit indices. A joint cognitive diagnostic model incorporating item responses and missing data mechanism was proposed by Shan and Wang for handling missing data not at random in the CDM framework. They demonstrate the practical value of the proposed method using the PISA 2015 computer-based mathematics data. Introducing the thinking of semi-supervised learning into CDM, Xue and Bradshaw proposed a semi-supervised learning-based diagnostic classification method using artificial neural networks. In both the simulation and real data study, they showed that the proposed method has some merit in improving the classification accuracy. The paper by Wang, Xin et al. proposed to train the BN

OPEN ACCESS

Edited and reviewed by:
Wenchao Ma,
University of Alabama, United States

***Correspondence:**
Tao Xin
xintao@bnu.edu.cn

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

Received: *13 March 2022*
Accepted: *15 March 2022*
Published: *19 April 2022*

Citation:
Xin T, Wang C, Chen P and Liu Y (2022) Editorial: Cognitive Diagnostic Models: Methods for Practical Applications. Front. Psychol. 13:895399. doi: 10.3389/fpsyg.2022.895399

first based on the ideal response pattern data and then to update the parameters of BN based on observed item response data using the EM or the GD algorithm. The simulation study and real data study showed the advantages of the proposed method compared to the BN training without adding the ideal response pattern data. A sequential hierarchical CDM was developed by Zhang and Wang and three absolute fit indices and five relative fit indices are used to assess model-data fit. Xu et al. proposed the slice-within-Gibbs sampler for estimating CDMs, and their simulation results confirm the viability of the sampler and show that the new method has the flexibility in incorporating a wider range of prior distributions than the Gibbs sampler.

Secondly, six articles provide use case examples of CDMs in real-world assessments, as well as demonstrate the merit and shortcoming of CDM and related model in a variety of educational and psychological research contexts. Ren et al. demonstrated that the finer-grained diagnostic feedback information from CDM can be used for targeted instruction to improve students' abilities effectively via an example with six attributes (sort, median, average, variance, weighted average, and mode) and the data distribution characteristic of the junior high school mathematics curriculum in China. Similarly, Huang et al. showed that CDMs are useful tool for developing and implementing a multi-level remedial teaching scheme, by taking electromagnetic induction as an example. Since many CDMs have been proposed, selecting the most suitable DCM for each item at the item level is a critical step in practical applications. Dong et al. compared several CDMs for the second language listening comprehension test at the test level and item level, their results verified that mixed-CDMs had a better model and person fit than the saturated GDINA model. This study provides useful insights into a better understanding of subskills and the underlying cognitive process for second language listening tests. To provide a reference for practitioners to develop parallel cognitive diagnostic tests for longitudinal learning, Tang and Zhan presented a detailed process for constructing the parallel rational number operations diagnostic tests. Three main phases of the development process are the Q-matrix and test item development, item quality monitoring, and test quality control. Liu and Bian argued that multidimensional item response theory (MIRT) model can be used for diagnostic purposes, and they compare the model-data fit of the MIRT with the reduced reparametrized unified model and generalized deterministic, noisy, and gate model for a reading comprehension test administered in China. They provide practitioners a suggestion of using the MIRT model for the diagnostic analysis for the reading test. To diagnose the strengths and weaknesses of the reading comprehension ability of the primary students, the Diagnostic Chinese Reading Comprehension Test was administered to a large population of students ($N = 21{,}466$) in grades 2–6 from 20 schools in a district of Changchun City, China in Li, Zhen et al. The essential components of the cognitive diagnostic assessment, attributes specification, test development, Q-Matrix validation, model comparison, reliabilities, validity, and skill profiles were reported in their study.

Thirdly, six articles focus on cognitive diagnostic computerized adaptive test and related fields. Wang, Zheng et al. proposed a new test assembly method of maximizing the minimum distance between latent classes by using mixed-integer linear programming in cognitive diagnosis, and the simulation results revealed that compared with the CDI test assembly and random test assembly, the new test assembly method had the highest accuracy rate in terms of attribute mastery pattern and attribute correct classification rates. Two item selection methods, maximum deviation global discrimination index and maximum limitation global discrimination index were developed by Li, Ma et al. to achieve better attribute coverage balance and item exposure control in the context of CD-CAT. The simulation results showed that the two new proposed methods outperformed other existing item selection methods. Similarly, Huijing et al. proposed two new item selection heuristic methods minimum parameters–information–distance method and minimum information–parameters–distance method for assembling multiple test forms, and they find that the two new methods yield better performance when the information curve of the item pool has a unimodal distribution. Tang et al. presented a simple and effective method, the theoretical construct validity to predict that the upper bound of the pattern match ratio in CDM tests through theoretical derivation and simulation study. They found that the TCV is related to the distribution of knowledge states and item categories and has nothing to do with the number of items. Wang, Tu et al. developed a hybrid optimal design for online estimation of item parameters and online calibration of the Q-matrix for new items that can be used to develop item bank effectively for cognitive diagnostic computerized adaptive testing. Three adaptive designs under different practical situations were investigated using simulation studies, the results show that the new optimal design performs better than the random design. Sun et al. investigated the influences of the calibration errors of item parameters on the variable-length cognitive diagnostic computerized adaptive testing in terms of measurement accuracy, average test length, and test efficiency. Results showed that calibration error has negative effect for the deterministic input, noisy "and" gate model and the reduced reparameterized unified model, but has less influenced for the compensatory reparameterized unified model.

In summary, this collection of papers in this Research Topic provides valuable tools, methods, insights, and examples for practitioners in obtaining diagnostic feedback information using CDMs or relevant methods.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the Cultural Experts, and Four Groups of Talented People Foundation of China and National Natural Science Foundation of China (grant nos. 31900794 and 32071093).

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 © 2022 Xin, Wang, Chen 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: 02 September 2020
doi: 10.3389/fpsyg.2020.02246

The Development of an Instrument for Longitudinal Learning Diagnosis of Rational Number Operations Based on Parallel Tests

*Fang Tang and Peida Zhan**

Department of Psychology, College of Teacher Education, Zhejiang Normal University, Jinhua, China

The precondition of the measurement of longitudinal learning is a high-quality instrument for longitudinal learning diagnosis. This study developed an instrument for longitudinal learning diagnosis of rational number operations. In order to provide a reference for practitioners to develop the instrument for longitudinal learning diagnosis, the development process was presented step by step. The development process contains three main phases, the Q-matrix construction and item development, the preliminary/pilot test for item quality monitoring, and the formal test for test quality control. The results of this study indicate that (a) both the overall quality of the tests and the quality of each item are good enough and that (b) the three tests meet the requirements of parallel tests, which can be used as an instrument for longitudinal learning diagnosis to track students' learning.

Keywords: learning diagnosis, longitudinal assessments, rational number operations, parallel tests, longitudinal cognitive diagnosis

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Wenyi Wang,
Jiangxi Normal University, China
Ren Liu,
University of California, Merced, 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 July 2020*
Accepted: *11 August 2020*
Published: *02 September 2020*

Citation:
Tang F and Zhan P (2020) The Development of an Instrument for Longitudinal Learning Diagnosis of Rational Number Operations Based on Parallel Tests. Front. Psychol. 11:2246. doi: 10.3389/fpsyg.2020.02246

INTRODUCTION

In recent decades, with the development of psychometrics, learning diagnosis (Zhan, 2020b) or cognitive diagnosis (Leighton and Gierl, 2007), which objectively quantifies students' current learning status, has drawn increasing interest. Learning diagnosis aims to promote students' learning according to diagnostic results which typically including diagnostic feedback and interventions. However, most existing cross-sectional learning diagnoses are not concerned about measuring growth in learning. By contrast, longitudinal learning diagnosis evaluates students' knowledge and skills (collectively known as latent attributes) and identifies their strengths and weaknesses over a period (Zhan, 2020b).

A complete longitudinal learning diagnosis should include at least two parts: an instrument for longitudinal learning diagnosis of specific content and a longitudinal learning diagnosis model (LDM). The precondition of the measurement of longitudinal learning is a high-quality instrument for longitudinal learning diagnosis. The data collected from the instrument for longitudinal learning diagnosis can provide researchers with opportunities to develop longitudinal LDMs that can be used to track individual growth over time. Additionally, in recent years, several longitudinal LDMs have been proposed, for review, see Zhan (2020a). Although the usefulness of these longitudinal LDMs in analyzing longitudinal learning diagnosis data has been evaluated through some simulation studies and a few applications, the development process of an instrument

for longitudinal learning diagnosis is rarely mentioned (cf. Wang et al., 2020). The lack of an operable development process of instrument hinders the application and promotion of longitudinal learning diagnosis in practice and prevents practitioners from specific fields to apply this approach to track individual growth in specific domains.

Currently, there are many applications use cross-sectional LDMs to diagnose individuals' learning status in the field of mathematics because the structure of mathematical attributes is relative clear to be identified, such as fraction calculations (Tatsuoka, 1983; Wu, 2019), linear algebraic equations (Birenbaum et al., 1993), and spatial rotations (Chen et al., 2018; Wang et al., 2018). Some studies also apply cross-sectional LDMs to analyze data from large-scale mathematical assessments (e.g., George and Robitzsch, 2018; Park et al., 2018; Zhan et al., 2018; Wu et al., 2020). However, most of these application studies use cross-sectional design and cannot track the individual growth of mathematical ability.

In the field of mathematics, understanding rational numbers is crucial for students' mathematics achievement (Booth et al., 2014). Rational numbers and their operations are one of the most basic concepts of numbers and mathematical operations, respectively. The fact that many effects are put into rational number teaching makes many students and teachers struggle to understand rational numbers (Cramer et al., 2002; Mazzocco and Devlin, 2008). The content of rational number operation is the first challenge that students encounter in the field of mathematics at the beginning of junior high school. Learning rational number operation is not only the premise of the subsequent learning of mathematics in junior high school but is also an important opportunity to cultivate students' interest and confidence in mathematics learning.

The main purpose of this study is to develop an instrument for longitudinal learning diagnosis, especially for the content of rational number operations. We present the development process step by step to provide a reference for practitioners to develop the instrument for longitudinal learning diagnosis.

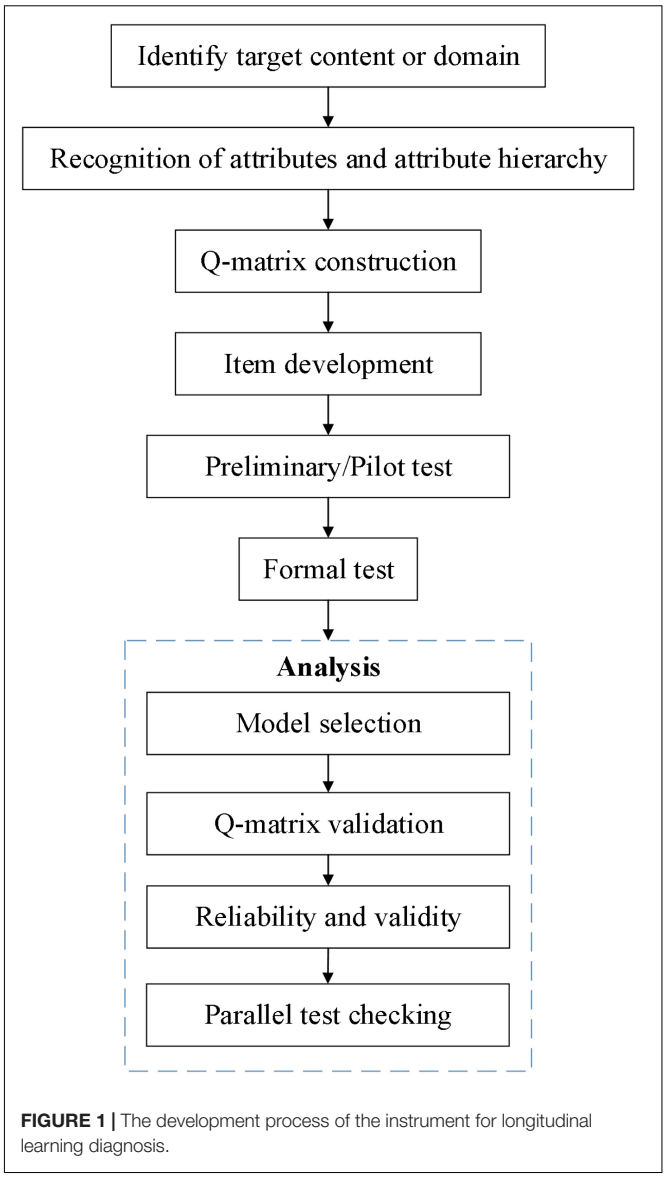


FIGURE 1 | The development process of the instrument for longitudinal learning diagnosis.

DEVELOPMENT OF THE INSTRUMENT FOR LONGITUDINAL LEARNING DIAGNOSIS

As the repeated measures design is not always feasible in longitudinal educational measurement, in this study, the developed instrument is a longitudinal assessment consisting of parallel tests. The whole development process is shown in **Figure 1**. In the rest of the paper, we describe the development process step by step.

Recognition of Attributes and Attribute Hierarchy

The first step in designing and developing a diagnostic assessment is recognizing the core attributes involved in the field of study (Bradshaw et al., 2014). In the analysis of previous studies, the confirmation of attributes mainly adopted the method of literature review (Henson and Douglas, 2005) and expert judgment (Buck et al., 1997; Roduta Roberts et al., 2014; Wu, 2019). This study used the combination of these two methods.

First, relevant content knowledge was extracted according to the analysis of mathematics curriculum standards, mathematics exam outlines, teaching materials and supporting books, existing provincial tests, and chapter exercises. By reviewing the literature, we find that the existing researches mainly focus on one or several parts of rational number operation. For example, fraction addition and subtraction is the most involved in existing researches (e.g., Tatsuoka, 1983; Wu, 2019). In contrast, it is not common to focus on the whole part of rational number operation in diagnostic tests. Ning et al. (2012) pointed out that rational number operation contains 15 attributes; however, such a larger number of attributes does not apply in practice.

TABLE 1 | Attribute framework of the rational number operation.

Label	Attribute	Description
A1	Rational number	Concepts and classifications
A2	Related concepts of the rational number	Opposite number, absolute value
A3	Number axis	Concept, number conversion, comparison of the size of numbers
A4	Addition and subtraction of rational numbers	Addition, subtraction, and addition operation rules
A5	Multiplication and division of rational numbers	Multiplication, involution, multiplication operation rule, division and reciprocal; Reduction of fractions to a common denominator
A6	Mixed operation of rational numbers	First involution, then multiplication and division, and finally addition and subtraction; if there are numbers in parentheses, calculate the ones in the parentheses first.

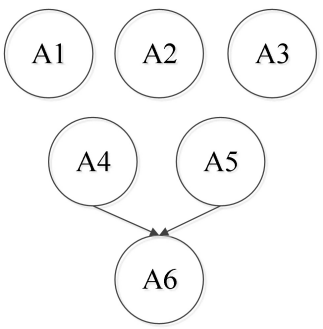


FIGURE 2 | Attribute hierarchy of the rational number operation. Note that A1 = rational number; A2 = related concepts of rational numbers; A3 = axis; A4 = addition and subtraction of rational numbers; A5 = multiplication and division of rational numbers; and A6 = mixed operation of rational numbers.

Second, according to the attribute framework based on the diagnosis of mathematics learning among students in 20 countries in the Third International Math and Science Study–Revised (Tatsuoka et al., 2004), the initial attribute framework and the corresponding attribute hierarchy (Leighton et al., 2004) of this study were determined after a discussion among six experts, including two frontline mathematics teachers who have more than 10 years' experience in mathematics education, two graduate students majoring in mathematics, and two graduate students majoring in psychometrics (see **Table 1** and **Figure 2**).

Third, a reassessment by another group of eight experts (frontline mathematics teachers) and the think-aloud protocol analysis (Roduta Roberts et al., 2014) were used to verify the rationality of the initial attribute framework and that of the corresponding attribute hierarchy. All experts agreed that the attributes and their hierarchical relationships were reasonable. In the think-aloud protocol analysis, six items were initially prepared according to the initial attribute framework and attribute hierarchy (see **Table 2**). Then, six seventh graders were selected according to above-average

TABLE 2 | Items in think-aloud protocol analysis (original items are written in Chinese).

Please **say out aloud** your thoughts when you solve the problem.

(1) Which one of the following statement about rational numbers is correct? ().

(A) Rational numbers can be divided into two categories: positive rational numbers and negative rational numbers

(B) The set of positive integers and the set of negative integers together constitute the set of integers

(C) Integers and fractions are collectively called rational numbers

(D) Positive numbers, negative numbers, and zeros are collectively called rational numbers

(2) Which rational number's inverse equals to itself? ().

(A) () 1 (B) −1 (C) 0 (D) 0 and 1

(3) On the number axis, point A indicates −1. Now A starts to move, first move 3 units to the left, then 9 units to the right, and 5 units to the left. At this time, what the number is point A indicates? ().

(A) −1 (B) 0 (C) 1 (D) 8

(4) Computing: $9 + (-13) - (-7) + (-5) =$

(5) Computing: $(-2) \times 14 \div (-\frac{1}{7}) \times (-1)^5 =$

(6) Computing: $(-\frac{2}{5}) \times (-\frac{3}{7}) - \frac{2}{7} - \frac{3}{5} \div (-\frac{7}{3}) =$

performance, gender balance, willingness to participate, and ability to express their thinking process (Gierl et al., 2008). The experimenter individually tested these students and recorded their responses; in the response process, the students were required to say aloud their problem-solving train of thought. Taking the responses of two students to item 6 as an example, **Figure 3** and **Table 3** present their problem-solving process and thinking process, respectively. Although different students used different problem-solving processes, they all used addition, subtraction, multiplication, and division to solve the items of the mixed operation of rational numbers. Therefore, mastering A4 and A5 are prerequisites to mastering A6, and they validate the rationality of the attribute hierarchy proposed by experts.

Finally, as presented in **Table 1**, the attributes of rational number operation fell into the following six categories: (A1) rational number, (A2) related concepts of rational numbers, (A3) axis, (A4) addition and subtraction of rational numbers, (A5) multiplication and division of rational numbers, and (A6) mixed operation of rational numbers. The six attributes followed a hierarchical structure (**Figure 2**), which indicates that A1–A3 are structurally independent and that A4 and A5 are both needed to master A6.

Q-Matrix Construction and Item Development

According to the attribute hierarchy, A4 and A5 are both needed to master A6. Therefore, the attribute patterns that contain A6 but lack either A4 or A5 are unattainable. Theoretically, there are 40 rather than $2^6 = 64$ attainable attribute patterns. Correspondingly, the initial Q-matrix (i.e., test blueprint) (Tatsuoka, 1983) was constructed based on these 40 permissible attribute patterns and with the following factors

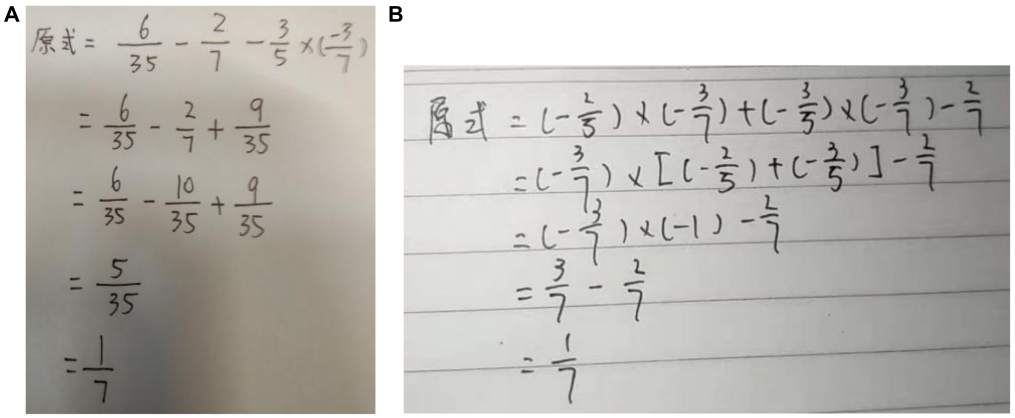


FIGURE 3 | (A,B) Problem-solving process of two students in the think-aloud protocol analysis. Note that in item 6, $(-\frac{2}{5}) \times (-\frac{3}{7}) - \frac{2}{7} - \frac{3}{5} \div (-\frac{7}{3})$with the required attribute pattern (000111).

TABLE 3 | The thinking process of two students in think-aloud protocol analysis.

Student 1:
Step 1: Read the item, and judge that the content knowledge investigated in this item is the mixed operation of rational numbers;
Step 2: Recall the rule for mixed operation of rational numbers: First power, then multiplication and division, final addition and subtraction; If there are parentheses, count them in parentheses first;
Step 3: Make sure multiply and divide first: $(-\frac{2}{5}) \times (-\frac{3}{7}) = \frac{6}{35}$, and change division by $(-\frac{7}{3})$ to multiply by$(-\frac{3}{7})$;
Step 4: Use multiplication: $(-\frac{3}{5}) \times (-\frac{3}{7}) = \frac{9}{35}$;
Step 5: Use addition, and get the answer:
Student 4:
Step 1: Read the item, and judge that the content knowledge investigated in this item is the mixed operation of rational numbers;
Step 2: Recall the rule for mixed operation of rational numbers: First power, then multiplication and division, final addition and subtraction; If there are parentheses, count them in parentheses first;
Step 3: Observe dividing by $(-\frac{7}{3})$ can be changed to multiplying by $(-\frac{3}{7})$, the multiplication distribution law can be used;
Step 4: Use the multiplication distribution law, put $(-\frac{3}{7})$outside of the parentheses, then $(-\frac{2}{5}) + (-\frac{3}{5}) = (-1)$ in the parentheses;
Step 5: Use subtraction, and get the answer.

Item 6: $(-\frac{2}{5}) \times (-\frac{3}{7}) - \frac{2}{7} - \frac{3}{5} \div (-\frac{7}{3})$*with required attribute pattern (000111).*

in mind: (a) the Q-matrix contains at least one reachability matrix for completeness (Ding et al., 2010); (b) each attribute is examined at least twice, and (c) the test time is limited to a teaching period of 40 min to ensure that students have a high degree of involvement. Finally, the test length was determined as 18, including 12 multiple-choice items and 6 calculation items (see **Figure 4**). Notice that all items are dichotomous scored in current study. To ensure that the initial item bank contains enough items, we prepared 4–5 items for each of the 18 attribute patterns contained in the initial Q-matrix. Finally, an initial item bank containing 80 items was formed.

Preliminary Test: Item Quality Monitoring

Participants

In the preliminary test, 296 students (145 males and 151 females) were conveniently sampled from six classes in grade seven of junior high school A[1].

Procedure

To avoid the fatigue effect, 80 items were divided into two tests (preliminary test I and preliminary test II, with 40 items in each test). All participants took part in the two tests. Each test lasted for 90 min, and the two tests were completed within 48 h.

Analysis

Item difficulty and discrimination were computed based on the classical test theory. The differential item functioning (DIF) was checked using the difR package (version 5.0) (Magis et al., 2018) in R software.

Results

A total of 296 students took the preliminary test. After data cleaning, 270 and 269 valid tests were collected in preliminary test I and preliminary test II, respectively. The effective rates of preliminary test I and preliminary test II were 91.22 and 91.19%, respectively. **Table 4** presents the basic sample information and descriptive statistics of the raw scores. The distribution of the raw scores for the two tests was the same.

Table 5 presents the average difficulty and the average discrimination of the preliminary test (the difficulty and discrimination of each item are presented in **Table 6**). In classical test theory, item difficulty (i.e., the pass rate) is equal to the ratio of the number of people who have a correct response to the total number of people, and item discrimination is equal to the difference between the pass rate of the upper 27% of the group

[1]Three schools were used in the complete study. In the instrument development, students in schools A and B participated in the preliminary test and the formal test, respectively; students in school C participate in the quasi-experiment.

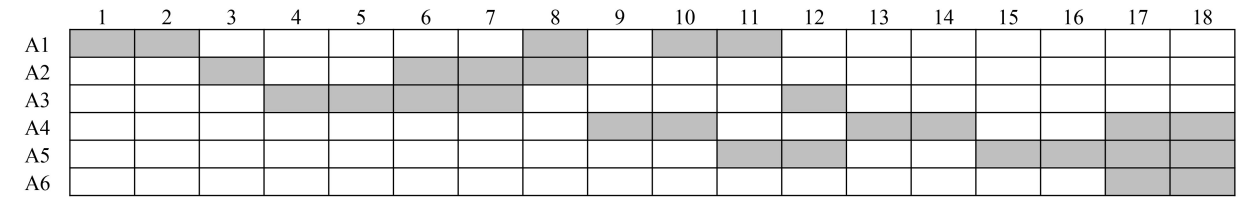


FIGURE 4 | Q-matrix, where blank means "0" and gray means "1." Note that A1 = rational number; A2 = related concepts of rational numbers; A3 = axis; A4 = addition and subtraction of rational numbers; A5 = multiplication and division of rational numbers; and A6 = mixed operation of rational numbers.

TABLE 4 | Basic sample information and descriptive statistics of raw scores in the preliminary test.

Preliminary test	Male	Female	Total	Average score
I	133	137	270	14.77 (8.59)
II	133	136	269	14.78 (8.33)

Note that each test has a full mark of 40. The standard deviation is indicated in parentheses.

TABLE 5 | Average difficulty and average discrimination of the preliminary tests (based on classical test theory).

Preliminary test	Average difficulty	Average discrimination
I	0.37 (0.15)	0.52 (0.18)
II	0.37 (0.14)	0.51 (0.19)

Note that the standard deviation is indicated in parentheses.

and that of the lower 27% of the group. In general, a high-quality test should have the following characteristics: (a) the average difficulty of the test is 0.5, (b) the difficulty of each item is between 0.2 and 0.8, and (c) the discrimination of each item is greater than 0.3. Based on the above three criteria, we deleted eight items in preliminary test I and seven items in preliminary test II.

Table 7 presents the results of the DIF testing of the preliminary tests. DIF is an important index to evaluate the quality of an item. If an item has a DIF, it will lead to a significant difference in the scores of two observed groups (male and female) in the case of a similar overall ability. In the preliminary tests, the Mantel-Haenszel method (Holland and Thayer, 1986) was used to conduct DIF testing. Male is treated as the reference group, and female is treated as the focal group. The results indicated that items 28 and 36 in preliminary test I had DIF, and no item in preliminary test II had DIF. According to item difficulty and discrimination in the above analysis, these two items were classified as items to be deleted.

By analyzing item difficulty, item discrimination, and DIF, 65 items finally remained (including 32 items in preliminary test I and 33 items in preliminary test II) to form the final item bank. Among them, there are 3–5 candidate items corresponding to each of the 18 attribute patterns in the initial Q-matrix. Furthermore, based on the initial Q-matrix, three learning diagnostic tests with the same Q-matrix were randomly extracted from the final item bank to form the instrument of the formal tests: formal test A, formal test B, and formal test C.

Formal Test: Q-Matrix Validation, Reliability and Validity, and Parallel Test Checking

It was possible that the initial Q-matrix was not adequately representative despite the level of care exercised. Thus, empirical validation of the initial Q-matrix was still needed to improve the accuracy of subsequent analysis (de la Torre, 2008). Although item quality was controlled in the preliminary test, it was necessary to ensure that these three tests, as instruments for longitudinal learning diagnosis, met the requirements of parallel tests. Only in this way could the performance of students at different time points be compared.

Participants

In the formal tests, 301 students (146 males and 155 females) were conveniently sampled from six classes in grade seven of junior high school B.

Procedure

All participants were tested simultaneously. The three tests (i.e., formal tests A, B, and C) were tested in turn. Each test lasted 40 min, and the three tests were completed within 48 h.

Analysis

Except for some descriptive statistics, the data in the formal test were mainly analyzed based on the LDMs using the CDM package (version 7.4-19) (Robitzsch et al., 2019) in R software. Including the model–data fitting, the empirical validation of the initial Q-matrix, the model parameter estimation, and the testing of reliability and validity were conducted. In the parallel test checking, the consistency of the three tests among the raw scores, the estimated item parameters, and the diagnostic classifications were calculated.

The deterministic-input, noisy "and" (DINA) model (Junker and Sijtsma, 2001), the deterministic-input, noisy "or" (DINO) model (Templin and Henson, 2006), and the general DINA (GDINA) model (de la Torre, 2011) were used to fit the data. In the model–data fitting, as suggested by Chen et al. (2013), the AIC and BIC were used for the relative fit evaluation, and the RMSEA, SRMSR, MADcor, and MADQ3 were used for the absolute fit evaluation. In the model parameter estimation, only the estimates of the best-fitting model were presented. In the empirical validation of the initial Q-matrix, the procedure suggested by de la Torre (2008) was used. In the model-based DIF checking, the Wald test (Hou et al., 2014) was used. In the

TABLE 6 | Item difficulty and discrimination of preliminary test (based on classical test theory).

Items	Difficulty		Discrimination	
	Preliminary test I	Preliminary test II	Preliminary test I	Preliminary test II
1	0.56	0.52	0.68	0.66
2	0.56	0.18	0.59	0.15
3	0.47	0.54	0.68	0.67
4	0.37	0.50	0.49	0.70
5	0.18	0.43	0.18	0.51
6	0.65	0.22	0.56	0.64
7	0.75	0.64	0.45	0.63
8	0.32	0.36	0.37	0.37
9	0.13	0.52	0.18	0.75
10	0.48	0.30	0.67	0.40
11	0.41	0.41	0.62	0.51
12	0.50	0.49	0.56	0.59
13	0.37	0.33	0.45	0.41
14	0.38	0.31	0.30	0.36
15	0.26	0.28	0.42	0.26
16	0.30	0.42	0.37	0.41
17	0.28	0.20	0.78	0.38
18	0.27	0.38	0.75	0.51
19	0.32	0.27	0.75	0.78
20	0.21	0.12	0.29	0.19
21	0.26	0.26	0.78	0.74
22	0.23	0.32	0.68	0.82
23	0.21	0.33	0.74	0.42
24	0.14	0.15	0.38	0.22
25	0.65	0.22	0.53	0.67
26	0.54	0.24	0.60	0.12
27	0.47	0.20	0.62	0.75
28	0.27	0.65	0.27	0.62
29	0.28	0.54	0.29	0.62
30	0.37	0.31	0.51	0.51
31	0.33	0.64	0.40	0.58
32	0.53	0.47	0.77	0.63
33	0.17	0.21	0.16	0.26
34	0.40	0.38	0.45	0.48
35	0.50	0.33	0.53	0.32
36	0.19	0.52	0.40	0.79
37	0.40	0.39	0.73	0.44
38	0.38	0.50	0.56	0.52
39	0.41	0.29	0.62	0.25
40	0.27	0.41	0.77	0.73

Items to be deleted including items 5, 9, 20, 24, 28, 29, 33, and 36 in preliminary test I, and items 2, 15, 20, 24, 26, 33, and 39 in preliminary test II.

testing of reliability and validity, the classification accuracy (P_a) and consistency (P_c) indices (Wang et al., 2015) were computed.

Results

Descriptive statistics of raw scores

A total of 301 students took the formal test. After data cleaning, the same 277 valid tests (including those from 135 males

TABLE 7 | Differential item functioning testing of preliminary test.

Items	Preliminary test I			Preliminary test II		
	p	deltaMH	Code	*p*	deltaMH	Code
1	0.9012	0.0296	A	0.9446	−0.0276	A
2	0.1318	1.2167	B	0.9412	−0.0313	A
3	0.9508	0.1292	A	0.9317	0	A
4	0.7133	0.3546	A	0.9368	0	A
5	0.9155	−0.1626	A	0.9365	0	A
6	0.4241	0.7871	A	0.9457	0	A
7	0.9055	−0.182	A	0.9448	−0.0289	A
8	0.2583	1.333	B	0.9368	0	A
9	0.6578	0.4168	A	0.9428	0	A
10	0.1922	−0.9753	A	0.9445	0.0841	A
11	0.8356	−0.2203	A	0.9482	0	A
12	0.9223	0.0304	A	0.9455	0	A
13	0.7281	−0.3246	A	0.9383	0	A
14	0.3409	−0.8385	A	0.9416	0	A
15	0.4766	−0.6529	A	0.9483	0	A
16	0.5684	0.8441	A	0.9343	0	A
17	0.8443	0.4158	A	0.9441	0	A
18	0.7296	0.4761	A	0.9131	0	A
19	0.8582	0.2451	A	0.9269	0	A
20	0.9649	0.2428	A	0.9131	0	A
21	0.3897	1.4042	B	0.9272	0	A
22	0.9251	0.5733	A	0.9442	0	A
23	0.7546	0.5935	A	0.9179	0	A
24	0.9732	0.072	A	0.9145	0	A
25	0.5205	−0.4839	A	0.9448	0	A
26	0.9167	0.0326	A	0.9233	0	A
27	0.2342	0.9039	A	0.8897	0	A
28	**0.0425**	−1.5031	C	0.9448	0	A
29	0.8248	−0.251	A	0.9466	0	A
30	0.6341	0.4249	A	0.9442	0	A
31	0.3118	−0.7995	A	0.9445	0	A
32	0.3405	−0.9153	A	0.9429	−0.0299	A
33	0.7845	−0.3828	A	0.9403	0	A
34	0.4017	0.6477	A	0.9466	−0.0264	A
35	0.2172	−0.9052	A	0.9457	0	A
36	**0.0365**	1.9793	C	0.937	0	A
37	0.3919	−0.746	A	0.9449	0	A
38	0.8351	−0.2381	A	0.9454	−0.0271	A
39	0.8637	−0.2109	A	0.9464	0	A
40	0.1209	1.8533	C	0.9386	0.1058	A

DIF items in bold ($p < 0.05$); A, B, and C are the codes of effect size (i.e., the absolute value of deltaMH), where A means negligible effect, B means moderate effect, and C means large effect.

TABLE 8 | Descriptive statistics of raw scores in the formal tests.

Formal test	Average score	Mode	Median	Minimum	Maximum
A	7.24 (4.95)	2	6	0	18
B	7.36 (5.03)	3	6	0	18
C	7.31 (4.98)	2 and 3	6	0	18

Note that the standard deviation is indicated in parentheses. The tests have a full mark of 18.

TABLE 9 | Relative and absolute model–data fit indices.

Formal test	Model	AIC	BIC	RMSEA	SRMSR	MADcor	MADQ3
A	DINA	4646.96	4918.76	0.038	0.057	0.041	0.055
	DINO	4822.67	5094.47	0.050	0.096	0.066	0.071
	GDINA	4635.77	4994.55	0.057	0.046	0.032	0.057
B	DINA	4843.47	5115.27	0.065	0.061	0.046	0.064
	DINO	4994.34	5266.14	0.048	0.094	0.069	0.067
	GDINA	4834.38	5193.15	0.064	0.054	0.039	0.063
C	DINA	4877.31	5149.11	0.041	0.070	0.048	0.062
	DINO	4975.42	5247.22	0.040	0.093	0.065	0.066
	GDINA	4822.55	5182.24	0.060	0.049	0.049	0.063

TABLE 10 | Revision suggestion based on the empirical validation of the initial Q-matrix.

Formal test	Item	Initial required attribute pattern						Revision suggestion					
		A1	A2	A3	A4	A5	A6	A1	A2	A3	A4	A5	A6
A	9	0	0	0	1	0	0	0	0	1	1	1	1
B	9	0	0	0	1	0	0	0	1	1	1	1	0
C	9	0	0	0	1	0	0	0	0	1	1	0	0

TABLE 11 | Classification accuracy and consistency indices based on the DINA model.

Attributes	Formal test A		Formal test B		Formal test C	
	P_a	P_c	P_a	P_c	P_a	P_c
A1	0.96	0.93	0.95	0.92	0.94	0.90
A2	0.93	0.88	0.98	0.97	0.97	0.94
A3	0.94	0.89	0.96	0.93	0.93	0.87
A4	1.00	1.00	1.00	1.00	1.00	1.00
A5	1.00	1.00	0.99	0.98	0.99	0.99
A6	1.00	1.00	1.00	0.99	1.00	0.99
Attribute pattern	0.85	0.74	0.89	0.84	0.85	0.77

TABLE 12 | Reliability of formal tests.

Formal test	Cronbach' α	Split-half reliability
A	0.887	0.907
B	0.889	0.923
C	0.886	0.915
r_{AB}		0.97**
r_{AC}		0.96**
r_{BC}		0.96**

*Split-half reliability is calculated according to the items of odd and even numbers; r_{AB} = the parallel-forms reliability of tests A and B; r_{AC} = the parallel-forms reliability of tests A and C; r_{BC} = the parallel-forms reliability of tests B and C; **$p < 0.01$.*

and 142 females) were collected from each of the three tests; the effective rate of the formal tests was 93.57%. **Table 8** presents the descriptive statistics of raw scores in the formal tests. The average, standard deviation, mode, median, minimum, and maximum of raw scores of the three tests were the same.

TABLE 13 | Item parameter estimates in formal tests.

Items	Formal test A		Formal test B		Formal test C	
	g	*s*	*g*	*s*	*g*	*s*
1	0.4927	0.0861	0.3972	0.0433	0.3849	0.0348
2	0.0155	0.1101	0.3044	0.1604	0.3588	0.1948
3	0.1046	0.0978	0.0009	0.0412	0.0771	0.0775
4	0.0796	0.0770	0.1431	0.0510	0.1248	0.1016
5	0.1721	0.3809	0.1590	0.1940	0.1702	0.3689
6	0.2260	0.4177	0.2739	0.3422	0.2373	0.4373
7	0.2774	0.1514	0.2868	0.0431	0.2667	0.0669
8	0.1785	0.2215	0.2924	0.2209	0.2915	0.3311
9	0.2827	0.1860	0.2984	0.1746	0.3089	0.1629
10	0.2676	0.3429	0.2605	0.3315	0.2747	0.2124
11	0.2921	0.3247	0.3427	0.3018	0.2739	0.2673
12	0.1314	0.3891	0.2270	0.2827	0.2387	0.2271
13	0.0001	0.0001	0.0001	0.0119	0.0001	0.0468
14	0.0001	0.0233	0.0001	0.0468	0.0001	0.0119
15	0.0443	0.0001	0.0224	0.0417	0.0310	0.0001
16	0.0201	0.0001	0.0201	0.1881	0.0256	0.1310
17	0.0371	0.0711	0.0510	0.2085	0.0464	0.1626
18	0.0093	0.0532	0.0099	0.0403	0.0140	0.0583
Mean	0.1462	0.1630	0.1717	0.1513	0.1736	0.1607

g = guessing parameter; s = slip parameter.

Model–data fitting

The parameters in an LDM can be interpreted only when the selected model fits the data. The fit indices presented in **Table 9** provide information about the data fit of three LDMs, namely DINA, DINO, and GDINA, to determine the best-fitting model. Absolute fit indices hold that values near zero indicate an absolute fit (Oliveri and von Davier, 2011; Ravand, 2016). The result indices indicated that all three models fitted the data

TABLE 14 | Diagnostic classifications of students in formal test.

Attribute pattern	Formal test A		Formal test B		Formal test C	
	Proportion	Number of students	Proportion	Number of students	Proportion	Number of students
000000	20.06%	56	29.13%	81	19.02%	53
100000	0.00%	0	0.00%	0	6.89%	19
010000	19.38%	54	15.00%	42	6.82%	19
110000	6.59%	18	4.64%	13	11.43%	32
001000	6.16%	17	2.61%	7	5.35%	15
101000	0.00%	0	0.43%	1	1.50%	4
011000	0.92%	3	0.86%	2	1.81%	5
111000	12.99%	36	12.72%	35	14.01%	39
000100	1.53%	4	0.92%	3	0.00%	0
100100	0.00%	0	0.47%	1	2.09%	6
010100	0.75%	2	1.59%	4	1.80%	5
110100	2.58%	7	1.63%	5	1.50%	4
001100	0.00%	0	0.19%	1	0.30%	1
101100	1.09%	3	0.87%	2	0.84%	2
111100	0.50%	1	0.49%	1	1.05%	3
100010	0.23%	1	0.29%	1	0.00%	0
010010	0.49%	1	0.77%	2	0.42%	1
110010	0.56%	2	1.16%	3	0.71%	2
011010	0.00%	0	0.00%	0	0.26%	1
111010	1.56%	4	1.34%	4	0.72%	2
100110	2.06%	6	1.85%	5	1.09%	3
010110	0.00%	0	0.21%	1	0.32%	1
110110	0.51%	1	0.54%	1	0.00%	0
101110	0.00%	0	0.32%	1	0.00%	0
111110	0.98%	3	0.62%	2	0.71%	2
101111	0.48%	1	1.55%	4	0.00%	0
111111	20.56%	57	19.83%	55	21.34%	59

Attribute patterns with 0 person in all three tests are omitted.

well. For relative fit indices, smaller values indicate a better fit. The DINA model was preferred based on the BIC, and the GDINA model was preferred based on the AIC. According to the parsimony principle (Beck, 1943), a simpler model is preferred if its performance is not significantly worse than that of a more complex model. Both AIC and BIC introduced a penalty for model complexity. However, as the sample size was included in the penalty in BIC, the penalty in BIC was larger than that in AIC. The DINA model was chosen as the best-fitting model given the small sample size of this study, which might not meet the needs of an accurate parameter estimation of the GDINA model, and the item parameters in the DINA model having more straightforward interpretations. Therefore, the DINA model was used for the follow-up model-based analyses.

Q-matrix validation

A misspecified Q-matrix can seriously affect the parameter estimation and the results of diagnostic accuracy (de la Torre, 2008; Ma and de la Torre, 2019). Notice that the Q-matrix validation can also be skipped when the model fits the data well. **Table 10** presents the revision suggestion based on the empirical validation of the initial Q-matrix. In all three tests, the revision suggestion was only for item 9. However, after the subjective and empirical judgment of the experts (Ravand, 2016), this revision suggestion was not recommended to be adopted. Let us take item 9 ("Which number minus 7 is equal to −10?") in formal test A as an example. Clearly, this item does not address the suggested changes in A3, A5, and A6. As the expert-defined Q-matrix was consistent with the data-driven Q-matrix, the initial Q-matrix was used as the confirmed Q-matrix in the follow-up analyses.

Reliability and validity

Classification accuracy (P_a) and consistency (P_c) are two important indicators for evaluating the reliability and validity of classification results. According to Ravand and Robitzsch (2018), values of at least 0.8 for the P_a index and 0.7 for the P_c index can be considered acceptable classification rates. As shown in **Table 11**, both pattern- and attribute-level classification accuracy and consistency were within the acceptable range. Additionally, Cronbach's α, split-half reliability, and parallel form reliability were also computed based on the raw scores (see **Table 12**). The attribute framework of this study was reassessed by several experts, and the Q-matrix was confirmed, indicating that the content validity and the structural validity of this study were good. To further verify the external validity, the correlation

between the raw score of each formal test and the raw score of a monthly exam (denoted as S; the content of this test is the chapter on "rational numbers") was computed ($r_{AS} = 0.95$, $p < 0.01$; $r_{BS} = 0.95$, $p < 0.01$; $r_{CS} = 0.94$, $p < 0.01$). The results indicated that the reliability and validity of all three tests were good.

Parallel test checking

To determine whether there were significant differences in the performance of the same group of students in the three tests, the raw scores, estimated item parameters (**Table 13**), and diagnostic classifications (**Table 14**) were analyzed by repeated measures ANOVA. The results indicated no significant difference in the raw scores [$F(2,552) = 1.054$, $p = 0.349$, $BF_{10} = 0.038$[2]], estimated guessing parameters [$F(2,34) = 1.686$, $p = 0.200$, $BF_{10} = 0.463$], estimated slip parameters [$F(2,34) = 0.247$, $p = 0.783$, $BF_{10} = 0.164$], and diagnostic classifications [$F(2,78) \approx 0.000$, $p \approx 1.000$, $BF_{10} = 0.078$] in the same group of students in the three tests.

As the three tests examined the same content knowledge, contained the same Q-matrix, had high parallel-forms reliability, and had no significant differences in the raw scores, estimated item parameters, and diagnostic classifications, they could be considered to meet the requirements of parallel tests.

CONCLUSION AND DISCUSSION

This study developed an instrument for longitudinal learning diagnosis of rational number operations. In order to provide a reference for practitioners to develop the instrument for longitudinal learning diagnosis, the development process was presented step by step. The development process contains three main phases, the Q-matrix construction and item development, the preliminary test for item quality monitoring, and the formal test for test quality control. The results of this study indicate that (a) both the overall quality of the tests and the quality of each item are good enough and that (b) the three tests meet the requirements of parallel tests, which can be used as an instrument for longitudinal learning diagnosis to track students' learning.

However, there are still some limitations of this study. First, to increase operability, only the binary attributes were adopted. As the binary attribute can only divide students into two categories (i.e., mastery and non-mastery), it may not meet the need for a multiple levels division of practical teaching objectives (Bloom et al., 1956). Polytomous attributes and the corresponding LDMs (Karelitz, 2008; Zhan et al., 2020) can be adopted in future studies. Second, the adopted instrument for longitudinal learning diagnosis was based on parallel tests. However, in practice, perfect parallel tests do not exist. In further studies, the anchor-item design (e.g., Zhan et al., 2019) can be adopted to develop an instrument for longitudinal learning diagnosis. Third, an appropriate Q-matrix is one of the key factors in learning diagnosis (de la Torre, 2008). However, the Q-matrix used in the instrument may not strictly meet the requirements of identification (Gu and Xu, 2019), which may affect the diagnostic classification accuracy.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

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' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

FT conducted data acquisition and analysis. PZ provided the idea, wrote the first draft, and revised the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No. 19YJC190025).

[2]The Bayes factor (BF_{10}) was calculated using the JASP software (Goss-Sampson, 2020) based on the Bayesian estimation. $BF_{10} = 0.038$ means that the current data are 0.038 times more likely to occur under the alternative hypothesis (H1) being true than under the null hypothesis (H0) being true. As suggested by Dienes (2014), BF_{10} less than 1, 1/3, and 1/10 represents weak, moderate, and strong evidence for the H0, respectively. By contrast, BF_{10} greater than 1, 3, and 10 represents weak, moderate, and strong evidence for the H1, respectively.

REFERENCES

Beck, L. W. (1943). The principle of parsimony in empirical science. *J. Philos.* 40, 617–633. doi: 10.2307/2019692

Birenbaum, M., Kelly, A. E., and Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *J. Res. Math. Educ.* 24, 442–459. doi: 10.2307/749153

Bloom, B. S., Englehart, M., Frust, E., Hill, W., and Krathwohl, D. (1956). *Taxonomy of Educational Objectives: the Classification of Educational Goals: Handbook I, Cognitive Domain*. New York, NY: Longmans, Green.

Booth, J. L., Newton, K. J., and Twiss-Garrity, L. K. (2014). The impact of fraction magnitude knowledge on algebra performance and learning. *J. Exp. Child Psychol.* 118, 110–118. doi: 10.1016/j.jecp.2013.09.001

Bradshaw, L., Izsák, A., Templin, J., and Jacobson, E. (2014). Diagnosing teachers ' understandings of rational numbers?: building a multidimensional test within the diagnostic classification framework. *Educ. Meas. Issues Pract.* 33, 2–14. doi: 10.1111/emip.12020

Buck, G., Tatsuoka, K., and Kostin, I. (1997). The subskills of reading: rule-space analysis of a multiple-choice test of second language reading comprehension. *Lang. Learn.* 47, 423–466. doi: 10.1111/0023-8333.00016

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Meas.* 50, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

Chen, Y., Culpepper, S. A., Wang, S., and Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Appl. Psychol. Meas.* 42, 5–23. doi: 10.1177/0146621617721250

Cramer, K. A., Post, T. R., and delMas, R. C. (2002). Initial fraction learning by fourth- and fifth-grade students: a comparison of the effects of using commercial curricula with the effects of using the rational number project curriculum. *J. Res. Math. Educ.* 33, 111–144. doi: 10.2307/749646

de la Torre, J. (2008). An empirically based method of q-matrix validation for the dina model: development and applications. *J. Educ. Meas.* 45, 343–362. doi: 10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Front. Psychol.* 5:781. doi: 10.3389/fpsyg.2014.00781

Ding, S., Yang, S., and Wang, W. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *J. Jiangxi Normal Univ.* 34, 490–494.

George, A. C., and Robitzsch, A. (2018). Focusing on interactions between content and cognition: a new perspective on gender differences in mathematical sub-competencies. *Appl. Meas. Educ.* 31, 79–97. doi: 10.1080/08957347.2017.1391260

Gierl, M. J., Wang, C., and Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the sat© . *J. Technol. Learn. Assess.* 6, 1–53.

Goss-Sampson, M. A. (2020). *Bayesian Inference in JASP: A Guide for Students*. Available online at: http://static.jasp-stats.org/Manuals/Bayesian_Guide_v0_12_2_1.pdf

Gu, Y., and Xu, G. (2019). Identification and estimation of hierarchical latent attribute models. *arXiv:1906.07869*.

Henson, R., and Douglas, J. (2005). Test construction for cognitive diagnosis. *Appl. Psychol. Meas.* 29, 262–277. doi: 10.1177/0146621604272623

Holland, P., and Thayer, D. (1986). Differential item performance and the Mantel-Haenszel procedure. *Paper Presented at the Annual Meeting of the Aerican Educational Research Association*, San Francisco, CA.

Hou, L., de la Torre, J., and Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: application of the wald test to investigate dif in the dina model. *J. Educ. Meas.* 51, 98–125. doi: 10.1111/jedm.12036

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

Karelitz, T. M. (2008). How binary skills obscure the transition from non-mastery to mastery. *Measurement: Interdisciplinary Research and Perspectives* 6, 268–272. doi: 10.1080/15366360802502322

Leighton, J. P., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge, MA: Cambridge University Press.

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* 41, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Ma, W., and de la Torre, J. (2019). An empirical q-matrix validation method for the sequential generalized dina model. *Br. J. Math. Stat. Psychol.* 73, 142–163. doi: 10.1111/bmsp.12156

Magis, D., Beland, S., and Raiche, G. (2018). *difR: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF) [R package version 5.0]*.

Mazzocco, M. M. M., and Devlin, K. T. (2008). Parts and "holes": Gaps in rational number sense among children with vs. without mathematical learning disabilities. *Dev. Sci.* 11, 681–691. doi: 10.1111/j.1467-7687.2008.00717.x

Ning, G., Gong, T., and Wu, H. (2012). Application of cognitive diagnosis technique in the evaluation of subject achievement. *Examin. Res.* 32, 21–29.

Oliveri, M. E., and von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychol. Test Assess. Model.* 53, 315–333.

Park, Y. S., Xing, K., and Lee, Y. S. (2018). Explanatory cognitive diagnostic models: incorporating latent and observed predictors. *Appl. Psychol. Meas.* 42, 376–392. doi: 10.1177/0146621617738012

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

Ravand, H., and Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educ. Psychol.* 38, 1255–1277. doi: 10.1080/01443410.2018.1489524

Robitzsch, A., Kiefer, T., George, A. C., and Uenlue, C. (2019). *CDM: Cognitive diagnosis modeling [R package version 7.4-19]*.

Roduta Roberts, M., Alves, C. B., Chu, M.-W., Thompson, M., Bahry, L. M., and Gotzmann, A. (2014). Testing expert-based versus student-based cognitive models for a grade 3 diagnostic mathematics assessment. *Appl. Meas. Educ.* 27, 173–195. doi: 10.1080/08957347.2014.905787

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

Tatsuoka, K. K., Corter, J. E., and Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in timss-r across a sample of 20 countries. *Am. Educ. Res. J.* 41, 901–926. doi: 10.3102/00028312041004901

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989x.11.3.287

Wang, S., Hu, Y., Wang, Q., Wu, B., Shen, Y., and Carr, M. (2020). The development of a multidimensional diagnostic assessment with learning tools to improve 3-D mental rotation skills. *Front. Psychol.* 11:305. doi: 10.3389/fpsyg.2020.00305

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

Wang, W., Song, L., Chen, P., Meng, Y., and Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *J. Educ. Meas.* 52, 457–476. doi: 10.1111/jedm.12096

Wu, H., Liang, X., Yürekli, H., Becker, B., Paek, I., and Binici, S. (2020). Exploring the impact of Q-Matrix specifications through a DINA model in a large-scale mathematics assessment. *J. Psychoeduc. Assess.* doi: 10.1177/0734282919867535

Wu, H.-M. (2019). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. *Educ. Psychol.* 39, 1218–1232. doi: 10.1080/01443410.2018.1494819

Zhan, P. (2020a). A Markov estimation strategy for longitudinal learning diagnosis: providing timely diagnostic feedback. *Educ. Psychol. Meas.* doi: 10.1177/0013164420912318 [Epub ahead of print].

Zhan, P. (2020b). Longitudinal learning diagnosis: minireview and future research directions. *Front. Psychol.* 11:1185. doi: 10.3389/fpsyg.2020.01185

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

Zhan, P., Jiao, H., Liao, D., and Li, F. (2019). A longitudinal higher-order diagnostic classification model. *J. Educ. Behav. Stat.* 44, 251–281. doi: 10.3102/1076998619827593

Zhan, P., Wang, W.-C., and Li, X. (2020). A partial mastery, higher-order latent structural model for polytomous attributes in cognitive diagnostic assessments. *J. Classif.* 37, 328–351. doi: 10.1007/s00357-019-09323-7

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 Tang and Zhan. 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 September 2020
doi: 10.3389/fpsyg.2020.02260

Estimating CDMs Using the Slice-Within-Gibbs Sampler

Xin Xu[1]*, *Jimmy de la Torre*[2], *Jiwei Zhang*[3], *Jinxin Guo*[1] and *Ningzhong Shi*[1]*

[1] Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China, [2] Faculty of Education, The University of Hong Kong, Pokfulam, Hong Kong, [3] Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, School of Mathematics and Statistics, Yunnan University, Kunming, China

In this paper, the slice-within-Gibbs sampler has been introduced as a method for estimating cognitive diagnosis models (CDMs). Compared with other Bayesian methods, the slice-within-Gibbs sampler can employ a wide-range of prior specifications; moreover, it can also be applied to complex CDMs with the aid of auxiliary variables, especially when applying different identifiability constraints. To evaluate its performances, two simulation studies were conducted. The first study confirmed the viability of the slice-within-Gibbs sampler in estimating CDMs, mainly including G-DINA and DINA models. The second study compared the slice-within-Gibbs sampler with other commonly used Markov Chain Monte Carlo algorithms, and the results showed that the slice-within-Gibbs sampler converged much faster than the Metropolis-Hastings algorithm and more flexible than the Gibbs sampling in choosing the distributions of priors. Finally, a fraction subtraction dataset was analyzed to illustrate the use of the slice-within-Gibbs sampler in the context of CDMs.

Keywords: the slice-within-Gibbs sampler, CDMs, DINA model, G-DINA model, Gibbs sampling, MH algorithm

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Peida Zhan,
Zhejiang Normal University, China
Chanjin Zheng,
East China Normal University, China

***Correspondence:**
Xin Xu
xux636@nenu.edu.cn
Ningzhong Shi
shinz@nenu.edu.cn

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

Received: *02 June 2020*
Accepted: *11 August 2020*
Published: *25 September 2020*

Citation:
Xu X, de la Torre J, Zhang J, Guo J and Shi N (2020) Estimating CDMs Using the Slice-Within-Gibbs Sampler. Front. Psychol. 11:2260. doi: 10.3389/fpsyg.2020.02260

1. INTRODUCTION

Cognitive diagnosis models (CDMs) aim to provide a finer-grained evaluation of examinees' attribute profiles. As psychometric tools, CDMs have been employed in both educational and non-educational contexts (Rupp and Templin, 2008; de la Torre et al., 2018). Thus far, several reduced and general CDMs have been proposed. Examples of the former are the *deterministic inputs, noisy "and" gate* (DINA; Junker and Sijtsma, 2001) model and *deterministic inputs, noisy "or" gate* (DINO; Templin and Henson, 2006) model; whereas examples of the latter are the generalized DINA (G-DINA; de la Torre, 2011) model, *log-linear* CDM (Henson et al., 2009), and *general diagnostic* model (GDM; von Davier, 2008). When applying CDMs, a fundamental issue is model identifiability of the Q-matrix. For different models, different identifiability conditions have been proposed, including strict identifiability (Liu et al., 2013; Chen et al., 2015; Xu, 2017) and milder identifiability (Chen et al., 2020; Gu and Xu, 2020).

Basically, in the CDM literature, two estimation methods were widely used. The first is the Expectation-Maximization (EM) algorithm within the frequentist framework (de la Torre, 2009, 2011; Huo and de la Torre, 2014; Chiu et al., 2016; George et al., 2016; Minchen et al., 2017; Kuo et al., 2018). However, the main motivation of CDMs is to identify the latent attribute profiles of examinees' and Bayesian methods are often more natural to reach the goal. The second most commonly used method is Markov chain Monte Carlo (MCMC) method (de la Torre and Douglas, 2004; Culpepper, 2015; Culpepper and Hudson, 2018; Zhan et al., 2018, 2019; Jiang and Carter, 2019). Usually, to use the MH algorithm, it is necessary to choose a proposal distribution that

can lead to optimal sampling efficiency. However, empirically determining the optimal proposal distribution can be challenging in practice. Culpepper (2015) first introduced the Gibbs sampling to the DINA model and Zhang et al. (2020) applied the Pólya-Gamma Gibbs sampling based on auxiliary variables to DINA model. Culpepper and Hudson (2018) introduced Bayesian method to the Reduced Reparameterized Unified Model (rRUM; DiBello et al., 1995; Roussos et al., 2007).

With the development of the identifiability, more complex restrictions need to be taken into account. How to estimate more general models incorporating to the corresponding identifiability conditions has been a technically appealing task. In this paper, a sampling method called the slice-within-Gibbs sampler is introduced, in which the identifiability constraints are easy to be imposed. The slice-within-Gibbs sampler can avoid the boring choices of tunning parameters in the MH algorithm and converges faster over the MH algorithm with misspecified proposal distributions. In addition, it has more flexibility over the Gibbs sampling in prior choices and can be easier to apply to more general models compared with the Pólya-Gamma Gibbs Sampling and the Gibbs sampling. In line with the original idea of the slice-within-Gibbs sampler, data would still be augmented with auxiliary variables to make sampling from complicated posterior densities feasible. Existing theoretical results on convergence and stability of the slice-within-Gibbs sampler guarantees that the method is equally applicable to psychometric models, in general, and CDMs, in particular. As such, this paper focuses mainly on demonstrating the usage, as well as evaluating the performance of the slice-within-Gibbs sampler in conjunction with CDMs.

The remainder of this paper is organized as follows. Section 2 provides an overview of CDMs, mainly the G-DINA and DINA models. A detailed slice sampler algorithm for the DINA model is presented in section 3, followed by some advantages of the algorithm. In section 4, two simulations are conducted to illustrate the feasibility of the sampler and its advantages over other MCMC methods. Section 5 contains an application of the slice-within-Gibbs sampler to fraction subtraction data, and section 6 provides a discussion of the findings and limitations of this work and possible future research directions.

2. OVERVIEW OF CDMs

Suppose there are a total of I examinees and J items with K required attributes in a test. Let Y_{ij} denote the binary response of examinee i to item j, and $\boldsymbol{Y} = \{Y_{ij}\}_{I\times J}$ be the response matrix. In CDMs, it is often assumed that the latent trait of examinees is quantified by $K-$dimensional vectors, called attribute profiles. That is, for ith examinee, the latent profile is $\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iK})$, where $\alpha_{ik} \in \{0, 1\}$ and $\alpha_{ik} = 1$ means that examinee i has mastered the kth attribute, whereas $\alpha_{ik} = 0$ otherwise. Therefore, there possibly exist $C = 2^K$ different attribute profile classes, denoted by $\boldsymbol{\alpha_c} = (\alpha_{c1}, \alpha_{c2}, \cdots, \alpha_{cK})$, $c = 1, 2, \cdots, C$. The association between items and attributes is specified by Q-matrix Q=$\{q_{jk}\}_{J\times K}$ (Tatsuoka, 1983), where $q_{jk} = 1$ means the kth attribute is required to answer jth item correctly, and $q_{jk} = 0$ otherwise.

CDMs model the item response Y_{ij} using the following Bernoulli distribution,

$$P(Y_{ij} = y_{ij}|\boldsymbol{\alpha}_i, \Omega_j) = f_{ij}^{y_{ij}} h_{ij}^{1-y_{ij}}, \tag{1}$$

where $f_{ij} = 1 - h_{ij} = P(Y_{ij} = 1|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \Omega_j)$ is the probability of answering item j correctly for examinee i with attribute pattern $\boldsymbol{\alpha}_c$, and Ω_j denotes the unknown parameter set of item j. The likelihood of the data can be written by obtaining the weighted sum across the different attribute profiles. More specifically, assuming an identically and independently distributed latent membership, $\pi_c = P(\boldsymbol{\alpha_i} = \boldsymbol{\alpha_c})$, the joint likelihood can be written as,

$$P(\boldsymbol{Y}|\boldsymbol{\Omega}, \boldsymbol{\pi}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\sum_{c=1}^{C}\pi_c P(Y_{ij} = y_{ij}|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \Omega_j). \tag{2}$$

2.1. The G-DINA Model

The G-DINA model is a saturated CDM that subsumes a number of reduced CDMs. In this model, $P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \Omega_j)$ in Equation (1) is expressed as a function of the main effects and interactions of the required attributes for each item. Following de la Torre (2011), let $K_j^* = \sum_{k=1}^{K} q_{jk}$ denote the number of required attributes for item j. For notational convenience, but without loss of generality, let the first K_j^* attributes be required for item j, and let $\boldsymbol{\alpha}_{ij}^* = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iK_j^*})$ be the reduced vector of $\boldsymbol{\alpha}_i$ associated with item j. The f_{ij} in the G-DINA model for item j is,

$$P\left(Y_{ij} = 1 \left| \boldsymbol{\alpha}_{ij}^*, \boldsymbol{\delta}_j \right.\right) = \delta_{j0} + \sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1}\delta_{jkk'}\alpha_{ik}\alpha_{ik'}\cdots + \delta_{j12\cdots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{ik}, \tag{3}$$

where δ_{j0} is the intercept; δ_{jk} is the main effect of α_{ik}; $\delta_{jkk'}$ is the two-way interaction effect of α_{ik} and $\alpha_{ik'}$; and $\delta_{j12\cdots K_j^*}$ is the K_j^*-way interaction effect of $\alpha_{i1}, \cdots, \alpha_{iK_j^*}$. Aside from the identity link, the G-DINA model can be expressed using log and logit links (de la Torre, 2011).

2.2. The DINA Model

The DINA model is one of most commonly used CDMs, and its f_{ij} is given by

$$P\left(Y_{ij} = 1 \left| \boldsymbol{\alpha}_{ij}^*, s_j, g_j \right.\right) = \begin{cases} 1 - s_j & \text{for } \boldsymbol{\alpha}_{ij}^* = \mathbf{1}, \\ g_j & \text{otherwise}, \end{cases} \tag{4}$$

where g_j and s_j are the guessing and slip parameters, and $\boldsymbol{\alpha}_{ij}^* = \mathbf{1} = (1, 1, \ldots, 1)^T$ denotes that examinee i has possessed all the required attributes of item j. In the DINA model, $\Omega_j = \{g_j, s_j\}$.

As many researchers have already noted, the DINA model is a special case of the G-DINA model. The former can be derived from the latter by setting $\delta_{j0} = g_j$, $\delta_{j0} + \delta_{j12\cdots,K_j^*} = 1 - s_j$, and remaining parameters to zero. Thus, in the DINA model, only the K_j^*-way interaction is taken into account, which indicates that the response is expected to be correct only when all the required attributes have been mastered.

2.3. Identifiability of Restricted Latent Class Models

For most common statistical inferences, the identifiability of the models is a precondition. To guarantee the identifiability when estimating CDMs, we follow a set of sufficient conditions presented by Xu (2017) for a class of restricted latent class models. Specifically, these CDMs need to satisfy the following assumptions:

(*i*) (Monotonicity relations) For any attribute profile $\boldsymbol{\alpha}_i{}'$,

$$\begin{aligned} P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \Omega_j, \boldsymbol{\alpha}_i \succeq \boldsymbol{q}_j) &\geq P(Y_{ij} = 1|\boldsymbol{\alpha}_i{}', \Omega_j) \\ &\geq P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \Omega_j, \boldsymbol{\alpha}_i = \mathbf{0}); \end{aligned} \tag{5}$$

and (*ii*) If $\boldsymbol{q}_j = \boldsymbol{e}_k$ for $k = 1, 2, \ldots, K$,

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \Omega_j, \boldsymbol{\alpha}_i = \mathbf{1}) > P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \Omega_j, \boldsymbol{\alpha}_i \not\succeq \boldsymbol{e}_k), \tag{6}$$

where $\boldsymbol{\alpha} \succeq \boldsymbol{q}$ holds if and only if $\alpha_k \geq q_k$ for any $k \in \{1, 2, \ldots, K\}$ and $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}$ means there exists at least one $k \in \{1, 2, \ldots, K\}$ such that $\alpha_k < q_k$; $\mathbf{0} = (0, 0, \ldots, 0)^T$; and $\boldsymbol{e_k}$ is a vector whose kth element is one and the rest elements are zero.

Both the G-DINA and DINA models are considered restricted latent class models. Specifically, for the DINA model, the above assumptions are equivalent to $1-s_j > g_j$. For the G-DINA model, the transformation is more complicated and will be discussed in section 3.2.

Identifiability in restricted latent class models satisfies the following sufficient conditions (Xu, 2017):

(C1) The Q-matrix is constructed such that

$$Q = \begin{pmatrix} I_K \\ I_K \\ Q' \end{pmatrix},$$

where I_K is a $K \times K$ identity matrix; and

(C2) For any item in Q', examinees who possess no required attributes have the lowest success probabilities. That is, $\min_{\boldsymbol{\alpha}_i \neq \mathbf{0}} P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \Omega_j) > P(Y_{ij} = 1|\boldsymbol{\alpha}_i = \mathbf{0}, \Omega_j)$, for $j > 2K$.

3. INTRODUCING THE SLICE-WITHIN-GIBBS SAMPLER FOR CDMs

In this section, we introduce the slice-within-Gibbs sampler as a method of estimating CDMs. Moreover, we list its advantages.

3.1. Using the Slice-Within-Gibbs Sampler to Estimate CDMs

First, the joint posterior distribution of model parameters $(\boldsymbol{\Omega}, \boldsymbol{\pi})$ could be written as,

$$P(\boldsymbol{\Omega}, \boldsymbol{\pi}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{\Omega}, \boldsymbol{\pi}) P(\boldsymbol{\Omega}, \boldsymbol{\pi}), \tag{7}$$

where $P(\boldsymbol{\Omega}, \boldsymbol{\pi})$ denotes the joint prior distribution.

Step 1: Sample the positive auxiliary variables U_{ij} and V_{ij} from the following posterior distribution,

$$\begin{aligned} U_{ij}|\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\pi} &\sim \text{Uniform}(0, f_{ij}), \text{ if } Y_{ij} = 1, \\ V_{ij}|\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\pi} &\sim \text{Uniform}(0, h_{ij}), \text{ if } Y_{ij} = 0. \end{aligned} \tag{8}$$

The joint posterior distribution $P(\boldsymbol{\Omega}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{V}|\boldsymbol{Y})$ is proportional to

$$\prod_{i=1}^{I}\prod_{j=1}^{J}\left[\mathrm{I}_{(Y_{ij}=1)}(Y_{ij})\mathrm{I}_{(0,f_{ij})}(U_{ij}) + \mathrm{I}_{(Y_{ij}=0)}(Y_{ij})\mathrm{I}_{(0,h_{ij})}(V_{ij})\right] P(\boldsymbol{\Omega}, \boldsymbol{\pi}).$$

Note that $P(\boldsymbol{\Omega}, \boldsymbol{\pi}|\boldsymbol{Y}) = \int_0^\infty \int_0^\infty P(\boldsymbol{\Omega}, \boldsymbol{\pi}, \boldsymbol{U}, \boldsymbol{V}|\boldsymbol{Y}) d\boldsymbol{U} d\boldsymbol{V}$, which means that considering the above posterior distribution is enough to estimate $(\boldsymbol{\Omega}, \boldsymbol{\pi})$; $\mathrm{I}(\cdot)$ denotes the indicator function, and $\mathrm{I}_{(Y_{ij}=1)}(Y_{ij}) = 1$ if $Y_{ij} = 1$, and $\mathrm{I}_{(Y_{ij}=1)}(Y_{ij}) = 0$ otherwise.

Step 2: Sample item parameters Ω_j, $j = 1, 2, \ldots, J$, from the following truncated distribution:

$$\Omega_j \left|\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{V}, \Omega_{-j}, \boldsymbol{\pi}\right. \sim P\left(\Omega_j\right) \mathrm{I}_{\left(\Omega_j^L < \Omega_j < \Omega_j^R\right)}(\Omega_j), \tag{9}$$

where Ω_j^L and Ω_j^R are derived from the identifiability restrictions, and inequalities $0 < U_{ij} < f_{ij}$, and $0 < V_{ij} < h_{ij}$. For example, in the DINA model, $s_j^L = \max\{\max_{i \in \nabla_j}\left\{V_{ij}\right\}, 0\}$, and $s_j^R = \min\{\min_{i \in \triangle_j}\left\{1 - U_{ij}\right\}, 1 - g_j\}$, where $\nabla_j = \{i \left| Y_{ij} = 0, \boldsymbol{\alpha}_{ij}^* = \mathbf{1}\right.\}$ and $\triangle_j = \left\{i \left| Y_{ij} = 1, \boldsymbol{\alpha}_{ij}^* = \mathbf{1}\right.\right\}$. In the G-DINA model,

$$\delta_{jk}^L = \max\left\{\max_{i \in \boldsymbol{\Gamma}_j}\left\{U_{ij} - \left(\delta_{j0} + \sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1}\delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{ik}\right)\right\}, \delta^{*L}\right\}, \text{ and}$$

$$\delta_{jk}^R = \min\left\{\min_{i \in \Pi_j}\left\{1 - V_{ij} - \left(\delta_{j0} + \sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1}\delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{ik}\right)\right\}, \delta^{*R}, 1 - \sum_{-jk}\boldsymbol{\delta_j}\right\},$$

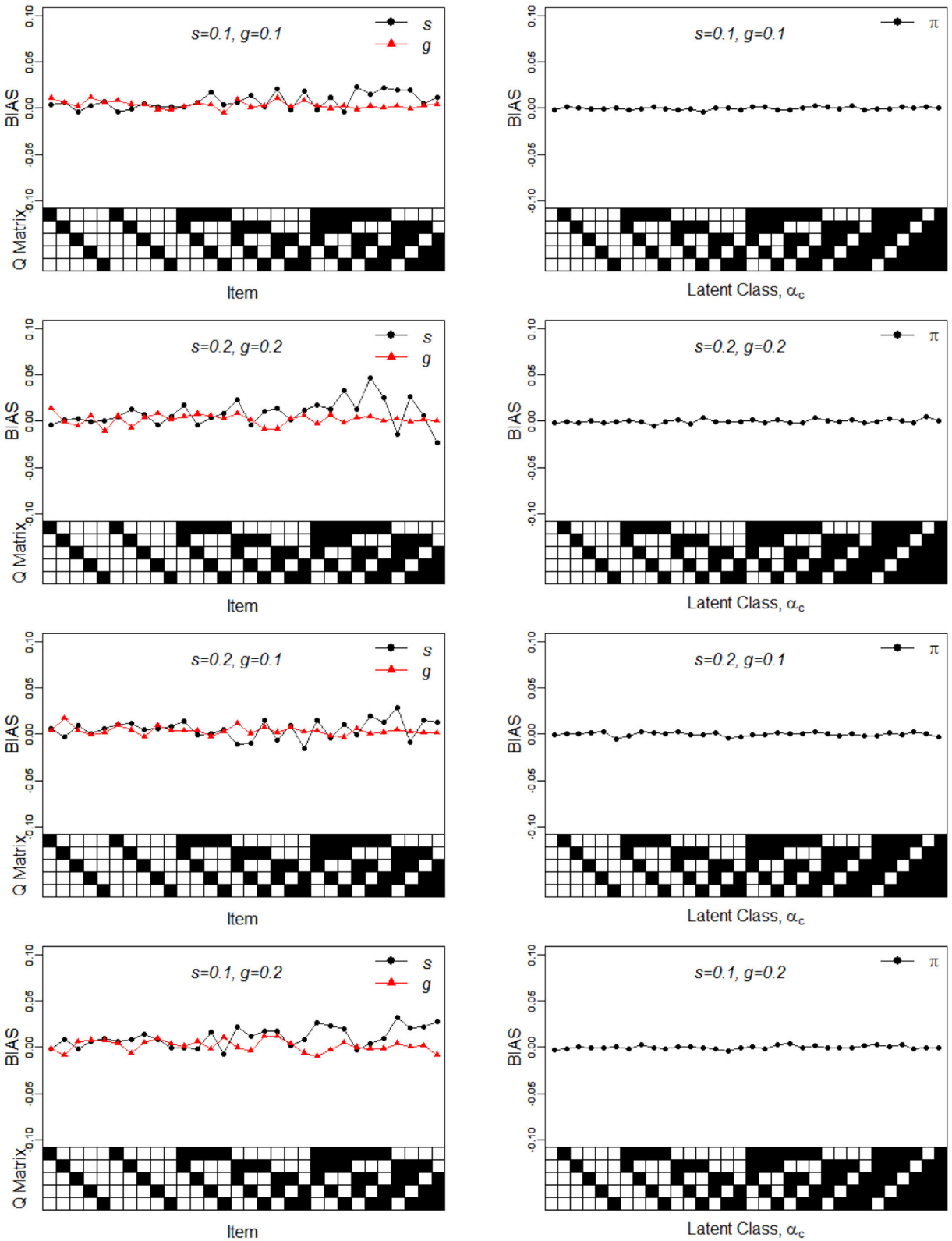


FIGURE 1 | Bias of the slip, guessing, and the latent membership parameters under four different noise levels. Given on the X-axis are the Q-matrix and $\boldsymbol{\alpha}_c$, where a black square denotes the presence of the attribute.

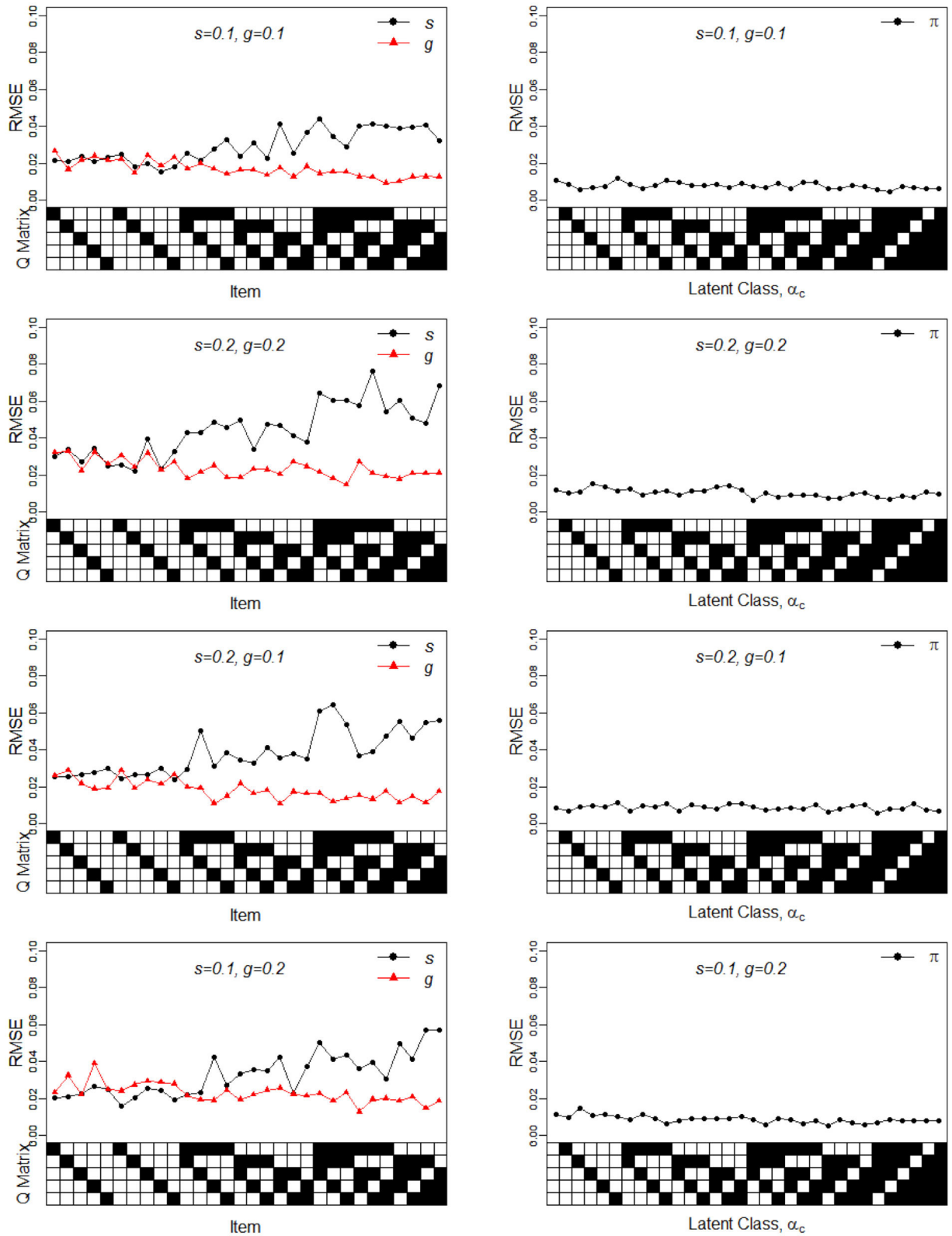


FIGURE 2 | RMSE of the slip, guessing, and the latent membership parameters under four different noise levels. Given on the X-axis are the Q-matrix and $\boldsymbol{\alpha}_c$, where a black square denotes the presence of the attribute.

where $\Pi_j = \{i | Y_{ij} = 0\}$, $\bar{\Pi}_j = \{i | Y_{ij} = 1\}$, and δ^{*L} and δ^{*R} are the lower and upper bounds, respectively, determined from the identifiability conditions of the restricted models.

Step 3: Update the latent membership probabilities $\boldsymbol{\pi}$ and the latent profile $\boldsymbol{\alpha_i}$. Following Huebner and Wang (2011) and Culpepper (2015), the prior of $\boldsymbol{\pi}$ is assumed to follow Dirichlet$(\varphi_0, \ldots, \varphi_0)$. The full conditional distribution of the latent class probabilities $\boldsymbol{\pi}$ can be written as

$$\boldsymbol{\pi} | \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C \sim \text{Dirichlet}\left(\varphi_0 + \sum_{i=1}^{I} \mathrm{I}_{(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_1)}(\boldsymbol{\alpha}_i), \ldots, \varphi_0 + \sum_{i=1}^{I} \mathrm{I}_{(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_1)}(\boldsymbol{\alpha}_i)\right).$$

In this process, $\boldsymbol{\alpha_i}$ is sampled from the distribution

$$\boldsymbol{\alpha}_i \left| \boldsymbol{Y}_i, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi} \right. \sim \text{Multinomial}\left(1, [\varsigma_{i1}, ..., \varsigma_{iC}]\right), \tag{10}$$

where

$$\varsigma_{ic} = P\left(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c | \boldsymbol{Y}_i, \boldsymbol{\Omega}, \boldsymbol{\pi}\right) = \frac{\pi_c \prod_{j=1}^{J} P\left(Y_{ij} \left| \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \Omega_j \right.\right)}{\sum_{c=1}^{C} \pi_c \prod_{j=1}^{J} P\left(Y_{ij} \left| \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \Omega_j \right.\right)}.$$

A number of differences exist in updating the item parameters using the MH algorithm, Gibbs sampling, and slice-within-Gibbs sampler. The MH algorithm samples the new value from a proposal distribution $p_{proposal}(\Omega_j)$. In this paper, we adopted truncated normal distributions as the proposal distributions. Within the Gibbs sampling framework, samples are drawn from the posterior distributions, which is a feature inherited by the slice-within-Gibbs sampler. For practicability, conjugate priors are normally employed for in the Gibbs sampling. In the "dina" 3 R package, for example, Culpepper (2015) used the Gibbs sampler to estimate the DINA model. In contrast, to make sampling more convenient and flexible to implement, the slice-within-Gibbs sampler transforms the posterior distributions of item parameters into a uniform distribution by introducing auxiliary variables. However, for updating the latent membership probabilities $\boldsymbol{\pi}$ and the latent profile $\boldsymbol{\alpha_i}$, the same formula was adopted by all the samplers.

3.2. About the Monotonicity Restrictions

When applying the slice-within-Gibbs sampler, the monotonicity restrictions are needed to cooperate with Step 3 for identifiability. For DINA model, it is easy to implement this constraint, that is, $s_j + g_j < 1$. However for other complex CDMs, it is a bit complicated. In this part, we present how to restrict parameters specially in G-DINA model.

In this part, we only took $K = 3$ as an example. When $K = 3$, there exist at most $2^K = 8$ parameters and corresponding C classes in G-DINA model. The inequalities (5) and (6) are actually equivalent to adopt the following inequality considering all combinations of the $q-$entries.

$$0 \leq \delta_0 \leq \begin{matrix} \delta_0 + \delta_1 \\ \delta_0 + \delta_2 \\ \delta_0 + \delta_3 \end{matrix} \leq \begin{matrix} \delta_0 + \delta_1 + \delta_2 + \delta_{12} \\ \delta_0 + \delta_1 + \delta_3 + \delta_{13} \\ \delta_0 + \delta_2 + \delta_3 + \delta_{23} \end{matrix} \leq \begin{matrix} \delta_0 + \delta_1 + \delta_2 + \delta_3 \\ +\delta_{12} + \delta_{13} + \delta_{23} + \delta_{123} \leq 1. \end{matrix} \tag{11}$$

Therefore, the corresponding bound can be imposed as follows:

1. Consider $\delta_0^{(t+1)} \in [\delta_0^L, \delta_0^R]$ and $\delta_0^L = 0$, $\delta_0^R = \min\{P(1,0,0), P(0,1,0), P(0,0,1)\}$, where $P(\boldsymbol{\alpha}) = P(Y_{ij} = 1 | \boldsymbol{\alpha})$.
2. Consider $\delta_1^{(t+1)}$ which is equivalent to consider $\delta^* = \delta_1^{(t+1)} + \delta_0^{(t+1)}$ and $\delta^* \in [\delta^{*L}, \delta^{*R}]$. And

$$\begin{cases} \delta^{*L} = \delta_0^{(t+1)}, \\ \delta^{*R} = \min\{P(1,1,0), P(1,0,1)\}. \end{cases}$$

3. Apply similar formula to other main-effect parameters.
4. Consider $\delta_{12}^{(t+1)}$ which is equivalent to consider $\delta^* = \delta_0^{(t+1)} + \delta_1^{(t+1)} + \delta_2^{(t+1)} + \delta_{12}^{(t+1)}$ and $\delta^* \in [\delta^{*L}, \delta^{*R}]$. And

$$\begin{cases} \delta^{*L} = \max\{P(1,0,0), P(0,1,0)\}, \\ \delta^{*R} = P(1,1,1). \end{cases}$$

3.3. Some Advantages of the Slice-Within-Gibbs Sampler

The MH algorithm typically relies heavily on the proposal distributions to achieve sampling efficiency. Under unidimensional cases, some researchers suggest that about 50% of candidates need to be accepted for an appropriate proposal distribution to be optimal. The probability of acceptance reduces to around 25% when sampling two- or three-dimensional parameters (Patz and Junker, 1999). For more complex CDMs, this probability needs to drop even more. Compared with MH algorithm, the slice-within-Gibbs sampler as an extension of the Gibbs sampler inherits the high efficiency of the latter. Specifically, the slice-within-Gibbs sampler avoids choosing a proposal distribution because the posterior acts as its proposal distribution. This gives the slice-within-Gibbs sampler acceptance probabilities equal to 1, which makes it highly efficient.

In contrast to the Gibbs sampler, the slice-within-Gibbs sampler has greater flexibility in choosing the prior distributions.

TABLE 1 | True Q-matrix for $K = 3$.

Item	α_1	α_2	α_3	Item	α_1	α_2	α_3
1	1	0	0	16	1	1	0
2	0	1	0	17	1	0	1
3	0	0	1	18	0	1	1
4	1	0	0	19	1	1	0
5	0	1	0	20	1	0	1
6	0	0	1	21	0	1	1
7	1	0	0	22	1	1	0
8	0	1	0	23	1	0	1
9	0	0	1	24	0	1	1
10	1	0	0	25	1	1	1
11	0	1	0	26	1	1	1
12	0	0	1	27	1	1	1
13	1	1	0	28	1	1	1
14	1	0	1	29	1	1	1
15	0	1	1	30	1	1	1

TABLE 2 | True parameters of the G-DINA model items.

Item	δ_{j0} δ_{j0}	δ_{j1} δ_{j1}	δ_{j2}	δ_{j12}	Item	δ_{j0} δ_{j0}	δ_{j1} δ_{j1}	δ_{j2} δ_{j2}	δ_{j12} δ_{j3}	δ_{j12}	δ_{j13}	δ_{j23}	δ_{j123}
1	0.10	0.70			16	0.20	0.10	0.15	0.40				
2	0.10	0.70			17	0.20	0.10	0.15	0.40				
3	0.10	0.70			18	0.20	0.30	0.30	-0.05				
4	0.10	0.70			19	0.20	0.30	0.30	-0.05				
5	0.10	0.70			20	0.20	0.30	0.30	-0.05				
6	0.10	0.70			21	0.20	0.20	0.20	0.00				
7	0.10	0.70			22	0.20	0.20	0.20	0.00				
8	0.10	0.70			23	0.20	0.20	0.20	0.00				
9	0.10	0.70			24	0.20	0.20	0.20	0.00				
10	0.10	0.70			25	0.20	0.10	0.10	0.10	0.05	-0.05	0.05	0.15
11	0.10	0.70			26	0.20	0.10	0.10	0.10	0.10	0.10	0.05	0.05
12	0.10	0.70			27	0.20	0.10	0.10	0.10	0.05	-0.05	0.05	0.15
13	0.20	0.10	0.15	0.40	28	0.20	0.10	0.10	0.10	0.05	-0.05	0.05	0.15
14	0.20	0.10	0.15	0.40	29	0.20	0.10	0.10	0.10	0.10	0.10	0.05	0.05
15	0.20	0.10	0.15	0.40	30	0.20	0.10	0.10	0.10	0.05	-0.05	0.05	0.15

TABLE 3 | Bias and RMSE for s, g, and π estimates under the DINA model.

	True			Bias			RMSE		
I	s	g	π	s	g	π	s	g	π
500	0.100	0.100	0.031	0.008	0.004	0.000	0.029	0.017	0.008
	0.200	0.100	0.031	0.011	0.002	0.000	0.033	0.023	0.009
	0.100	0.200	0.031	0.006	0.004	0.000	0.038	0.018	0.009
	0.200	0.200	0.031	0.009	0.002	0.000	0.044	0.024	0.010
1,000	0.100	0.100	0.031	0.003	0.001	0.000	0.022	0.012	0.006
	0.200	0.100	0.031	0.005	0.001	0.000	0.023	0.016	0.006
	0.100	0.200	0.031	0.004	0.002	0.000	0.030	0.013	0.006
	0.200	0.200	0.031	0.003	0.001	0.000	0.031	0.017	0.007
3,000	0.100	0.100	0.031	0.001	0.000	0.000	0.012	0.007	0.003
	0.200	0.100	0.031	0.003	0.000	0.000	0.013	0.010	0.004
	0.100	0.200	0.031	0.001	0.000	0.000	0.017	0.008	0.004
	0.200	0.200	0.031	0.001	0.000	0.000	0.019	0.011	0.005

TABLE 4 | Bias and RMSE for s, g and π estimates under the negatively correlated DINA model parameters.

	Bias			RMSE		
I	s	g	π	s	g	π
500	0.009	0.002	0.000	0.028	0.014	0.008
1,000	0.004	0.002	−0.000	0.018	0.011	0.006
3,000	0.002	0.000	0.000	0.011	0.006	0.003

TABLE 5 | Bias of δ and π estimates under the G-DINA model.

I	K_j^*	Intercept	One-way	Two-way	Three-way	π
500	1	0.007	−0.006	—	—	
	2	0.008	−0.001	-0.007	—	0.000
	3	0.017	0.021	−0.044	0.053	
1,000	1	0.003	−0.003	—	—	
	2	0.003	−0.000	−0.004	—	0.000
	3	0.008	0.006	−0.016	0.020	
3,000	1	0.001	−0.001	—	—	
	2	0.001	0.000	−0.001	—	0.000
	3	0.003	−0.001	0.000	−0.001	

TABLE 6 | RMSE of δ and π estimates under the G-DINA model.

I	K_j^*	Intercept	One-way	Two-way	Three-way	π
500	1	0.022	0.035	—	—	
	2	0.039	0.057	0.084	—	0.015
	3	0.059	0.070	0.115	0.182	
1,000	1	0.015	0.024	—	—	
	2	0.028	0.042	0.061	—	0.012
	3	0.040	0.051	0.080	0.122	
3,000	1	0.009	0.015	—	—	
	2	0.016	0.025	0.037	—	0.006
	3	0.023	0.034	0.051	0.074	

Although highly efficient, finding easy-to-use conjugate prior distributions renders the use of the Gibbs sampler challenging in practice. However, this is not an issue with the slice-within-Gibbs sampler - its efficiency is not affected by the choice of prior distributions. Even if misspecified priors are adopted, it can obtain satisfactory results.

Thus, the slice-within-Gibbs sampler not only has a relatively high convergence rate, but also overcomes the dependence on

the conjugate prior. Moreover, based on Theorem 7 in Mira and Tierney (2002), it can easily be shown that the slice-within-Gibbs sampler when used with CDMs is uniformly ergodic because f_{ij} is bounded by 1. However, it should be noted that a few other MCMC algorithms exhibit this robust property (Mira and Tierney, 1997; Roberts and Rosenthal, 1999).

4. SIMULATION STUDY

In this section, two simulation studies were conducted to evaluate the performance of the slice-within-Gibbs sampler in the CDM context. Simulation 1 was designed mainly to examine the extent the slice-within-Gibbs sampler can accurately recover the parameters of the DINA model and G-DINA models; Simulation 2 was designed to document the advantages of the slice-within-Gibbs sampler over the MH algorithm and Gibbs sampling in estimating the DINA model.

4.1. Simulation Study 1

4.1.1. Design

In Simulation Study 1, the number of attributes for the DINA and G-DINA models was fixed to $K = 5$ and $K = 3$, respectively, whereas the number of items was set to $J = 30$. The Q-matrices for the DINA model given in **Figures 1, 2** and for the G-DINA model given in **Table 1** were designed to ensure the identifiability of restricted latent class models.

In this context, the number of examinees I is set as 500, 1000, 3000. As for item parameters of DINA model, five different conditions were considered. Following Huebner and Wang (2011) and Culpepper (2015), four noise levels (i.e., item qualities) were considered: (1) a low noise level - $s_j = g_j = 0.1$; (2) a high noise level - $s_j = g_j = 0.2$; (3) the slip parameter was higher than the guessing parameter - $s_j = 0.2$, $g_j = 0.1$; and (4) the guessing parameter was higher than the slip parameter - $s_j = 0.1$, $g_j = 0.2$. A fifth condition was considered, where, as in Zhan et al. (2018), the negative correlation between the item parameters based on the empirical data was taken into account. Specifically, the guessing and slip parameters were generated from the following: $(\text{logit}(g_j), \text{logit}(s_j)) \sim N(\begin{pmatrix}-2.564\\-1.995\end{pmatrix}, \begin{pmatrix}1.233 & -0.415\\-0.415 & 0.571\end{pmatrix})$. Under this distribution, the mean guessing and slip parameters were 0.096 and 0.103, respectively; the corresponding maxima were 0.365 and 0.484, respectively. The true parameters of the G-DINA model are listed in **Table 2**. Finally, the latent class membership probabilities $\boldsymbol{\pi}$ were set to be equal for the different latent classes.

In this simulation study, all priors were set to be non-informative. With respect to the item parameters, priors of the slip and guessing parameters to the DINA model were set to be Uniform(0, 1), whereas $P(\boldsymbol{\delta}) \propto 1$ in the support set was assumed for the G-DINA model.

Two criteria were used to evaluate quality of the parameter recovery, namely, the bias and root mean squared error (RMSE) of $\boldsymbol{s}$, $\boldsymbol{g}$, $\boldsymbol{\delta}$, $\boldsymbol{\pi}$ across 25 replications. In both simulation studies, the slice-within-Gibbs sampler was iterated 20,000 times for each replication, where the first 10,000 iterations were discarded as burn-in.

To evaluate the convergence, four chains started at overdispersed starting values were run. The potential scale reduction factor (PSRF) $\widehat{R}$ (Brooks and Gelman, 1998) was computed using the R package "coda" (Plummer et al., 2006). A value of $\widehat{R}$ less than 1.1 (Brooks and Gelman, 1998) was used as the criterion for chain convergence.

4.1.2. Results

It was verified that the number of iterations and burn-in were sufficient for the chain to converge. For example, **Figure 3** shows the $\widehat{R}$ in G-DINA model for sample size $I = 1000$ that all the parameters came down to 1.1 at the 7266th iteration.

Table 3 shows the parameter recovery results of the slice-within-Gibbs sampler under the DINA model, **Table 4** shows the parameter recovery under the condition with negatively correlated item parameters, and **Tables 5**, **6** the results under the G-DINA model across different sample sizes and item qualities.

For the smallest sample size (i.e., $I = 500$), the maximum absolute bias of the item parameter estimates was 0.011 and 0.053 for the DINA and G-DINA models; the RMSE was below 0.044 and 0.182 for the DINA and G-DINA models, respectively. With the exception of the higher-order interaction terms when $K_j^* = 3$, these results indicate that satisfactory estimates can be obtained for the DINA and G-DINA models using the slice-within-Gibbs sampler even with sample size as small as $I = 500$. As the table shows, the performance of the slice-within-Gibbs sampler improved as the number of examinees increased. When $I = 3,000$, the absolute bias and RMSE of all the item parameters

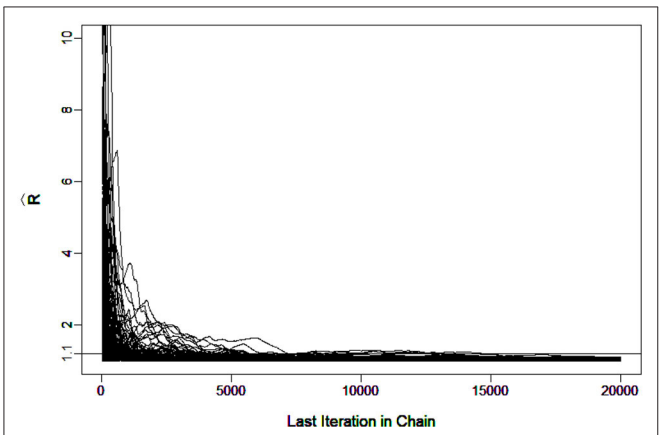


FIGURE 3 | Trace Plots of $\widehat{R}$ under G-DINA model for $I = 1,000$.

TABLE 7 | Bias and RMSE of the slice-within-Gibbs sampler and MH algorithm.

	Slice sampler		MH - Case 1		MH - case 2	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
s	0.005	0.027	0.006	0.026	0.013	0.035
g	0.002	0.016	0.004	0.018	0.001	0.020
π	0.000	0.008	0.000	0.008	0.000	0.008

were smaller, and their maximum values dropped to 0.003 and 0.019, respectively, for the DINA model, and to 0.003 and 0.074, respectively, for the G-DINA model. For the condition where the item parameters were negatively correlated, the average bias and RMSE were comparable to those obtained under the low-noise level condition. Finally, for the latent membership probabilities, all the parameters can be estimated extremely accurately (i.e., bias is 0.00) for both models. Moreover, the maximum RMSEs at $I = 500$ were 0.010 and 0.015 for the DINA and G-DINA models, respectively, and improved with larger sample sizes.

To better understand the properties of the slice-within-Gibbs sampler, **Figures 1**, **2** show the detailed results for $I = 500$ size under DINA model. Consistent with the results in Culpepper (2015) and de la Torre (2009), which were obtained using different estimation algorithms, worse results were obtained for items that required more attributes. The deterioration in the quality of item parameter estimates as the number of required attributes increased can be clearly observed in **Figure 2**, which displays the RMSE of the estimates. It should be noted that the guessing parameter estimates did in fact slightly improve with more required attributes; however, the improvement did not compensate for the stark deterioration in the slip parameter estimates. These results underscore that fact that, given a fixed same sample size, the quality of item parameter estimates of the DINA model can affected by of the number of required attributes. Finally, **Figures 1**, **2** indicate that item quality had only a small impact on the recovery on the individual latent class membership probabilities.

In sum, the results of Simulation Study 1 indicates that the slice-within-Gibbs sampler can provide accurate estimates of the DINA and G-DINA model parameter estimates. Moreover, it can provide results consistent with those of previously implemented algorithms.

4.2. Simulation Study 2

This simulation study had two-fold goals: (1) to compare the efficiency of the slice-within-Gibbs sampler to that of MH algorithm; and (2) to compare the slice-within-Gibbs sampler and Gibbs sampler in terms their flexibility in specifying the priors. For this study, the MH algorithm, Gibbs sampling and slice sampler were compared in the context of DINA model.

4.2.1. Design

The simulated data contained $I = 500$ examinees, $J = 30$ items and $K = 5$ attributes. All the slip and guessing parameters were set to 0.1, and the Q-matrix given in **Figure 1** was used.

For the MH algorithm, there exist infinite choices of proposal distributions. For demonstration purposes, this simulation study only considered the following two cases of the proposal distributions.

- Case 1: A larger step between iterations, where $s_j \sim N(s_j^{(t)}, 1)$, and $g_j \sim N(g_j^{(t)}, 1)$; and
- Case 2: A smaller step between iterations, where $s_j \sim N(s_j^{(t)}, 0.001)$, a $g_j \sim N(g_j^{(t)}, 0.001)$.

For the Gibbs sampling, the Beta family distributions were the conjugate priors of the items parameters. Following Culpepper (2015), only the conjugate prior Beta (1, 1) was considered.

For the slice-within-Gibbs sampler, both conjugate and non-conjugate priors were considered. Below are the two cases of the priors and their specific instances.

- Case 3: For conjugate priors, Beta $(1, 1)$, Beta $(1, 2)$ and Beta $(2, 2)$ were used; and
- Case 4: For non-conjugate priors, $N(0, 1)\, I_{(0,1)}(x)$, $N(2, 1)\, I_{(0,1)}(x)$, Uniform$(0, 2)\, I_{(0,1)}(x)$, and Exp $(1)\, I_{(0,1)}(x)$ were used.

TABLE 8 | Bias and RMSE of the slice-within-Gibbs sampler and Gibbs algorithm.

	Slice sampler - Case 3		Slice sampler - Case 4		Gibbs	
	Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
s	0.009	0.031	0.008	0.030	0.009	0.031
g	0.002	0.017	0.002	0.017	0.002	0.017
π	−0.000	0.008	0.000	0.008	0.000	0.008

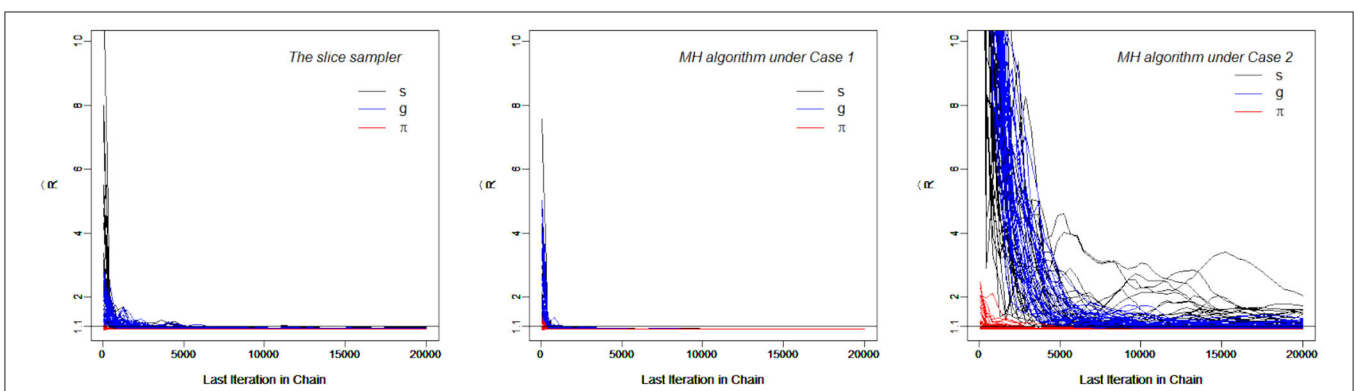


FIGURE 4 | The trace Plots of $\widehat{R}$ for the slice-within-Gibbs sampler and MH algorithms in Simulation Study 2.

As in Simulation Study 1, bias and RMSE were calculated to evaluate the quality of the parameter estimates. Similarly the PSRF was computed to evaluate convergence.

4.2.2. Results

Table 7 presents the recovery results of the slice-within-Gibbs sampler with uniform prior and MH algorithm under Cases 1 and Case 2. The results show that the accuracy of the MH algorithm parameter estimates was greatly influenced by the variance of proposal distribution. Specifically, the parameter estimates under Case 2 were worse than those under Case 1, which indicates that, for this particular condition, a smaller step between iterations was not a good as a larger step. It is also noteworthy that, despite the use of a uniform prior, the slice-within-Gibbs sampler provided estimates that were as good as, if not better than estimates obtained using the MH algorithm under Case 1.

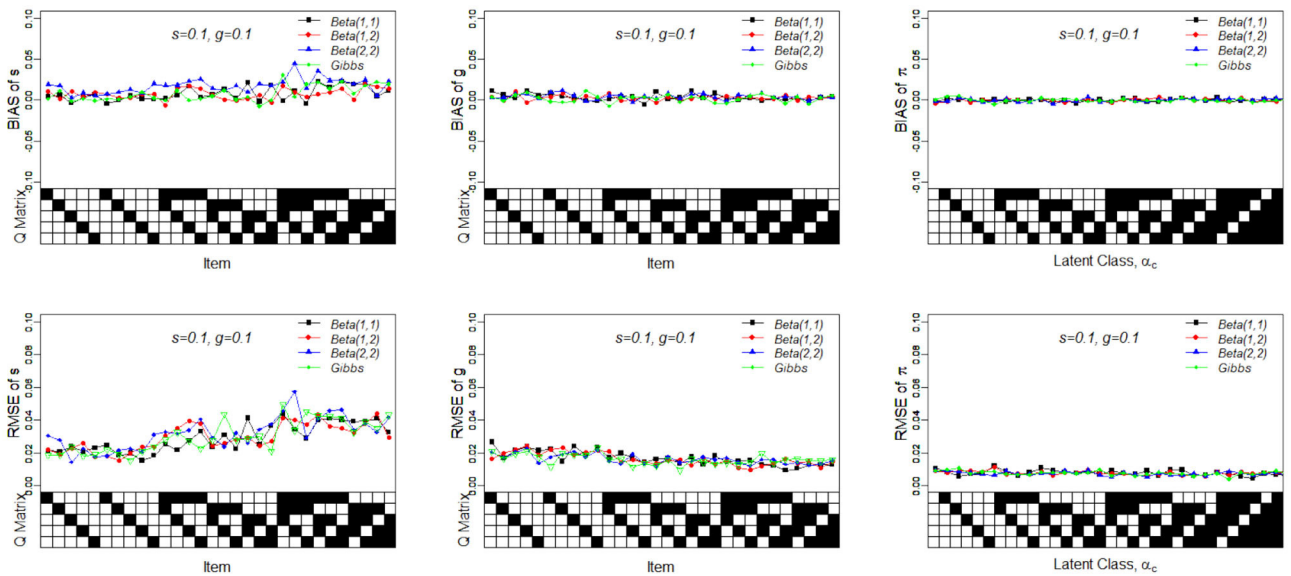


FIGURE 5 | Bias and RMSE of slip, guessing, and the latent membership parameters based on different conjugate priors. Given on the X-axis are the Q-matrix and $\boldsymbol{\alpha_c}$, where a black square denotes the presence of the attribute.

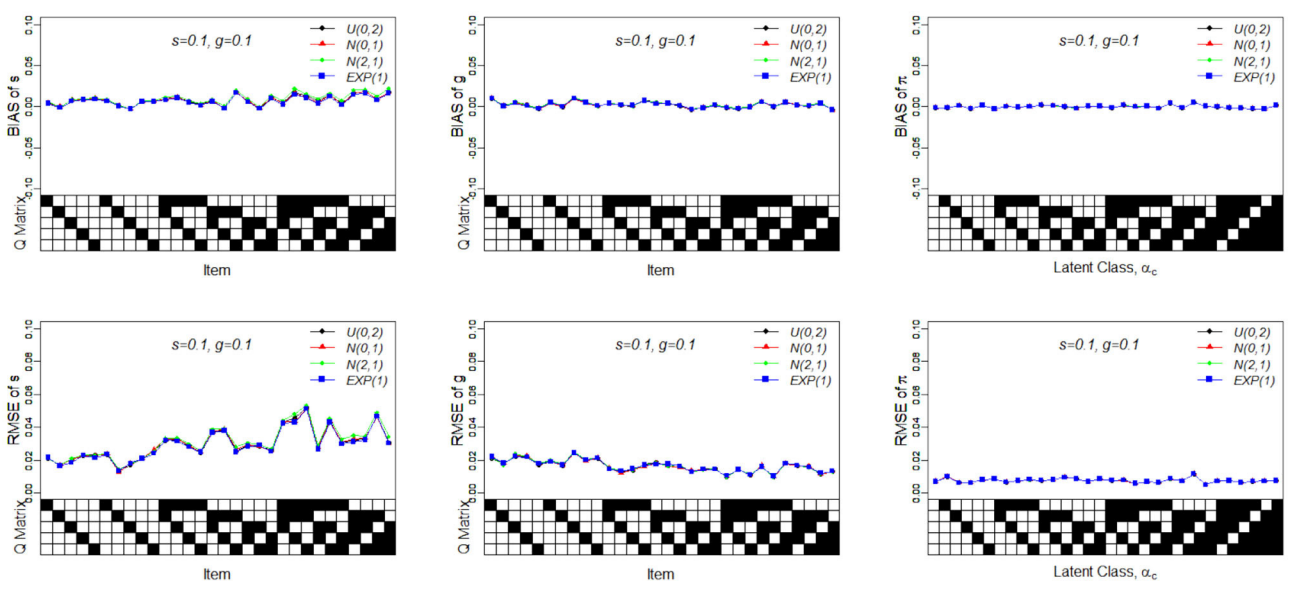


FIGURE 6 | Bias and RMSE of slip, guessing and the latent membership parameters based on different non-conjugate priors. Given on the X-axis are the Q-matrix and $\boldsymbol{\alpha_c}$, where a black square denotes the presence of the attribute.

Figure 4, which contains the $\widehat{R}$s for the slice-within-Gibbs sampler and MH algorithms across different iterations, shows the differing convergence rates of the two methods. As can be seen from the figure, $\boldsymbol{\pi}$ converged at the fastest rate, followed by g. For the MH algorithm, Case 1 converged faster than Case 2 - Case 1 reached convergence by the 1000th iteration, whereas Case 2 did not even reach convergence for some parameters. This indicates that the variance of the proposal distribution in Case 2 was too small to sufficiently explore the posterior distribution. In comparison, all the parameters estimated using the slice-within-Gibbs sampler reached convergence by the 2000th iteration.

TABLE 9 | Convergence rates of the Gibbs and slice samplers.

	Convergence observed	
Sampler	**Iterations**	**Time (min)**
Slice	1,573	6.006
Gibbs	258	4.654

TABLE 10 | The Q-matrix of the fraction subtraction data.

Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0
2	1	1	1	1	0
3	1	0	0	0	0
4	1	1	1	1	1
5	0	0	1	0	0
6	1	1	1	1	0
7	1	1	1	1	0
8	1	1	0	0	0
9	1	0	1	0	0
10	1	0	1	1	1
11	1	0	1	0	0
12	1	0	1	1	0
13	1	1	1	1	0
14	1	1	1	1	1
15	1	1	1	1	0

Table 8 presents the recovery results of the slice-within-Gibbs sampler with Beta $(1, 2)$ prior under Case 3 and $N(0, 1)\, I_{(0,1)}(x)$ prior under Case 4. **Figure 5** shows the bias and RMSE of the slice-within-Gibbs sampler under Case 3 (i.e., conjugate priors) and the Gibbs sampler under Beta $(1, 1)$. It can be seen that the slice-within-Gibbs sampler performed similarly to the Gibbs sampler, particularly for g_j and π_c. Although the estimates of s_j had a larger variability across the four priors, none of them was uniformly the best across the 30 items. The figure also shows that the slice-within-Gibbs sampler provided comparable results under the family of beta priors. Finally, regardless of the Beta priors used, the bias and RMSE of s_j were always higher than those of g_j and π_c, which is consistent with the previous results.

Figure 6 presents the recovery of the slice-within-Gibbs sampler under Case 4. It should be noted that the Gibbs sampler does not work under these specific priors. In contrast, the slice-within-Gibbs sampler can also be applied with different non-conjugate, even misspecified priors. The figures shows that the biases of s_j, g_j, and π_c were close to zero, and the corresponding RMSEs were below 0.05. Despite the use of non-conjugate priors, these results were almost the same those obtained using the Gibbs sampler under Beta $(1, 1)$.

Table 9 compares the convergence rate of the Gibbs and slice-within-Gibbs samplers. Specifically, the simulated data based on the DINA model used $I = 500$ examinees, $J = 30$ items, $K = 5$ attributes and the Q-matrix in **Figure 1**. For comparison purposes, two criteria were used to evaluate the convergence

TABLE 11 | The EAP of the latent class parameters under G-DINA model.

Latent classes					$\widehat{\pi}$		Latent classes					$\widehat{\pi}$	
α_1	α_2	α_3	α_4	α_5	**EAP**	**SE**	α_1	α_2	α_3	α_4	α_5	**EAP**	**SE**
0	0	0	0	0	0.005	0.004	1	1	1	0	0	0.138	0.011
1	0	0	0	0	0.005	0.003	1	1	0	1	0	0.014	0.006
0	1	0	0	0	0.004	0.003	1	1	0	0	1	0.003	0.003
0	0	1	0	0	0.079	0.011	1	0	1	1	0	0.001	0.003
0	0	0	1	0	0.110	0.010	1	0	1	0	1	0.006	0.003
0	0	0	0	1	0.004	0.004	1	0	0	1	1	0.013	0.005
1	1	0	0	0	0.006	0.004	0	1	1	1	0	0.003	0.002
1	0	1	0	0	0.035	0.007	0	1	1	0	1	0.001	0.001
1	0	0	1	0	0.009	0.006	0	1	0	1	1	0.006	0.001
1	0	0	0	1	0.003	0.002	0	0	1	1	1	0.001	0.001
0	1	1	0	0	0.006	0.006	1	1	1	1	0	0.118	0.013
0	1	0	1	0	0.002	0.002	1	1	1	0	1	0.083	0.008
0	1	0	0	1	0.002	0.002	1	1	0	1	1	0.001	0.001
0	0	1	1	0	0.003	0.002	1	0	1	1	1	0.001	0.001
0	0	1	0	1	0.001	0.001	0	1	1	1	1	0.001	0.001
0	0	0	1	1	0.002	0.002	1	1	1	1	1	0.335	0.015

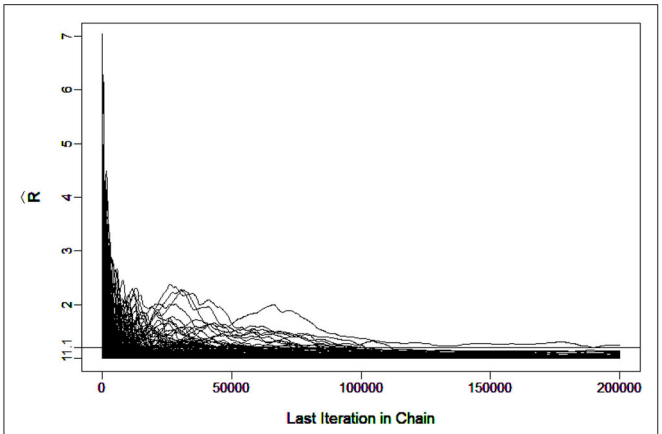


FIGURE 7 | Trace Plots of $\widehat{R}$ for the real data in the GDINA model.

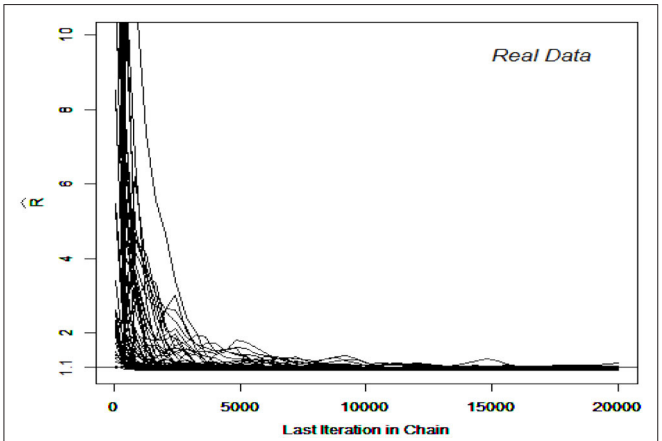


FIGURE 8 | Trace Plots of $\widehat{R}$ for the real data in the DINA model.

rates, namely, the iterations at which all the parameters reached convergence and the time to reach 20,000 iterations. Based on 100 replications, **Table 11** shows that the Gibbs sampler converged much earlier and was about 1.29 times faster than the slice-within-Gibbs sampler.

Overall, the results of Simulation Study 2 indicate that, depending on the proposal distribution, the slice-within-Gibbs sampler can be dramatically more or slightly less efficient than the MH algorithm. However, the MH algorithm is advantageous only to the extent that the proposal distribution is optimal, whereas the slice-within-Gibbs sampler can be implemented with a wide range of prior distributions. Similarly, although the slice and Gibbs samplers are comparable, the former, unlike the latter, is not restricted to the use of conjugate priors.

5. EMPIRICAL EXAMPLE

5.1. Data

The empirical example involved fraction subtraction data previously analyzed by Tatsuoka (1990), Tatsuoka (2002) and de la Torre (2009). The data analyzed here consisted of responses of 536 students to 15 fraction subtraction items. The five attributes measured by the test were: α_1 subtracting basic fractions; α_2 reducing and simplifying; α_3 separating whole from fraction; α_4 borrowing one from whole; and α_5 converting whole to fraction. The corresponding Q-matrix is given in **Table 10**.

5.2. Methods and Results

The DINA and G-DINA models were fitted to the data and the corresponding parameters estimated using the slice-within-Gibbs sampler, with monotonicity constraints imposed to stabilize the estimates due to the relatively small sample size. Incidentally, the Gibbs sampler was not considered for these data due to the difficulty in finding conjugate priors that can also accommodate the monotonicity constraints. The estimates based on the expected a posteriori (EAP) and the corresponding standard errors (SEs) were computed for DINA and G-DINA models. Finally, the deviance information criterion (DIC) was employed to select between the two models. **Figures 7**, **8** show the $\widehat{R}$ for the G-DINA and DINA analyses of the empirical data, respectively. In addition to the convergence of the chains, the figures also show that the DINA model converged faster than the G-DINA model for these data.

In terms of DIC, a model with smaller DIC is to be preferred (Spiegelhalter et al., 2002). In fitting the fraction subtraction data, the DICs of the DINA and G-DINA models were 27719.86 and 27017.43, respectively, which indicates that the G-DINA model provided a better fit to data. Thus, only results pertaining to the G-DINA model are presented below.

Table 11 contains the EAP estimates of the latent membership parameters, $\widehat{\pi}_c$, and their corresponding SEs under the G-DINA model. The eight latent classes with the largest memberships were: $\pi(1,1,1,1,1) = 0.335$, $\pi(1,1,1,0,0) = 0.138$, $\pi(1,1,1,1,0) = 0.118$, $\pi(0,0,0,1,0) = 0.110$, $\pi(1,1,1,0,1) = 0.083$, $\pi(0,0,1,0,0) = 0.079$, $\pi(1,0,1,0,0) = 0.035$, and $\pi(1,1,0,1,0) = 0.014$. They accounted for over 91% of the latent class memberships. In terms of individual attribute mastery, α_1 through α_5 had the following prevalences: 0.771, 0.723, 0.812, 0.620, and 0.463, which makes α_3 and α_5 the easiest and most difficult attributes to master, respectively. It can be noted that latent classes which showed mastery of all but one of the three easiest attributes to master, as in (0,1,1,1,1), (1,0,1,1,1), and (1,1,0,1,1), had the lowest latent class memberships. In this example, it can be noted that latent classes with the largest class memberships also had the larger SEs.

Table 12 gives the G-DINA model estimates of the fraction subtraction items in term of the latent group success probability $P(\boldsymbol{\alpha}_{ij}^*)$. The item parameter estimates clearly show why the G-DINA model was preferred over the DINA model. For the DINA model to provide a satisfactory fit to the data, all parameters and except the intercept and the highest-order interaction effect must be equal to zero. This was not the case with, say, items that require two attributes (i.e., items 8, 9, and 12). For these items, $P(00) < P(10)$ and $P(00) < P(01)$ indicating that the main effects are not equal to zero. The remaining multi-attribute items also indicate that the conjunctive assumption of the DINA model was not tenable. As a rough measure of item discrimination,

TABLE 12 | Results of the fraction subtraction data analysis under G-DINA model using the slice-within-Gibbs sampler and EM algorithm.

		$P(\alpha_{ij}^*)$															
		0	1														
		00	10	01	11												
		000	100	010	001	110	101	011	111								
		0000	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1011	1101	0111	1111
		00000	10000	01000	00100	00010	00001	11000	10100	10010	10001	01100	01010	01001	00110	00101	00011
Method	Item	11100	11010	11001	10110	10101	10011	01110	01101	01011	00111	11110	11101	11011	10111	01111	11111
	1	0.103	0.877														
	2	0.029	0.170	0.054	0.064	0.353	0.301	0.530	0.527	0.110	0.426	0.496	0.655	0.650	0.823	0.632	0.922
	3	0.511	0.969														
	4	0.003	0.026	0.008	0.009	0.017	0.069	0.064	0.148	0.058	0.400	0.025	0.045	0.151	0.072	0.140	0.144
		0.315	0.126	0.533	0.277	0.513	0.484	0.189	0.291	0.280	0.275	0.549	0.718	0.674	0.673	0.498	0.939
	5	0.370	0.853														
	6	0.003	0.056	0.009	0.009	0.037	0.237	0.217	0.409	0.024	0.167	0.156	0.425	0.676	0.630	0.353	0.840
	7	0.011	0.043	0.036	0.045	0.073	0.118	0.453	0.643	0.099	0.155	0.328	0.547	0.723	0.908	0.419	0.963
	8	0.119	0.823	0.857	0.957												
The slice	9	0.060	0.904	0.755	0.964												
sampler	10	0.006	0.097	0.027	0.031	0.147	0.332	0.202	0.561	0.326	0.335	0.253	0.716	0.736	0.700	0.584	0.952
	11	0.055	0.828	0.757	0.940												
	12	0.006	0.060	0.061	0.023	0.236	0.662	0.735	0.935								
	13	0.006	0.324	0.047	0.046	0.023	0.451	0.494	0.657	0.110	0.204	0.254	0.605	0.747	0.783	0.388	0.863
	14	0.001	0.012	0.006	0.006	0.007	0.011	0.050	0.058	0.046	0.069	0.017	0.042	0.056	0.051	0.060	0.063
		0.182	0.130	0.213	0.174	0.227	0.268	0.166	0.186	0.221	0.220	0.433	0.457	0.523	0.564	0.474	0.891
	15	0.002	0.011	0.012	0.009	0.017	0.074	0.067	0.301	0.032	0.354	0.324	0.278	0.660	0.677	0.597	0.910
	1	0.088	0.880														
	2	0.000	0.000	0.000	0.096	0.298	0.000	0.419	0.811	0.096	0.298	0.318	0.433	0.871	0.811	0.366	0.908
	3	0.513	0.970														
	4	0.000	0.000	0.000	0.000	0.031	0.000	0.014	0.000	0.031	0.235	0.000	0.031	0.000	0.031	0.000	0.031
		0.587	0.031	0.819	0.319	0.396	0.908	0.031	0.000	0.345	0.031	0.908	0.823	0.908	0.908	0.346	0.908
	5	0.327	0.839														
	6	0.000	0.000	0.000	0.000	0.025	0.100	0.104	0.025	0.000	0.025	0.025	0.300	0.641	0.833	0.025	0.833
	7	0.000	0.000	0.000	0.019	0.051	0.000	0.019	0.335	0.110	0.066	0.051	0.586	0.972	0.768	0.332	0.972
	8	0.000	0.853	0.953	0.953												
EM	9	0.032	0.912	0.721	0.971												
algorithm	10	0.000	0.000	0.000	0.000	0.000	0.000	0.168	0.751	0.215	0.000	0.945	0.712	0.945	0.945	0.945	0.945
	11	0.011	0.783	0.770	0.952												
	12	0.000	0.052	0.039	0.000	0.052	0.641	0.930	0.930								
	13	0.000	0.000	0.054	0.047	0.000	0.647	0.248	0.362	0.090	0.066	0.047	0.857	0.837	0.857	0.457	0.857
	14	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		0.265	0.098	0.000	0.000	0.301	0.245	0.000	0.000	0.000	0.000	0.879	0.740	0.245	0.497	0.001	0.879
	15	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.122	0.000	0.000	0.704	0.630	0.914	0.914

$\Delta_j = P_j(\mathbf{1}) - P_j(\mathbf{0})$, was computed. All but two items had $\Delta_j > 0.70$, and the average item discrimination was $\bar{\Delta}_j = 0.829$. These results indicate that the fraction subtraction items are highly discriminating.

For comparison purpo ses, the EM estimates of the same items were also obtained using the R package "GDINA" (Ma and de la Torre, 2020), and are given **Table 12**. It can be noted that for one-attribute items (i.e., items 1 and 5), the slice-within-Gibbs sampler and EM estimates are highly comparable. However, for multi-attribute items, the estimates can be quite disparate, except for $P_j(\mathbf{1})$ was comparable across the two methods. The difference could be due to the small sample size relative to the complexity of the G-DINA model for the multi–attribute items. To better understand the behavior of the slice-within-Gibbs sampler and EM algorithm vis-a-vis the fraction subtraction data, we conducted a simulation study where data were generated based on the parameters obtained using the slice-within-Gibbs sampler. **Figure 9** shows the mean absolute error (MAE) of the two estimates across 100 replications. The figure indicates that the slice-within-Gibbs sampler had smaller mean absolute biases

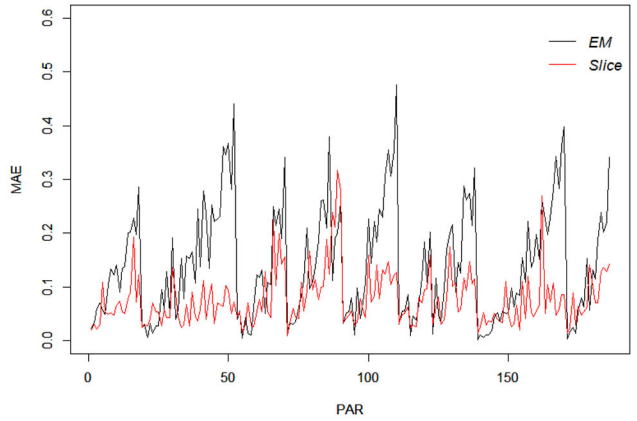


FIGURE 9 | Mean absolute bias of the slice-within-Gibbs sampler and EM estimates of data simulated based on the fraction subtraction data.

for most of the parameters, thus, a more reliable method for the fraction subtraction data.

6. DISCUSSION

In this work, the slice-within-Gibbs sampler is introduced as a method of estimating CDMs. Unlike the MH algorithm, the slice-within-Gibbs sampler obviates the need to choosing an optimal proposal distribution; unlike Gibbs sampler, the slice-within-Gibbs sampler has the flexibility to work with a wider range of prior distributions. As shown in the simulation studies and empirical example, it can be used to estimate complex CDMs, such as the G-DINA model. Thus, the slice-within-Gibbs sampler provides an alternative and viable estimation procedure in the context of CDMs.

Based on the results of **Table 9**, additional work is needed to speed up the implementation of the slice-within-Gibbs sampler for researchers to be able to fully take advantage of the flexibility of the sampler to estimate a wide range CDMs.

In the present work, only two CDMs (i.e., DINA and G-DINA models) were employed to illustrate the slice-within-Gibbs sampler. However, the slice-within-Gibbs sampler can be easily extended to other CDMs (e.g., additive CDM, GDM), attribute structure (e.g., higher-order CDMs; de la Torre and Douglas, 2004), and potentially to CDMs that incorporate various types of covariates.

Finally, it should be noted that other MCMC sampling procedures that use auxiliary variables are currently available. One such procedure is the Hamiltonian Monte Carlo (Neal, 2011; Duane et al., 1987, HMC) algorithm. The HMC algorithm is based on the *Hamiltonian dynamics*, and has a physical interpretation and can provide useful intuitions. As an extension of the MH algorithm, it exploits the gradient information to draw samples from the posterior. Because HMC algorithm often provides a large move with acceptance rates close to one, its efficiency is higher than that of the MH algorithm. Future research should systematically compare the performance of the slice-within-Gibbs sampler and the HMC algorithm, as well as other auxiliary-variable sampling procedures, in estimating CDMs.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://CRAN.R-project.org/package=CDM.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

XX and JG worked on the technical details of the article. XX and JZ completed the writing, with the support of JT. JT and JZ provided a few suggestions on the focus and direction of the research. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the National Natural Science Foundation of China (11871141). XX (CSC No. 201806620023) and JG (CSC No. 201906620083) were sponsored by the China Scholarship Council as joint Ph.D. students.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.02260/full#supplementary-material

REFERENCES

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Chen, Y., Culpepper, S., and Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika* 85, 121–153. doi: 10.1007/s11336-019-09693-2

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification models. *J. Am. Stat. Assoc.* 110, 850–866. doi: 10.1080/01621459.2014.934827

Chiu, C.-Y., Köhn, H.-F., and Wu, H.-M. (2016). Fitting the reduced rum with mplus: a tutorial. *Int. J. Test.* 16, 331–351. doi: 10.1080/15305058.2016.1148038

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *J. Educ. Behav. Stat.* 40, 454–476. doi: 10.3102/1076998615595403

Culpepper, S. A., and Hudson, A. (2018). An improved strategy for Bayesian estimation of the reduced reparameterized unified model. *Appl. Psychol. Meas.* 42, 99–115. doi: 10.1177/0146621617707511

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *J. Educ. Behav. Stat.* 34, 115–130. doi: 10.3102/1076998607309474

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. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

de la Torre, J., van der Ark, L. A., and Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Meas. Eval. Counsel. Dev.* 51, 281–296. doi: 10.1080/07481756.2017.1327286

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cogn. Diagnost. Assess.* 361389.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte Carlo. *Phys. Lett. B* 195, 216–222. doi: 10.1016/0370-2693(87)91197-X

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., and Unlu, A. (2016). The r package CDM for cognitive diagnosis models. *J. Stat. Softw.* 74, 1–24. doi: 10.18637/jss.v074.i02

Gu, Y., and Xu, G. (2020). Partial identifiability of restricted latent class models. *Ann. Statist.* 4, 2082–2107. doi: 10.1214/19-AOS1878

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74:191. doi: 10.1007/s11336-008-9089-5

Huebner, A., and Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educ. Psychol. Meas.* 71, 407–419. doi: 10.1177/0013164410388832

Huo, Y., and de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Appl. Psychol. Meas.* 38, 464–485. doi: 10.1177/0146621614533986

Jiang, Z., and Carter, R. (2019). Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via stan. *Behav. Res. Methods* 51, 651–662. doi: 10.3758/s13428-018-1069-9

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

Kuo, B.-C., Chen, C.-H., and de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Appl. Psychol. Meas.* 42, 179–191. doi: 10.1177/0146621617722791

Liu, J., Xu, G., and Ying, Z. (2013). Theory of the self-learning q-matrix. *Bernoulli* 19:1790. doi: 10.3150/12-BEJ430

Ma, W., and de la Torre, J. (2020). Gdina: an R package for cognitive diagnosis modeling. *J. Stat. Softw.* 93, 1–26. doi: 10.18637/jss.v093.i14

Minchen, N. D., de la Torre, J., and Liu, Y. (2017). A cognitive diagnosis model for continuous response. *J. Educ. Behav. Stat.* 42, 651–677. doi: 10.3102/1076998617703060

Mira, A., and Tierney, L. (1997). On the use of auxiliary variables in Markov chain monte Carlo sampling. *Scand. J. Stat.*

Mira, A., and Tierney, L. (2002). Efficiency and convergence properties of slice samplers. *Scand. J. Stat.* 29, 1–12. doi: 10.1111/1467-9469.00267

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, eds S. Brooks, A. Gelman, G. Jones, and X. L. Meng (Boca Raton: CRC Press), 113–162. doi: 10.1201/b10905-6

Patz, R. J., and Junker, B. W. (1999). A straightforward approach to Markov chain monte Carlo methods for item response models. *J. Educ. Behav. Stat.* 24, 146–178. doi: 10.3102/10769986024002146

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* 6, 7–11.

Roberts, G. O., and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 61, 643–660. doi: 10.1111/1467-9868.00198

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 Assessment for Education: Theory and Applications*, eds J. Leighton and M. Gierl, (Cambridge University Press), 275–318. doi: 10.1017/CBO9780511611186.010

Rupp, A. A., and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Measurement* 6, 219–262. doi: 10.1080/15366360802490866

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 Stat. Methodol.* 64, 583–639. doi: 10.1111/1467-9868.00353

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *J. R. Stat. Soc. Ser. C Appl. Stat.* 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. Educ. Meas.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K. K. (1990). "Toward an integration of item-response theory and cognitive error diagnosis," in *Diagnostic Monitoring of Skill and Knowledge Acquisition*, eds N. Frederiksen, R. Glaser, A. Lesgold, and M. Shafto (Hillsdale, NJ: Erlbaum), 453–488.

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

von Davier, M. (2008). A general diagnostic model applied to language testing data. *Brit. J. Math. Stat. Psychol.* 61, 287–307. doi: 10.1348/000711007X193957

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *Ann. Stat.* 45, 675–707. doi: 10.1214/16-AOS1464

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

Zhan, P., Jiao, H., Man, K., and Wang, L. (2019). Using jags for bayesian cognitive diagnosis modeling: a tutorial. *J. Educ. Behav. Stat.* 44, 473–503. doi: 10.3102/1076998619826040

Zhang, Z., Zhang, J., Lu, J., and Tao, J. (2020). Bayesian estimation of the dina model with Pólya-gamma Gibbs sampling. *Front. Psychol.* 11:384. doi: 10.3389/fpsyg.2020.00384

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 Xu, de la Torre, Zhang, Guo and Shi. 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 October 2020
doi: 10.3389/fpsyg.2020.579018

On the Sequential Hierarchical Cognitive Diagnostic Model

Xue Zhang[1] *and Juntao Wang*[2]*

[1] *China Institute of Rural Education Development, Northeast Normal University, Changchun, China,* [2] *School of Mathematics and Statistics, Northeast Normal University, Changchun, China*

Model data fit plays an important role in any statistical analysis, and the primary goal is to detect the preferred model based on certain criteria. Under the cognitive diagnostic assessment (CDA) framework, a family of sequential cognitive diagnostic models (CDMs) is introduced to handle polytomously scored data, which are attained by answering constructed-response items sequentially. The presence of attribute hierarchies, which can provide useful information about the nature of attributes, will help understand the relation between attributes and response categories. This article introduces the sequential hierarchical CDM (SH-CDM), which adapts the sequential CDM to deal with attribute hierarchy. Furthermore, model fit analysis for SH-CDMs is assessed using eight model fit indices (i.e., three absolute fit indices and five relative fit indices). Two misfit sources were focused; that is, misspecifying attribute structures and misfitting processing functions. The performances of those indices were evaluated *via* Monte Carlo simulation studies and a real data illustration.

Keywords: sequential hierarchical cognitive diagnostic model, polytomous response data, attribute hierarchy, processing function, model fit

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Peida Zhan,
Zhejiang Normal University, China
Wenchao Ma,
University of Alabama, United States

***Correspondence:**
Juntao Wang
wangjt566@nenu.edu.cn

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

Received: *01 July 2020*
Accepted: *04 September 2020*
Published: *07 October 2020*

Citation:
Zhang X and Wang J (2020) On the Sequential Hierarchical Cognitive Diagnostic Model. Front. Psychol. 11:579018. doi: 10.3389/fpsyg.2020.579018

INTRODUCTION

Cognitive diagnostic assessment (CDA) has gained widespread use since its introduction, as it can provide fine-grained feedback through pinpointing the presence or absence of multiple fine-grained skills or attributes (Leighton and Gierl, 2007; Templin and Bradshaw, 2013) based on some *cognitive diagnostic models* (CDMs). Many different names according to their different connotations (Rupp et al., 2010; Ma, 2017) can be applied to refer to the CDM, in which the *diagnostic classification model* (DCM; Rupp et al., 2010) is most widely applied.

For dichotomously scored items, a number of CDMs can be found in literature, among others, the deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker and Sijtsma, 2001) model, the deterministic inputs, noisy "or" gate (DINO; Templin and Henson, 2006) model, and the additive CDM (A-CDM; de la Torre, 2011) were most widely used. Furthermore, three most general CDMs, the general diagnostic model (GDM; von Davier, 2008), the log-linear CDM (LCDM; Henson et al., 2009), and the generalized deterministic input noisy and gate model (GDINA; de la Torre, 2011), were proposed to better understand and handle the above models. Specifically, the GDINA model is equivalent to the LCDM when the logit link is used, and the GDM is a general version of both of them.

For polytomously scored items, which yield graded responses with ordered categories or nominal responses, a few models have been developed, such as the GDM for graded response (von Davier, 2008), the nominal response diagnostic model (NRDM; Templin et al., 2008), the partial credit DINA model (de la Torre, 2010), the polytomous LCDM (Hansen, 2013), the sequential CDM (sCDM; Ma and de la Torre, 2016), and the diagnostic tree model (DTM; Ma, 2019). Among them, only sCDM and DTM can model the possible relation between attributes and response categories. Furthermore, Liu and Jiang (2020) proposed the rating scale diagnostic model (RSDM), which was a special version of the NRDM with fewer parameters. Culpepper (2019) presented an exploratory diagnostic framework for ordinal data.

On the other hand, attribute dependencies often occur in practical applications, instead of that all attributes are independent for each examinee. To this end, four different types of attribute hierarchies (i.e., linear, convergent, divergent, and unstructured) were considered to reflect attribute dependencies (Gierl et al., 2007). An example of different types of attribute hierarchies is shown in **Figure 1**. For an external shape, a directed acyclic graph (DAG) is used to express the attribute hierarchy; and for an internal organization, all possible attribute profiles are provided. Let α_1, α_2, α_3, α_4 denote four attributes measured by a CDA. Take the linear structure as an example, α_k is the prerequisite of α_{k+1} (k = 1, 2, 3), as a result, the number of all possible attribute patterns is 5, which is less than $2^4 = 16$. To model attribute hierarchy, Templin and Bradshaw (2014) proposed a hierarchical diagnostic classification model (HDCM). Zhan et al. (2020) proposed a sequential higher-order DINA model with attribute hierarchy to handle the higher-order and hierarchical structures simultaneously using the sequential tree. Interested readers can refer to Rupp et al. (2010) and von Davier and Lee (2019) for detailed information.

It is a central concern to assess global-level fit (i.e., model fit) in the psychometric area. Model fit analysis can be evaluated by two aspects: absolute fit analysis, which assesses how well a given model reproduces the sample data directly; and relative fit analysis, which recommends a better-fit model through comparing at least two candidates. Under the CDA framework, the sources of model misfit include inaccurate item response function (IRF), Q-matrix misspecification, misspecifying attribute pattern structures (Han and Johnson, 2019), abnormal response behaviors (e.g., rapid guessing, and cheating), local item dependence, and so on.

Regarding absolute fit assessment, for the sequential GDINA model, Ma and de la Torre (2019) proposed category-level model selection criteria based on the Wald test and the likelihood ratio (LR) test to identify different IRFs; and Ma (2020) used limited-information indices [i.e., M_{ord} and standardized mean square root of squared residual (SRMSR)] to detect model misspecification or Q-matrix misspecification. On the other hand, Lim and Drasgow (2019) proposed the Mantel–Haenszel (MH) chi-square statistic to detect latent attribute misspecification in non-parametric cognitive diagnostic methods.

In terms of relative fit assessment, Sen and Bradshaw (2017) compared the Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted BIC (aBIC), and found that they performed poorly to differentiate CDMs and multidimensional item response theory (MIRT) models for dichotomous response data.

Furthermore, some researches focused on the performances of both absolute fit indices (AFIs) and relative fit indices (RFIs) for dichotomous CDMs. Under Bayesian framework, Sinharay and Almond (2007) used Bayesian residuals, which are based on the individuals' raw scores, as AFI and deviance information criterion (DIC) as RFI to distinguish different measurement models. Chen et al. (2013) used both AFIs [i.e., abs(*fcor*), residuals based on the proportion correct of individual items and the log-odds ratios of item pairs] and RFIs (i.e., -2LL, AIC, and BIC) to identify model and/or Q-matrix misspecifications. Hu et al. (2016) investigated the usefulness of AIC, BIC, and consistent AIC (CAIC) as RFIs, and limited information root mean square error of approximation (RMSEA2), abs(*fcor*), and $\max(\chi^2_{jj'})$ as AFIs to detect model or Q-matrix misspecification. Lei and Li (2016) detected model and Q-matrix misspecifications using AFIs [i.e., RMSEA, mean of absolute values of Q3 statistic (MADQ3), mean of absolute values of pairwise item covariance residuals (MADres), mean of absolute deviations in observed and expected correlations (MADcor), and mean of all item pair $\chi^2_{jj'}$ statistics ($M\chi^2_{jj'}$)] and RFIs (i.e., AIC and BIC). Empirically, Han and Johnson (2019) assessed global-level fit for dichotomous CDMs using both RFIs (i.e., AIC, BIC, and aBIC) and AFIs [i.e., M_2, RMSEA2, and maximum of all item pair $\chi^2_{jj'}$ statistics ($\max(\chi^2_{jj'})$)] through a real data illustration.

As no prior studies have analyzed the CDM fit for polytomous response data with attribute hierarchy, the current study concentrates on a novel model, named as the sequential hierarchical CDM (SH-CDM), and assesses model-data fit. This study provides important evidence and insight to support the usefulness of SH-CDMs in the future. The remainder of this article is listed as follows. First, we introduce the sequential hierarchical CDMs and review different model fit indices. Next, simulation studies and a real data illustration are provided to evaluate the performances of those indices. Finally, we end with some concluding remarks. **Supplementary Appendix A** and **Supplementary Material** are provided to complement the detailed information and simulation results.

MATERIALS AND METHODS

Model Description

In this section, SH-CDMs are introduced based on Templin and Bradshaw (2014) and Ma and de la Torre (2016). The sCDM (Ma and de la Torre, 2016) is a special version of SH-CDM with non-hierarchical attribute structure, and the HDCM (Templin and Bradshaw, 2014) is a special version of SH-CDM when all response data are scored dichotomously.

To analyze the polytomously scored data from constructed-response items, the sCDM is built upon a sequential process model. As a result, it can provide the detailed problem-solving

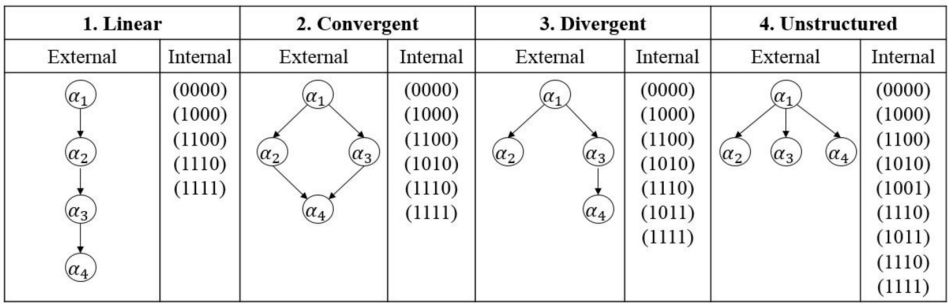


FIGURE 1 | Types of hierarchical attribute structures. For the convergent structure, mastering α_4 needs to master both α_2 and α_3 (Zhan et al., 2020), as well as α_1.

procedures to support subsequent inference. For a constructed-response test with J items, assuming item j (j = 1, 2, …, J) involves H_j tasks that need to be solved sequentially; therefore, if the *1*st task is failed, the score should be 0; if the first h ($0 < h < H_j$) tasks are successful but the ($h + 1$)th task is failed, the score should be h; else if all tasks are successful, the score should be H_j. To this end, Ma and de la Torre (2016) proposed a $\sum_{j=1}^{J} H_j$ –by–K category-level matrix, named as the Q_c-matrix, the element of which is a binary indicator of whether the task of the corresponding item measures this attribute. In this article, only the restricted Q_c-matrix is considered. Mathematically, let $Q_c = \{q_{jh,k}\}$, $q_{jh,k} = 1$ if attribute k (k = 1, 2, … K) is measured by item j for task h (h = 1, 2, …, H_j), otherwise, $q_{jh,k} = 0$.

Let the processing function, $S_{jh}(\vec{\alpha}_g)$, be the probability of examinees with attribute pattern $\vec{\alpha}_g$ answering the first h tasks of item j correctly given answering the first (h-*1*)th tasks correctly, where g = 1, 2, …, G, and G denotes the total number of latent classes. Notating $\vec{\alpha}^*_{l,jh}$ as the reduced attribute profile of examinee l (l = 1, 2, …, N), which contains all the required attributes of item j for task h. Then, $S_{jh}\left(\vec{\alpha}_g\right) = S(\vec{\alpha}^*_{l,jh})$. In this article, the DINA model, DINO model, A-CDM, and GDINA model are considered, which represent the conjunctive, disjunctive, additive, and general condensation rules, respectively. Hereafter, only the identity link is considered for the GDINA model, which is equivalent to the logit link[1]. Assuming that all categories share the same condensation rule, let K^*_{jh} be the total number of required attributes by item j for task h; the expressions for processing functions are summarized in **Table 1**.

Furthermore, the existence of attribute hierarchy will reduce the model complexity of IRF. To deal with attribute hierarchies, the HDCM was proposed by Templin and Bradshaw (2014). As the HDCM is nested within more general CDMs, the SH-CDM is introduced by combining it with the sCDM. In the SH-CDM, the influence of attribute hierarchy is reflected in the processing function. Take the SH-GDINA model as an example, assuming four linear attributes (**Figure 1**) are measured in the test and three attributes are required by item j for task h, $S\left(\vec{\alpha}^*_{l,jh}\right) = \phi_{jh,0} + \phi_{jh,k_1}\alpha_{lk_1} + \phi_{jh,k_1k_2}\alpha_{lk_1}\alpha_{lk_2} + \phi_{jh,k_1k_2k_3}\alpha_{lk_1}\alpha_{lk_2}\alpha_{lk_3}$, where the subscript of k denotes the order of the required attributes.

[1]A simulation check has been done to test the equivalence using the GDINA package; results indicated there was no significant difference between the identity link and the logit link.

TABLE 1 | Summary of $S(\vec{\alpha}^*_{l,jh})$ for different processing functions.

Processing function	Formula	Notations
DINA	$S\left(\vec{\alpha}^*_{l,jh}\right) = \phi_{jh,0} + \phi_{jh,12\ldots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{lk}$	**Examinee parameters**: α_{lk} is the kth attribute of examinee l.
DINO	$S\left(\vec{\alpha}^*_{l,jh}\right) = \phi_{jh,0} + \phi_{jh,k}\alpha_{lk}$	**Item parameters**: $\phi_{jh,0}$ is the intercept; $\phi_{jh,k}$ is the main effect of α_{lk}; $\phi_{jh,kk'}$ is the two-way interaction effect of α_{lk} and $\alpha_{lk'}$; $\phi_{jh,12\ldots K^*_{jh}}$ is the K^*_{jh} –way interaction effect of $\vec{\alpha}^*_{l,jh}$
A-CDM	$S\left(\vec{\alpha}^*_{l,jh}\right) = \phi_{jh,0} + \sum_{k=1}^{K^*_{jh}} \phi_{jh,k}\alpha_{lk}$	
GDINA	$S\left(\vec{\alpha}^*_{l,jh}\right) = \phi_{jh,0} + \sum_{k=1}^{K^*_{jh}} \phi_{jh,k}\alpha_{lk} + \sum_{k'=k+1}^{K^*_{jh}} \sum_{k=1}^{K^*_{jh}-1} \phi_{jh,kk'}\alpha_{lk}\alpha_{lk'} + \ldots + \phi_{jh,12\ldots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{lk}$	

To ensure joint identifiability of the HDCMs, Gu and Xu (2019) restricted that the sparsified version of Q-matrix had at least three entries of "1"s in each column and Q-matrix can be rearranged as $Q = (\frac{Q^0}{Q^*})$, where Q^0 was equivalent to a K-by-K identity matrix $\mathbf{I}_K$ under the attribute hierarchy and the densified version of Q^* contained K distinct column vectors. In addition, if Q-matrix was constrained to contain an $\mathbf{I}_K$, the HDCMs were identified. The SH-CDM identifiability shares the same restrictions as those mentioned above. Interested reader can refer to Gu and Xu (2019) for further discussion on model identifiability.

Model Fit Indices

To assess global-level fit of CDMs, both absolute fit assessment and relative fit assessment are done to identify adequate-fit models and select the best-fit model, respectively. To ensure the comparability of different AFIs, SRMSR, 100*MADRESIDCOV, and MADcor are considered as they share the same rule. To choose the best-fit model among all candidates, five widely used RFIs [i.e., AIC, BIC, the second-order information criterion (AICc), aBIC, and CAIC] are evaluated and compared.

Absolute Fit Index

The SRMSR (Maydeu-Olivares, 2013) is a measure of pairwise correlations. As a standardized statistic, test length has few influences on the performance of SRMSR. The SRMSR can be calculated as

$$\text{SRMSR} = \sqrt{\frac{2}{J(J-1)}\sum_{j<j'}\left(r_{jj'}-\hat{r}_{jj'}\right)^2}, \quad (1)$$

where $r_{jj'}$ and $\hat{r}_{jj'}$ denote the observed and predicted pairwise item correlations, respectively. The model with smaller SRMSR will be identified as a good fit one.

The mean absolute deviation (MAD) is a fundamental statistic to calculate the last two AFIs mentioned before, which measures the discrepancy between the observed item conditional probabilities of success and the predicted ones. The RESIDCOV denotes the residual covariance of pairwise items. Then, we can obtain the mean of absolute deviations of residual covariances (MADRESIDCOV; McDonald and Mok, 1995) by replacing the conditional probabilities in MAD by the RESIDCOVs. The MADRESIDCOV measures the discrepancy between observed and predicted pairwise item residual covariance (RESIDCOV). Let,

$$\gamma_{jj'} \triangleq \text{RESIDCOV}_{jj'} = \frac{n_{jj',11}n_{jj',00}-n_{jj',10}n_{jj',01}}{n^2} - \frac{e_{jj',11}e_{jj',00}-e_{jj',10}e_{jj',01}}{n^2}, \quad (2)$$

Then, we can obtain

$$\text{MADRESIDCOV} = \frac{\sum_{j=1}^{J}\sum_{j\neq j'}\left|\gamma_{jj'}-\hat{\gamma}_{jj'}\right|}{J(J-1)}. \quad (3)$$

100*MADRESIDCOV is used equivalently, as the magnitude of MADRESIDCOV is usually small. The MADcor (DiBello et al., 2007) is the mean of absolute deviations in observed and expected correlations of pairwise items. The MADcor equals $\frac{2}{J(J-1)}\sum_{j<j'}\left|r_{jj'}-\hat{r}_{jj'}\right|$, where $r_{jj'}$ and $\hat{r}_{jj'}$ have the same meanings as those in SRMSR. For MAD-type indices, a smaller value (i.e., value near to zero) denotes better fit.

Relative Fix Index

Different types of information criteria are calculated with respect to the penalty term, the expressions of which are presented in **Table 2**. AIC (Akaike, 1974) and BIC (Schwarz, 1978) are most widely applied since their introductions. The second-order information criterion (AICc; Sugiura, 1978) was derived to deal with the small ratio of sample size to estimated number of parameters case (i.e., less than 40; Burnham and Anderson, 2002). As the sample size gets large, AICc converges to AIC. The aBIC (Sclove, 1987) modified BIC by adjusting the sample size term to handle the small sample size case well. The CAIC was proposed by Bozdogan (1987), in which then penalty terms include both the order of the model and the sample size. The candidate model with smaller RFI is recommended.

TABLE 2 | Formulas of different relative fit indices.

Index	Formula	Notation
AIC	$AIC = -2\ln(LL) + 2P$	
BIC	$BIC = -2\ln(LL) + P\ln(N)$	*LL*: likelihood
AICc	$AICc = AIC + [2P(P+1)]/(N-P-1)$	*P*: the effective number of parameters
aBIC	$aBIC = -2\ln(LL) + P\ln[(N+2)/24]$	*N*: sample size
CAIC	$CAIC = -2\ln(LL) + (\ln(N)+1)P = BIC - P$	

AIC, Akaike information criterion; BIC, Bayesian information criterion; aBIC, adjusted BIC; CAIC, consistent AIC.

Simulation Studies

The simulation studies aim (a) to examine parameter recovery for SH-CDMs and (b) to compare the performances of different AFIs and RFIs for SH-CDMs. Two different sources of misfit are considered: the first type of misfit is due to attribute structures misspecification, and the second type of misfit relies on different processing functions. To this end, three simulation studies are conducted: (I) to examine whether parameters of SH-CDMs can be recovered well; (II) to investigate the performances of model fit indices to identity attribute structures; and (III) to investigate their performances to detect the processing function misspecification, respectively.

The simulation conditions are summarized in **Table 3**. More details will be given in the simulation design sections. In this article, the GDINA R package (Ma and de la Torre, 2020) was used to estimate different SH-CDMs and assess the model fit. The source code including the computation of indices, which were not provided in the GDINA R package, was provided in **Supplementary Appendix A**. The mapping matrix method (Tutz, 1997) and the expected *a posteriori* (EAP) method, which are the default methods in the GDINA package for sequential CDMs, were used to estimate item parameters and attribute profiles, respectively. Five hundred replications were conducted for each condition.

Study Design I

In this study, the GDINA model was chosen as the processing function, as its generality. Attribute profiles were generated from the uniform structure; that is, all the possible latent classes shared the same probability. Sample size was 1,000. A 24-item test, in

TABLE 3 | Summary of simulation conditions.

Factors	Studies I and II	Study III
Sample size (*N*)	1,000	1,000; 3,000
Attribute structure	Non-hierarchical; linear; convergent; divergent; unstructured	
Generation processing function	GDINA	DINA, DINO, A-CDM, GDINA
Calibration processing function	GDINA	DINA, DINO, A-CDM, GDINA
Item quality	High: *U*(0.1, 0.2), low: *U*(0.2, 0.3)	
Attribute generation	Uniform structure	
Test length (*J*)	24	
Number of attributes (*K*)	4	
Number of categories (*G*)	4	

TABLE 4 | Summary of parameter recovery when CM = GM.

Model	Low item quality					High item quality				
	AAR	PAR	RMSE	bias	RAbias	AAR	PAR	RMSE	bias	RAbias
S1	0.9008	0.6742	0.1683	0.0000	6.8400	0.9787	0.9226	0.1336	0.0001	2.0298
S2	0.9623	0.8593	0.1842	0.0688	4.6691	0.9915	0.9664	0.2448	0.0961	4.5692
S3	0.9514	0.8240	0.1998	0.0655	4.4051	0.9904	0.9626	0.2552	0.0906	12.9081
S4	0.9518	0.8248	0.1715	0.0390	6.1866	0.9899	0.9604	0.2176	0.0544	3.7261
S5	0.9422	0.7938	0.1801	0.0396	10.9168	0.9903	0.9624	0.2242	0.0556	6.3453

AIC, Akaike information criterion; BIC, Bayesian information criterion; aBIC, adjusted BIC; CAIC, consistent AIC.

TABLE 5 | Summary of parameter recovery when data were fitted by S1.

GM	Low item quality					High item quality				
	AAR	PAR	RMSE	bias	RAbias	AAR	PAR	RMSE	bias	RAbias
S2	0.9475	0.8154	0.4556	−0.0004	13.6676	0.9862	0.9486	0.4808	−0.0001	8.2845
S3	0.9490	0.8169	0.4077	0.0000	9.6069	0.9894	0.9595	0.4383	−0.0002	9.2714
S4	0.9369	0.7796	0.4300	−0.0002	11.2261	0.9837	0.9385	0.4452	0.0000	30.6335
S5	0.9255	0.7426	0.4159	−0.0001	13.1993	0.9825	0.9343	0.4057	0.0001	6.0283

GM, data generation model; AAR, attribute-wise agreement rate; PAR, pattern-wise agreement rate; RMSE, root mean square error; RAbias, average relative absolute bias.

TABLE 6 | Correct detection rates of RFIs for different hierarchical attribute structures.

GM	CM	Low item quality					High item quality				
		AIC	BIC	aBIC	CAIC	AICc	AIC	BIC	aBIC	CAIC	AICc
S2	S1	0	0	0	0	0	0	0	0	0	0
	S2	**0.99**	**1**	**1**	**1**	**1**	**0.998**	**1**	**1**	**1**	**1**
	S3	0.01	0	0	0	0	0.008	0	0	0	0
	S4	0	0	0	0	0	0	0	0	0	0
	S5	0	0	0	0	0	0	0	0	0	0
S3	S1	0	0	0	0	0	0	0	0	0	0
	S2	0	0	0	0	0	0	0	0	0	0
	S3	**0.99**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	S4	0.01	0	0	0	0	0	0	0	0	0
	S5	0	0	0	0	0	0	0	0	0	0
S4	S1	0	0	0	0	0	0	0	0	0	0
	S2	0	0	0	0	0	0	0	0	0	0
	S3	0	0	0	0	0	0	0	0	0	0
	S4	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	S5	0	0	0	0	0	0	0	0	0	0
S5	S1	0	0	0	0	0	0	0	0	0	0
	S2	0	0	0	0	0	0	0	0	0	0
	S3	0	0	0	0	0	0	0	0	0	0
	S4	0	0	0	0	0	0	0	0	0	0
	S5	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

GM, data generation model; CM, calibration model; RFI, relative fit index; AIC, Akaike information criterion; BIC, Bayesian information criterion; aBIC, adjusted BIC; CAIC, consistent AIC. The bold value means the largest value in a specific condition.

which there were 20 four-category items and four dichotomously scored items, was used. For four-category items, four attributes were measured totally and no more than three attributes were required by each item. Without loss of generality, the last four items were dichotomously scored with an identity sub$Q$-matrix to ensure model identifiability. Two manipulated factors included attribute structure (non-hierarchical, linear, convergent, divergent, and unstructured) and item quality (high and low).

TABLE 7 | CDRs of AFIs for different hierarchical attribute structures.

GM	CM	Low item quality			High item quality		
		MADcor	100MAD RESIDCOV	SRMSR	MADcor	100MAD RESIDCOV	SRMSR
S2	S1	**0.282**	0.226	**0.270**	**0.292**	0.210	0.304
	S2	0.206	**0.238**	0.228	0.248	**0.306**	**0.224**
	S3	0.150	0.162	0.174	0.128	0.150	0.136
	S4	0.178	0.164	0.150	0.134	0.128	0.124
	S5	0.184	0.210	0.178	0.198	0.206	0.192
S3	S1	**0.296**	0.260	**0.308**	**0.302**	0.266	**0.298**
	S2	0.006	0.008	0	0	0	0
	S3	0.280	**0.312**	0.274	0.294	**0.314**	0.292
	S4	0.196	0.226	0.230	0.212	0.200	0.224
	S5	0.222	0.194	0.188	0.192	0.220	0.186
S4	S1	**0.426**	**0.364**	**0.396**	**0.410**	0.37	**0.420**
	S2	0	0	0	0	0	0
	S3	0	0	0	0	0	0
	S4	0.314	0.360	0.326	0.306	**0.33**	0.296
	S5	0.260	0.276	0.278	0.284	0.30	0.284
S5	S1	**0.510**	0.446	0.496	0.454	0.42	0.458
	S2	0	0	0	0	0	0
	S3	0	0	0	0	0	0
	S4	0.002	0.014	0.004	0	0	0
	S5	0.488	**0.540**	**0.500**	**0.546**	**0.58**	**0.542**

GM, data generation model; CM, calibration model; CDR, correct detection rate; AFI, absolute fit index; SRMSR, standardized mean square root of squared residual. The bold value means the largest value in a specific condition.

For each data set, item parameters and attribute profiles were simulated separately and the *Q*-matrix was kept consistent.

The item parameter recovery was calculated in terms of average root mean square error (RMSE), average bias, and average relative absolute bias (RAbias), and the classification accuracies were examined using pattern-wise agreement rate (PAR) and attribute-wise agreement rate (AAR).

Simulation Result I

Hereafter, for convenience, let S1 = the SH-GDINA model with non-hierarchical attribute structure; S2 = the SH-GDINA model with linear attribute structure; S3 = the SH-GDINA model with convergent attribute structure; S4 = the SH-GDINA model with divergent attribute structure; S5 = the SH-GDINA model with unstructured attribute structure.

Table 4 summarizes the estimation accuracy and precision of SH-GDINA models. For different attribute structures, attribute profiles in high item quality cases could be recovered better than in the corresponding low item quality cases. The estimation accuracy of item parameters had a similar trend except for convergent attribute structure cases, although the values of RMSE and bias in low item quality cases were smaller than those in high item quality cases, which is because true item parameters' values in the low cases were smaller, and the smaller true values led to larger RAbiases.

Furthermore, the SH-GDINA model with non-hierarchical attribute structure, which was the most general one among these models, was used to fit response data generated by different SH-GDINA models. The parameter recovery is summarized in **Table 5**. Compared with results shown in **Table 4, AAR** and **PAR** were smaller, and RMSE and RAbias were larger. It appears that specifying the attribute hierarchy can significantly improve the estimation accuracy and precision, which supports the introduction of SH-CDMs.

Study Design II

In this study, the same simulation settings as the simulation study I were considered. Correct detection rates (CDRs) were used to evaluate the performances of different indices. For AFIs, as there is no sufficient evidence for the cutoff values of these AFIs to support model fit assessment, the CDR is calculated as the rate of the smallest values for all replications. Meanwhile, the box plot of AFIs is also provided to compare the performances of different AFIs intuitively.

Simulation Result II

A popular rule for AIC (Burnham and Anderson, 2002) is that a difference of 2 or less is considered negligible and a difference exceeding 10 constitutes strong support. In this article, the same rule is used for all the RFIs. For the non-hierarchical attribute structure, all AFIs of the data generation model had the smallest values, and all RFIs recommended the true model with strong support. For hierarchical attribute structures, **Tables 6**, **7** provide CDRs of RFIs and AFIs, respectively.

As shown in **Table 6**, all RFIs could select the true model with a probability larger than 0.99. Regarding the effectiveness of detecting distinguished models with similar RFIs, we calculated the rates of the differences between RFI values of two candidates, which were smaller than 10, and named as *the indistinguishable proportion*. When response data were generated by S2, the

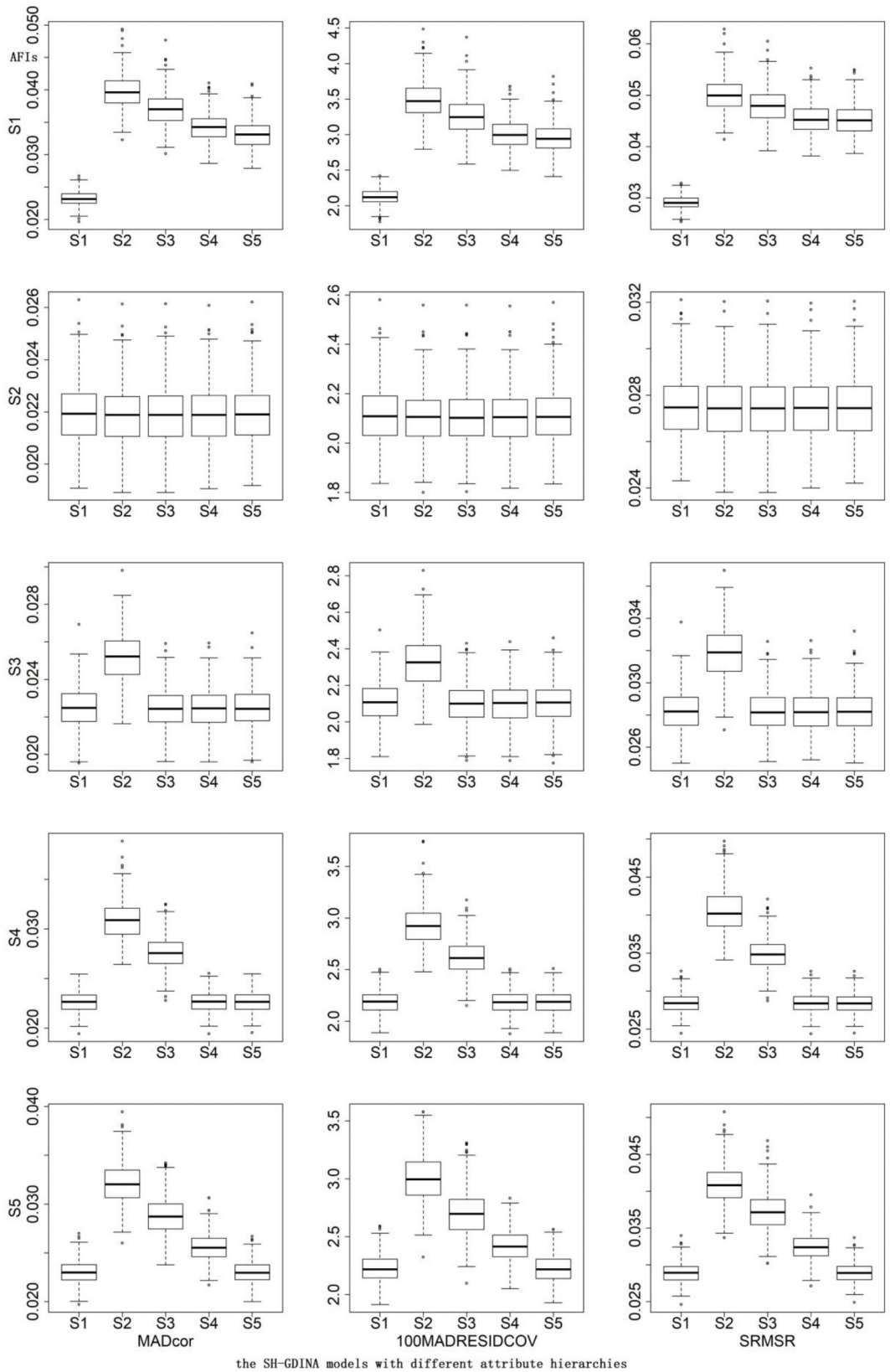


FIGURE 2 | Box plots of absolute fit indices (AFIs) in low item quality cases of simulation study II.

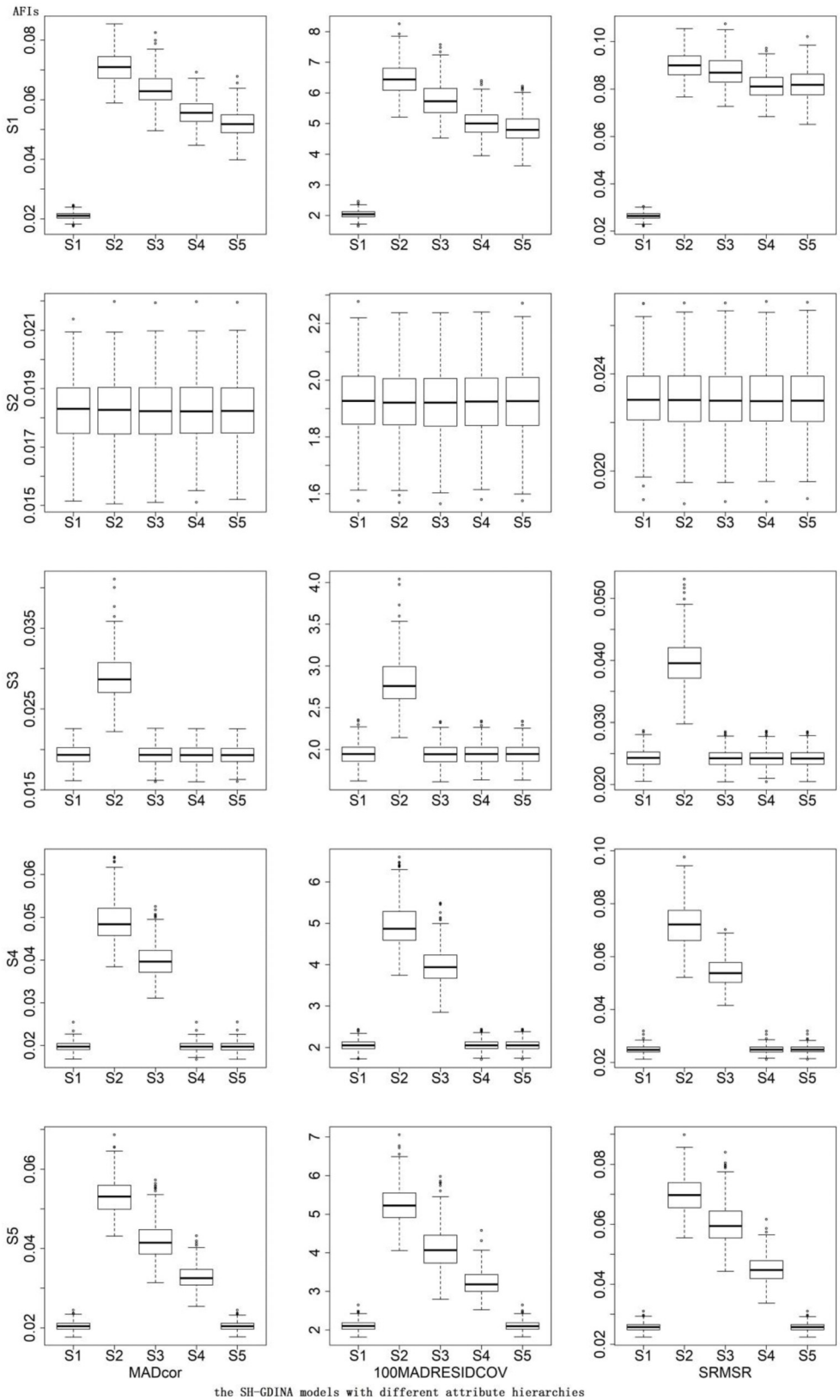


FIGURE 3 | Box plots of absolute fit indices (AFIs) in high item quality cases of simulation study II.

TABLE 8 | Selected results of CDRs of RFIs for different models.

Condition	CM	AIC	BIC	aBIC	CAIC	AICc
Non-hierarchical structure Low item quality, *N* = 1,000 GM = M4	M1	0	0	0	0	0
	M2	0	0	0	0	0
	M3	0	**1**	**0.502**	**1**	0.134
	M4	**1**	0	0.498	0	**0.866**
Non-hierarchical structure Low item quality, *N* = 3,000 GM = M4	M1	0	0	0	0	0
	M2	0	0	0	0	0
	M3	0	**0.76**	0	**0.978**	0
	M4	**1**	0.24	**1**	0.022	**1**
Convergent structure Low item quality, *N* = 1,000 GM = M4	M1	0	0	0	0	0
	M2	0	0	0	0	0
	M3	0	**0.642**	0.05	**0.822**	0.014
	M4	**1**	0.358	**0.95**	0.178	**0.986**
Divergent structure Low item quality, *N* = 1,000 GM = M4	M1	0	0	0	0	0
	M2	0	0	0	0	0
	M3	0	**0.888**	0.02	**0.988**	0
	M4	**1**	0.112	**0.98**	0.012	**1**
Unstructured structure Low item quality, *N* = 1,000 GM = M4	M1	0	0	0	0	0
	M2	0	0	0	0	0
	M3	0	**0.894**	0.002	**0.998**	0
	M4	**1**	0.106	**0.998**	0.002	**1**
Unstructured structure High item quality, *N* = 3,000 GM = M4	M1	0	0	0	0	0
	M2	0	0	0	0	0
	M3	0	**0.652**	0.002	**0.938**	0
	M4	**1**	0.348	**0.998**	0.062	**1**

GM, data generation model; CM, calibration model; CDR, correct detection rate; RFI, relative fit index; AIC, Akaike information criterion; BIC, Bayesian information criterion; aBIC, adjusted BIC; CAIC, consistent AIC. The bold value means the largest value in a specific condition.

indistinguishable proportions of AIC to differentiate S3 were 4.8% and 27.2% for the high item quality case and the low item quality case, respectively; and the indistinguishable proportion of AICc was 0.2% for the high item quality case. To generate response data using S3, for the high item quality case, the indistinguishable proportion of AIC to differentiate S4 was 3.8%; for the low item quality case, the indistinguishable proportion of AIC to differentiate S4 was 29%, among them 1% of the time AIC could not differentiate S3, S4, and S5. For the case in which data were generated by S4, the indistinguishable proportion of AIC to differentiate S4 and S5 was less than 6%. Other cases could be differentiated well.

In terms of AFIs (**Table 7**), S1 mostly had the smallest AFIs. The CDRs of different AFIs were similar. According to the box plots of AFIs (**Figures 2**, **3**), when generating data using S2, it was very hard to differentiate these models. Similarly, it was difficult to distinguish S3 from S1 and S5, S4 from S1 and S5, or S5 from S1. High item quality led to large values of AFIs. It appears that RFIs outperform AFIs for SH-GDINA models, and RFIs distinguish SH-GDINA models with convergent attribute structures from models with unstructured attribute structures with difficulty.

Simulation Design III

In this study, attribute generation method, test length, and test structure were kept the same as those in simulation studies I and II. The manipulated factors included sample size (1,000 and 3,000), attribute structure (non-hierarchical, linear, convergent, divergent, and unstructured), item quality (high and low), and data generation/calibration model (SH-DINA, SH-DINO, SH-ACDM, and SH-GDINA). To evaluate the performances of different indices, the CDRs were reported. Hence, item parameters and the *Q*-matrix were fixed for each condition, and attribute profiles were simulated separately.

Simulation Result III

For convenience, let M1 = the SH-DINA model, M2 = the SH-DINO model, M3 = the SH-ACDM, and M4 = the SH-GDINA model.

Relative Fit Index

Limited by the space, we only provided the results when the probabilities of true model selection by different RFIs were different or incorrect in **Table 8**. The whole results (i.e., CDRs of both RFIs and AFIs, box plots of AFIs) were provided in the **Supplementary Material**. The different performances were reflected in the SH-ACDM and the SH-GDINA model. Under the conditions mentioned in **Table 8, AIC** and **AICc** always chose the true model, BIC and CAIC always chose the alternative model, and aBIC mostly selected the true model. Besides, for linear attribute structures, when the data generation model was the SH-ACDM, all RFIs recommended the SH-GDINA model with a probability larger than 0.98. In other cases, all RFIs could

TABLE 9 | CDRs of AFIs for different models with *linear* attribute structures.

GM	CM	Low item quality			High item quality		
		MADcor	100MAD RESIDCOV	SRMSR	MADcor	100MAD RESIDCOV	SRMSR
				N = 1,000			
M1	M1	0.158	0.142	0.114	0.176	0.144	0.126
	M2	0	0	0	0	0	0
	M3	**0.53**	**0.526**	**0.528**	**0.42**	**0.472**	**0.454**
	M4	0.312	0.332	0.358	0.404	0.384	0.42
M2	M1	0	0	0	0	0	0
	M2	0.082	0.082	0.056	0.042	0.044	0.036
	M3	0.444	0.438	0.454	**0.536**	**0.522**	**0.514**
	M4	**0.474**	**0.48**	**0.49**	0.422	0.434	0.45
M3	M1	0	0	0	0	0	0
	M2	0	0	0	0	0	0
	M3	**0.554**	**0.548**	**0.558**	**0.554**	**0.592**	**0.54**
	M4	0.446	0.452	0.442	0.446	0.408	0.46
M4	M1	0	0	0	0	0	0
	M2	0	0	0	0	0	0
	M3	0.46	0.492	0.492	**0.562**	**0.568**	**0.556**
	M4	**0.54**	**0.508**	**0.508**	0.438	0.432	0.444
				N = 3,000			
M1	M1	0.146	0.158	0.114	0.154	0.162	0.128
	M2	0	0	0	0	0	0
	M3	**0.438**	**0.44**	**0.488**	**0.464**	**0.44**	**0.508**
	M4	0.416	0.402	0.398	0.382	0.398	0.364
M2	M1	0	0	0	0	0	0
	M2	0.078	0.072	0.048	0.078	0.072	0.048
	M3	0.422	0.438	0.46	0.422	0.438	0.46
	M4	**0.5**	**0.49**	**0.492**	**0.5**	**0.49**	**0.492**
M3	M1	0	0	0	0	0	0
	M2	0	0	0	0	0	0
	M3	0.492	0.462	0.48	**0.502**	0.498	**0.53**
	M4	**0.508**	**0.538**	**0.52**	0.498	**0.502**	0.47
M4	M1	0	0	0	0	0	0
	M2	0	0	0	0	0	0
	M3	**0.558**	0.49	**0.554**	**0.562**	**0.542**	**0.544**
	M4	0.442	**0.51**	0.446	0.438	0.458	0.456

GM, data generation model; CM, calibration model; CDR, correct detection rate; AFI, absolute fit index; SRMSR, standardized mean square root of squared residual. The bold value means the largest value in a specific condition.

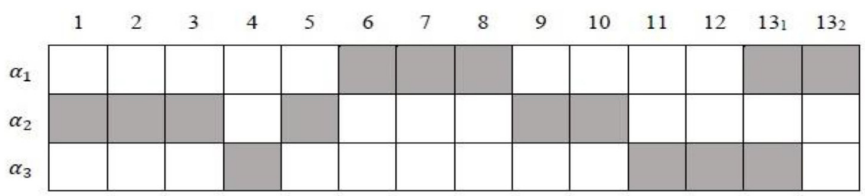


FIGURE 4 | The Q_c-matrix for Booklet 1 data from TIMSS 2007. 13_1 and 13_2 denote the *1*st and *2*nd tasks of item 13, respectively.

select the data generation model as the better fitting one with a probability larger than 0.93.

When data were generated by M4 with non-hierarchical attribute structures, in low item quality cases, the indistinguishable proportions of aBIC and AICc to differentiate M3 and M4 were about 10% for a small sample size; for a large sample size, BIC could not differentiate M3 and M4 5.4% of the time and the indistinguishable proportion of CAIC was 1%.

For linear attribute structures, AIC could not distinguish M1 from M3 or M4 for one or two replications under different conditions. When the generation model was M2, the indistinguishable proportions of AIC with high item quality were similar to those in the "GM = M1" case. When response

data were generated by M3 or M4, all RFIs could not differentiate M3 and M4.

For convergent attribute structures, AIC could not distinguish M2 from M4 in no more than one replication under different conditions. M3 and M4 could not be distinguished well. When generating response data using M3, the indistinguishable proportions of AIC to differentiate M4 ranged from 45% to 69%, the corresponding indistinguishable proportions of AICc ranged from 3% to 27%, and BIC was not able to differentiate M3 from M4 11% of the time for the high item quality case with small sample size. When "GM = M4," for the low item quality case with the large sample size, aBIC was not able to differentiate M4 and M3 32% of the time, and the indistinguishable proportion of CAIC was 8%; the indistinguishable proportions of AIC, BIC, aBIC, CAIC, and AICc for the low item quality case with the small sample size were about 1.6%, 39%, 7.8%, 30.2%, and 7.6%, respectively; and for the high item quality case with the small sample size, the indistinguishable proportions of BIC and aBIC were 11.4% and 24.8%, respectively.

For divergent attribute structures, in one replication, AIC could not distinguish M1 from M4. Furthermore, AIC was not able to distinguish M3 from M4 about 8.4% of the time, and the corresponding indistinguishable proportion of AICc was 2%. When data were generated by M4, the indistinguishable proportions of BIC and CAIC for high item quality with the small sample size were 7.2% and 14%; for low item quality, the corresponding proportions of BIC, aBIC, CAIC, and AICc with the small sample size were 7.6%, 5.2%, 2.8%, and 1.4%, respectively, and those of BIC and CAIC with the large sample size were 2.2% and 5.6%, respectively.

For unstructured attribute structures, distinguishing M3 from M4 using AIC would fail about 1.4% of the time, and the indistinguishable proportions of BIC, aBIC, and CAIC

TABLE 10 | Summary of hierarchical attribute structures.

#structure	DAG	#structure	DAG
2	$\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$	7	$\alpha_2 \rightarrow \alpha_3,\ \alpha_1$
3	$\alpha_1 \rightarrow \alpha_3 \rightarrow \alpha_2$	8	$\alpha_3 \rightarrow \alpha_2,\ \alpha_1$
4	$\alpha_3 \rightarrow \alpha_1 \rightarrow \alpha_2$	9	$\alpha_3 \rightarrow \alpha_1,\ \alpha_2$
5	$\alpha_1 \rightarrow \alpha_2,\ \alpha_3$	10	$\alpha_1 \rightarrow \alpha_2,\ \alpha_1 \rightarrow \alpha_3$
6	$\alpha_1 \rightarrow \alpha_3,\ \alpha_2$	11	$\alpha_3 \rightarrow \alpha_1,\ \alpha_3 \rightarrow \alpha_2$

DAG, directed acyclic graph.

TABLE 11 | Model fit indices of SH-GDINA models for TIMSS data.

#structure	MADcor	100MAD RESIDCOV	SRMSR	AIC	BIC	aBIC	CAIC	AICc
1	0.039	0.698	**0.062**	7904.80	8063.86	7946.41	8100.86	7910.36
2	**0.038**	**0.687**	0.063	**7899.01**	**8036.58**	**7935.00**	**8068.58**	**7903.14**
3	0.039	0.692	0.063	7899.95	8037.51	7935.93	8069.51	7904.08
4	0.039	0.703	0.062	7899.34	8036.90	7935.32	8068.90	7903.47
! 5	0.039	0.701	**0.062**	7901.89	8052.35	7941.25	8087.35	7906.85
6	**0.038**	0.690	0.063	7902.96	8049.12	7941.19	8083.12	7907.63
7	**0.038**	0.691	0.063	7902.46	8052.92	7941.82	8087.92	7907.42
8	0.039	0.703	0.063	7901.72	8052.18	7941.08	8087.18	7906.68
9	0.039	0.703	**0.062**	7902.13	8048.29	7940.36	8082.29	7906.80
10	**0.038**	**0.687**	0.063	7901.01	8042.87	7938.12	8075.87	7905.41
11	0.039	0.698	**0.062**	7904.80	8063.86	7946.41	8100.86	7910.36

The recommended attribute structures by RFIs were highlighted in gray. SRMSR, standardized mean square root of squared residual; AIC, Akaike information criterion; BIC, Bayesian information criterion; aBIC, adjusted BIC; CAIC, consistent AIC; AICc, second-order information criterion. The bold value means the smallest value of an index.

TABLE 12 | Summary of estimated proportion of each latent class.

#structure	000	100	010	001	110	101	011	111
1	0.487	0.030	0.001	0.035	0.014	0.000	0.021	0.412
2	0.535	0.033	–	–	0.016	–	–	0.417
3	0.535	0.041	–	–	–	0.000	–	0.424
4	0.517	–	–	0.052	–	0.000	–	0.431
5	0.489	0.034	–	0.047	0.012	0.000	–	0.418
6	0.531	0.032	0.007	–	0.016	0.000	–	0.415
7	0.514	0.028	0.004	–	0.017	–	0.029	0.408
8	0.483	0.038	–	0.042	–	0.000	0.019	0.419
9	0.510	–	0.003	0.039	–	0.000	0.025	0.423
10	0.535	0.033	–	–	0.016	0.000	–	0.417
11	0.511	–	–	0.040	–	0.000	0.025	0.424

"–" denotes the latent class is impermissible.

to differentiate M4 from M3 were about 11%, 2%, and 3%, respectively. Overall, AICc can select the true model and almost distinguish it from others for all manipulated conditions.

Absolute Fit Index

As M4 almost had the smallest values of AFIs for all conditions except the linear attribute structure cases, only the results of these cases were presented in **Table 9**. Both small sample size and high item quality led to larger values of AFIs. For linear attribute structures, it was hard to differentiate M3 and M4, when to generate data using M4, the difference between AFIs of M3 and M4 decreased as item quality became low. For other structures, the trends were similar. This observation indicates that AFIs cannot easily differentiate SH-CDMs that differ by processing functions.

Real Data Illustration

Data Source

This example application is from the TIMSS 2007 eighth-grade mathematics assessment, which is from Booklet 1 that measured three attributes (Lee et al., 2013): Attribute 1 (α_1): whole numbers and integers; Attribute 2 (α_2): fractions, decimals, and percentages; and Attribute 3 (α_3): data analysis and probability. There are 12 dichotomously scored items and one three-category item answered by 544 students from the United States.

The Q_c-matrix based on the works by Lee et al. (2013) and Ma (2019) is presented in **Figure 4**. According to general knowledge about numeric, α_2 cannot be the prerequisite of α_1. Hence, including the non-hierarchical attribute structure (#structure = 1), 11 different attribute structures are analyzed. The detailed DAGs of different hierarchical attribute structures are shown in **Table 10**. Without loss of generality, the GDINA model is chosen as the processing function in this section because, in simulation studies, we noticed that it is not easy to differentiate the SH-GDINA model from others. This dataset is analyzed using the GDINA R package (Ma and de la Torre, 2020).

RESULTS

Table 11 shows the comparison among different structures using both AFIs and RFIs. The smallest values are in boldface. Different indices performed similarly to assess model-data fit that is because most of the items required only one attribute. From an item-level perspective, if only one attribute is required by one item, there is no difference among candidate models with different attribute structures.

On the other hand, Wang and Lu (2020) proposed two predetermined cutoff values (i.e., 0.025 and 0.05) of estimated proportions of latent classes to identify the labels for estimated latent classes. The estimated proportions of latent classes were shown in **Table 12**. All estimated proportions of (000), (100), (001), and (111) were larger than 0.025, and except (000) and (111), only (001) modeled using *4*th structure had an estimated proportion larger than 0.05. It is not enough to distinguish different structures, which may be because the sample size of this dataset was small.

CONCLUSION

In order to avoid possible misleading conclusion, model-data fit must be thoroughly assessed before drawing the model-based inference. Although there are abundant research examining model fit assessment for CDMs, there is a lack of an effective guidance on how to deal with polytomously scored items with hierarchical attribute structures, and the aim of the present study is to fill in this gap. In this paper, we developed a sequential hierarchical cognitive diagnostic model to handle polytomous response data with hierarchical attribute structures and further evaluated model-data fit using both absolute fit indices and relative fit indices.

Across all simulation conditions, the SH-CDM can be recovered well, and aBIC and AICc are recommended for the SH-CDMs due to their high CDRs and acceptable distinguishable proportions. To distinguish different attribute structures for SH-GDINA models, RFIs outperform AFIs. Furthermore, AFIs used in this study are inappropriate to differentiate processing functions of the SH-CDM.

This study was the first attempt at assessing global-level fit of hierarchical CDMs and polytomous response data. However, the results are limited to SH-CDMs using the same condensation rules; future research pertaining to mixture measurement model and different condensation rules for different tasks in one item should be expanded to enhance the practicability of SH-CDMs. Also, it is necessary to extend the study to deal with sparse *Q*-matrix with large *K*. In addition, local-level (i.e., item-level) fit should be further examined to complement global fit analysis. On the other hand, as smaller values (close to zero) of AFIs indicate a good model-data fit, it would be worthwhile to identify the corresponding cutoff values using the resampling technique.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://timss.bc.edu/TIMSS2007/idb_ug.html.

AUTHOR CONTRIBUTIONS

XZ provided original thoughts and completed the writing of this article. JW provided key technical support. Both authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Fundamental Research Funds for the Central Universities (Grant 2412020QD004) and the Key Laboratory of Applied Statistics of MOE.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.579018/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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/bf02294361

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer.

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Measur.* 50, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

Culpepper, S. A. (2019). An exploratory diagnostic model for ordinal responses with binary attributes: identifiability and estimation. *Psychometrika* 84, 921–940. doi: 10.1007/s11336-019-09683-4

de la Torre, J. (2010). The partial-credit DINA model. *Paper Presented at the International Meeting of the Psychometric Society*, Athens, GA.

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. doi: 10.1016/s0169-7161(06)26031-0

Gierl, M. J., Leighton, J. P., and Hunka, S. M. (2007). "Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills," in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. P. Leighton and M. J. Gierl (Cambridge: Cambridge University Press), 242–274. doi: 10.1017/cbo9780511611186.009

Gu, Y., and Xu, G. (2019). Identification and estimation of hierarchical latent attribute models. *arXiv* [Preprint], Available online at: https://arxiv.org/abs/1906.07869v1 (accessed July 5, 2020).

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

Han, Z., and Johnson, M. S. (2019). "Global-and item-level model fit indices," in *Handbook of Diagnostic Classification Models*, eds M. von Davier, and Y. S. Lee (Cham: Springer), 265–285. doi: 10.1007/978-3-030-05584-4_13

Hansen, M. (2013). *Hierarchical Item Response Models for Cognitive Diagnosis*, Doctoral thesis, University of California, Los Angeles, CA.

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Hu, J., Miller, M. D., Huggins-Manley, A. C., and Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnosis models. *Intern. J. Test.* 16, 119–141. doi: 10.1080/15305058.2015.1133627

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

Lee, Y.-S., Johnson, M. S., Park, J. Y., Sachdeva, R., Zhang, J., and Waldman, M. (2013). A MDS Approach for investigating student's cognitive weakness and strength on the TIMSS 2007 mathematics assessment. *Paper Presented at the Annual Meeting of the American Educational Research Association*, San Francisco, CA.

Lei, P. W., and Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Appl. Psychol. Measur.* 40, 405–417. doi: 10.1177/0146621616647954

Leighton, J., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

Lim, Y. S., and Drasgow, F. (2019). Conditional independence and dimensionality of cognitive diagnostic models: a test for model fit. *J. Classif.* 36, 295–305. doi: 10.1007/s00357-018-9287-5

Liu, R., and Jiang, Z. (2020). A general diagnostic classification model for rating scales. *Behav. Res. Methods* 52, 422–439. doi: 10.3758/s13428-019-01239-9

Ma, W. (2017). *A Sequential Cognitive Diagnosis Model for Graded Response: Model Development, Q-Matrix Validation, and Model Comparison*. Doctoral thesis, Rutgers University-School of Graduate Studies, New Brunswick, NJ.

Ma, W. (2019). A diagnostic tree model for polytomous responses with multiple strategies. *Br. J. Math. Statist. Psychol.* 72, 61–82. doi: 10.1111/bmsp.12137

Ma, W. (2020). Evaluating the fit of sequential G-DINA model using limited-information measures. *Appl. Psychol. Measur.* 44, 167–181. doi: 10.1177/0146621619843829

Ma, W., and de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *Br. J. Math. Statist. Psychol.* 69, 253–275. doi: 10.1111/bmsp.12070

Ma, W., and de la Torre, J. (2019). Category-level model selection for the sequential G-DINA model. *J. Educ. Behav. Statist.* 44, 45–77. doi: 10.3102/1076998618792484

Ma, W., and de la Torre, J. (2020). GDINA: an R package for cognitive diagnosis modeling. *J. Statist. Softw.* 93, 1–26. doi: 10.14689/ejer.2019.80.9

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measur. Interdiscipl. Res. Perspect.* 11, 71–137. doi: 10.1080/15366367.2013.831680

McDonald, R. P., and Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multiv. Behav. Res.* 30, 23–40. doi: 10.1207/s15327906mbr3001_2

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

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

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343. doi: 10.1007/bf02294360

Sen, S., and Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Appl. Psychol. Measur.* 41, 422–438. doi: 10.1177/0146621617695521

Sinharay, S., and Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educ. Psychol. Measur.* 67, 239–257. doi: 10.1177/0013164406292025

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Commun. Statist. Theor. Methods A* 7, 13–26. doi: 10.1080/03610927808827599

Templin, J., and Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *J. Classific.* 30, 251–275. doi: 10.1007/s00357-013-9129-4

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. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989x.11.3.287

Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., and Ahmed, M. (2008). Cognitive diagnosis models for nominal response data. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, New York, NY.

Tutz, G. (1997). "Sequential models for ordered responses," in *Handbook of modern item response theory*, eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 139–152. doi: 10.1007/978-1-4757-2691-6_8

von Davier, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Statist. Psychol.* 61, 287–307. doi: 10.1348/000711007x193957

von Davier, M., and Lee, Y. S. (2019). *Handbook of Diagnostic Classification Models*. Cham: Springer International Publishing.

Wang, C., and Lu, J. (2020). Learning attribute hierarchies from data: two exploratory approaches. *J. Educ. Behav. Statist.* [Epub ahead of print].

Zhan, P., Ma, W., Jiao, H., and Ding, S. (2020). A sequential higher order latent structural model for hierarchical attributes in cognitive diagnostic assessments. *Appl. Psychol. Measur.* 44, 65–83. doi: 10.1177/0146621619832935

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 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

METHODS
published: 30 November 2020
doi: 10.3389/fpsyg.2020.564707

Cognitive Diagnosis Modeling Incorporating Item-Level Missing Data Mechanism

Na Shan [1,2]* *and Xiaofei Wang* [2,3]

[1] School of Psychology, Northeast Normal University, Changchun, China, [2] Key Laboratory of Applied Statistics of the Ministry of Education, Northeast Normal University, Changchun, China, [3] School of Mathematics and Statistics, Northeast Normal University, Changchun, China

The aim of cognitive diagnosis is to classify respondents' mastery status of latent attributes from their responses on multiple items. Since respondents may answer some but not all items, item-level missing data often occur. Even if the primary interest is to provide diagnostic classification of respondents, misspecification of missing data mechanism may lead to biased conclusions. This paper proposes a joint cognitive diagnosis modeling of item responses and item-level missing data mechanism. A Bayesian Markov chain Monte Carlo (MCMC) method is developed for model parameter estimation. Our simulation studies examine the parameter recovery under different missing data mechanisms. The parameters could be recovered well with correct use of missing data mechanism for model fit, and missing that is not at random is less sensitive to incorrect use. The Program for International Student Assessment (PISA) 2015 computer-based mathematics data are applied to demonstrate the practical value of the proposed method.

Keywords: cognitive diagnosis, item-level, missing data, missing data mechanism, cognitive diagnosis model

OPEN ACCESS

Edited by:
Tao Xin,
Beijing Normal University, China

Reviewed by:
Pasquale Anselmi,
University of Padua, Italy
Wenchao Ma,
University of Alabama, United States

***Correspondence:**
Na Shan
dann285@nenu.edu.cn

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

Received: *22 May 2020*
Accepted: *28 October 2020*
Published: *30 November 2020*

Citation:
Shan N and Wang X (2020) Cognitive Diagnosis Modeling Incorporating Item-Level Missing Data Mechanism. Front. Psychol. 11:564707. doi: 10.3389/fpsyg.2020.564707

1. INTRODUCTION

Cognitive diagnosis has recently received increasing concern in psychological and educational assessment, which can provide fine-grained classifications and diagnostic feedback for respondents from their performance on test items (Leighton and Gierl, 2007; Rupp et al., 2010). It is a useful tool to identify students' mastery status of different latent skills based on their responses to test items and to evaluate patients' presence of mental disorders based on their responses to diagnostic questions. More specifically, cognitive diagnosis has been used to study fraction subtraction (de la Torre and Douglas, 2004), language proficiency (von Davier, 2008; Chiu and Köhn, 2016), psychological disorders (Templin and Henson, 2006; Peng et al., 2019), and so forth.

Various cognitive diagnosis models (CDMs), also called diagnosis classification models, have been developed, such as the deterministic inputs, noisy "and" gate (DINA) model (Macready and Dayton, 1977; Junker and Sijtsma, 2001) and the deterministic inputs, noisy "or" gate (DINO) model (Templin and Henson, 2006). Most CDMs are parametric and model the probability of item response as a function of latent attributes. The simplicity and interpretability make the parametric CDMs popular in practice. More general CDMs, such as the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) and the generalized DINA model (de la Torre, 2011), assume a more flexible relationship between the item responses and latent attributes. Moreover, higher-order latent trait models for cognitive diagnosis (de la Torre and Douglas, 2004) have been introduced to link the correlated latent traits by a general high-order ability.

Multiple items are often used for cognitive diagnosis. When respondents choose to answer some but not all items, item-level missing data occur (Chen et al., 2020). Respondents may refuse to answer items that they deem too difficult, quit the test early because it is too long, or just skip items because of carelessness. Missing data lead to loss of information and may result in biased conclusions (Glas and Pimentel, 2008; Köhler et al., 2015; Kuha et al., 2018). Many studies have employed a complete case analysis that only use subjects without missing data (e.g., Xu and Zhang, 2016; Chen et al., 2017; Zhan et al., 2019a). In this case, the subjects with missing data cannot receive any diagnostic feedback and, more importantly, may produce biased results when subjects with complete data are systematically different from those with missing data (Pan and Zhan, 2020).

In the literature, three missing data mechanisms should be distinguished: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976; Little and Rubin, 2020). The MCAR holds if the probability of missingness is independent of both the observed and unobserved responses, whereas the MAR holds if the probability of missingness is independent of the unobserved responses given the observed responses. If either of these two conditions cannot be satisfied, i.e., the probability of missingness depends on the unobserved responses, the MNAR occurs. If the missing data mechanism is MCAR or MAR, unbiased estimation can be obtained from the observed data; if the missing data mechanism is MNAR, a model for the missing data mechanism should be included to obtain valid estimations of the primary parameters. Limited approaches have been proposed in CDMs incorporating missing data mechanisms. Ömür Sünbül (2018) considered MCAR and MAR in the DINA model. Recently, Ma et al. (2020) have used the sequential process model to accommodate omitted items due to MNAR, where the first internal task, i.e., making the decision to either skip the item or respond to it, is assumed to be affected by a latent categorical variable representing the response tendency. As stated by Heller et al. (2015), CDMs have connections to knowledge space theory (KST), which has been developed by Doignon and Falmagne (1985) (see also Doignon and Falmagne, 1999; Falmagne and Doignon, 2011). de Chiusole et al. (2015) and Anselmi et al. (2016) have developed models for the analysis of MCAR, MAR, and MNAR data in the framework of KST. In their work, the MCAR holds if the missing response pattern is independent of the individual's knowledge state (i.e., the collection of all items that an individual is capable of solving in a certain disciplinary domain) and of the observed responses; the MAR holds if the missing response pattern is conditionally independent of the knowledge state given the observed responses; and the MNAR holds if the missing response pattern depends on the knowledge state. In CDMs, the attribute profile (i.e., the collection of all attributes that an individual masters in a certain disciplinary domain) is similar to the knowledge state in KST. In our study, an additional latent categorical variable is introduced to indicate missingness propensity. This latent categorical variable affects the probability of missingness for each item and, in the meantime, may affect the latent attributes through the influence of the general high-order ability. The missing data mechanism can then be incorporated into cognitive diagnosis. Similar ideas can be seen in Holman and Glas (2005), Rose et al. (2015), and Kuha et al. (2018).

In this paper, we propose a joint cognitive diagnosis modeling including a higher-order latent trait model for item responses and a missingness propensity model for item-level missing data mechanism. We take the DINA model as an example for illustration because of its popular use, and the latent traits are linked by a general high-order ability. For a flexible specification of the missingness propensity model, the latent missingness propensity is represented by a categorical variable. The MNAR holds if the distribution of the general high-order ability depends on the latent classes, whereas the MCAR holds if the distribution of the general high-order ability is independent of the latent classes.

The rest of this paper is organized as follows. Section 2 presents the proposed joint model for item responses and item-level missing data mechanism. The Bayesian approach is then developed for model parameter estimation using JAGS. In section 3, simulation studies are conducted to compare the performance of the parameter recovery under different missing data mechanisms. Real data analysis using the PISA 2015 computer-based mathematics data is given in section 4. Some concluding remarks are given in section 5.

2. JOINT MODELING INCORPORATING ITEM-LEVEL MISSING DATA MECHANISM

We consider N subjects taking a test of I items, and there are K latent attributes to be evaluated. Let Y_{ni} be the response for subject $n(n = 1, \cdots, N)$ to item $i(i = 1, \cdots, I)$. Let R_{ni} be a missingness indicator corresponding to Y_{ni}, where $R_{ni} = 1$ if Y_{ni} is observed and $R_{ni} = 0$ if Y_{ni} is missing.

2.1. The Missingness Propensity Model

Let ξ_n denote the latent missingness propensity for subject n. ξ_n is unobserved and has C categories, which we refer to as latent missingness classes. The missingess probability of $(R_{n1}, \cdots, R_{nI})$ is given by

$$p(R_{n1}, \cdots, R_{nI}) = \sum_{c=1}^{C} \Big\{ \prod_{i=1}^{I} p(R_{ni}|\xi_n = c) \Big\} p(\xi_n = c), \quad (1)$$

where $p(R_{n1}, \cdots, R_{nI})$ is the joint probability of $(R_{n1}, \cdots, R_{nI})$, $p(R_{ni}|\xi_n = c)$ is the conditional probability of R_{ni} given $\xi_n = c$ and $p(\xi_n = c)$ is the probability of $\xi_n = c$.

The latent missingness propensity ξ_n is specified as the categorical distribution

$$\xi_n \sim \text{Categorical}(\boldsymbol{\pi}), \quad (2)$$

where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_C)$ and $\Sigma_{c=1}^{C} \pi_c = 1$. The conditional probability of R_{ni} is specified as the logistic function

$$\text{logit}\{p(R_{ni} = 1|\xi_n = c)\} = \tau_{0i} + \sum_{c=2}^{C} \tau_{ci} \xi_{(c)}, \quad (3)$$

where τ_{0i} and $\tau_{ci}(c = 2, \cdots, C)$ are the intercept and slope parameters, and $\xi_{(c)}(c = 2, \cdots, C)$ are dummy variables for the missingness classes. A positive slope parameter τ_{ci} is assumed, which means that the missingness probability reduces in the latter missingness class compared to the first class. Denote $\boldsymbol{\tau}_0 = (\tau_{01}, \cdots, \tau_{0I})$ and $\boldsymbol{\tau}_c = (\tau_{c1}, \cdots, \tau_{cI})$ for $c = 2, \cdots, C$. The idea of introducing latent variables to model the non-response mechanism have been proposed previously by, for example, Lin et al. (2004) and Hafez et al. (2015). The missingness propensity model is identifiable when $C \leq 2^I/(1 + I)$.

2.2. The High-Order DINA Model

The DINA model describes the probability of item response as a function of latent attributes as follows:

$$p(Y_{ni} = 1) = g_i + (1 - s_i - g_i) \prod_{k=1}^{K} \alpha_{nk}^{q_{ik}}, \tag{4}$$

where $p(Y_{ni} = 1)$ is the probability of a correct response for subject n to item i; s_i and g_i are the slipping and guessing probability for item i respectively, and $1 - s_i - g_i$ is the item discrimination index for item i (IDI_i; de la Torre, 2008), α_{nk} is the kth ($k = 1, \cdots, K$) latent attribute for subject n, with $\alpha_{nk} = 1$ if subject n masters attribute k and $\alpha_{nk} = 0$ otherwise. The Q matrix is a $I \times K$ matrix with binary entries q_{ik} (Tatsuoka, 1983). For each i and k, $q_{ik} = 1$ indicates that attribute k is required to answer item i correctly and $q_{ik} = 0$ otherwise.

Equation (4) can be reparameterized, and it is called the reparameterized DINA model (DeCarlo, 2011) as

$$\text{logit}(p(Y_{ni} = 1)) = \beta_i + \delta_i \prod_{k=1}^{K} \alpha_{nk}^{q_{ik}}, \tag{5}$$

where $\beta_i = \text{logit}(g_i)$ and $\delta_i = \text{logit}(1 - s_i) - \text{logit}(g_i)$ are called item intercept and interaction parameter, respectively.

As stated in the literature (de la Torre and Douglas, 2004; Zhan et al., 2018a), attributes in a test are often correlated, and a higher-order structure for the attributes can be formulated by

$$\text{logit}(p(\alpha_{nk} = 1)) = \gamma_k \theta_n - \lambda_k, \tag{6}$$

where $p(\alpha_{nk} = 1)$ is the probability of subject n's mastery of attribute k; θ_n is a general (higher-order) ability for subject n; and γ_k and λ_k are the slope and intercept parameter for attribute k. Denote $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_K)$. Following Zhan et al. (2018a), a positive slope parameter is assumed, which means the probability of mastery of attribute k increases as the general ability θ_n grows. Including a higher-order structure for cognitive diagnosis can not only reduce the number of model parameters for correlated latent attributes but also obtain an assessment for subjects' overall ability.

The distribution of the general ability θ_n may be affected by the latent missingness classes of ξ_n, and we suppose that

$$\theta_n | \xi_n = c \sim N(\mu_c, \sigma_c^2), \tag{7}$$

for $n = 1, \cdots, N$ and $c = 1, \cdots, C$. If μ_c or σ_c^2 $(c = 1, \cdots, C)$ may vary between different classes, the missing data mechanism is MNAR, and we set $\mu_1 = 0$ and $\sigma_1^2 = 1$ for model identification. If μ_c and σ_c^2 remain unchanged for different c, the probability of missingness does not depend on the responses and MCAR holds, and this is where we set $\theta_n \sim N(0, 1)$ for identification.

2.3. Bayesian Parameter Estimation

The parameters of the proposed model can be estimated using the Bayesian MCMC approach. JAGS (version 4.2.0; Plummer, 2015) and the R2jags package (Su and Yajima, 2020) in R (R Core Team, 2019) were used for estimation, and the JAGS code can be found in the **Supplementary Material**. The priors of the model parameters are given below. For the majority of the parameters, the conjugate priors are used. The priors and the hyper priors for the item parameters are assigned the same as those given in Zhan et al. (2018a), please find them for details. Moreover, the noninformative prior is used for Dirichlet distribution. The (truncated) normal distribution and the inverted gamma distribution priors are chosen to obtain dispersed values for each corresponding parameter.

$$(\pi_1, \cdots, \pi_C) \sim \text{Dirichlet}(1, \cdots, 1),$$

$$\mu_c \sim N(0, 4), \quad \sigma_c^2 \sim \text{InvGamma}(1, 1), \quad (c = 2, \cdots, C),$$

$$\gamma_k \sim N(0, 4)I(\gamma_k > 0), \quad \lambda_k \sim N(0, 4), \quad (k = 1, \cdots, K),$$

$$\tau_{0i} \sim N(0, 4), \quad \tau_{ci} \sim N(0, 4)I(\tau_{ci} > 0), \quad (i = 1, \cdots, I; c = 2, \cdots, C),$$

$$\begin{pmatrix} \beta_i \\ \delta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \end{pmatrix}, \Sigma_{\text{item}} \right), \quad (i = 1, \cdots, I).$$

The hyper priors are specified:

$$\mu_\beta \sim N(-2.197, 2),\ \mu_\delta \sim N(4.394, 2)I(\mu_\delta > 0),\ \Sigma_{\text{item}} \sim \text{InvWishart}(\boldsymbol{I}, 2),$$

where $\boldsymbol{I}$ is a 2×2 identity matrix. In this case, the mean guessing and slipping probabilities are approximately equal to 0.1.

In this paper, the number of latent missingess classes C is taken as fixed. In fact, C can be selected by some information criterion, for example, deviance information criterion (DIC; Spiegelhalter et al., 2002), which can result in a statistically optimal number. In practice, it is efficient to determine C beforehand, using latent class analysis (Linzer and Lewis, 2011) just for the missingness indicators. In the following simulation studies, the number of latent missingess classes C is fixed to 2 for simplicity. For the values of $C > 2$, the results are similar and we report some results for $C = 3$ in the **Supplementary Material**.

3. SIMULATION STUDY

Two simulation studies were conducted to evaluate the empirical performance of the proposed method. Simulation 1 aimed to examine the parameter recovery using the Bayesian MCMC

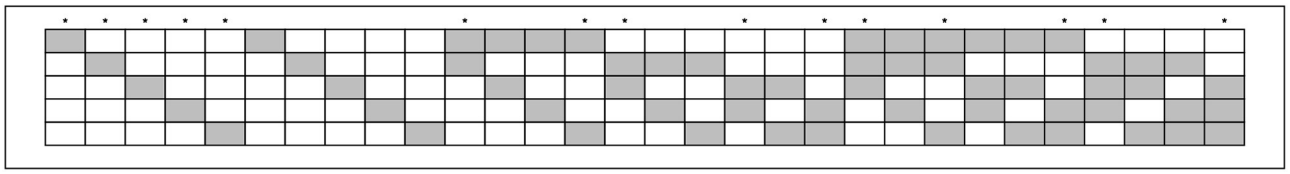


FIGURE 1 | *K*-by-*I Q* matrix for simulation study 1. Blank means "0," gray means "1"; "*" denotes items used in $I = 15$ conditions; K = the number of attributes; I = test length.

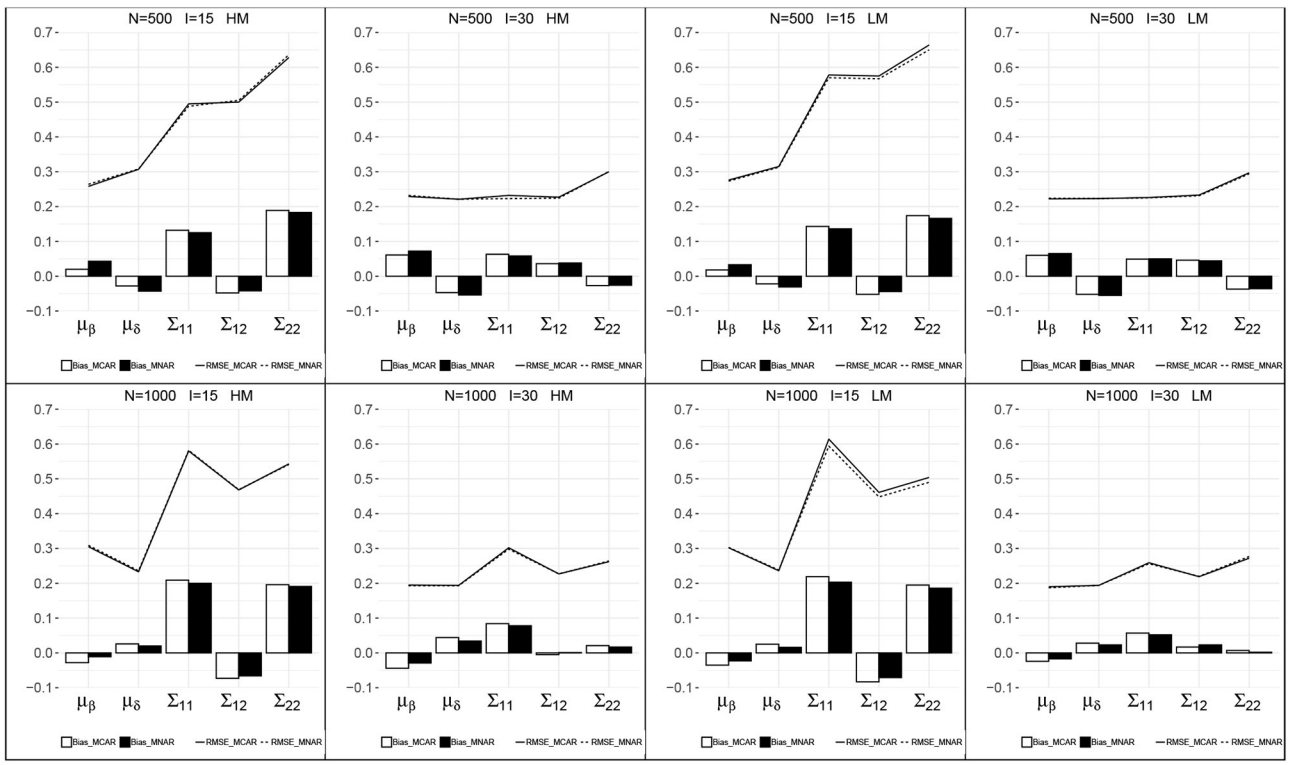


FIGURE 2 | Recovery of the parameters in the item mean vector and the item covariance matrix. HM, high missingness; LM, low missingness.

algorithm when the simulated data were generated under MNAR. Under different conditions, models with MCAR and MNAR were fitted to the simulated data, respectively, where the distribution of the general higher-order ability was unrelated or related to the latent missingness classes in MCAR or MNAR. In Simulation 2, our purpose was to study the sensitivity of incorrect use of MNAR for model fit. Parameter recovery related to diagnostic classification are reported here. The other parameters about the missing data mechanisms can be recovered well but not reported here since they are not our primary interest.

3.1. Simulation Study 1

In Simulation 1, three factors were manipulated, including (*a*) sample sizes (N) at two levels of 500 and 1,000; (*b*) test length (I) at two levels of 15 and 30; and (*c*) the probability of missingness for each item, high missingness (HM) and low missingness (LM). Five attributes ($K = 5$) were measured and the simulated Q matrices for two test length $I = 15$ and $I = 30$ were given in **Figure 1**, which were used in Zhan et al. (2018b). Most of the model parameters were assigned by referring to the real data analysis presented in Zhan et al. (2018a). Specifically, $\boldsymbol{\pi} = (0.3, 0.7)$, representing unequal probabilities for each latent class; $\tau_{2i} = 1.2$ for all items, $\boldsymbol{\tau}_0 = (0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0)$ for high missingness with $I = 15$, $\boldsymbol{\tau}_0 = (0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5, 0.0, 0.5)$ for high missingness with $I = 30$, $\boldsymbol{\tau}_0 = (1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0)$ for low missingness with $I = 15$ and $\boldsymbol{\tau}_0 = (1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5, 1.0, 1.5)$ for low miss -ingness with $I = 30$, corresponding to the missingness probability for each item between 0.22 and 0.31 in the case of high missingness and between 0.1 and 0.22 in

the case of low missingness; $\mu_2 = 1.0$ and $\sigma_2^2 = 0.5$; $\gamma_k = 1.5$ for all attributes and $\boldsymbol{\lambda} = (-1.0, -0.5, 0.0, 0.5, 1.0)$, indicating moderate correlations between attributes; $\mu_{\text{item}} = (\mu_\beta, \mu_\delta) = (-2.197, 4.394)$ and

$$\Sigma_{\text{item}} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} 1.0 & -0.8 \\ -0.8 & 1.0 \end{pmatrix},$$

which was the mean vector and the covariance matrix for a bivariate normal distribution generating β_i and δ_i. Other assignments for the model parameters, for example, equal probabilities for each latent class as those used in Ma et al. (2020), also make sense, and the results are similar to the above parameter settings.

In each of the eight conditions, models with MCAR and MNAR were fitted to the simulated data, respectively. Thirty replications were implemented for each fitted model. Pilot runs showed that the algorithm converged using 20,000 iterations with a burn-in phase of 10,000 iterations. The convergence of the chains was monitored by multivariate potential scale reduction factor, which were < 1.1 (Gelman and Rubin, 1992). The bias and the root mean square error (RMSE) of the Bayesian estimates were computed to assess the parameter recovery. For evaluating the classification of each attribute and attribute profiles, the attribute correct classification rate (ACCR) and the pattern correct classification rate (PCCR) were computed. More formally, if the kth ($k = 1, \cdots, K$) latent attribute for subject n ($n = 1, \cdots, N$), denoted as $\alpha_{k,n}$, is estimated by $\hat{\alpha}_{k,n}$, ACCR for the kth latent attribute and PCCR can be expressed as ACCR $= \sum_{n=1}^{N} I(\hat{\alpha}_{k,n} = \alpha_{k,n})/N$ and PCCR $= \sum_{n=1}^{N}[\prod_{k=1}^{K} I(\hat{\alpha}_{k,n} = \alpha_{k,n})]/N$, respectively. Here $I(\hat{\alpha}_{k,n} = \alpha_{k,n})$ is an indicator function such that $I(\hat{\alpha}_{k,n} = \alpha_{k,n}) = 1$ if $\hat{\alpha}_{k,n} = \alpha_{k,n}$ and zero otherwise.

Figure 2 presents the recovery of the item mean vector and the item covariance matrix for the models with MCAR and MNAR in all eight conditions. First, the patterns of the item parameter recovery were similar between MCAR and MNAR, especially the RMSEs of MCAR and MNAR were close in each condition. The item mean vector can be well recovered, as the bias were small and the RMSEs were relatively smaller than those of the item covariance matrix. The bias of the item mean vector under MNAR were smaller than those under MCAR at a higher sample size $N = 1,000$. The RMSEs of the item mean vector decreased as test length increased. The bias and RMSEs of the item covariance matrix were relatively higher, and decreased as test length increased. The sample size had no consistent impact on the estimates of the item covariance matrix. Moreover, the missingness probabilities had little impact on the item parameter recovery.

Table 1 summarizes the item parameter recovery for models with MCAR and MNAR in high missingness conditions. The results for low missingness are similar and presented in **Supplementary Figure 1**. The mean absolute value (MAV) of the bias and RMSE are reported. In each condition, the MAVs of the bias and RMSE under MCAR and MNAR were close. Detailed information about the recovery of each item parameter with MCAR and MNAR was similar and not reported here.

TABLE 1 | Summary of the item parameters for high missingness conditions.

		β				δ			
		Bias		**RMSE**		**Bias**		**RMSE**	
I	***N***	**MAR**	**MNAR**	**MAR**	**MNAR**	**MAR**	**MNAR**	**MAR**	**MNAR**
15	500	0.058	0.067	0.398	0.400	0.062	0.070	0.500	0.502
	1,000	0.056	0.047	0.355	0.355	0.063	0.057	0.407	0.407
30	500	0.053	0.054	0.319	0.318	0.061	0.059	0.406	0.405
	1,000	0.043	0.037	0.241	0.239	0.049	0.045	0.306	0.304

Table 2 summarizes the estimation of general ability parameter in high missingness conditions. The results for low missingness are similar and presented in **Supplementary Figure 2**. The MAV and Range of the bias and RMSE are reported. The correlation between the true and the estimated general abilities is also given. The MAVs of the bias under MCAR were much higher than those under MNAR, and the bias under MCAR were all negative, which can be seen from their Ranges. The MAVs of the RMSEs under MCAR were also higher than those under MNAR. The correlations under MNAR were higher than those under MAR. From the above results, we find that when data are generated with MNAR, incorrect use of missing data mechanism could lead to biased estimation of the general abilities.

Figures 3, **4** presents the recovery of higher-order parameters between the attributes and the general ability in each condition. For the attribute slope parameters, the bias was closer to zero and the RMSEs were relatively small. For the attribute intercept parameters, their recovery under MNAR were good with small bias and RMSEs. Under MCAR, the bias and RMSEs for attribute intercept parameters were large, with absolute values >1.0, and all the bias were negative.

Figure 5 shows the correct classification rates for each attribute and attribute profiles in all conditions. All ACCRs were > 0.90 and all PCCRs were > 0.70, which indicated a good recovery of the mastery status of attributes. ACCRs and PCCRs raised as test length increased and changed little as sample size increased. The correct classification rates of MCAR and MNAR were very close in each condition. These results may be caused by the fact that the impact of the missing data mechanism on the latent attributes is indirect through the general high-order ability, and we assume the same model for item response under both missing data mechanisms.

3.2. Simulation Study 2

The aim of Simulation 2 was to empirically examine the sensitivity of incorrect use of MNAR for model fit. In this study, the simulated data were generated with MCAR, i.e., the distribution of the general higher-order ability is independent of the latent missingness classes, and we set $\theta_n \sim N(0, 1)$ for identification. The other settings were the same as those used for $N = 500$, $J = 15$ and low missingness in Simulation 1, which is a weak condition studied in Simulation Study 1.

TABLE 2 | Summary of the general ability for high missingness conditions.

		Bias				RMSE				Cor	
		MAV		Range		MAV		Range			
I	N	MCAR	MNAR	MCAR	MNAR	MCAR	MNAR	MCAR	MNAR	MCAR	MNAR
15	500	0.702	0.094	(−1.130, −0.311)	(−0.342, 0.387)	0.966	0.656	(0.680, 1.286)	(0.382, 1.012)	0.691	0.782
	1,000	0.699	0.095	(−1.036, −0.291)	(−0.283, 0.366)	0.965	0.647	(0.674, 1.335)	(0.347, 1.045)	0.690	0.728
30	500	0.707	0.090	(−1.066, −0.328)	(−0.283, 0.382)	0.943	0.610	(0.610, 1.272)	(0.403, 0.873)	0.730	0.766
	1,000	0.707	0.085	(−1.078, −0.344)	(−0.404, 0.323)	0.939	0.588	(0.649, 1.230)	(0.346, 0.899	0.742	0.776

MAV, mean absolute value; Range = (minimum, maximum).

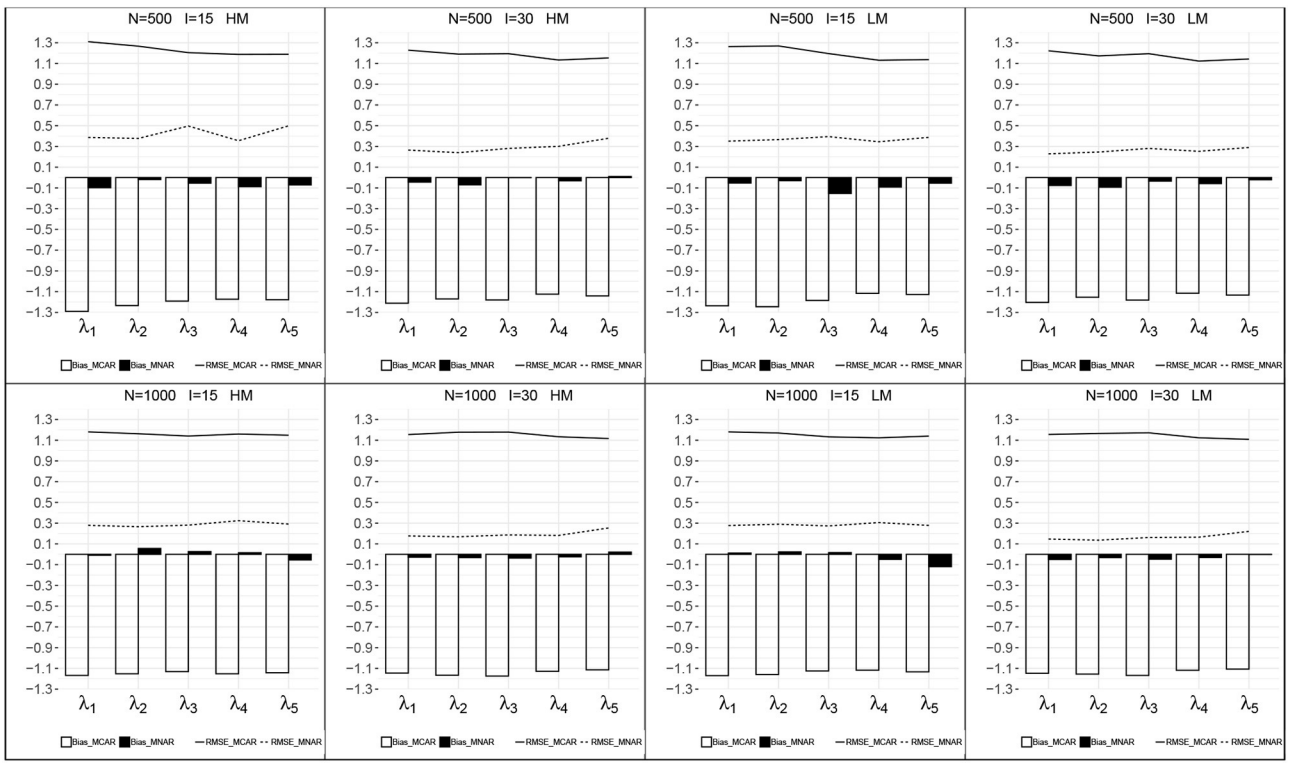


FIGURE 3 | Recovery of the attribute intercept parameters. HM, high missingness; LM, low missingness.

Table 3 presents the bias and RMSE for the item mean vector, the item covariance matrix and the higher-order structure parameters. The recovery of the item parameter mean vector and covariance matrix were similar under MCAR and MNAR. For the higher-order structure parameters, unlike the results in Simulation Study 1, the recovery of the parameters under MNAR were as good as those under MCAR.

Table 4 summarizes the recovery of the item and the general ability parameter, where the mean, standard deviation, minimum, and maximum of the bias and RMSE are reported. It also shows the correct classification rates for each attribute and attribute profiles. The results under MCAR and MNAR were similar, with mean bias close to zero and approximately equal correct classification rates for the recovery of each attribute and attribute profile. From **Tables 3**, **4**, we found that when the data were generated under MCAR, the missing data mechanism used in the model fit had little impact on the results for diagnostic classification in our framework.

4. REAL DATA ANALYSIS

To illustrate the application of the proposed method, the PISA 2015 computer-based mathematics data were used. The data include item scores and response times for each item, and we only select dichotomous item scores for illustration. Four attributes belonging to the mathematical content knowledge were evaluated, i.e., change and relationship (α_1), quantity (α_2),

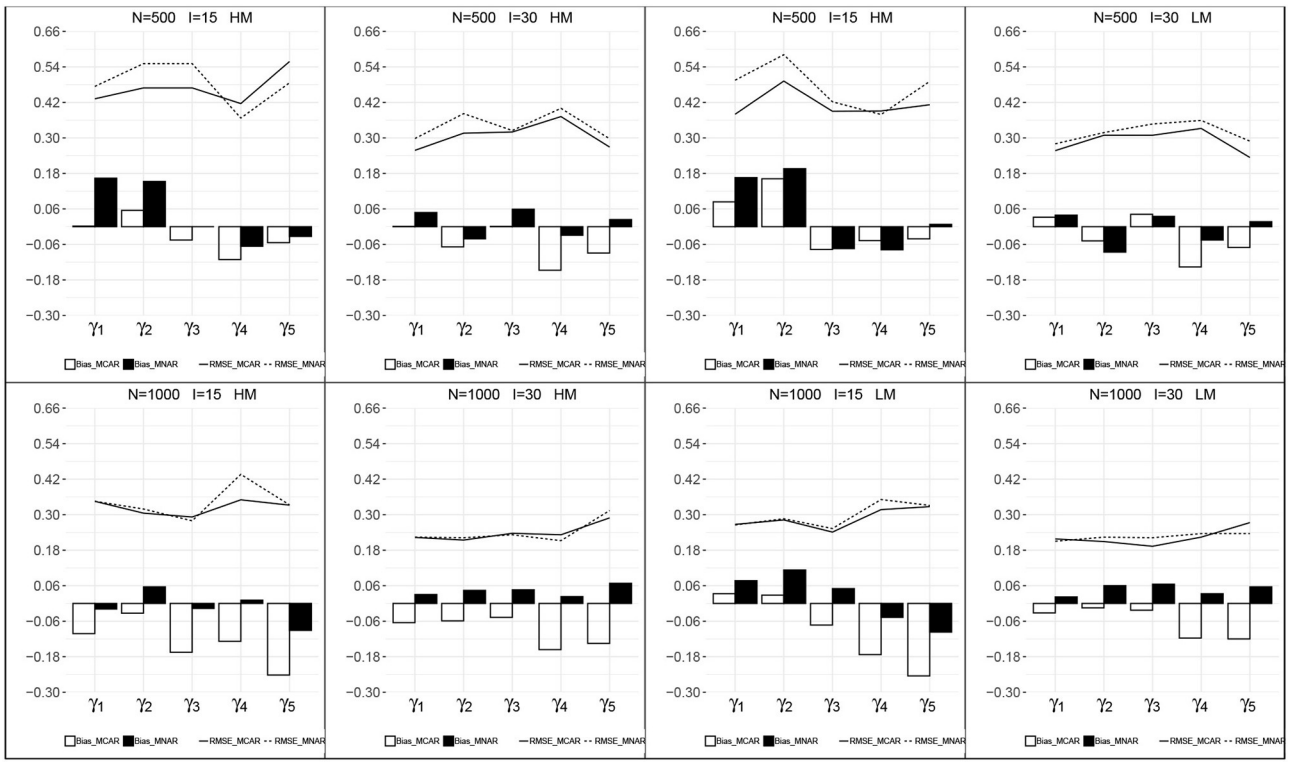


FIGURE 4 | Recovery of the attribute slope parameters. HM, high missingness; LM, low missingness.

space and shape (α_3) and uncertainty and data (α_4). $I = 9$ items with dichotomous responses were selected, with items IDs CM033Q01, CM474Q01, CM155Q01, CM155Q04, CM411Q01, CM411Q02, CM803Q01, CM442Q02, and CM034Q01. The Q matrix was shown in **Table 5**. Item responses with code 0 (no credit), code 1 (full credit), and code 9 (noresponse) were considered here. The responses "nonresponse" (code 9) were treated as missing data and might be due to a MNAR mechanism. $N = 758$ test-takers from Albania with responses 0, 1, and 9 to each of the nine items were used for analysis. The missing rate (i.e., the proportion of "nonresponse") for each item ranged from 0 to 14.78%. The models with missing data mechanisms MCAR and MNAR were both fitted to the data. The number of latent missingess classes $C = 2$ was determined by latent class analysis. Then, the analysis was specified in the same way as the simulation study. The DIC was applied to compare model fit for models under different missing data mechanisms.

The DIC values under MCAR and MNAR were 11454.2 and 10406.3, respectively, which indicated that MNAR was preferred with a lower DIC value. We were only interested in the results concerning diagnostic classification. **Table 6** reports the estimated parameters and its corresponding standard deviations for the item mean vector, the item covariance matrix, and the attribute intercept and slope parameters. The results for the item mean vector and covariance matrix were similar under different missing data mechanisms, and the estimated Σ_{12} was -1.069, which indicated that items with a higher intercept corresponded to a lower interaction. μ_β was estimated to be -2.033, which means that the mean guessing probability was approximate 0.12. All estimated attribute intercept and slope parameters were positive. The estimations for the item covariance matrix and the attribute intercept and slope parameters were poor, which were consistent with previous studies (de la Torre and Douglas, 2004; Zhan et al., 2018a). **Table 7** reports the estimated item parameters. All δ_i were positive, which means that all items satisfy $g_i < 1-s_i$. Only β_i for CM033Q01 was positive, which means that the guessing probability of this item is higher than 0.5.

Though there were 16 possible attribute patterns for four attributes, 15 attribute patterns except (0101) were found in the estimated attributes under both MCAR and MNAR. **Figure 6** presents the top five most frequent attribute patterns in the real data under MCAR and MNAR. The most prevalent attribute pattern was (0000) and the second most prevalent pattern was (1111) under both MCAR and MNAR, where the corresponding proportions were slightly different. The third and the fourth most prevalent patterns under MCAR and MNAR were reverse. The above results indicate that the missing data mechanisms have some influences on the estimated attribute patterns.

5. CONCLUSIONS

When multiple items are used to classify subjects' mastery status of latent attributes, it is almost inevitable that item-level

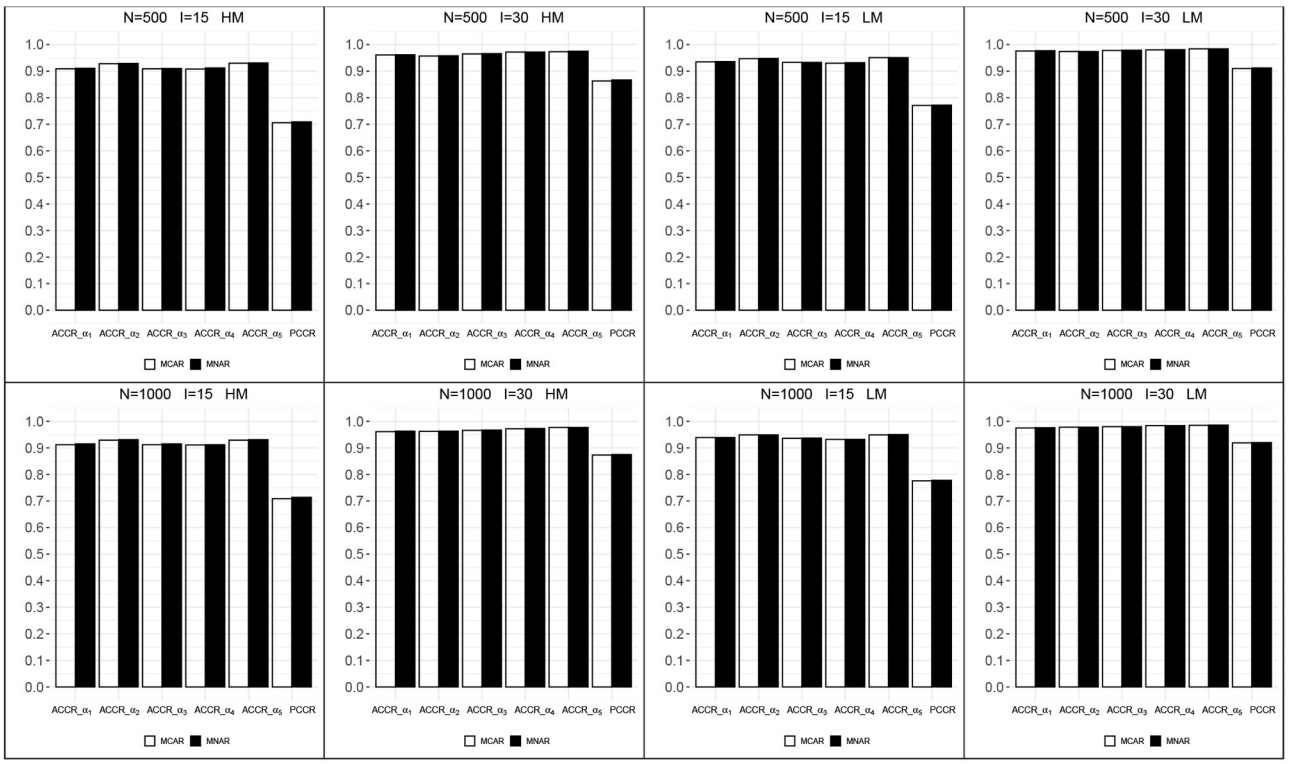


FIGURE 5 | The correct classification rates for each attribute and attribute profiles. HM, high missingness; LM, low missingness.

TABLE 3 | Recovery of the item mean vector, the item covariance matrix, the attribute slope, and intercept.

	Index	μ_β	μ_δ	Σ_{11}	Σ_{12}	Σ_{22}	λ_1	λ_2	λ_3	λ_4	λ_5	γ_1	γ_2	γ_3	γ_4	γ_5
MAR	Bias	0.043	−0.005	0.080	0.028	0.158	0.001	−0.083	0.025	0.046	0.055	0.092	0.096	0.136	0.095	0.114
	RMSE	0.265	0.263	0.394	0.357	0.454	0.271	0.221	0.232	0.278	0.230	0.487	0.552	0.438	0.372	0.367
MNAR	Bias	0.045	−0.008	0.084	0.020	0.168	−0.052	−0.122	−0.028	0.019	0.024	0.070	0.062	0.104	0.067	0.095
	RMSE	0.263	0.261	0.399	0.368	0.469	0.346	0.280	0.299	0.342	0.317	0.494	0.624	0.462	0.384	0.386

missing data will occur. It is possible that the missing data mechanism is related to the item responses, and without very strong assumptions, item-level non-response can be thought to depend on some latent variables. Motivated by this idea, we have proposed a joint modeling method incorporating item responses and missing data mechanism for cognitive diagnosis. A latent categorical variable is employed to describe the latent missingness propensity, which can avoid distributional assumptions and result in a more flexible model. Then, the latent missingness classes are linked to each item missingness indicator by the logistic function. Applying the hierarchical modeling framework, the general higher-order ability's distribution is affected by the latent missingness class in the case of MNAR and is independent of the latent missingness class in the case of MCAR.

A Bayesian MCMC method is used to estimate the model parameters under the missing data mechanisms MCAR and MNAR. The simulation study demonstrates that when the data are generated under MNAR, the estimated general ability and the attribute intercept parameters have higher bias if an incorrect missing data mechanism is used for model fit; when the data are generated under MCAR, the results between different missing data mechanisms do not have much difference. Similar results about the impact of the missing data mechanisms have been found by de Chiusole et al. (2015) in the framework of KST. The PISA 2015 computer-based mathematics data are used to explore the magnitude and direction of item and person parameters, and the results support the MNAR in the real data analysis.

Our proposed method can be further investigated in several aspects. First, the proposed model has good performance with DINA model used for illustration, and the joint modeling method can be extended to other types of CDMs for further studies. Second, this study assumes that each latent attribute is a binary variable. When polytomous attributes are involved in CDMs (Zhan et al., 2019b), modified higher-order CDMs could be utilized in the framework of our model. Third, multiple sources of data about a subject behavior, for example, response time and other process data, can be combined to build up a more

TABLE 4 | Summary of the item parameter, general ability, and attributes.

	Index		β	δ	θ		α_1	α_2	α_3	α_4	α_5
MAR	Bias	Mean	0.036	0.054	0.100	ACCR	0.863	0.891	0.876	0.885	0.920
		SD	0.040	0.062	0.126	PCCR	0.595				
		Min.	−0.030	−0.077	−0.388						
		Max.	0.110	0.138	0.464						
	RMSE	Mean	0.343	0.510	0.683						
		SD	0.177	0.131	0.097						
		Min.	0.175	0.401	0.431						
		Max.	0.853	0.909	1.060						
	Cor.				0.722						
MNAR	Bias	Mean	0.028	0.027	0.111	ACCR	0.864	0.891	0.875	0.885	0.921
		SD	0.045	0.067	0.129	PCCR	0.594				
		Min.	−0.041	−0.091	−0.425						
		Max.	0.112	0.141	0.476						
	RMSE	Mean	0.340	0.509	0.716						
		SD	0.171	0.124	0.097						
		Min.	0.171	0.399	0.462						
		Max.	0.823	0.871	1.049						
	Cor.				0.720						

SD, standard deviation; Min., minimum; Max., maximum; Cor., correlation between true and estimated values.

TABLE 5 | The *Q* matrix in the real data.

Attribute	CM033Q01	CM474Q01	CM155Q01	CM155Q04	CM411Q01	CM411Q02	CM803Q01	CM442Q02	CM034Q01
α_1	0	0	1	1	0	0	0	0	0
α_2	1	0	0	0	0	0	0	0	1
α_3	0	1	0	0	1	0	0	1	0
α_4	0	0	0	0	0	1	1	0	0

TABLE 6 | Estimates and standard errors of the parameters for the real data.

	DIC	Index	μ_β	μ_δ	Σ_{11}	Σ_{12}	Σ_{22}	λ_1	λ_2	λ_3	λ_4	γ_1	γ_2	γ_3	γ_4
MCAR	11454.2	Est.	−2.033	2.548	3.383	−1.069	1.451	0.677	2.051	1.421	1.888	3.843	3.795	4.208	3.259
		SD	0.587	0.426	2.341	1.411	1.195	0.688	0.845	0.710	0.866	1.082	0.891	0.990	0.912
MNAR	10406.3	Est.	−2.073	2.419	3.313	−1.105	1.269	2.828	2.653	3.055	3.852	2.261	1.787	2.099	1.675
		SD	0.572	0.391	2.280	1.305	1.087	0.966	0.901	0.994	1.317	0.862	0.778	0.683	0.702

Est., Estimated values; SD, standard deviation of the posterior distribution.

TABLE 7 | Estimates and standard errors of the parameters for the real data.

	Par.	033Q01	474Q01	155Q01	155Q04	411Q01	411Q02	803Q01	442Q02	034Q01
MCAR	β_i	0.155	−0.501	−0.665	−1.445	−1.737	−1.355	−4.227	−4.587	−2.945
		(0.139)	(0.156)	(0.289)	(0.185)	(0.256)	(0.166)	(0.811)	(0.846)	(0.412)
	δ_i	2.447	1.589	3.342	1.588	2.508	0.788	3.458	3.219	2.775
		(0.588)	(0.256)	(0.678)	(0.262)	(0.333)	(0.304)	(0.824)	(0.869)	(0.456)
MNAR	β_i	0.019	−0.579	−0.676	−1.452	−1.761	−1.321	−4.073	−4.543	−3.347
		(0.151)	(0.145)	(0.245)	(0.164)	(0.244)	(0.139)	(0.795)	(0.740)	(0.575)
	δ_i	2.081	1.655	3.068	1.588	2.316	0.733	3.308	3.042	2.881
		(0.432)	(0.233)	(0.500)	(0.228)	(0.295)	(0.277)	(0.809)	(0.747)	(0.570)

Par., estimated parameter; the first two letters for item IDs (i.e., "CM") are omitted to save space.

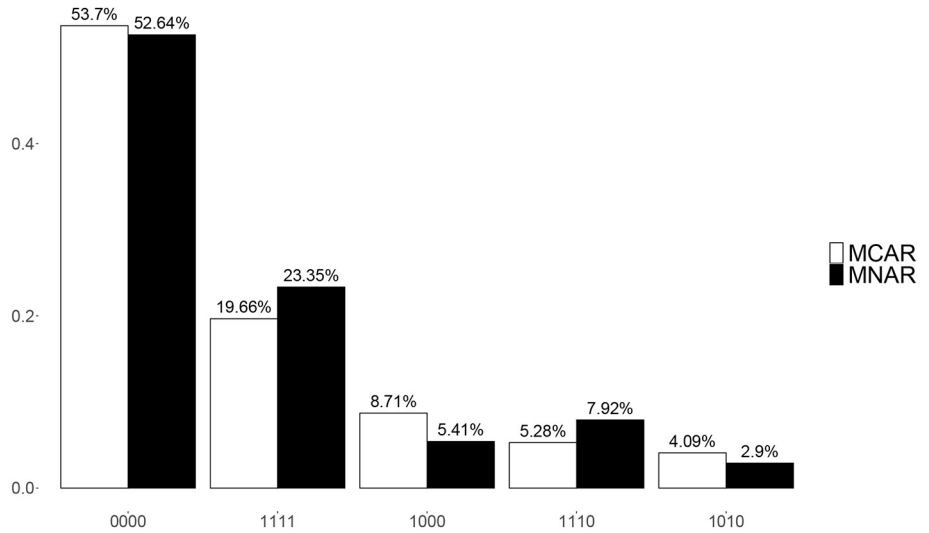


FIGURE 6 | Posterior mixing proportions for top 5 most frequent attribute patterns under MCAR and MNAR in the real data.

general model, which could provide a more comprehensive reflection of individual behavior. Finally, the number of latent missingness classes can be varied. For selecting it, Akaike information criterion (AIC), Bayesian information criterion (BIC) or DIC, can be utilized to compare different models under various choices.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://www.oecd.org/pisa.

AUTHOR CONTRIBUTIONS

NS provided original idea of the paper. XW provided the technical support. Both authors contributed to the article and approved the submitted version.

FUNDING

The research of NS was supported in part by the National Natural Science Foundation of China (grant number 11871013) and the Project of the Educational Department of Jilin Province of China (grant number 2016315). The research of XW was supported in part by the Fundamental Research Funds for the Central Universities (grant number 2412019FZ030) and Jilin Provincial Science and Technology Development Plan funded Project (grant number 20180520026JH).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.564707/full#supplementary-material

REFERENCES

Anselmi, P., Robusto, E., Stefanutti, L., and de Chiusole, D. (2016). An upgrading procedure for adaptive assessment of knowledge. *Psychometrika* 81, 461–482. doi: 10.1007/s11336-016-9498-9

Chen, L., Savalei, V., and Rhemtulla, M. (in press). Two-stage maximum likelihood approach for item-level missing data in regression. *Behav. Res. Methods*. doi: 10.3758/s13428-020-01355-x

Chen, Y., Li, X., Liu, J., and Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika* 82, 660–692. doi: 10.1007/s11336-016-9545-6

Chiu, C. Y., and Köhn, H. F. (2016). The reduced RUM as a logit model: parameterization and constraints. *Psychometrika* 81, 350–370. doi: 10.1007/s11336-015-9460-2

de Chiusole, D., Stefanutti, L., Anselmi, P., and Robusto, E. (2015). Modeling missing data in knowledge space theory. *Psychol. Methods* 20, 506–522. doi: 10.1037/met0000050

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: development and applications. *J. Educ. Meas.* 45, 343–362. doi: 10.1111/j.1745-3984.2008.00069.x

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. A. (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. Psych. Meas.* 35, 8–26. doi: 10.1177/0146621610377081

Doignon, J. P., and Falmagne, J. C. (1985). Spaces for the assessment of knowledge. *Int. J. Man Mach. Stud.*, 23, 175–196. doi: 10.1016/S0020-7373(85)80031-6

Doignon, J. P., and Falmagne, J. C. (1999). *Knowledge Spaces*. Berlin; Heidelberg; New York, NY: Springer-Verlag.

Falmagne, J. C., and Doignon, J. P. (2011). *Learning Spaces: Interdisciplinary Applied Mathematics*. Berlin; Heidelberg: Springer.

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

Glas, C. A. W., and Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educ. Psychol. Meas.* 68, 907–922. doi: 10.1177/0013164408315262

Hafez, M. S., Moustaki, I., and Kuha, J. (2015) Analysis of multivariate longitudinal data subject to nonrandom dropout. *Struct. Equat. Model.* 22, 193–201. doi: 10.1080/10705511.2014.936086

Heller, J., Stefanutti, L., Anselmi, P., and Robusto, E. (2015). On the link between cgnitive diagnostic models and knowledge space theory. *Psychometrika* 80, 995–1019. doi: 10.1007/s11336-015-9457-x

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Holman, R., and Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *Br. J. Math. Stat. Psychol.* 58, 1–17. doi: 10.1348/000711005X47168

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

Köhler, C., Pohl, S., and Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores: evaluation of item response theory models for nonignorable omissions. *Educ. Psychol. Meas.* 75, 850–874. doi: 10.1177/0013164414561785

Kuha, J., Katsikatsou, M., and Moustaki, I. (2018). Latent variable modelling with non-ignorable item non-response: multigroup response propensity models for cross-national analysis. *J. R. Stat. Soc. A* 181, 1169–1192. doi: 10.1111/rssa.12350

Leighton J., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge, UK: Cambridge University Press.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics* 60, 295–305. doi: 10.1111/j.0006-341X.2004.00173.x

Linzer, D. A., and Lewis, J. B. (2011). poLCA: an R package for polytomous variable latent class analysis. *J. Stat. Softw.* 42, 1–29. doi: 10.18637/jss.v042.i10

Little, R. J., and Rubin, D. B. (2020). *Statistical Analysis With Missing Data, 3rd Edn*. New York, NY: John Wiley and Sons, Inc.

Ma, W., Jiang, Z., and Schumacker, R. E. (2020). "Modeling omitted items in cognitive diagnosis models [Roundtable Session]," in *AERA Annual Meeting* (San Francisco, CA).

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

Ömür Sünbül, S. (2018). The impact of different missing data handling methods on DINA model. *Int. J. Eval. Res. Educ.* 7, 77–86. doi: 10.11591/ijere.v1i1.11682

Pan, Y., and Zhan, P. (2020). The impact of sample attrition on longitudinal learning diagnosis: a prologue. *Front. Psychol.* 11:1051. doi: 10.3389/fpsyg.2020.01051

Peng, S., Wang, D., Gao, X., Cai, Y. and Tu, D. (2019). The CDA-BPD: retrofitting a traditional borderline personality questionnaire under the cognitive diagnosis model framework. *J. Pac. Rim. Psychol.* 13:e22. doi: 10.1017/prp.2019.14

Plummer, M. (2015). *JAGS Version 4.0.0 User Manual*. Lyon. Retrieved from: https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rose, N., von Davier, M., and Nagengast, B. (2015). Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psychol. Test Assess. Model.* 57, 472–498.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* 63, 581–592. doi: 10.1093/biomet/63.3.581

Rupp, A., Templin, J., and Henson, R. (2010) *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

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. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Su, Y. S., and Yajima, M. (2020). *R2jags: Using R to Run 'JAGS'*. R package version 0.6-1. Retrieved from: https://CRAN.R-project.org/package=R2jags.

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

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

von Davier, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* 61, 287–301. doi: 10.1348/000711007X193957

Xu, G., and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika* 81, 625–649. doi: 10.1007/s11336-015-9471-z

Zhan, P., Jiao, H., and Liao, D. (2018a). Cognitive diagnosis modelling incorporating item response times. *Br. J. Math. Stat. Psychol.* 71, 262–286. doi: 10.1111/bmsp.12114

Zhan, P., Jiao, H., Liao, D., and Li, F. (2019a). A longitudinal higher-order diagnostic classification model. *J. Educ. Behav. Stat.* 44, 251–281. doi: 10.3102/1076998619827593

Zhan, P., Jiao, H., Liao, M., and Bian, Y. (2018b). Bayesian DINA modeling incorporating within-item characteristic dependency. *Appl. Psych. Meas.* 43, 143–158. doi: 10.1177/0146621618781594

Zhan, P., Wang, W. C., and Li, X. (2019b). A partial mastery, higher-order latent structural model for polytomous attributes in cognitive diagnostic assessments. *J. Classif.* 37, 328–351. doi: 10.1007/s00357-019-09323-7

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 Shan 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: 02 December 2020
doi: 10.3389/fpsyg.2020.575141

The Impact of Item Calibration Error on Variable-Length Cognitive Diagnostic Computerized Adaptive Testing

Xiaojian Sun[1,2], Yanlou Liu[3], Tao Xin[4]* and Naiqing Song[1,2]*

[1] School of Mathematics and Statistics, Southwest University, Chongqing, China, [2] Southwest University Branch, Collaborative Innovation Center of Assessment for Basic Education Quality, Chongqing, China, [3] China Academy of Big Data for Education, Qufu Normal University, Qufu, China, [4] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

Calibration errors are inevitable and should not be ignored during the estimation of item parameters. Items with calibration error can affect the measurement results of tests. One of the purposes of the current study is to investigate the impacts of the calibration errors during the estimation of item parameters on the measurement accuracy, average test length, and test efficiency for variable-length cognitive diagnostic computerized adaptive testing. The other purpose is to examine the methods for reducing the adverse effects of calibration errors. Simulation results show that (1) calibration error has negative effect on the measurement accuracy for the deterministic input, noisy "and" gate (DINA) model, and the reduced reparameterized unified model; (2) the average test lengths is shorter, and the test efficiency is overestimated for items with calibration errors; (3) the compensatory reparameterized unified model (CRUM) is less affected by the calibration errors, and the classification accuracy, average test length, and test efficiency are slightly stable in the CRUM framework; (4) methods such as improving the quality of items, using large calibration sample to calibrate the parameters of items, as well as using cross-validation method can reduce the adverse effects of calibration errors on CD-CAT.

Keywords: variable-length cognitive diagnostic computerized adaptive testing, calibration errors, classification accuracy, test efficiency, cognitive diagnosis assessment

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Wenyi Wang,
Jiangxi Normal University, China
Chen-Wei Liu,
National Taiwan Normal University, Taiwan

****Correspondence:***
Xiaojian Sun
sun.xiaojian@outlook.com
Tao Xin
xintao@bnu.edu.cn

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

Received: *22 June 2020*
Accepted: *04 November 2020*
Published: *02 December 2020*

Citation:
Sun X, Liu Y, Xin T and Song N (2020) The Impact of Item Calibration Error on Variable-Length Cognitive Diagnostic Computerized Adaptive Testing. Front. Psychol. 11:575141. doi: 10.3389/fpsyg.2020.575141

INTRODUCTION

Cognitive diagnostic assessment (CDA) has attracted lots of attention because of its advantage that can provide the strengths and weaknesses of examinees for specific content domains rather than just provide an overall score to indicate the position of one examinee relative to others (Leighton and Gierl, 2007). One of the research areas of CDA is cognitive diagnostic computerized adaptive testing (CD-CAT; Cheng, 2009). Compared with paper-and-pencil (P&P) test, CD-CAT can generate suitable tests that match examinees' latent attribute profiles to produce similar even higher measurement accuracy (Chen et al., 2012; Mao and Xin, 2013; Chang, 2015). Meanwhile, the generated tests have lesser items than the P&P test (Weiss, 1982).

The CD-CAT system has some important components such as item bank, item selection method, psychometric model, and terminal rule, among which item selection method is the key element (Chang, 2015), whereas the item bank is the fundamental factor. A large number of items should

be included in the item bank to make sure that all possible latent attribute profiles can be covered. Before the implementation of CD-CAT, the parameters of items should be calibrated; a commonly adopted method is using a limited number of examinees to calibrate these items (Doebler, 2012; van der Linden and Glas, 2000). For now, most studies treated the calibrated parameters of items as their true values, and items were chosen based on their calibrated parameters (Patton et al., 2013). However, studies showed that using a small number of examinees to calibrate items would produce large calibration errors (Şahin and Weiss, 2015). In other words, the estimated parameters of items in the item bank are different from their optimal parameters because of the existence of calibration errors. Researchers have demonstrated that the estimation of examinees' latent traits would be biased systematically if the calibration errors of items are ignored both in item response theory (IRT)–based CAT and CD-CAT (e.g., Doebler, 2012; Patton et al., 2013; Cheng et al., 2015; Huang, 2018).

Patton et al. (2013) examined the impacts of capitalization on chance on classification accuracy, recovery of ability, and test length in IRT-based variable-length CAT. Results showed that test information would be spuriously overestimated when the calibration sample size is small. In addition, the average test length (ATL) for small calibration sample size was shorter than for large calibration sample size, which indicated that small calibration sample size caused the tests to terminate prematurely (Patton et al., 2013). Huang (2018) investigated the influences of calibration errors on the attribute classification accuracy and measurement precision of attribute mastery classification by using three simulation studies in the context of fixed-length CD-CAT. The author found that calibration errors had negative effect on the classification accuracy and test information when the Deterministic Input Noisy Output "and" Gate (DINA; Junker and Sijtsma, 2001) model was used, while the effect was small for the compensatory reparameterized unified model (CRUM; Templin, unpublished) because of its additive characteristics. In addition, the author also found that using high-quality items, larger sample size, or increasing test length could mitigate the risks of calibration error and increase the measurement accuracy (Huang, 2018).

Although studies conducted by Patton et al. (2013) and Huang (2018) have examined the effects of standard error (SE) on the measurement accuracy and test efficiency for IRT-based variable-length CAT and fixed-length CD-CAT, respectively, the effect of SE on variable-length CD-CAT is still unclear. Compared with fixed-length CD-CAT, the posterior probability of attribute profile was commonly used as the termination rule in variable-length CD-CAT, which became more complicated because the estimation of posterior probability mainly depended on the parameters of items, and the measurement accuracy of attribute profile was strongly related to the posterior probability (Hsu et al., 2013; Hsu and Wang, 2015). The measurement accuracy would be misleadingly estimated due to inaccurate estimates of the posterior probability. For instance, the tests would be terminated prematurely if the spuriously high posterior probability was obtained, as in IRT-based CAT (Huang, 2018). Meanwhile, differed from IRT-based variable-length CAT, which used the conditional SE of ability as the termination rule, the variable-length CD-CAT could not use the SE of attribute profile as the termination rule directly because of the non-computability of the SE. Therefore, the purpose of this study is to investigate the impact of the SE of parameter calibration on the measurement of variable-length CD-CAT.

The rest of the article is organized as follows. First, three kinds of CDMs, which are the DINA model, reduced reparameterized unified model (RRUM; Hartz, 2002), and CRUM, are introduced briefly. Second, an item selection method used in the current study is described. Third, a simulation study is conducted to examine the performance of variable-length CD-CAT under different levels of calibration error. Lastly, discussions and conclusions are provided.

INTRODUCTION OF CDMS

The DINA Model

The DINA model is commonly adopted in CD-CAT framework because of its simplicity (Cheng, 2009; Chen et al., 2012; Mao and Xin, 2013). The model contains two item parameters, which are the guessing and slipping parameters, and can be written as

$$P\left(Y_{ij}=1|\eta_{ij}\right)=\left(1-s_j\right)^{\eta_{ij}}g_j^{1-\eta_{ij}},$$

where $\eta_{ij}=\prod_{k=1}^{K}\left(\alpha_{ik}\right)^{q_{jk}}$ is the ideal response, which is equal to 1 if the ith examinee masters all attributes that item j required, and is equal to 0 if at least one of the required attributes of item j is missing; K is the number of attributes; α_{ik} is the mastery or deficiency of the kth attribute for the ith examinee; q_{jk} is the element of the Q-matrix; g_j and s_j are the guessing and slipping parameters for item j, respectively.

The DINA model tends to classify examinees into two classes for each item. Specifically, some are those who master all attributes, and others are those who lack of at least one attribute that the item requires, respectively.

The RRUM

Compared with the DINA model, the RRUM can classify examinees into more than two classes for each item by using different probabilities for different attribute profiles. The RRUM has also attracted considerable attention in CD-CAT in recent years (e.g., Dai et al., 2016; Huebner et al., 2018; Wang et al., 2020) and can be expressed as

$$P\left(Y_{ij}=1|\alpha_i\right)=\pi_j*\prod_{k=1}^{K}r_{jk}^{*(1-\alpha_{ik})q_{jk}},$$

where π_j^*, the baseline parameter, refers to the probability of correct response to item j when individuals have mastered all attributes that the item requires; r_{jk}^*, the penalty parameter, denotes the reduction in the probability of correct response to item j when an individual lacks attribute k. Both π_j^* and r_{jk}^* range from 0 to 1.

The CRUM

Both of the DINA model and the RRUM are non-compensatory CDMs; the CRUM, on the contrary, is the compensatory model. The probability of correctly answering an item is defined as the addition of intercept and main effect of attributes that the item requires by using the logit as the link function (Templin, unpublished). The item response function of the CRUM can be written as

$$P\left(Y_{ij}=1|\alpha_i^*\right)=\frac{\exp\left(\delta_{j0}+\sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{ik}q_{jk}\right)}{1+\exp\left(\delta_{j0}+\sum_{k=1}^{K_j^*}\delta_{jk}\alpha_{ik}q_{jk}\right)},$$

where δ_{j0} is the intercept, which refers to the probability of correctly answering item j when none of the required attributes is mastered; δ_{jk} is the main effect of attribute k, representing the change in probability when attribute k is mastered, and K_j^* is the number of attributes that item j requires.

ITEM SELECTION METHOD

A number of item selection methods have been developed in CD-CAT, for instance, the Kullback–Leibler (KL) and the Shannon entropy (SHE; Xu et al., 2003) strategies, the posterior-weighted KL (PWKL; Cheng, 2009) strategy, the modified PWKL (MPWKL) and the generalized DINA (GDINA; de la Torre, 2011) model discrimination index (GDI) strategies (Kaplan et al., 2015), and the mutual information (MI; Wang, 2013) strategy. Among these strategies, both of MI and MPWKL strategies can produce high measurement accuracy when the test length is short; in addition, the MPWKL strategy performs better than the traditional WPKL method (Kaplan et al., 2015). The MPWKL strategy is commonly used in CD-CAT (e.g., Huang, 2018; Huebner et al., 2018; Wang et al., 2020); therefore, it will be adopted in current study. The expression of the MPWKL for item j can be formulated as

$$MPWKL_{ij}=\sum_{d=1}^{2^K}\left\{\sum_{c=1}^{2^K}\left[\sum_{x=0}^{1}\log\left(\frac{P\left(Y_{ij}=x|\alpha_d\right)}{P\left(Y_{ij}=x|\alpha_c\right)}\right)P\left(Y_{ij}=x|\alpha_d\right)\pi\left(\alpha_c|Y_{i,n-1}\right)\right]\pi\left(\alpha_d|Y_{i,n-1}\right)\right\},$$

where $P\left(Y_{ij}=x|\alpha\right)$ is the item response function, $Y_{i,n-1}$ is the response vector for the first n—1 items for the ith examinee, and $\pi\left(\alpha|Y_{i,n-1}\right)$ is the posterior probability for attribute profile α; the rests are defined as above. Items with the largest MWPKL index will be selected during the implementation of CD-CAT.

SIMULATION STUDY

Independent Variables

Several factors are manipulated in this study, including model type, quality of the item bank, magnitude of calibration error, and termination rule. Specifically, three CDMs, which are the DINA model, RRUM, and CRUM, are adopted in this study, and these models have been used in previous studies (e.g., Chen et al., 2012; Mao and Xin, 2013; Wang, 2013; Huang, 2018; Huebner et al., 2018; Wang et al., 2020). In addition, three levels are manipulated for the quality of the item bank, which are high-, low-, and mix-item quality (i.e., the mix of high- and low-quality items), respectively, and the corresponding item parameters are listed in **Table 1**. All these settings are modified from previous studies (Chen et al., 2012; Huang, 2018; Wang et al., 2020). As for the magnitude of calibration error, similar with Huang (2018), four levels are adopted in this study, while the specific values are different. A pilot study shows that when the models are the DINA model and the RRUM in CD-CAT, the standard deviations (SDs) of the calibration errors are smaller than 0.1 for different calibration sample sizes conditional on high-quality items. Considering that the SD might be larger than 0.1 when the quality of items is low, the SD values of calibration errors are 0, 0.1, 0.2, and 0.3, which are named true value, small, median, and large errors, respectively. The mean value is set as 0 for all the calibration errors. Consistent with previous studies (Hsu et al., 2013; Hsu and Wang, 2015), the posterior probability of attribute profile is used as the termination rule, and the termination criteria are 0.7 and 0.8, respectively.

TABLE 1 | Simulation design of the current study.

Independent variable		Values
Model		DINA, RRUM, CRUM
Calibration error		0, 0.1, 0.2, 0.3
Termination rule		0.7, 0.8
Item quality	High	DINA: $s, g \sim U$ (0.05, 0.25)
		RRUM: $\pi^* \sim U$ (0.75, 0.95); $r^* \sim U$ (0.05, 0.40)
		CRUM: $\delta_0 \sim U$ (−3, −1.1); $P\left(Y_j=1\right) \sim U$ (0.85, 1.0)
	Low	DINA: $s, g \sim U$ (0.25, 0.45)
		RRUM: $\pi^* \sim U$ (0.55, 0.75); $r^* \sim U$ (0.15, 0.50)
		CRUM: $\delta_0 \sim U$ (−1.1, −0.2); $P\left(Y_j=1\right) \sim U$ (0.6, 0.85)
	Mix	DINA: $s, g \sim U$ (0.05, 0.45)
		RRUM: $\pi^* \sim U$ (0.55, 0.95); $r^* \sim U$ (0.05, 0.50)
		CRUM: $\delta_0 \sim U$ (−3, −0.2); $P\left(Y_j=1\right) \sim U$ (0.6, 1.0)

$P\left(Y_j=1\right)$ refers to the probability of correctly answering item j.

Control Variables

Five variables are fixed in current study. Specifically, 5 attributes and 300 items, which are commonly adopted in empirical and simulation studies (e.g., Liu et al., 2013; Mao and Xin, 2013; Huang, 2018; Huebner et al., 2018; Wang et al., 2020), are used in this study. The generation of Q-matrix is consistent with Chen et al. (2012). Specifically, three basic Q-matrices, which include items measuring one, two, and three attribute(s), respectively, are generated, namely Q_1, Q_2, and Q_3, respectively. Then, the final Q-matrix is constructed by merging 20 basic matrix Q_1, 14 basic matrix Q_2, and 6 basic matrix Q_3. In addition, the number of examinees is set as 2,000, which is consistent with Mao and Xin (2013) and Huang (2018). The generation of the attribute profile is consistent with Cheng (2009), which assumes that the probability that each examinee masters each attribute is

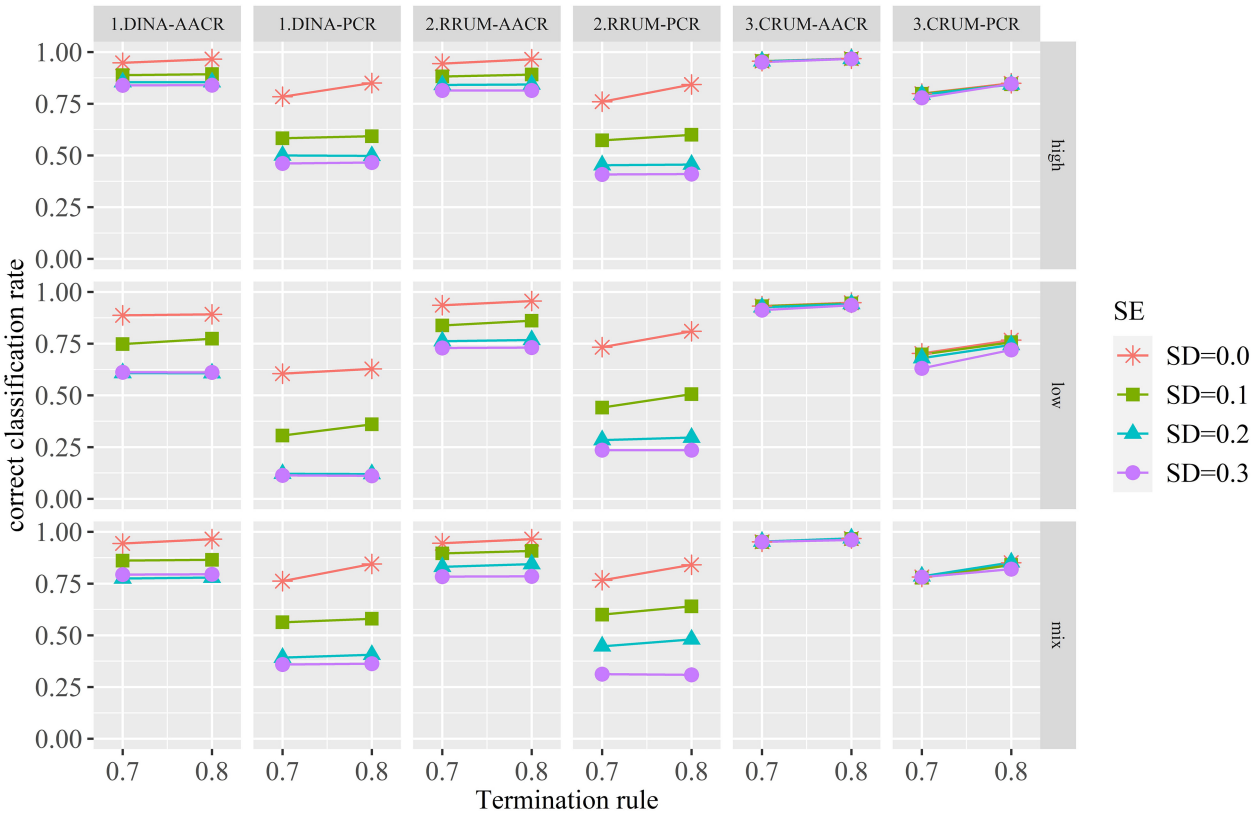


FIGURE 1 | The correct classification rates for all conditions.

0.5. Meanwhile, the item selection method is MPWKL, and the maximum number of items of the test is 30, which are adopted in previous studies (Huebner et al., 2018).

There are $3 \times 3 \times 4 \times 2 = 72$ conditions in total, among which the termination rule is within-group variable, and the rest are between-group variables. Twenty-five replications are generated for each condition to reduce the sample error. The simulation study is implemented in R software, and the codes are available upon request from the corresponding author.

Measurement Criteria

Two indices are used to evaluate the measurement accuracy of attribute classification, which include pattern correct classification rate (PCR) and average attribute correct classification rate (AACR). These two indices both range from 0 to 1, and larger value indicates better performance. They can be expressed as

$$PCR = \sum_{i=1}^{N} I\left(\hat{\alpha}_i = \alpha_i\right) \Big/ N,$$

$$AACR = \sum_{i=1}^{N}\sum_{k=1}^{K} I\left(\hat{\alpha}_{ik} = \alpha_{ik}\right) \Big/ (N \times K),$$

where N is the number of examinees; $I(\cdot)$ is the indicator function, which equals 1 if $\hat{\alpha}_i = \alpha_i (\text{or} \hat{\alpha}_{ik} = \alpha_{ik})$ is true, and vice versa; $\hat{\alpha}_i$ is the estimate of attribute profile, and α_i is the true attribute profile.

In addition to the evaluation of the measurement accuracy, the relative measurement precision (RMP) index is also used to evaluate the measurement precision of the estimated attribute profile (Huang, 2018). Moreover, based on the study conducted by Patton et al. (2013), the relative test efficiency (RTE) is used to evaluate the test efficiency in CD-CAT. These two indices can be written as

$$RMP = \frac{\sum_{i=1}^{N} \max\limits_{c=1,2,\cdots,2^K} \left(\pi\left(\alpha_c | Y_i, \hat{\delta}\right)\right)}{\sum_{i=1}^{N} \max\limits_{c=1,2,\cdots,2^K} \left(\pi\left(\alpha_c | Y_i, \delta\right)\right)} \text{ and}$$

$$RTE = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J_i} \widehat{CDI}_j}{\sum_{i=1}^{N}\sum_{j=1}^{J_i} CDI_j},$$

where $\hat{\delta}$ and δ are vectors including all estimated and true item parameters that examinee i answers, respectively. CDI is the cognitive diagnostic index and proposed by Henson and Douglas (2005), which can be regarded as the test information for each examinee in current study. The details can be found in Henson and Douglas (2005).

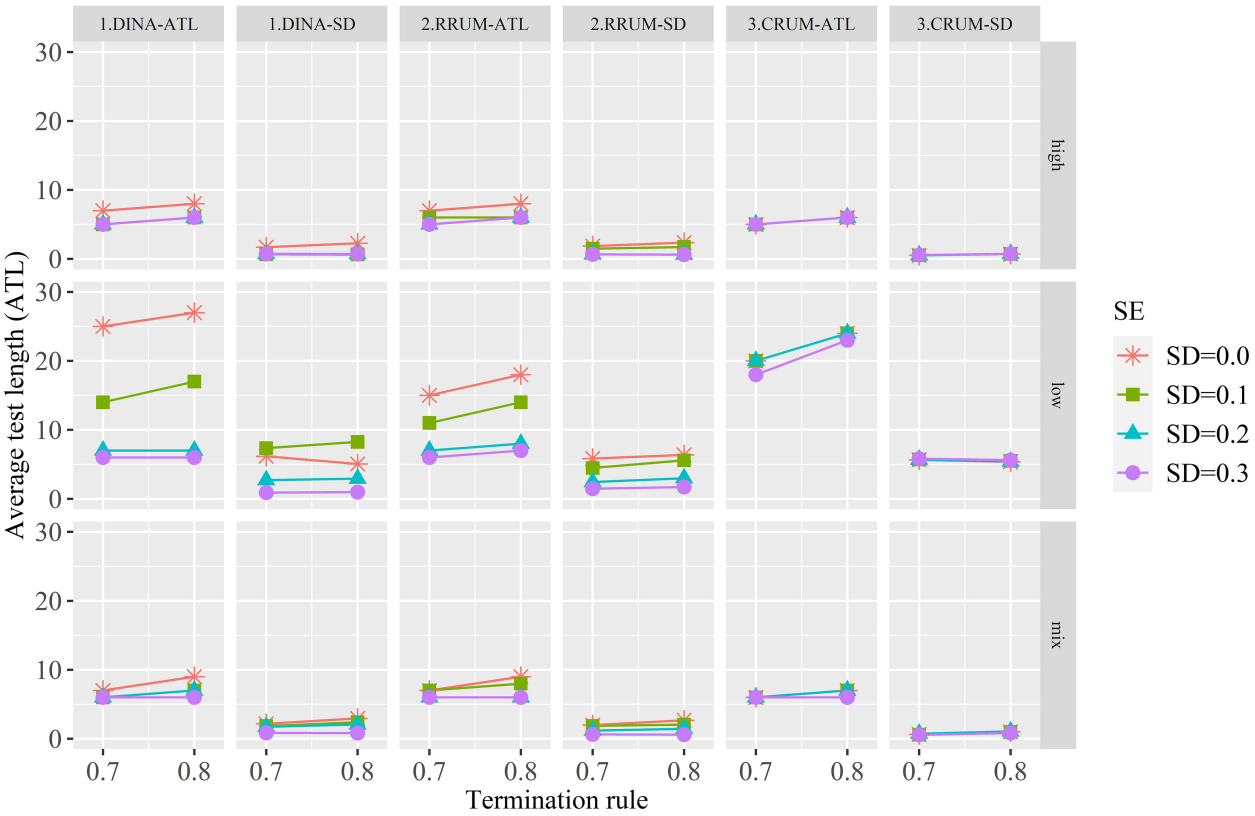


FIGURE 2 | The average test length and the corresponding SDs for all conditions.

RESULTS

Figure 1 presents the estimated AACR and PCR for each condition. Some meaningful findings can be summarized from the figure. First, the calibration error has negative effect on the AACR and PCR for the DINA model and RRUM irrespective of the item quality and termination rule. With the increase in SD of calibration error, the AACR and PCR decrease for these two models. The effect of calibration error on CRUM can be ignorable, especially for high- and mix-item quality. The differences between the true value ($SD = 0$) and calibration errors are small enough. Second, for the DINA model, the differences of AACR and PCR between median ($SD = 0.2$) and large ($SD = 0.3$) are small, regardless of the item quality and termination rule. The same results are found for the RRUM under high- and low-item quality, whereas the differences are large when the item bank is of mixed quality. Third, the quality of the item bank has positive effect on PCR, regardless of model type and termination rule. Items with high quality produce higher PCRs than low-quality items. Fourth, termination rule has a positive effect on PCR for the CRUM, whereas the effects are complicated for the DINA model and RRUM. When calibration error is zero, termination rule has positive effect on PCR, while the positive effect of termination rule on PCR tends to decrease with the increase of calibration error. Fifth, the PCRs and AACRs are stable for all conditions, and the SDs of PCR for the DINA model, RRUM, and CRUM range from 0.005 to 0.013, 0.008 to 0.015, and 0.007 to 0.013, respectively. Meanwhile, the SDs of AACR range from 0.006 to 0.049, 0.006 to 0.044, and 0.004 to 0.016, respectively, for these three CDMs.

Figure 2 shows the ATL and the corresponding SD. Some meaningful points are summarized. First, the true value ($SD = 0$) produces the longest ATL for the DINA model and RRUM, and with the increase of calibration error, the ATL, in general, tends to decrease, especially for the low-quality item. The CRUM, on the contrary, produces similar ATL irrespective of calibration error. Second, the quality of the item bank has an effect on ATL. Items with high and mixed quality tend to produce shorter ATL than low-quality items, and high- and mixed-quality items produce similar ATL. Third, termination rule has positive effect on ATL, especially for true and small calibration error. The ATLs are relatively stable for median and large calibration error. These results indicate that items with calibration error tend to terminate prematurely for the DINA model and RRUM. The CRUM, on the contrary, is not affected by calibration error.

Figure 3 depicts the RMPs with different calibration errors. The RMPs are greater than one in all conditions. Specifically, small calibration error produces the smallest RMPs for the DINA model and RRUM irrespective of the item quality and termination rule. The large calibration error, on the contrary, produces the largest RMPs. In other words, items with calibration error can produce spuriously high measurement precision,

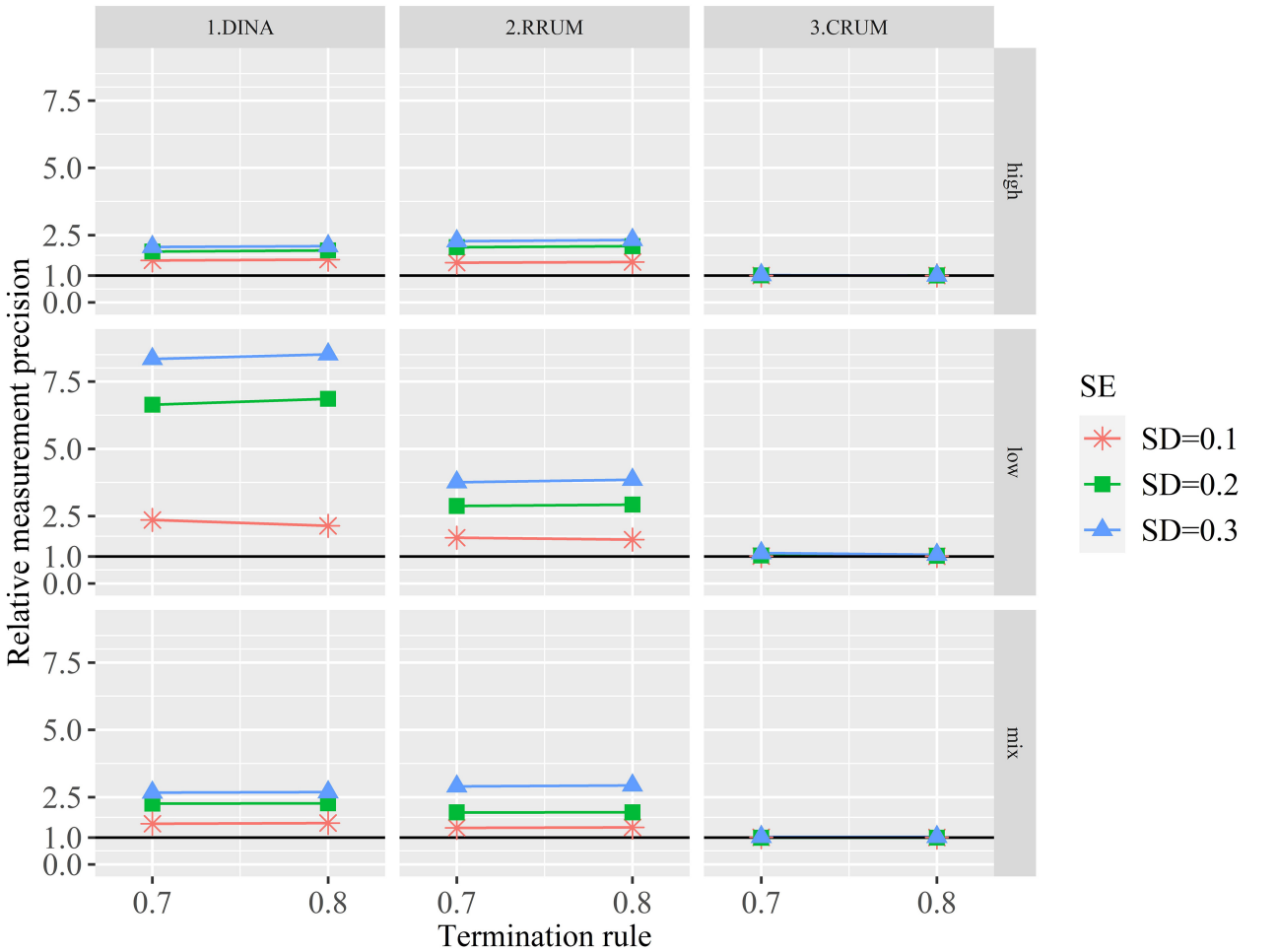


FIGURE 3 | The relative test efficiency with different calibration errors.

indicating that the posterior probability that an examinee belongs to a specific attribute profile is overestimated substantially, especially for low-quality items. By contrast, the RMPs are closed to one for the CRUM, regardless of the item quality and termination rule, suggesting that the CRUM can produce similar measurement precision between true and calibration error–affected items.

Table 2 lists the RTE for each calibration error. The results are similar to those for the RMP. Overall, the smaller the calibration error is, the better the RTE is for the DINA model and RRUM, regardless of the item quality and termination rule. On the contrary, the calibration error has ignorable effect on the RTE for the CRUM. In addition, the calibration error has more serious effect on the DINA model than the RRUM; the RTE indices are larger than those for the RRUM.

DISCUSSION

CD-CAT can diagnose the strengths and weaknesses of examinees with fewer items and provide feedback immediately (Chen et al., 2012; Mao and Xin, 2013; Chang, 2015; Wang et al., 2020). Before the implementation of CD-CAT, parameters of items in the bank need to be calibrated. However, a limited number of examinees are commonly used to calibrate the item parameters to maintain the security of the item bank; therefore, calibration errors exist in these items (van der Linden and Glas, 2000; Doebler, 2012; Patton et al., 2013; Huang, 2018). The current study investigates the impacts of calibration error on variable-length CD-CAT, as well as examines the methods for reducing the adverse effects of calibration errors. Simulation study indicates that calibration error has substantial effect on the classification accuracy, the ATL, measurement precision, and test efficiency. Meanwhile, several factors are found that can be used to reduce the adverse effects of calibration errors.

Results show that for the DINA mode and the RRUM, with the increase of calibration error, the correct classification rates are decreasing, suggesting that the calibration error should not be ignored when CD-CAT is used to diagnose the mastery of examinees, or the results will be inaccuracy, which is consistent with previous studies (Patton et al., 2013; Huang, 2018). At the same time, compared with true item parameter, the ATL becomes

TABLE 2 | The relative test efficiency for each calibration error.

Model	Item quality	Calibration error	ε = 0.7		ε = 0.8	
			M	*SD*	*M*	*SD*
DINA	Low	0.1	2.456	0.016	2.299	0.029
		0.2	18.782	0.200	19.528	0.259
		0.3	45.504	0.703	45.149	0.460
	High	0.1	5.149	0.034	5.117	0.039
		0.2	6.752	0.050	6.777	0.047
		0.3	8.442	0.064	8.408	0.049
	Mix	0.1	4.008	0.026	3.998	0.035
		0.2	5.160	0.055	5.346	0.053
		0.3	11.294	0.069	11.251	0.076
RRUM	Low	0.1	1.778	0.015	1.683	0.013
		0.2	6.112	0.061	6.008	0.072
		0.3	9.882	0.100	9.617	0.093
	High	0.1	3.386	0.018	3.444	0.019
		0.2	5.968	0.040	5.937	0.035
		0.3	6.120	0.028	6.134	0.039
	Mix	0.1	3.112	0.025	2.885	0.025
		0.2	5.005	0.036	5.106	0.034
		0.3	6.802	0.060	6.766	0.082
CRUM	Low	0.1	1.019	0.002	1.017	0.002
		0.2	1.041	0.004	1.026	0.004
		0.3	1.076	0.006	1.054	0.008
	High	0.1	1.007	0.000	1.005	0.000
		0.2	1.013	0.001	1.006	0.001
		0.3	1.065	0.001	0.983	0.001
	Mix	0.1	1.007	0.000	1.013	0.000
		0.2	0.997	0.001	0.978	0.001
		0.3	1.098	0.002	1.059	0.001

shorter for these error-affected item parameters. In other words, when true item parameters are used to select candidate items, to obtain the prespecified posterior probability of attribute profile, more items are needed. However, fewer items are sufficient to obtain the prespecified posterior probability for these error-affected item parameters. Such results indicate that the error-affected item parameters would terminate the test prematurely for examinees, and the similar results were found in IRT-based CAT conducted by Patton et al. (2013).

In addition, the results based on the RMP demonstrate that calibration error tends to produce spuriously high posterior probability that an examinee belongs to a specific attribute profile, which can be used to explain why the test is terminated prematurely for error-affected item parameters. In variable-length CD-CAT, the posterior probability is mainly used as the termination rule, which is adopted in current study as well. When items with calibration error are used in the test, the overestimated posterior probability is obtained; therefore, it is easy to meet the prespecified termination criteria; consequently, the test would be ended. Huang (2018) also obtained the similar result in the context of fix-length CD-CAT. Moreover, the results of the RTE indicate that the calibration error-affected items produce overestimated test information. This finding is consistent with Patton et al. (2013), but inconsistent with Huang (2018). Huang (2018) found that larger calibration errors produce lower CDI values. One possible explanation for this inconsistence is that the ways of calculating the CDI are different. In current study, the CDI values are calculated based on the same items, but the parameters are different, while both of the items and item parameters are different in Huang's (2018) study.

The results based on the CRUM show that calibration error, in general, has ignorable effects on the attribute classification accuracy, ATL, measurement precision, and test efficiency, regardless of the item quality and termination rule. The classifications of attribute and attribute profile are relatively higher for the CRUM than for the DINA model and the RRUM. The similar results are found in Huang's (2018) study. In addition, item quality has a negative effect on ATL for the CRUM, which is low-quality item that produces the longest test length. The differences of ATL between high and mixed quality are comparable.

In summary, compared with the DINA and RRUM, the presence of calibration error does not affect the measurement accuracy of the CRUM, and the possible reason can be attributed to the additive characteristic of this model (Huang, 2018). For instance, some parameters may be overestimated, whereas others may be underestimated for a specific item, and at this point, the summon of these parameters may be consistent with the true value because of the counterbalance of the overestimated and underestimated parameters. The DINA model and RRUM, on the contrary, do not possess the additive characteristic. Specifically, the probability of answering correctly involves only one parameter (the slipping or guessing) for the DINA model; thus, the result would be inconsistent with the true value if the calibration error is considered. Meanwhile, the characteristic of the RRUM is multiplicative property rather than additive property; therefore, the multiplication of the baseline and penalty parameters contaminated by calibration error may produce the final result that is different from the true value of the item.

According to the results of the simulation study, some factors can be summarized to reduce the adverse effect of calibration error on variable-length CD-CAT. First, considering that there is negative relationship between calibration sample size and calibration error of items (Huang, 2018), therefore increasing the sample size of calibration should be a feasible way to achieve this purpose. Second, compared with the low- and mixed-quality items, the differences between contaminated by calibration error and true item parameters are relatively smaller conditional on the high-quality items, which means improving the quality of the item bank is another way to achieve this purpose. In addition to these factors, previous studies indicated that increasing the test length could also relieve the negative effect of the calibration error (Patton et al., 2013; Huang, 2018). Moreover, the cross-validation method, which is used by Patton et al. (2013) in the CAT framework, can also reduce the adverse effect of calibration error. Results based on four conditions, including the cross combination of the quality of the item bank (low and high quality) and the SD of calibration error (0.2 and 0.3), show that using the cross-validation method can produce

better performance of the measurement accuracy, RMP, and RTE than the method without using cross-validation. Furthermore, Huang (2018) also found that using the item exposure control method, such as the Sympson and Hetter on line method with freeze (SHOF method), could also relieve the negative impact of calibration error. The item selection strategy with the SHOF method produces slightly lower classification accuracy than strategy without SHOF method, while the values of the RMP produced by strategy with the SHOF method are close to one for low-quality items (Huang, 2018).

Although some promising results are obtained in this study, some possible directions can be investigated for future studies. For instance, the complexity of Q-matrix has an important effect on classification accuracy and is not considered in this study. Future study can be conducted to explore the performance of this factor. In addition, the setting of calibration error may not represent the empirical situation sufficiently; therefore, using different calibration sample size to calibrate the item parameters may be an option for future study. Moreover, only MPWKL strategy is used in the current study; other item selection strategies such as MI and SHE strategies should be adopted in the future. Furthermore, all CDMs used in the current study are special cases of the GDINA model and the log-linear cognitive diagnostic model (LCDM; Henson et al., 2009). Compared with the GDINA model and LCDM, the CDMs used in this study lack flexibility and have more restrictions (Rupp et al., 2010). Therefore, general CDMs such as the GDINA model and LCDM should be investigated in the future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

XS, TX, and NS proposed the original concept and designed the fundamental study of this study. XS and YL wrote the simulation study code and organized the article. All authors contributed to the manuscript revision.

FUNDING

This work was supported by the Cultural Experts, and "Four Groups of Talented People" Foundation of China and National Natural Science Foundation of China (Grant Nos. 31900794 and 32071093).

REFERENCES

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika* 80, 1–20. doi: 10.1007/s11336-014-9401-5

Chen, P., Xin, T., Wang, C., and Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika* 77, 201–222. doi: 10.1007/S11336-012-9255-7

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

Cheng, Y., Patton, J. M., and Shao, C. (2015). a-stratified computerized adaptive testing in the presence of calibration error. *Educ. Psychol. Meas.* 75, 260–283. doi: 10.1177/0013164414530719

Dai, B., Zhang, M., and Li, G. (2016). Exploration of item selection in dual-purpose cognitive diagnostic computerized adaptive testing: based on the RRUM. *Appl. psychol. Meas.* 40, 625–640. doi: 10.1177/0146621616666008

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

Doebler, A. (2012). The problem of bias in person parameter estimation in adaptive testing. *Appl. Psychol. Meas.* 36, 255–270. doi: 10.1177/0146621612443304

Hartz, S. (2002). *A Bayesian Framework for The Unified Model for Assessing Cognitive Abilities: Blending Theory with Practice*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Champaign, IL.

Henson, R., and Douglas, F. (2005). Test constriction for cognitive diagnosis. *Appl. Psychol. Meas.* 29, 262–277. doi: 10.1177/0146621604272623

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Hsu, C. L., and Wang, W. C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *J. Educ. Meas.* 52, 125–143. doi: 10.1111/jedm.12069

Hsu, C.-L., Wang, W.-C., 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

Huang, H. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *J. Classif.* 35, 437–465. doi: 10.1007/s00357-018-9265-y

Huebner, A., Finkelman, M. D., and Weissman, A. (2018). Factors affecting the classification accuracy and average length of a variable-length cognitive diagnostic computerized test. *J. Comput. Adapt. Test.* 6, 1–14. doi: 10.7333/1802-060101

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., de la Torre, J., and Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Appl. Psychol. Meas.* 39, 167–188. doi: 10.1177/0146621614554650

Leighton, J., and Gierl, M. (eds) (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., and Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *J. Classif.* 30, 152–172. doi: 10.1007/s00357-013-9128-5

Mao, X., and Xin, T. (2013). The application of the Monte Carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Appl. Psychol. Meas.* 37, 482–496. doi: 10.1177/0146621613486015

Patton, J., Cheng, Y., Yuan, K.-H., and Diao, Q. (2013). The influence of calibration error on variable-length computerized adaptive testing. *Appl. Psychol. Meas.* 37, 24–40. doi: 10.1177/0146621612461727

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guilford Press.

Şahin, A., and Weiss, D. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educ. Sci. Theory. Pract.* 15, 1585–1595. doi: 10.12738/estp.2015.6.0102

van der Linden, W. J., and Glas, C. A. (2000). Capitalization on item calibration error in adaptive testing. *Appl. Meas. Educ.* 13, 35–53. doi: 10.1207/s15324818ame1301_2

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, Y., Sun, X., Chong, W., and Xin, T. (2020). Attribute discrimination index-based method to balance attribute coverage for short-length cognitive diagnostic computerized adaptive testing. *Front. Psychol.* 11:224. doi: 10.3389/fpsyg.2020.00224

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Appl. Psychol. Meas.* 6, 473–492. doi: 10.1177/014662168200600408

Xu, X., Chang, H., and Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. *Paper presented at the annual meeting of the American Educational Research Association*, (Chicago, IL: AERA).

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 Sun, Liu, Xin and Song. 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: 20 January 2021
doi: 10.3389/fpsyg.2020.618336

A Semi-supervised Learning-Based Diagnostic Classification Method Using Artificial Neural Networks

Kang Xue[1]* *and Laine P. Bradshaw*[2]

[1] NWEA, Portland, OR, United States, [2] Department of Educational Psychology, University of Georgia, Athens, GA, United States

The purpose of cognitive diagnostic modeling (CDM) is to classify students' latent attribute profiles using their responses to the diagnostic assessment. In recent years, each diagnostic classification model (DCM) makes different assumptions about the relationship between a student's response pattern and attribute profile. The previous research studies showed that the inappropriate DCMs and inaccurate Q-matrix impact diagnostic classification accuracy. Artificial Neural Networks (ANNs) have been proposed as a promising approach to convert a pattern of item responses into a diagnostic classification in some research studies. However, the ANNs methods produced very unstable and unappreciated estimation unless a great deal of care was taken. In this research, we combined ANNs with two typical DCMs, the deterministic-input, noisy, "and" gate (DINA) model and the deterministic-inputs, noisy, "or" gate (DINO) model, within a semi-supervised learning framework to achieve a robust and accurate classification. In both simulated study and real data study, the experimental results showed that the proposed method could achieve appreciated performance across different test conditions, especially when the diagnostic quality of assessment was not high and the Q-matrix contained misspecified elements. This research study is the first time of applying the thinking of semi-supervised learning into CDM. Also, we used the validating test to choose the appropriate parameters for the ANNs instead of using typical statistical criteria.

Keywords: cognitive diagnostic classification, artificial neural networks, semi-supervised learning, machine learning, co-training algorithm

OPEN ACCESS

Edited by:
Tao Xin,
Beijing Normal University, China

Reviewed by:
Alexander Robitzsch,
IPN-Leibniz Institute for Science and Mathematics Education, Germany
Dandan Liao,
American Institutes for Research, United States

***Correspondence:**
Kang Xue
kang.xue@nwea.org

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

Received: *16 October 2020*
Accepted: *21 December 2020*
Published: *20 January 2021*

Citation:
Xue K and Bradshaw LP (2021) A Semi-supervised Learning-Based Diagnostic Classification Method Using Artificial Neural Networks. Front. Psychol. 11:618336. doi: 10.3389/fpsyg.2020.618336

1. INTRODUCTION

The purpose of cognitive diagnostic modeling (CDM; Templin and Henson, 2006) or diagnostic measurement is to provide students' skill/knowledge/attributes mastery status (mastery or non-mastery) through their responses to items from carefully designed assessments. Because of the ability to provide educators diagnostic feedback from students' assessment results, CDM has been the focus of much research in the last decade. Various types of diagnostic classification models (DCMs), such as the deterministic inputs, noisy and gate (DINA; Junker and Sijtsma, 2001), the reparametrized unified model/fusion model (RUM; Hartz, 2002), and the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), are designed based on different cognitive theories or assumptions about the relationship between a student's response pattern and attribute profile.

A principal research question of the previous research studies in CDM is which model better describes the data. When analysing a particular assessment dataset, selecting inappropriate DCMs (model misspecification) impacts the classification accuracy and parameter estimation. For example, when the attributes measured by an assessment are non-compensatory, which indicates that non-mastery on one attribute cannot be compensated by mastery on another attribute, selecting a compensatory model will decrease the performance of classification and measurement. DINA model and DINO (Templin and Henson, 2006) model achieved worse fit than did the other more relaxed DCMs, such as G-DINA (DeCarlo, 2011), LCDM, and RUM because both DINA and DINO might be too restrictive to reflect actual students' knowledge status (Yamaguchi and Okada, 2018). Some recent research studies (Chiu and Köhn, 2019; Yamaguchi and Okada, 2020; Zhan, 2020) started to apply the non-compensatory or conjunctive DCM, DINA model, and the compensatory or disjunctive DCM, DINO model, to build up a more general item response function (IRF) for CDM. However, these methods still require pre-data analysis procedure and assumptions of IRF to determine the hyperparameters contained in the mixture (or hybrid) CDM.

A Q-matrix indicates the relationship between items and attributes in an assessment. Q-matrices are often carefully designed by assessment experts, whereas some existing research and their experimental results have shown that Q-matrices constructed by content experts do not always reflect the relationship precisely and may require empirically-driven modifications (Bradshaw et al., 2014; Tjoe and de la Torre, 2014). In CDM, the diagnostic quality of an item indicates the discriminating power of the item to determine the success of the diagnosis. The item with high discriminating refers to that students who have mastered the attributes required by the item are expected to have a high probability of responding to the item correctly, while students who have not are expected to have a low probability. Items with low discriminating power compromise the accuracy of the estimate of student attribute mastery. In the previous research studies, the performances of all DCMs are sensitive to either the diagnostic quality of items or the accuracy of Q-matrices (Kunina-Habenicht et al., 2012; Liu et al., 2017).

Because of the increase of data size and development of computational power, artificial neural networks (ANNs; Goodfellow et al., 2016) have been proposed as an attractive approach to convert a pattern of item responses into a diagnostic classification (Cui et al., 2016; Guo et al., 2017; Paulsen, 2019; Xue, 2019). An ANN is a computational system inspired by biological neural systems for information processing in animals' brains. An ANN is built on inputs being translated to outputs through a series of neuron layers. It consists of three types of layers: an input layer, hidden layer(s), and an output layer. Each layer consists of a number of neurons (or nodes), and each node is connected to the nodes in the next layer. Each layer (except for the input layer) uses the output of its previous layer as the input. Supervised learning ANNs were applied in some research studies (Cui et al., 2016; Guo et al., 2017; Paulsen, 2019). To train the supervised learning ANNs, the ideal response patterns were set as the input layer and the associated attribute profiles as the output layer. Cui et al. (2016) hypothesized DINA model with both slipping and guessing equalling to 0 to synthesize ideal responses to train a multilayer perceptron (MLP). The experimental results showed that the classification accuracy of the supervised learning ANNs was not appreciated even in the simulated study. Another disadvantage of applying supervised learning ANNs for CDM is how to create the ideal response patterns using a DCM because both DCM and parameters are difficult to hypothesize. In addition to supervised learning ANNs, Cui et al. (2016) used one type of unsupervised learning ANNs, self-organizing map (SOM), to classify test-takers into different latent groups for CDM. One disadvantage of the unsupervised learning ANNs is that some further data analysis approaches are required to label the clusters. For example, although cluster analysis can place test-takers into different latent groups, *post hoc* techniques are required to discern the attributes from these latent groups. To do cluster labeling, Xue (2018) proposed a modified autoencoder network with a sparsely connected decoder explained the code layer outputs by using a part of the Q-matrix information. However, in both research studies, the unsupervised learning ANNs cannot yield comparable classification results compared with the DCMs, especially when the diagnostic quality of the assessment was not high. In addition, the ANNs methods produced very unstable and unappreciated estimation unless a great deal of care was taken to conduct sensitivity analyses (Briggs and Circi, 2017).

Regarding the disadvantages in supervised leaning ANNs and unsupervised learning ANNs, in this research, semi-supervised learning thinking is introduced to provide reasonable labels for ANN training and provide accurate and robust classification under different test conditions. In the machine learning field, semi-supervised learning (Zhu, 2005) concerns the study of how computers and natural systems learn in the presence of both labeled and unlabeled data, and it is somewhere between supervised learning and unsupervised learning. The research goal of semi-supervised learning is to understand how combining labeled and unlabeled data change the learning behavior, and design algorithms that take advantage of such a combination. Semi-supervised learning is a great interest in a wide range of applications, such as image search (Fergus et al., 2009), natural language parsing (Liang, 2005), and speech analysis (Liu and Kirchhoff, 2014) because the labeled data is scarce or expensive.

In this research, we firstly applied the semi-supervised learning thinking into the ANNs-based CDM. Unlike the hybrid CDM research studies, which used DINA and DINO models in a mixture CDM, in this research, DINA and DINO models were contained in a semi-supervised learning framework to improve the accuracy and consistency of the ANN's classification. In the following sections, we will first briefly introduce the Co-Training method, which is the semi-supervised learning method we used in this framework. Then, we will describe the structure of the ANNs will. Additionally, we will illustrate the experimental results under simulated experiments to compare the proposed method and five different DCMs. Lastly, we will outline the

benefits and challenges of this methodology are summarized and future research.

2. METHOD

2.1. Co-training Methods of Using DINA Model and DINO Model

As one typical semi-supervised learning method, Co-Training (Nigam and Ghani, 2000) methods use a pair of classifiers with separate views of the data to iteratively learn and generate additional training labels. Like the self-training scheme, Co-Training is a wrapper method and widely applicable to many tasks. Co-Training bears a strong resemblance to the self-training scheme because each classifier uses its most confident predictions on unlabeled instances to teach itself. Two classifiers operate on different views of one observation, and the success of CoTraining depends on the following two assumptions (Zhu and Goldberg, 2009): (1) each view alone is sufficient to make good classifications, given enough labeled data; (2) the two views are conditionally independent given the class label.

Inspired by the typical Co-Training method, in this research, we chose the DINA model and DINO model as two classifiers to operate on different views of one response pattern to an item. The DINA model is a non-compensatory, or conjunctive DCM means that a lack of one attribute cannot be compensated by the mastery of another attribute measured by an item. For each item, the DINA model classifies students into two groups: those who have mastered all the attributes required by the item and those who have not. The jth item response probability of the ith student can be written as:

$$P(y_{ij} = 1|\xi_{ij}, s_j, g_j) = (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}} \quad (1)$$

where $\xi_{ij} = 1$ indicates the ith student has mastered all required attributes of jth item, and $\xi_{ij} = 0$ refers to non-mastery status; s_j and g_j are the slipping parameter and guessing parameter of the jth item.

In contrast to the DINA model, the DINO model is a compensatory or disjunctive DCM, which means that a non-mastery on one latent attribute can be compensated for by a mastery status on another attribute. The jth item response probability of the ith student can be written as:

$$P(y_{ij} = 1|\omega_{ij}, s_j, g_j) = (1 - s_j)^{\omega_{ij}} g_j^{1-\omega_{ij}} \quad (2)$$

where the latent response $\omega_{ij} = 1$ indicates that the ith student has mastered at least one attribute measured by jth item, and $\omega_{ij} = 0$ indicates the absence of all required attributes. Like DINA, s_j and g_j are the slipping parameter and guessing parameter of the jth item.

The reason for selecting the DINA model and the DINO model is to hold the two assumptions of successfully applying Co-Training. First, in an assessment, either the DINA model or the DINO model can be the correct model for different items. For example, both the DINA and DINO models are the correct models for a simple structure item, which only measures a single attribute. Thus, using either the DINA model or the DINO model is sufficient to make accurate classification results. Second, the DINA model and DINO model's item response functions are represented based on different assumptions on the relationship between response patterns and attribute profiles. When the true latent class labels of students are known, for one item, the students can be divided into two groups, DINA-type and DINO-type, respectively. Considering the local independence (Wang and Douglas, 2015), test-takers' item responses from these two groups are statistically independent conditional on the true latent class labels.

In this paper, given the response data and Q-matrix, the DINA model and the DINO model were fitted. For an individual test-taker, we use two labels c_{DINA} and c_{DINO}. c_{DINA} was the estimated latent class under the assumption of using the DINA model, and c_{DINO} was the estimated latent class under the assumption of using the DINO model. c_{DINA} and c_{DINO} could be either the same or different. In this research, the One-Hot encoding method (Harris and Harris, 2015) was applied to the integer encoding c_{DINA} and c_{DINO} to create two new One-Hot representation vectors $\boldsymbol{c}_{DINA} = \{c^k_{DINA}\}$ and $\boldsymbol{c}_{DINA} = \{c^{k'}_{DINA}\}$. c^k_{DINA} and $c^{k'}_{DINA} \in \{0, 1\}$, and $\sum_k c^k_{DINA} = \sum_{k'} c^{k'}_{DINA} = 1$. In machine learning, a One-Hot is a group of bits among which the legal combinations of values are only those with a single 1 bit and all the others 0 bits. For example, if there are 4 latent classes, the integer encoding labels 1, 2, 3, and 4 are converted to One-Hot encoding [0001], [0010], [0100], and [1000], respectively.

2.2. Semi-supervised Learning ANN for Diagnostic Classification

As shown in **Figure 1**, the proposed semi-supervised learning ANN consisted of four parts: the input layer, two hidden layers, class layer, and the output layer. The number of nodes (the circles in **Figure 1**) on the input layer was equal to the number of items contained in the assessment. The number of nodes on the class layer was equal to the number of latent classes. To establish the relationship between the input and class nodes, we used two hidden layers (i.e., hidden layer 1 and hidden layer 2) to convert observed response patterns to latent classes. The numbers of nodes at these two hidden layers are 200 and 100. We use the Rectified linear unit (ReLU; Goodfellow et al., 2016) as the activation function for these two hidden layers and softmax function as the activation function for the class layer. In deep learning field, ReLU is widely used because the mathematical form of ReLU is very simple and efficient, and RelU can avoid a small derivative causing vanishing gradient problem. Softmax function is used for a multi-classification problem in ANNs. Since the number of nodes at the two hidden layers could be viewed as a hyperparameters of ANNs, we selected the two numbers (i.e., 200 and 100) for three reasons: (1) deep learning provides information-theoretically optimal approximation of a very wide range of functions and function classes used in mathematical signal processing (Grohs et al., 2019); (2) Lu et al. (2017) showed a universal approximation theorem for width-bounded ReLU networks: width-$(d + 4)$ ReLU networks, where d is the input dimension, are universal approximators; (3) based on the validation test in our previous research studies using ANNs for

psychometrics (Xue, 2018, 2019; Xue et al., 2020), these two values could achieve a balance between efficiency and accuracy.

In the supervised learning ANNs in CDM, only a single label was used for each observation. For example, when only using DINA classification as labels, the supervised learning ANN was used to train the standard softmax regression or a sigmoid regression (Pang et al., 2020) inputs to outputs without taking into account incorrect labels.The incorrect labels will impact the prediction performance of the ANNs for supervised learning ANNs. In contrast, the output layer in our proposed semi-supervised learning ANN consisted of three parts. The first part (output 1 or $\boldsymbol{Y}_1$) corresponded to the DINA model classification, the second part (output 2 or $\boldsymbol{Y}_2$) corresponded to the DINO model classification, and the third part was the reconstructed response pattern ($\tilde{\boldsymbol{X}}$). The total number of output nodes was equal to two times of the number of hidden classes plus the number of items. For example, given an assessment with 30 items that measured a total of 4 attributes, the input layer $\boldsymbol{X}$ consisted of 30 input nodes (30 items), the class layer $\boldsymbol{t}$ consisted of 16 nodes ($2^4 = 16$ latent classes), and the output layer $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \tilde{\boldsymbol{X}}\}$ consisted of 62 nodes ($16 + 16 + 30$).

Let $\boldsymbol{X} \in \{0,1\}^I$ be the response patterns (I is the number of items), $\boldsymbol{c}_{DINA}$ and $\boldsymbol{c}_{DINO}$ be the One-Hot encoding of the DINA class labels and DINO class labels, respectively. Then we introduced into our ANN model the "true" latent class label (as opposed to the DINA and DINO class labels) as a latent multinomial variable $\boldsymbol{t} \in \{0,1\}^C$, $\sum_j^C t_j = 1$, where C is the number of latent classes. Like $\boldsymbol{c}_{DINA}$ and $\boldsymbol{c}_{DINO}$, $\boldsymbol{t}$ was also a One-Hot encoding label for each response pattern. The output of the class layers (or the input of the output layer) of our ANN was the posterior over $\boldsymbol{t}$ using the softmax regression. The ith element of $\boldsymbol{t}$ can be represented as:

$$P(t_j = 1|\boldsymbol{X}) = \frac{\tilde{P}(t_j = 1|\boldsymbol{X})}{\sum_{j'=1}^{C} \tilde{P}(t_{j'} = 1|\boldsymbol{X})} = \frac{\phi_j(\boldsymbol{X}; \boldsymbol{w}_1)}{\sum_{j'=1}^{C} \phi_{j'}(\boldsymbol{X}; \boldsymbol{w}_1)} \tag{3}$$

where $\tilde{P}$ denotes the unnormalized probability distribution, $\Phi = \{\phi_j(\boldsymbol{X}; \boldsymbol{w}_1)\}$, $j \in \{1, \ldots, C\}$ indicates the calculation from the input layer to class layer's output, and $\phi_j(\boldsymbol{X}; \boldsymbol{w}_1)$ indicates the jth node's values on the class layer, the computation of $\phi_j(\boldsymbol{X}; \boldsymbol{w}_1)$ is as follows:

$$\phi_j(\boldsymbol{X}; \boldsymbol{w}_1) = \sigma\Big(\sum_{m=1}^{100} \sum_{n=1}^{200} w_{mj}^{H_2 \boldsymbol{t}} \max(w_{nm}^{H_1 H_2} \max(w_{in}^{XH_1} X_i, 0), 0)\Big) \tag{4}$$

where $\sigma(\cdot)$ is the softmax function, $\max(\cdot, 0)$ is the ReLU function, $\boldsymbol{w}_1 = \{\{w_{in}^{XH_1}\}, \{w_{nm}^{H_1 H_2}\}, \{w_{mj}^{H_2 \boldsymbol{t}}\}\}$ indicate the all the weights of the ANNs from the input layer to the class layer. $\{w_{in}^{XH_1}\}$ is the weights between input layer X and first hidden layer H_1; $\{w_{nm}^{H_1 H_2}\}$ is the weights between the first hidden layer H_1 and second hidden layer H_2; $\{w_{mj}^{H_2 \boldsymbol{t}}\}$ is the weights between the second hidden layer H_2 and the class layer. $\boldsymbol{w}_1$ needs to be estimated in the training of ANNs. Given the true label $\boldsymbol{t}$, the output 1 (DINA model classification) and output 2 (DINO model classification) can be modeled using another softmax with logits as follows:

$$\begin{aligned} logit(P(\boldsymbol{Y}_1|\boldsymbol{X})) &= logit(P(c_{DINA}^k = 1|\boldsymbol{X})) = \sum_{j=1}^{C} w_{jk}^{t Y_1} t_j \\ logit(P(\boldsymbol{Y}_2|\boldsymbol{X})) &= logit(P(c_{DINO}^{k'} = 1|\boldsymbol{X})) = \sum_{j=1}^{C} w_{jk'}^{t Y_2} t_j \end{aligned} \tag{5}$$

where the weights $w_{jk}^{tY_1}$ and $w_{jk'}^{tY_2}$ learn the log-probability of the "true" label j as DINA class label k (the kth class in DINA classification) and as DINO class label k' (the k'th class in DINA classification), respectively. Thus, in the proposed ANN, the joint relationship between input layer x and the kth node of $\boldsymbol{Y}_1$ and k'th node of $\boldsymbol{Y}_2$ can be represented as follows:

$$\begin{aligned} P(\boldsymbol{Y}_1, \boldsymbol{Y}_2|\boldsymbol{X}) &= P(c_{DINA}^k = 1, c_{DINO}^{k'} = 1|\boldsymbol{X}) \\ &= \sum_{j=1}^{C} P(c_{DINA}^k = 1, c_{DINO}^{k'} = 1, t_j = 1|\boldsymbol{X}) \\ &= \sum_{j=1}^{C} P(c_{DINA}^k = 1|t_j = 1) \\ &\quad P(c_{DINO}^{k'} = 1|t_j = 1) P(t_j = 1|\boldsymbol{X}) \end{aligned} \tag{6}$$

where $P(t_j = 1|\boldsymbol{X})$, $P(c_{DINO}^{k'} = 1|t_j = 1)$, $P(c_{DINA}^k = 1|t_j = 1)$ are defined in Equations (3) and (5).

In addition to the difference between Co-Training labels and $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, we also added a regularization term, $H(\boldsymbol{X}, \tilde{\boldsymbol{X}})$, to encourage the classification to be perceptually consistent. $\boldsymbol{X}$ is the observed response pattern, and $\tilde{\boldsymbol{X}}$ is the reconstructed response pattern corresponding to the estimated latent class. The $\tilde{\boldsymbol{X}}$ can be calculated from the true label $\boldsymbol{t}$ as:

$$\tilde{X}_i = \sum_{j=1}^{C} w_{ji}^{t\tilde{X}} t_j \tag{7}$$

where $w_{ij}^{t\tilde{X}}$ is the connection weights between jth class layer node and ith reconstructed output node. The weights between class layer and output layer of the ANNs is $\boldsymbol{w_2} = \{\{w_{jk}\}, \{w_{jk'}\}, \{w_{ji}\}\}$. We could perform training via stochastic gradient descent (SGD; Bottou and Bousquet, 2007) to minimize the following cost function:

$$\{\boldsymbol{w}\} = \arg\min\{H(\boldsymbol{Y}_1, \boldsymbol{c}_{DINA}) + H(\boldsymbol{Y}_2, \boldsymbol{c}_{DINO}) + \lambda H(\boldsymbol{X}, \tilde{\boldsymbol{X}})\} \tag{8}$$

where $\{\boldsymbol{w}\} = \{\boldsymbol{w_1}, \boldsymbol{w_2}\}$ indicates all the weights of the ANNs to be estimated, $H(\cdot, \cdot)$ is the cross-entropy to calculate the difference between $\boldsymbol{Y}_1$ and One-Hot DINA labels $\boldsymbol{c}_{DINA}$, and the difference between $\boldsymbol{Y}_2$ and One-Hot DINO labels $\boldsymbol{c}_{DINO}$, and the difference between observed response pattern $\boldsymbol{X}$ and reconstructed response pattern $\tilde{\boldsymbol{X}}$. λ is a scaling parameter which was determined through a validation test (Xue et al., 2020).

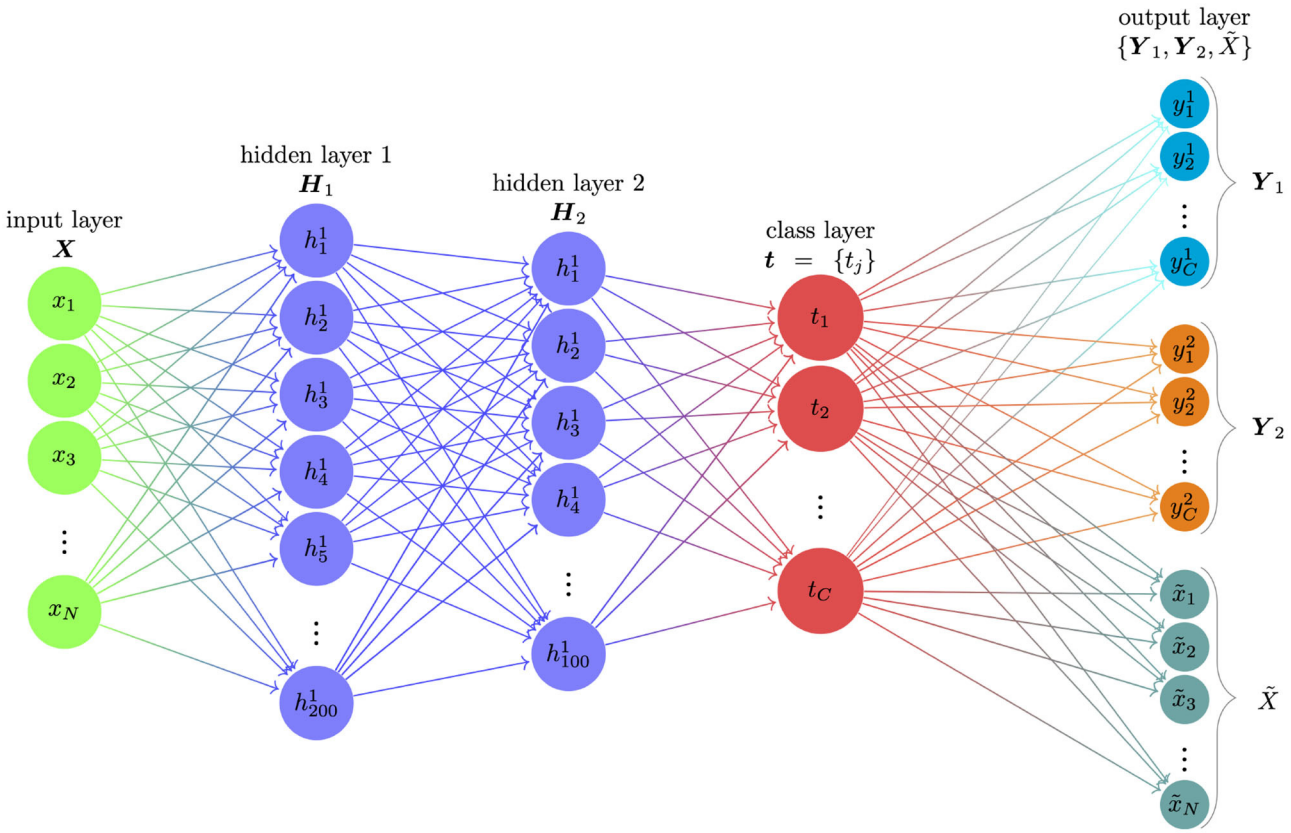


FIGURE 1 | The structure of the proposed semi-supervised learning ANN. The proposed semi-supervised learning ANN consisted of one input layer, two hidden layers, one class layer and one output layer. The input $\boldsymbol{X}$ is response pattern of a test-taker; $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are two hidden layers; $\boldsymbol{t}$ indicates the class layer; and the output layer consists of three parts: $\boldsymbol{Y}_1$ is the classification under DINA assumption; $\boldsymbol{Y}_2$ is the classification under DINO assumption; $\tilde{\boldsymbol{X}}$ is the reconstructed response pattern of the test-taker.

Because of the large number of parameters contained in the deep learning structure, the random initialization of parameters may impact the optimization when the training sample size is not large enough. Thus, one concern of using ANNs for CDM is that using the feature extracted by deep learning through a single training is risky or sensitive to the starting points of the parameters (Briggs and Circi, 2017). Cui et al. (2016) only set a maximum number of iterations (e.g., 10,000) to stop training the supervised learning ANN in their research study. We applied two methods to deal with this issue. The first method was the early stopping, which is a simple, effective, and widely used approach to avoid overtraining the ANNs. The early stopping method is used to train on the training dataset but to stop training at the point when performance on a validation dataset starts to degrade. In addition, through the validating, we determined the scaling parameter in Equation (10). In our method, the whole data set was divided into two parts: the training dataset consisted of 80% observations, and the validating dataset consisted of the rest 20% observations. The second method was that we conducted 100 ANN trainings individually, produced a probability of latent class for each training, and then averaged the 100 probabilities as the final probability of the latent class for each test-taker.

3. EXPERIMENTAL STUDY

The aims of the experiment were (1) to examine the attribute profile estimation and classification accuracy of the proposed method under different test factors which are expected to affect the estimates' accuracy, and (2) to compare the proposed method with the performance of five DCMs: the DINA, DINO, G-DINA (De La Torre, 2011), LCDM (Henson et al., 2009), and RUM (Hartz, 2002). Thus, we conducted a simulation study under different assessment conditions with a variety of fixed factors and four manipulated factors.

3.1. Method

3.1.1. Manipulated Factors

Using item by latent class matrix, we manipulated three assessment factors in the data generation for the simulation, including the number of items (20 or 30), number of attributes (three or four), and test diagnostic quality (high or mixed). When estimating the conditions, we also manipulated the Q-matrix accuracy (100 and 90% correct) as another factor expected to impact classification accuracy.

3.1.1.1. Test Length and Number of Attributes

The number of items (20 or 30) and the number of attributes were selected to reflect the current real assessment applications, which often contained between 20 to 30 items and measured three or four attributes [e.g., MELAB data (Li and Suen, 2013); DTMR data (Bradshaw et al., 2014)]. For three attributes, we generated 20 items, and for four attributes, 20 and 30 items were generated, respectively. The three Q-matrices (i.e., 20 items measured 3 attributes, 20 items measured 4 attributes, and 30 items measured 4 attributes) for these conditions are shown in **Supplementary Tables 1**–**3**, respectively.

3.1.1.2. Test Diagnostic Quality

Item discriminating power is another factor impact performance of DCMs shown in previous research studies (e.g., Cui et al. 2016; Roussos et al. 2005). The item discriminating power d_i is calculate as $d_i = p(x = 1|\alpha_1) - p(x = 1|\alpha_0)$. α_0 is the attribute pattern where none of the attributes measured by the ith item are mastered, and α_1 is the attribute pattern where all attributes measured by the ith item are mastered. If $d_i > 0.3$, the Item i is a highly discriminating item, and if $0 < d_i \leq 0.3$, the Item i is a lowly discriminating item. In the assessments with high diagnostic quality, all items are of high discriminating power; in the assessments with mixed diagnostic quality, 50% items are of high discriminating power, and 50% items are of low discriminating power.

3.1.1.3. Accuracy of Q-Matrix

Since the Q-matrices constructed by content experts do not always reflect the relationship precisely and may require empirically-driven modifications (Bradshaw et al., 2014; Tjoe and de la Torre, 2014), two levels of Q-matrix accuracy were also created for DCMs model fitting and Co-Training methods: 100% accuracy indicated that the Q-matrix were completely known; 90% accuracy indicated that 10% of elements in each Q-matrix were incorrect. We mis-specified the 10% elements in Q-matrix randomly drawing a Q-matrix entries and changing its value, with the constraint that each item must measure at least one attribute (i.e., a randomly drawn value of "1" for a simple structure item could not be changed to "0"). Such constrain makes there is no all zero q-vector (e.g., [0, 0, 0], [0, 0, 0, 0]) in Q-matrix.

3.1.2. Generating Item Response Probabilities

Sample sizes of 1,000 were used for all conditions. The true class probabilities of correct response for the items in the item pools were simulated using the logic of a DCM with respect to the Q-matrix defining the item-class relationships and the probabilities following monotonicity constraints across non-equivalence classes on an item (i.e., masters of all attributes measured by the item having a higher probability of correct response than masters of a proper subset of these attributes; masters of no attributes measured by the item having a lower probability of correct response than masters of a proper subset of these attributes), but did not follow a particular existing DCM item response function (e.g., the LCDM or DINA function). Current DCM item response functions constrain the item response probabilities to be equal within all equivalence classes; our simulated data did not. Item-based equivalence classes are latent classes that have the same attribute profile, or the same pattern of mastery, for all attributes that are measured by the item. Conversely, item-based non-equivalence classes differ on the mastery status of one or more attributes measured by the item.

We simulated data using a general $I \times C$ item by latent class matrix (Xu and Zhang, 2016) according to DCM logic (i.e., defining latent classes by attribute profiles and specifying item-latent class relationships by the Q-matrix) without the specific mathematic representation of the item response function:

$$\Pi = \begin{bmatrix} \pi_{1,1} & \pi_{1,2} & \dots & \pi_{1,C} \\ \pi_{2,1} & \pi_{2,2} & \dots & \pi_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{I,1} & \pi_{I,2} & \dots & \pi_{I,C} \end{bmatrix}. \tag{9}$$

where the conditional probability that students in lth latent class answer ith item correctly $P(x_i = 1|c) = \pi_{i,c}$, which is also known as item response probability (IRP) for each class. I indicated the number of items, C indicated the number of latent classes.

We denote π_{i,α_0}, π_{i,α_1}, and π_{i,α_p} as the IRPs for non-mastery group, mastery group, and partial mastery group, respectively. The mastery group contained students who mastered all of the attributes required by ith item, the partial mastery group contains students who only mastered a proper subset of attributes required by ith item, and the non-mastery group contained students who mastered none of the attributes required by ith item.

As shown in **Table 1**, when simulating response patterns to high discrimination items for the mastery group π_{i,α_1} were drawn from a uniform distribution $U[0.65, 0.9]$; for the non-mastery group π_{i,α_0} were drawn from a uniform distribution $U[0.15, 0.35]$; and for the partial mastery group π_{i,α_p} were drawn from a uniform distribution $U[0.4, 0.6]$. These draws yielded an average item discrimination value of 0.530 in 3 highly discriminating assessments. When simulating response patterns to low discrimination items, for the non-mastery group π_{i,α_0} were drawn from a uniform distribution $U[0.2, 0.4]$; for partial mastery group π_{i,α_p} were drawn from a uniform distribution $U[\pi_{i,\alpha_0}, \pi_{i,\alpha_0} + 0.2]$; lastly for the mastery group (students who mastered all the attributes required by ith item) π_{i,α_1} were based on a uniform distribution $U[\pi_{i,\alpha_p}, \pi_{i,\alpha_0} + 0.3]$ for complex items and $U[\pi_{i,\alpha_0}, \pi_{i,\alpha_0} + 0.3]$ for simple items. This yielded an average item discrimination value of 0.387 in three mixed discriminating assessments.

By drawing true item parameters in this way, the $\pi_{i,c}$s in our simulated data differs from IRPs simulated from the LCDM in that partial mastery classes with the same attribute pattern with respect to the measured attributes on a given item (the partial mastery item-based equivalence classes) have different true item response probabilities. The item response probabilities for these classes are, however, drawn from the same uniform distribution, so while they may be different values, they will be in the same range. Taking Item 10 that measures Attribute 1 and Attribute 2 as an example (as shown in **Supplementary Table 4**), Classes C2, C3, C6, and C7 are all partial mastery classes with respect to this

TABLE 1 | The table of selecting $\pi_{i,c}$ for item by class matrix.

Latent groups	High discrimination	Low discrimination
Non-mastery π_{i,a_0}	**U**[.15, .35]	**U**[.20, .40]
Partial-mastery π_{i,α_p}	**U**[.40, .60]	**U**[$\pi_{i,a_0}, \pi_{i,a_0} + .15$]
Mastery π_{i,a_1}	**U**[.65, .90]	$U[\pi_{i,\alpha_p}, \pi_{i,a_0} + 0.30]$, Complex items $U[\pi_{i,a_0}, \pi_{i,a_0} + 0.30]$, Simple items

For each item, π_{i,a_0}, π_{i,a_1}, and π_{i,α_p} indicate the $\pi_{i,c}$ for non-mastery group, mastery group, and partial mastery group, respectively.

item: Class C2 and C6 both have mastered Attribute 1 but not Attribute 2, and Class C3 and C7 has both mastered Attribute 2 and not Attribute 1. Under the LCDM, Class C2 and C6 would have the same IRP, while Class C3 and C7 would have the same IRP; under our generating model, the IRP for all four classes were drawn from the same interval, but the draws were different, resulting in, Class C2 having an IRP of 0.509, Class C3 having an IRP of 0.519, Class C6 having an IRP of 0.458, and Class C7 having an IRP of 0.429 (see **Supplementary Table 4**). For non-mastery equivalence classes and mastery equivalence classes, the true model did constrain draws to be equal within the interval (i.e., Class C1 and C5 have IRP values of 0.33 and Class C4 and C8 have IRP values of 0.891). Only for partial mastery item-based equivalence classes were they allowed to differ. The purpose of allowing this difference was to add some noise in the data while still controlling the item discrimination level (IRP of mastery group minus IRP of non-mastery group).

The values in the item by latent class matrix Π for the 6 item pools are shown in **Supplementary Tables 4–9**, respectively. These appendices showed that the DCMs primary monotonicity assumptions held. Namely, the mastery group has the greatest IRP, the non-mastery group has the lowest IRP, and the IRP of partial mastery groups lie between them. These appendices show this simulation procedure firstly held that $0.3 \leq d_i < 0.75$ for high discrimination items and $0.3 < d_i < 0.75$ for low discrimination items; it also again shows the DCM monotonicity assumptions that the mastery group has a greater IRP than the non-mastery group held.

3.1.3. Estimation

In our simulated study, as a comparison, five types of widely used DCMs were introduced as baselines to evaluate the diagnostic classification performance of the proposed framework. DINA and DINO models were selected as two baselines because they were the two classifiers used for Co-Training method. In addition, we chose three more general models, the G-DINA with identity link function (De La Torre, 2011), the LCDM with the logit link function (Henson et al., 2009), and the RUM (Hartz, 2002).

Results were analyzed in terms of classification accuracy of the five DCMs and proposed method under 12 different test conditions. Since in the proposed method, a validation test was introduced for early stop in the training procedure to avoid overtraining, the whole data set was divided to two parts: training dataset which contains 80% observations; and validating dataset which contains 20% observations. In the results shown in **Tables 2–4**, we list three types of the results of using the proposed ANN method:

1. ANN: the classification results of applying the trained ANN structure to the whole dataset containing training set and validation set;
2. ANN*: the classification results of applying the trained ANN structure to the training dataset;
3. ANN**: the classification results of applying the trained ANN structure to the validating dataset.

The data simulation and five DCMs were conducted using the "CDM" package (George et al., 2016) in R. The proposed semi-supervised learning ANN was conducted using the "tensorflow" library (Pang et al., 2020) in Python. In the experimental study, we conducted 100 replications. In each replication, new response patterns were created based on the fixed values in the item by latent class matrices in **Supplementary Tables 4–9**.

3.2. Results

First, we tested the effects of the four assessment factors of test length, number of attributes, test diagnostic quality, and Q-matrix accuracy on the attribute profile and classification accuracy for the proposed method. Then we compared the proposed method to the five DCMs, under 12 different test conditions. Results are given in **Tables 2–4**.

3.2.1. Classification Accuracy and Four Assessment Factors

We first focus on results for the proposed method. As mentioned in the Estimation session, ANN, ANN* and ANN** in **Tables 2–4** indicate the classification accuracy on whole dataset (including training set and validating set), the training set and validating set, respectively. Results show that the proposed method (ANN) works reasonably well and has classification accuracy values >70% under 6 out of 12 assessment conditions (condition 1, 2, 3, 4, 9, and 10) when applying the trained ANN to the whole data set (i.e., ANN). Condition 1–4 are all four test conditions for the assessment measures 3 attributes using 20 items with either highly diagnostic quality or mixed diagnostic quality. Condition 9 and 10 are the two test conditions for assessment measures 4 attributes using 30 items with highly diagnostic quality. Results show classification accuracy increased in expected ways for the proposed method. Namely, average classification accuracy increases from 0.670 to 0.722 as test length increases from 20 to 30 for assessments measure 4 attributes (there is only one test length of assessment that measures 3 attributes); when the number of attribute measured decreases from 4 to 3 in assessment with 20 items, the average classification accuracy increases from 0.670 to 0.834; when the test diagnostic quality increases from mixed to high, the average classification accuracy increases from 0.621 to 0.736; and when the accuracy of Q-matrix increases from 90 to 100%, the average accuracy increases slightly from 0.675 to 0.682. In addition, we can see that ANN* always achieves the best performance with average classification accuracy 0.692, ANN** always achieves the worst performance with average

TABLE 2 | Comparison of classification rates for three attributes using 20 items.

Test condition	Methods	Quality	Q-matrix accuracy	Attribute 1	Attribute 2	Attribute 3	Class
1	DINA	High	100%	0.949 (0.00)	0.864 (0.02)	0.957 (0.01)	0.778 (0.02)
	DINO			0.953 (0.01)	0.871 (0.02)	0.952 (0.02)	0.784 (0.04)
	LCDM			0.96 (0.00)	0.917 (0.00)	0.957 (0.00)	0.842 (0.01)
	G-DINA			0.96 (0.00)	0.917 (0.00)	0.957 (0.01)	0.842 (0.00)
	RUM			0.953 (0.01)	0.91 (0.00)	0.958 (0.00)	0.827 (0.00)
	ANN			0.956 (0.01)	0.915 (0.01)	0.957 (0.01)	0.834 (0.02)
	ANN*			0.962 (0.01)	0.921 (0.01)	0.964 (0.02)	0.851 (0.02)
	ANN**			0.945 (0.01)	0.901 (0.02)	0.942 (0.01)	0.818 (0.03)
2	DINA		90%	0.944 (0.00)	0.824 (0.01)	0.957 (0.00)	0.741 (0.02)
	DINO			0.946 (0.01)	0.852 (0.01)	0.944 (0.01)	0.757 (0.02)
	LCDM			0.956 (0.00)	0.897 (0.00)	0.958 (0.00)	0.819 (0.00)
	G-DINA			0.956 (0.00)	0.897 (0.00)	0.958 (0.01)	0.819 (0.00)
	RUM			0.949 (0.00)	0.879 (0.01)	0.958 (0.00)	0.794 (0.01)
	ANN			0.955 (0.01)	0.900 (0.02)	0.958 (0.02)	0.821 (0.02)
	ANN*			0.962 (0.01)	0.910 (0.02)	0.969 (0.02)	0.831 (0.03)
	ANN**			0.945 (0.01)	0.881 (0.02)	0.932 (0.04)	0.807 (0.04)
3	DINA	Mixed	100%	0.875 (0.00)	0.859 (0.01)	0.914 (0.00)	0.693 (0.01)
	DINO			0.863 (0.01)	0.864 (0.00)	0.896 (0.01)	0.665 (0.01)
	LCDM			0.879 (0.01)	0.884 (0.00)	0.913 (0.00)	0.712 (0.01)
	G-DINA			0.879 (0.00)	0.884 (0.00)	0.913 (0.00)	0.712 (0.00)
	RUM			0.873 (0.01)	0.9 (0.00)	0.917 (0.01)	0.724 (0.00)
	ANN			0.883 (0.01)	0.884 (0.02)	0.915 (0.01)	0.720 (0.01)
	ANN*			0.892 (0.01)	0.896 (0.01)	0.929 (0.02)	0.730 (0.02)
	ANN**			0.868 (0.01)	0.878 (0.02)	0.911 (0.01)	0.704 (0.02)
4	DINA		90%	0.878 (0.01)	0.85 (0.01)	0.906 (0.00)	0.676 (0.02)
	DINO			0.869 (0.00)	0.861 (0.00)	0.908 (0.00)	0.679 (0.01)
	LCDM			0.878 (0.00)	0.85 (0.00)	0.918 (0.00)	0.685 (0.01)
	G-DINA			0.877 (0.00)	0.85 (0.01)	0.918 (0.00)	0.684 (0.00)
	RUM			0.877 (0.00)	0.85 (0.01)	0.915 (0.00)	0.685 (0.01)
	ANN			0.874 (0.01)	0.888 (0.02)	0.908 (0.02)	0.704 (0.02)
	ANN*			0.889 (0.01)	0.901 (0.01)	0.923 (0.01)	0.719 (0.01)
	ANN**			0.867 (0.04)	0.871 (0.04)	0.890 (0.03)	0.683 (0.03)

ANN indicate the attribute profile estimation using the proposed method on whole data set. ANN indicate the attribute profile estimation using the proposed method on the training data set. ANN** indicate the attribute profile estimation using the proposed method on the validation data set.*

classification accuracy 0.661, and ANN falls between ANN* and ANN** with average classification accuracy 0.678. The reason is that the parameters of ANN structure were trained based on the training dataset but not considered the validation dataset.

Next, we examine the results for the five DCMs. Results show that DINA model has classification accuracy values >70% under 2 out of 12 assessment conditions (condition 1 and 2); DINO model has classification accuracy values >70% under 2 out of 12 assessment conditions (condition 1 and 2); G-DINA has classification accuracy values >70% under 5 out of 12 test conditions (condition 1, 2, 3, 9, and 10); LCDM has classification accuracy values >70% under 5 out of 12 test conditions (condition 1, 2, 3, 9, and 10); and RUM has classification accuracy values >70% under 5 out of 12 test conditions (condition 1, 2, 3, 9, and 10). Condition 1 and 2 are two tests (high and mixed diagnostic quality) with 20 items measures 3 attributes and the Q-matrix accuracy is 100%; condition 3 is a test with high diagnostic quality consists of 20 items to measure 3 attribute but the Q-matrix accuracy is 90%; condition 9 and 10 are two tests (high and mixed diagnostic quality) with 30 items measures 4 attributes and the Q-matrix accuracy is 100%. We could also notice that the G-DINA and LCDM achieved almost the same classification results because the only difference between G-DINA and LCDM in the CDM::gdina() is the link function. We chose "identity" function for G-DINA and "logit" function for LCDM. In addition, like the proposed method, results show classification accuracy increased in expected way for the 5 DCMs. Namely, accuracy increases as test length increases, as the number of attribute measured decreases, as the test diagnostic quality increases, and as the accuracy of Q-matrix increases.

TABLE 3 | Comparison of classification rates for four attributes using 20 items.

Test condition	Methods	Quality	Q-matrix accuracy	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Class
5	DINA	High	100%	0.908 (0.02)	0.924 (0.03)	0.79 (0.02)	0.893 (0.02)	0.591 (0.03)
	DINO			0.909 (0.04)	0.928 (0.05)	0.858 (0.02)	0.899 (0.03)	0.653 (0.04)
	LCDM			0.918 (0.01)	0.929 (0.02)	0.858 (0.00)	0.919 (0.01)	0.67 (0.01)
	G-DINA			0.918 (0.01)	0.929 (0.01)	0.858 (0.01)	0.919 (0.00)	0.67 (0.01)
	RUM			0.923 (0.02)	0.921 (0.01)	0.853 (0.02)	0.917 (0.01)	0.664 (0.03)
	ANN			0.919 (0.01)	0.925 (0.01)	0.858 (0.03)	0.922 (0.04)	0.67 (0.03)
	ANN*			0.931 (0.02)	0.942 (0.01)	0.870 (0.03)	0.941 (0.01)	0.691 (0.03)
	ANN**			0.909 (0.03)	0.918 (0.02)	0.861 (0.01)	0.912 (0.04)	0.655 (0.03)
6	DINA		90%	0.909 (0.04)	0.922 (0.04)	0.74 (0.02)	0.886 (0.02)	0.56 (0.03)
	DINO			0.903 (0.04)	0.924 (0.02)	0.852 (0.03)	0.879 (0.04)	0.621 (0.04)
	LCDM			0.904 (0.01)	0.922 (0.00)	0.824 (0.01)	0.887 (0.01)	0.616 (0.01)
	G-DINA			0.904 (0.01)	0.922 (0.01)	0.824 (0.01)	0.887 (0.02)	0.616 (0.01)
	RUM			0.905 (0.02)	0.922 (0.02)	0.8 (0.02)	0.884 (0.01)	0.599 (0.03)
	ANN			0.912 (0.04)	0.923 (0.01)	0.862 (0.03)	0.89 (0.02)	0.648 (0.03)
	ANN*			0.924 (0.02)	0.931 (0.01)	0.877 (0.01)	0.901 (0.02)	0.657 (0.02)
	ANN**			0.903 (0.01)	0.917 (0.02)	0.853 (0.03)	0.883 (0.02)	0.632 (0.03)
7	DINA	Mixed	100%	0.854 (0.01)	0.836 (0.03)	0.824 (0.02)	0.851 (0.03)	0.503 (0.02)
	DINO			0.863 (0.02)	0.817 (0.04)	0.854 (0.02)	0.816 (0.04)	0.484 (0.04)
	LCDM			0.867 (0.01)	0.823 (0.01)	0.855 (0.02)	0.84 (0.03)	0.509 (0.01)
	G-DINA			0.867 (0.01)	0.824 (0.02)	0.855 (0.04)	0.84 (0.03)	0.51 (0.01)
	RUM			0.878 (0.03)	0.831 (0.02)	0.856 (0.03)	0.837 (0.04)	0.522 (0.03)
	ANN			0.864 (0.04)	0.842 (0.02)	0.857 (0.03)	0.859 (0.02)	0.531 (0.02)
	ANN*			0.879 (0.01)	0.855 (0.03)	0.870 (0.02)	0.871 (0.01)	0.550 (0.02)
	ANN**			0.853 (0.03)	0.839 (0.02)	0.826 (0.02)	0.850 (0.04)	0.504 (0.05)
8	DINA		90%	0.856 (0.04)	0.826 (0.02)	0.744 (0.01)	0.854 (0.01)	0.448 (0.02)
	DINO			0.854 (0.02)	0.817 (0.02)	0.855 (0.01)	0.851 (0.04)	0.503 (0.05)
	LCDM			0.865 (0.00)	0.817 (0.01)	0.776 (0.02)	0.844 (0.01)	0.469 (0.01)
	G-DINA			0.865 (0.02)	0.817 (0.01)	0.776 (0.00)	0.844 (0.01)	0.469 (0.01)
	RUM			0.864 (0.03)	0.821 (0.01)	0.855 (0.04)	0.84 (0.01)	0.509 (0.03)
	ANN			0.852 (0.02)	0.871 (0.02)	0.855 (0.03)	0.852 (0.01)	0.542 (0.04)
	ANN*			0.869 (0.03)	0.883 (0.02)	0.867 (0.01)	0.870 (0.00)	0.558 (0.03)
	ANN**			0.850 (0.04)	0.851 (0.02)	0.855 (0.02)	0.854 (0.03)	0.512 (0.05)

ANN indicate the attribute profile estimation using the proposed method on whole data set. ANN indicate the attribute profile estimation using the proposed method on the training data set. ANN** indicate the attribute profile estimation using the proposed method on the validation data set.*

3.2.2. Comparison Classification With 5 DCMs

Simulation results indicated that when using the proposed ANN, the classification rates were higher than rates from the DINA and DINO models, the two initial classifiers used in Co-Training. Compared to DINA and DINO models, at the attribute level, the average improvements of classification using ANN was 0.0218 and 0.0140, and at the class level (i.e., attribute profiles level), the average improvements were 0.0589 and 0.0432. Compared to the general models LCDM and G-DINA, which often achieved the best performance in classification, the performance of ANN was also better than these two methods. The improvements at the attribute level were 0.0056 and 0.0055 compared with LCDM and G-DINA models, respectively. At the class level, the improvements were 0.0130 and 0.0132.

The simulated study also indicated that when the Q-matrix became less accurate, the classification accuracy for each method dropped at both attribute level and latent class level when holding other test assessment factors. When the Q-matrix accuracy decreased to 90% accurate, at the attribute level, the average reductions of classification accuracy were 0.0071, 0.0055, 0.0114, 0.0114, 0.0095, and 0.0038 corresponding to DINA, DINO, LCDM, G-DINA, RUM, and our ANN methods, respectively. At the attribute pattern level, the average accuracy reductions were 0.0163, 0.0138, 0.0298, 0.0302, 0.0243, and 0.0075 for DINA, DINO, LCDM, G-DINA, RUM and, our ANN methods, respectively. From this observation, we could find that firstly, the relaxed models (LCDM, G-DINA, and RUM) were more sensitive to the accuracy of Q-matrix; secondly, the proposed ANN was more robust to the noise within the Q-matrix compared to the five DCMs.

TABLE 4 | Comparison of classification rates for 4 attributes using 30 items.

Test condition	Methods	Quality	Q-matrix accuracy	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Class
9	DINA	High	100%	0.937 (0.04)	0.938 (0.03)	0.814 (0.06)	0.892 (0.02)	0.641 (0.05)
	DINO			0.942 (0.01)	0.941 (0.08)	0.854 (0.11)	0.902 (0.08)	0.681 (0.12)
	LCDM			0.947 (0.01)	0.949 (0.00)	0.873 (0.02)	0.925 (0.02)	0.732 (0.02)
	G-DINA			0.947 (0.00)	0.949 (0.01)	0.873 (0.02)	0.925 (0.03)	0.732 (0.02)
	RUM			0.948 (0.02)	0.945 (0.00)	0.872 (0.04)	0.917 (0.03)	0.719 (0.05)
	ANN			0.949 (0.01)	0.944 (0.02)	0.872 (0.03)	0.916 (0.02)	0.722 (0.03)
	ANN*			0.955 (0.01)	0.952 (0.01)	0.880 (0.02)	0.935 (0.01)	0.741 (0.02)
	ANN**			0.942 (0.02)	0.940 (0.02)	0.860 (0.04)	0.903 (0.03)	0.711 (0.04)
10	DINA		90%	0.934 (0.03)	0.94 (0.02)	0.853 (0.04)	0.853 (0.03)	0.64 (0.04)
	DINO			0.935 (0.02)	0.924 (0.03)	0.855 (0.08)	0.874 (0.03)	0.644 (0.06)
	LCDM			0.948 (0.00)	0.946 (0.01)	0.858 (0.02)	0.92 (0.03)	0.708 (0.04)
	G-DINA			0.948 (0.02)	0.946 (0.01)	0.859 (0.03)	0.92 (0.03)	0.709 (0.03)
	RUM			0.945 (0.01)	0.945 (0.01)	0.869 (0.02)	0.915 (0.01)	0.713 (0.03)
	ANN			0.952 (0.02)	0.948 (0.01)	0.873 (0.02)	0.916 (0.01)	0.723 (0.02)
	ANN*			0.960 (0.02)	0.954 (0.02)	0.890 (0.01)	0.926 (0.01)	0.733 (0.02)
	ANN**			0.935 (0.03)	0.940 (0.04)	0.860 (0.02)	0.902 (0.04)	0.703 (0.04)
11	DINA	Mixed	100%	0.903 (0.03)	0.876 (0.03)	0.801 (0.01)	0.882 (0.02)	0.56 (0.02)
	DINO			0.911 (0.03)	0.884 (0.05)	0.858 (0.06)	0.858 (0.04)	0.586 (0.07)
	LCDM			0.912 (0.03)	0.886 (0.02)	0.857 (0.02)	0.88 (0.02)	0.616 (0.03)
	G-DINA			0.912 (0.02)	0.886 (0.01)	0.858 (0.02)	0.88 (0.01)	0.617 (0.02)
	RUM			0.9 (0.02)	0.884 (0.01)	0.858 (0.02)	0.871 (0.03)	0.592 (0.03)
	ANN			0.91 (0.01)	0.889 (0.02)	0.862 (.01)	0.881 (0.01)	0.616 (0.02)
	ANN*			0.916 (0.02)	0.898 (0.01)	0.869 (.02)	0.900 (0.01)	0.623 (0.02)
	ANN**			0.905 (0.02)	0.881 (0.03)	0.850 (.02)	0.881 (0.03)	0.605 (0.03)
12	DINA		90%	0.908 (0.03)	0.887 (0.03)	0.847 (0.01)	0.876 (0.03)	0.603 (0.02)
	DINO			0.906 (0.03)	0.883 (0.07)	0.852 (0.08)	0.836 (0.07)	0.566 (0.09)
	LCDM			0.908 (0.02)	0.891 (0.01)	0.863 (0.03)	0.868 (0.01)	0.605 (0.02)
	G-DINA			0.908 (0.01)	0.891 (0.02)	0.863 (0.03)	0.868 (0.01)	0.605 (0.02)
	RUM			0.905 (0.02)	0.891 (0.01)	0.864 (0.03)	0.861 (0.03)	0.602 (0.03)
	ANN			0.909 (0.01)	0.885 (0.02)	0.859 (0.01)	0.871 (0.02)	0.61 (0.02)
	ANN*			0.921 (0.01)	0.903 (0.02)	0.869 (0.01)	0.878 (0.01)	0.624 (0.01)
	ANN**			0.901 (0.03)	0.889 (0.02)	0.850 (0.01)	0.857 (0.03)	0.603 (0.03)

ANN indicate the attribute profile estimation using the proposed method on whole data set. ANN indicate the attribute profile estimation using the proposed method on the training data set. ANN** indicate the attribute profile estimation using the proposed method on the validation data set.*

Besides, high item discriminating was a positive impact on the classification accuracy of all six methods. When the discrimination of items decreased (from high to mixed), the classification rate dropped 0.0301, 0.0383, 0.0458, 0.0458, 0.0392, and 0.0397 for DINA, DINO, LCDM, G-DINA, RUM, and our ANN at the attribute level. The reductions were 0.0780, 0.1095, 0.1318, 0.1318, 0.1137, and 0.1158 for DINA, DINO, LCDM, G-DINA, RUM, and our ANN at the latent class level. The reason that our ANN method dropped more than DINA, DINO, and RUM (only at the attribute level) was that when the items were high discriminating, the improvement of classification rate using our ANN was more significant than using mixed discriminating items. Even though the performance of our ANN at both the attribute level and the latent class level was the best among the six diagnostic classification methods.

4. CONCLUSION

The purpose of this research is to solve two problems that exist in current supervised learning ANN methods and unsupervised learning ANNs: the supervised learning method requires ideal response pattern to train the model; the classification accuracy of unsupervised learning methods was not as good as DCMs. We designed a novel semi-supervised learning ANN to do diagnostic classification and evaluated the performances of the proposed method through a simulation study. In the proposed framework, we combined ANN with a semi-supervised learning method, the Co-Training method. To hold the two assumptions of successfully applying Co-Training, we used two DCMs, DINA, and DINO models, as the two classifiers.

In the simulated study, we compared the proposed method with five widely used DCMs, DINA, DINO, LCDM, G-DINA,

and RUM. By varying the four assessment factors (item discrimination, Q-matrix accuracy, number of attributes, and items) which impact the performance of DCMs, the comparison results indicated some advantages of the proposed method.

The first advantage is that the proposed ANN method achieved comparable performance compared with the five DCMs even under the ideal assessment condition (high diagnostic quality and 100% Q-matrix accuracy). It means that the proposed ANN method could be used for providing reasonable cognitive diagnostic classification result without an appropriate DCM for an assessment.

The second advantage is that proposed ANN was robust to the Q-matrix mis-specification because the classification rate dropped less than the other five DCMs when the Q-matrix accuracy decreased to 90% accuracy. This advantage make the proposed method can be used for real large scale assessment because the Q-matrix of a large number of items can hardly be guaranteed to be 100% accurate.

The last advantage is that although the classification rates of the proposed method dropped more than DINA and DINO when the item discriminating power reduced, the proposed method was still more robust to the item discriminating reduction than the general DCMs. In other words, the proposed method finds a trade-off between classification accuracy and robustness to the noise.

Generally, the proposed method could demonstrated the ability to provide a reasonably accurate classification results which can be used for either providing diagnostic classification. In addition, the classification can be used to determine the relationship between items and latent class. Then, the relationship can help researchers to choose the appropriate DCM to fit the data and estimate both personal variable and item variables.

5. DISCUSSION

Although the study demonstrates promise for using the proposed semi-supervised learning artificial neural networks, there are still some limitations. One concern of this study is that the current analysis only focused on the classification rate but not consider the item parameters, which are very important to provide appropriate item matching students' ability in an computer adaptive test or online adaptive learning environment. Another concern of this study is that the missing response was not considered in the proposed ANN. In the simulation, we assumed that all test-takers responded all items, but in real assessment, the missingness is a very common issue in CDM. The last concern is that although we introduced the validating test for early stop to avoid over training, this research did not evaluate the prediction performance of the proposed method. The reason is that in current CDM area, the research studies focus on explaining data not doing prediction on a new dataset. With regard to these three concerns, there will be three future research topics.

The first future study is that the classification results could be used to determine the item parameters to evaluate item discriminating power among students' mastery level for specific attributes or determine the relationship between items and attributes to explore the attribute structures. An appropriate difficulty that matches a student's momentary attribute profile is expected to encourage the student to complete the item.

The second future research direction is to convert the dichotomous response patterns to polychotomous response patterns by considering missing values into the input response pattern. Then a multiclass classification algorithm is applied to classify the latent classes by considering the missing values even the missingness is related to the latent class (i.e., non-ignorable missingness).

The last future research is to evaluate the prediction performance of the artificial neural network based cognitive diagnostic classification method, and compare the performance with the DCMs in doing prediction on new dataset, although DCMs are proposed to interpret the current dataset (i.e., training data). With regard to the knowledge in educational data mining (EDM), the prediction will consist of two directions: (1) how is the model's performance on predicting new test-takers' latent variables; (2) how is the model's performance on estimating new item's characteristics. For different directions, the ANN based method will be built up using different architecture.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR'S NOTE

We designed a novel semi-supervised learning ANN to do diagnostic classification and evaluated the proposed method's performances through a simulation study. This research study is the first time applying the thinking of semi-supervised learning into CDM using artificial neural networks. The results show that even without an appropriate theoretical DCM, the proposed method can demonstrate the ability to provide comparable classification results compared with the theoretical DCMs. It means that the proposed ANN method could provide reasonable cognitive diagnostic classification results without an appropriate TDCM for an assessment. Besides, compared with the theoretical DCMs, the proposed method can be used for real large scale assessment because it is more robust to noisy assessment factors (e.g., inaccurate Q-matrix, low discriminative items).

AUTHOR CONTRIBUTIONS

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

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.618336/full#supplementary-material

REFERENCES

Bottou, L., and Bousquet, O. (2007). The tradeoffs of large scale learning. *Adv. Neural Inform. Process. Syst.* 20, 161–168.

Bradshaw, L., Izsák, A., Templin, J., and Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: building a multidimensional test within the diagnostic classification framework. *Educ. Measure.* 33, 2–14. doi: 10.1111/emip.12020

Briggs, D. C., and Circi, R. (2017). Challenges to the use of artificial neural networks for diagnostic classifications with student test data. *Int. J. Test.* 17, 302–321. doi: 10.1080/15305058.2017.1297816

Chiu, C.-Y., and Köhn, H.-F. (2019). Consistency theory for the general nonparametric classification method. *Psychometrika* 84, 830–845. doi: 10.1007/s11336-019-09660-x

Cui, Y., Gierl, M., and Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: an artificial neural network approach. *Educ. Psychol.* 36, 1065–1082. doi: 10.1080/01443410.2015.1062078

De La Torre, J. (2011). The generalized Dina model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The Dina model, classification, latent class sizes, and the q-matrix. *Appl. Psychol. Meas.* 35, 8–26. doi: 10.1177/0146621610377081

Fergus, R., Weiss, Y., and Torralba, A. (2009). "Semi-supervised learning in gigantic image collections," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 522–530.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., and Ünlü, A. (2016). The r package cdm for cognitive diagnosis models. *J. Stat. Softw.* 74, 1–24. doi: 10.18637/jss.v074.i02

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Grohs, P., Perekrestenko, D., Elbrächter, D., and Bölcskei, H. (2019). Deep neural network approximation theory. *arXiv [Preprint]*. arXiv:1901.02220.

Guo, Q., Cutumisu, M., and Cui, Y. (2017). "A neural network approach to estimate student skill mastery in cognitive diagnostic assessments," in *Proceedings of the 10th International Educational Data Mining Conference*. doi: 10.7939/R35H7C71D

Harris, S., and Harris, D. (2015). *Digital Design and Computer Architecture: Arm Edition*. San Francisco, CA: Morgan Kaufmann. doi: 10.1016/B978-0-12-800056-4.00006-6

Hartz, S. M. (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, Champaign, IL, United States.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74:191. 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

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *J. Educ. Meas.* 49, 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Li, H., and Suen, H. K. (2013). Constructing and validating a q-matrix for cognitive diagnostic analyses of a reading test. *Educ. Assess.* 18, 1–25. doi: 10.1080/10627197.2013.761522

Liang, P. (2005). *Semi-supervised learning for natural language* (Ph.D. thesis). Massachusetts Institute of Technology, Cambridge, MA, United States.

Liu, R., Huggins-Manley, A. C., and Bradshaw, L. (2017). The impact of q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educ. Psychol. Meas.* 77, 220–240. doi: 10.1177/0013164416645636

Liu, Y., and Kirchhoff, K. (2014). "Graph-based semi-supervised acoustic modeling in DNN-based speech recognition," in *2014 IEEE Spoken Language Technology Workshop (SLT)* (South Lake Tahoe, CA), 177–182. doi: 10.1109/SLT.2014.7078570

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). "The expressive power of neural networks: a view from the width," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 6231–6239.

Nigam, K., and Ghani, R. (2000). "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the Ninth International Conference on Information and Knowledge Management* (McLean, VA), 86–93. doi: 10.1145/354756.354805

Pang, B., Nijkamp, E., and Wu, Y. N. (2020). Deep learning with tensorflow: a review. *J. Educ. Behav. Stat.* 45, 227–248. doi: 10.3102/1076998619872761

Paulsen, J. (2019). *Examining cognitive diagnostic modeling in small sample contexts* (Ph.D. thesis). Indiana University, Bloomington, IN, United States.

Roussos, L., Henson, R., and Jang, E. (2005). *Simulation Study Evaluation of the Fusion Model System Stepwise Algorithm*. ETS Project Report. Princeton, NJ: Educational Testing Service.

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

Tjoe, H., and de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: an application of a cognitive diagnosis modeling framework. *Math. Educ. Res. J.* 26, 237–255. doi: 10.1007/s13394-013-0090-7

Wang, S., and Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika* 80, 85–100. doi: 10.1007/s11336-013-9372-y

Xu, G., and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika* 81, 625–649. doi: 10.1007/s11336-015-9471-z

Xue, K. (2018). "Non-model based attribute profile estimation with partial q-matrix information for cognitive diagnosis using artificial neural network," in *Proceedings of the 11th International Conference on Educational Data Mining* (Buffalo, NY).

Xue, K. (2019). "Computational diagnostic classification model using deep feedforward network based semi-supervised learning," in *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Deep Learning for Education* (Anchorage, AK).

Xue, K., Leite, W., and Huggins-Manley, A. C. (2020). "Semi-supervised learning method for adjusting biased item difficulty estimates caused by nonignorable missingness under 2PL-IRT model," in *Proceedings of the 13th International Conference on Educational Data Mining* (Ifrane).

Yamaguchi, K., and Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PLoS ONE* 13:e188691. doi: 10.1371/journal.pone.0188691

Yamaguchi, K., and Okada, K. (2020). Hybrid cognitive diagnostic model. *Behaviormetrika* 47, 497–518. doi: 10.1007/s41237-020-00111-x

Zhan, P. (2020). Deterministic-inputs, noisy mixed modeling for identifying coexisting condensation rules. doi: 10.31234/osf.io/qwx2m

Zhu, X., and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. Synthesis lectures on artificial intelligence and machine learning, San Rafael, CA: Morgan & Claypool Publishers. doi: 10.2200/S00196ED1V01Y200906AIM006

Zhu, X. J. (2005). *Semi-Supervised Learning Literature Survey*. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

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 Xue and Bradshaw. 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 January 2021
doi: 10.3389/fpsyg.2020.621251

Sample Size Requirements for Applying Diagnostic Classification Models

Sedat Sen[1]* *and Allan S. Cohen*[2]

[1] Educational Sciences Department, Faculty of Education, Harran University, Sanliurfa, Turkey, [2] Educational Psychology Department, College of Education, University of Georgia, Athens, GA, United States

Results of a comprehensive simulation study are reported investigating the effects of sample size, test length, number of attributes and base rate of mastery on item parameter recovery and classification accuracy of four DCMs (i.e., C-RUM, DINA, DINO, and LCDMREDUCED). Effects were evaluated using bias and RMSE computed between true (i.e., generating) parameters and estimated parameters. Effects of simulated factors on attribute assignment were also evaluated using the percentage of classification accuracy. More precise estimates of item parameters were obtained with larger sample size and longer test length. Recovery of item parameters decreased as the number of attributes increased from three to five but base rate of mastery had a varying effect on the item recovery. Item parameter and classification accuracy were higher for DINA and DINO models.

Keywords: diagnostic classification models, cognitive diagnostic models, sample size, item recovery, classification accuracy

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Ren Liu,
University of California, Merced, United States
Mark D. Reckase,
Michigan State University, United States
Lokman Akbay,
Istanbul University-Cerrahpasa, Turkey

***Correspondence:**
Sedat Sen
sedatsen06@gmail.com

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

Received: *25 October 2020*
Accepted: *24 December 2020*
Published: *25 January 2021*

Citation:
Sen S and Cohen AS (2021) Sample Size Requirements for Applying Diagnostic Classification Models. Front. Psychol. 11:621251. doi: 10.3389/fpsyg.2020.621251

INTRODUCTION

Diagnostic classification models (DCMs), also known as cognitive diagnostic models (CDMs), can be viewed as restricted versions of general latent class models (Rupp and Templin, 2008). These models provide one way of classifying respondents into different diagnostic states. They are computationally intensive and generally require use of iterative algorithms to obtain estimates of model parameters. Both general and specific DCMs have been proposed in the educational and psychological measurement literature. Examples of Specific DCMs include deterministic inputs, noisy "and" gate (DINA; Haertel, 1989; Junker and Sijtsma, 2001), deterministic inputs, noisy "or" gate (DINO; Templin and Henson, 2006), noisy-input, deterministic "and" gate (NIDA), and the compensatory reparameterized unified model (C-RUM; Hartz, 2002). General DCMs include the log-linear cognitive diagnostic model (LCDM; Henson et al., 2009), the general diagnostic model (GDM; von Davier, 2005), and the generalized DINA (G-DINA; de la Torre, 2011) model. de la Torre (2011) and von Davier (2014) have shown that these three general models are equivalent.

The LCDM specifies the conditional probability that examinee j with attribute pattern α_c provides a correct answer to item i as

$$P\left(X_{ic}=1|\boldsymbol{\alpha}_c\right)=\frac{\exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^{\mathrm{T}}\mathrm{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right)}{1+\exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^{\mathrm{T}}\mathrm{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right)} \quad (1)$$

where $\lambda_{i,0}$ represents the intercept, that is the logit of a correct response for an examinee who has not mastered any of the attributes required by item *i*. $\boldsymbol{\lambda}_i^T h\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)$ is the kernel function as shown below:

$$\boldsymbol{\lambda}_{\mathrm{i}}^{\mathrm{T}} \mathrm{h}\left(\boldsymbol{\alpha}_{\mathrm{c}}, \mathbf{q}_{\mathrm{i}}\right) = \sum_{\mathrm{a}=1}^{\mathrm{A}} \lambda_{\mathrm{i},1,(\mathrm{a})} \alpha_{\mathrm{ca}} \mathrm{q}_{\mathrm{ia}} + \sum_{\mathrm{a}=1}^{\mathrm{A}-1} \sum_{\mathrm{a}'>1}^{\mathrm{A}} \lambda_{\mathrm{i},2,\left(\mathrm{a},\mathrm{a}'\right)} \alpha_{\mathrm{ca}} \alpha_{\mathrm{ca}'} \mathrm{q}_{\mathrm{ia}} \mathrm{q}_{\mathrm{ia}'} + \ldots \quad (2)$$

The $\mathbf{q}_i$ represents the *i*th row vector of the Q-matrix (Tatsuoka, 1983) that consists of 0 and 1 to indicate an item *i* gives information about the presence of an attribute *a* ($a = 1, \ldots,$ A). That is, $\mathbf{q}_{ia} = 1$ when item *i* requires attribute *a* for correct response and 0 otherwise. The vector $\boldsymbol{\alpha}_c = (\alpha_{c1}, \ldots, \alpha_{cA})^{\mathrm{T}}$ includes the attribute mastery pattern that belongs to latent class *c*. The total number of possible latent classes in a DCM equals 2^{A} where A is the number of attributes that should be mastered to respond to an item correctly. For instance, an item requiring three attributes yields eight latent classes. A full LCDM can have different item parameters including intercept [e.g., $\lambda_{1,0}$], main [e.g., $\lambda_{1,1(2)}$], and interaction [e.g., $\lambda_{1,2(3,4)}$]. Other DCMs such as DINA, DINO, and C-RUM models can be obtained from the full LCDM (see Equation 1) by modifying different parameter constraints.

Suppose that a number of items measuring arithmetic ability (addition, subtraction, multiplication, and division) were included on the test. Some of the items may require only addition (Attribute 1), and some may require only subtraction (Attribute 2). However, some of the items can require two attributes (multiplication and division) at the same time. Suppose Item 1 (for example, 5×8/4 =) requires having two attributes: multiplication (Attribute 3) and division (Attribute 4). The item response function for Item 1 can be written as

$$\begin{aligned} &\mathrm{P}\left(\mathrm{X}_{1\mathrm{c}}=1|\alpha_{\mathrm{c}3}, \alpha_{\mathrm{c}4}\right) \\ &= \frac{\exp(\lambda_{1,0}+\lambda_{1,1(3)}\alpha_{\mathrm{c}3}+\lambda_{1,1(4)}\alpha_{\mathrm{c}4}+\lambda_{1,2(3,4)}\alpha_{\mathrm{c}3}\alpha_{\mathrm{c}4})}{1+\exp(\lambda_{1,0}+\lambda_{1,1(3)}\alpha_{\mathrm{c}3}+\lambda_{1,1(4)}\alpha_{\mathrm{c}4}+\lambda_{1,2(3,4)}\alpha_{\mathrm{c}3}\alpha_{\mathrm{c}4})}, \quad (3) \end{aligned}$$

where α_{c3} and α_{c4} are latent variables for Attribute 3 and 4, respectively. This model includes one intercept ($\lambda_{1,0}$), two main effects [$\lambda_{1,1(3)}$ and $\lambda_{1,1(4)}$], and a two-way interaction effect [$\lambda_{1,2(3,4)}$] between two attributes. The latent predictor variables are combined by a series of linear modeling effects that can result in compensatory or non-compensatory DCMs. These specific DCMs do not include all of the terms in the item response function. Each DCM makes different assumptions about mastery of attributes and their effects on the item response. For example, the DINA model requires mastery of both attributes to be able to correctly respond to this item. Thus, the item response function for Item 1 includes an intercept and a two-way interaction term as shown below:

$$\mathrm{P}\left(\mathrm{X}_{1\mathrm{c}}=1|\alpha_{\mathrm{c}3}, \alpha_{\mathrm{c}4}\right) = \frac{\exp(\lambda_{1,0}+\lambda_{1,2(3,4)}\alpha_{\mathrm{c}3}\alpha_{\mathrm{c}4})}{1+\exp(\lambda_{1,0}+\lambda_{1,2(3,4)}\alpha_{\mathrm{c}3}\alpha_{\mathrm{c}4})} \quad (4)$$

The DINO model, on the other hand, functions differently than the DINA, as it requires the mastery of at least one attribute for a correct response. Students mastering either Attribute 3 or Attribute 4 can get this item correct. Thus, the item response function for this item can be written as

$$\mathrm{P}\left(\mathrm{X}_{1\mathrm{c}}=1|\alpha_{\mathrm{c}3}, \alpha_{\mathrm{c}4}\right) = \frac{\exp\left(\lambda_{1,0}+\lambda_{\mathrm{aA}}\right)}{1+\exp\left(\lambda_{1,0}+\lambda_{\mathrm{A}}\right)} \quad (5)$$

where $\lambda_A = \lambda_{1,1(3)}\alpha_{c3} + \lambda_{1,1(4)}\alpha_{c4} - \lambda_{1,2(3,4)}\alpha_{c3}\alpha_{c4}$ (see also Rupp et al., 2010, p. 163) As shown in the equation, the DINO model includes two parameters ($\lambda_{1,0}$ and λ_A). Finally, consider Item 1 for the C-RUM. The C-RUM can be considered an LCDM without an interaction effect as it includes only intercept and main effects as shown below:

$$\mathrm{P}\left(\mathrm{X}_{1\mathrm{c}}=1|\alpha_{\mathrm{c}3}, \alpha_{\mathrm{c}4}\right) = \frac{\exp(\lambda_{1,0}+\lambda_{1,1(3)}\alpha_{\mathrm{c}3}+\lambda_{1,1(4)}\alpha_{\mathrm{c}4})}{1+\exp(\lambda_{1,0}+\lambda_{1,1(3)}\alpha_{\mathrm{c}3}+\lambda_{1,1(4)}\alpha_{\mathrm{c}4})} \quad (6)$$

Slipping and guessing parameters are typically used in DCMs to describe item characteristics. Slipping parameter is used for a situation when a respondent who mastered all the required attributes for an item but fails to answer the item correctly, and guessing parameter refers to a situation when a respondent who lacks at least one of the required attributes for an item correctly answers the item. These parameters can be obtained using the intercept, main, and interaction terms presented in Equations (3)–(6). For instance, $\frac{\exp(\lambda_{1,0})}{1+\exp(\lambda_{1,0})}$ can be used to estimate guessing (g_i) and $\frac{\exp(\lambda_{1,0}+\lambda_{1,2(3,4)})}{1+\exp(\lambda_{1,0}+\lambda_{1,2(3,4)})}$ to estimate the one minus slipping ($1 - s_i$) parameter in DINA model (Rupp et al., 2010). Slipping parameter can be defined differently for other DCMs (see also Rupp et al., 2010).

As is the case with many statistical models, model parsimony is an important consideration in DCM selection, such that the simpler model is generally preferred over the more complex model. More complex DCMs require larger sample sizes to yield accurate estimates and more reduced DCMs can usually be estimated accurately with smaller sample sizes. Reduced models also typically provide for more straightforward interpretations and higher correct classification rates than more saturated models, particularly when the sample sizes are small.

Complexities which require larger samples tend to increase with the numbers of attributes and items. Sessoms and Henson (2018) have shown that 61% of the studies of DCMs have used sample sizes >1,000 and 31% have used sample sizes of 1,000–2,000. Some research has been reported with samples as small as 44 (Jang et al., 2015) and 96 (Im and Yin, 2009). Results of these latter studies have been reported with low or negative item discrimination values.

One concern with respect to sample size is that there is as yet little information reported on use of DCMs with smaller samples although a number of studies have been conducted on the effect of sample sizes on different aspects of diagnostic models. For instance, Akbay (2016) showed that the non-parametric cognitive diagnosis approach (Chiu and Douglas, 2013) performs as well as the CDM based empirical Bayes estimation method for attribute classification in the presence of small sample sizes such as 250, 500, and 1,000. Sünbül and Kan (2016) investigated the effect of

several factors including number of attributes and sample size (i.e., 200, 500, 1,000, and 5,000) on model fit, item recovery, and classification accuracy of the DINA model. The number of attributes and sample size had positive effects on the model estimates. Lei and Li (2016) investigated the performance of several model-fit indices for selecting model and on Q-matrix design under four sample size levels (500, 1,000, 2,000, and 4,000). Results indicated that performance of fit indices appeared to increase as the sample size increased. Tzou and Yang (2019) also compared the performance of model fit indices in CDMs using small sample sizes (i.e., 50, 75, 100, and 200) and showed that AIC (Akaike, 1974) performed better than other indices. Similarly, Hu et al. (2016) evaluated model fit for CDMs using sample sizes of 200, 500, and 1,000 and showed that performances of the three relative fit statistics AIC, BIC (Schwarz, 1978), and CAIC (Bozdogan, 1987) improved when sample size increased. Başokcu (2014) found classification accuracy increased as the number of attributes (1–5) decreased and sample size increased in DINA and G-DINA models. Similarly de la Torre et al. (2010) showed that sample size increase from 1,000 to 4,000 reduced the bias in item parameter estimates.

In another sample size related simulation study, Cui et al. (2012) showed that the asymptotic normal theory of classification consistency index and classification accuracy index can be applied with small sample sizes (100, 500 and 1,000) for attribute classifications in DINA model. Paulsen (2019) also investigated the effect of very small sample sizes including 25, 50, 150, and 1,000 simulated respondents on three CDMs (DINA, non-parametric cognitive diagnosis, and the supervised artificial neural network models) by focusing on characteristics such as model performances and model fit. Results of that study showed that those three models were able to estimate examinee classifications at even the smallest sample size. Galeshi and Skaggs (2016) conducted a simulation study using C-RUM under different sample size levels including 50, 100, 500, 1,000, 5,000, and 10,000, and showed that attribute classification was effected by different combinations of sample size and test length.

Choi et al. (2010) found that relative model fit indices were able to detect the correct DCM with samples of 200 or more. Rojas et al. (2012) found attribute classification accuracy of DINA, DINO, A-CDM, and G-DINA models was more accurate when test length and sample size were large. Previous simulation studies have focused on a limited number of models under the assumption that test items measured a common underlying model. However, this may not necessarily be the case as each item may also be designed to reflect a specific DCM. Thus, a more comprehensive simulation study is needed to examine the effects of small sample size on the classification accuracy and on parameter estimates when the test items reflect different model structures.

The present study was designed to investigate the effects of sample size on estimation and accuracy of parameter estimates and on classification as a function of sample size and for different types of DCMs. This study investigated the performance of specific DCMs under a set of practical testing conditions.

MATERIALS AND METHODS

Simulation Study Design

Parameter recovery and classification accuracy for the LCDMs were assessed under several simulated conditions. In this regard, five factors were manipulated: sample size (50, 100, 200, 300, 400, 500, 1,000, and 5,000), test length (12, 24, and 36 items), number of attributes (3 and 5), base rate (0.25 and 0.50), and generating model (the reduced LCDM, the DINA model, the DINO model, and the C-RUM). A total of 100 replications were simulated for each condition using the maximum likelihood estimation algorithm as implemented in the Mplus 8.4 software package (Muthén and Muthén, 1998–2019).

Constant Factors

For purposes of this study, tetrachoric correlation between each pair of attributes, item quality, and Q-matrices were held constant across simulation conditions. The tetrachoric correlation between each pair of attributes was set to be 0.70. This is within the typical range of correlations for subdomains in national and international educational assessments (Kunina-Habenicht et al., 2012).

Medium level item quality was used to simulate items that were better at separating masters and non-masters of the measured attributes. This was achieved with medium level item discrimination based on the difference in the probability of a correct response for two groups of students (i.e., item discrimination value of $0.60 = 0.85–0.25$). In this study, two different Q-matrix specifications were used for models with three and five attributes. The Q-matrix specification used in this study was intended to reflect the kinds of Q-matrices used in previous simulation studies with DCMs (e.g., Kunina-Habenicht et al., 2012; de la Torre and Chiu, 2016). The Q-matrix of the models with three attributes and 36 items used in this study is presented in **Table 1**. The first 12 rows of the Q-matrix in **Table 1** were used for the 12-item models and the first 24 rows of the Q-matrix in **Table 1** were used for the 24-item models. The Q-matrix of the models with 5 attributes and 36 items is presented in **Table 2**. The 12-item and 24-item models include the first 12 and 24 rows of this table, respectively.

Manipulated Factors

The simulation study had five manipulated factors including sample size, test length, number of attributes, base rate, and generating models. The sample sizes were 50, 100, 200, 300, 400, 500, 1,000, and 5,000 simulated examinees. These values were selected to represent a range of sample sizes from very small (50) to large (5,000). The number of respondents were selected to comply with studies reported in the DCM literature. Rojas et al. (2012), for example, used 100, 200, 400, 800, and 1,600 and Başokcu (2014) used 30, 50, 100, 200, and 400 for the sample size conditions. In this study, we extended sample sizes to include 5,000 simulated examinees.

Test length included 12, 24, and 36 items. These were intended to simulate small, medium, and long test lengths. For a math test, for example, it usually takes about 1 ½ min per multiple-choice item. For a 36-item test, this would actually be 48 min, which

TABLE 1 | Q-Matrix for conditions with three attributes.

	Attribute					Attribute			
Item	**1**	**2**	**3**	**Item type***	**Item**	**1**	**2**	**3**	**Item type***
1	1	0	0	LCDM**	19	1	0	0	DINO**
2	0	1	0	DINA**	20	0	1	0	C-RUM**
3	0	0	1	DINO**	21	0	0	1	LCDM**
4	1	1	0	C-RUM	22	1	1	0	DINA
5	1	0	1	LCDM	23	1	0	1	DINO
6	0	1	1	DINA	24	0	1	1	C-RUM
7	1	0	0	DINO**	25	1	0	0	LCDM**
8	0	1	0	C-RUM	26	0	1	0	DINA**
9	0	0	1	LCDM**	27	0	0	1	DINO**
10	1	1	0	DINA	28	1	1	0	C-RUM
11	1	0	1	DINO	29	1	0	1	LCDM
12	0	1	1	C-RUM	30	0	1	1	DINA
13	1	0	0	LCDM**	31	1	0	0	DINO**
14	0	1	0	DINA	32	0	1	0	C-RUM**
15	0	0	1	DINO**	33	0	0	1	LCDM**
16	1	1	0	C-RUM	34	1	1	0	DINA
17	1	0	1	LCDM	35	1	0	1	DINO
18	0	1	1	DINA	36	0	1	1	C-RUM

Item type specified for LCDMREDUCED model across all items.

***The parameterization of four DCMs are the same for items requiring only 1 attribute.*

TABLE 2 | Q-Matrix for conditions with five attributes.

	Attribute						Attribute				
Item	**1**	**2**	**3**	**4**	**5**	**Item**	**1**	**2**	**3**	**4**	**5**
1	1	0	0	0	0	19	1	0	1	0	0
2	0	1	0	0	0	20	1	0	0	1	0
3	0	0	1	0	0	21	1	0	0	0	1
4	0	0	0	1	0	22	0	1	1	0	0
5	0	0	0	0	1	23	0	1	0	1	0
6	1	1	0	0	0	24	0	1	0	0	1
7	1	0	1	0	0	25	1	0	0	0	0
8	1	0	0	1	0	26	0	1	0	0	0
9	1	0	0	0	1	27	0	0	1	0	0
10	0	1	1	0	0	28	0	0	0	1	0
11	0	1	0	1	0	29	0	0	0	0	1
12	0	1	0	0	1	30	1	1	0	0	0
13	1	0	0	0	0	31	1	0	1	0	0
14	0	1	0	0	0	32	1	0	0	1	0
15	0	0	1	0	0	33	1	0	0	0	1
16	0	0	0	1	0	34	0	1	1	0	0
17	0	0	0	0	1	35	0	1	0	1	0
18	1	1	0	0	0	36	0	1	0	0	1

would be a relatively long amount of time for most students up to and including high school age. These test lengths were set to be multiples of four to produce items based on four different models in the reduced LCDM model. The numbers of items included in this study are typical of test lengths observed in real tests such as the TIMSS and PISA tests. For example, TIMSS assessment items are grouped into a series of item blocks, with ∼10–14 items in each block at the fourth grade and 12–18 items at the eighth grade

TABLE 3 | Item-to-profile table for item 4.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
α_c	[0,0,0]	[0,0,1]	[0,1,0]	[0,1,1]	[1,0,0]	[1,0,1]	[1,1,0]	[1,1,1]
LCDM	$\lambda_{4,0}$	$\lambda_{4,0}$	$\lambda_{4,0} + \lambda_{4,1(2)}$	$\lambda_{4,0} + \lambda_{4,1(2)}$	$\lambda_{4,0} + \lambda_{4,1(1)}$	$\lambda_{4,0} + \lambda_{4,1(1)}$	$\lambda_{4,0} + \lambda_{4,1(1)} + \lambda_{4,1(2)} + \lambda_{4,2(1,2)}$	$\lambda_{4,0} + \lambda_{4,1(1)} + \lambda_{4,1(2)} + \lambda_{4,2(1,2)}$
Mplus label	t4_1	t4_1	t4_2	t4_2	t4_3	t4_3	t4_4	t4_4

TABLE 4 | Mplus syntax specifications for item 4 (Q-matrix entry 110).

Full LCDM	C-RUM	DINA	DINO
NEW (l4_0 l4_12 l4_11 l4_212);	NEW (l4_0 l4_12 l4_11);	NEW (l4_0 l4_e);	NEW (l4_0 l4_e);
t4_1 = –(l4_0);	t4_1 = –(l4_0);	t4_1 = –(l4_0);	t4_1 = –(l4_0);
t4_2 = –(l4_0 + l4_12);	t4_2 = –(l4_0 + l4_12);	t4_2 = –(l4_0);	t4_2 = –(l4_0 + l4_e);
t4_3 = –(l4_0 + l4_11);	t4_3 = –(l4_0 + l4_11);	t4_3 = –(l4_0);	t4_3 = –(l4_0 + l4_e);
t4_4 = –(l4_0 + l4_11 + l4_12 + l4_212);	t4_4 = –(l4_0 + l4_11 + l4_12);	t4_4 = –(l4_0 + l4_e);	t4_4 = –(l4_0 + l4_e);
! Order constrains	! Order constrains	! Order constrains	! Order constrains
l4_12 > 0;	l4_12 > 0;	l4_e > 0;	l4_e > 0;
l4_11 > 0;	l4_11 > 0;		
l4_212 > –l4_11;			
l4_212 > –l4_12;			

level. Similarly, von Davier et al. (2019) notes that the number of items administered in each assessment cycle of PISA consisted of 28 items for reading in 2003 and 35 items for mathematics assessment in 2009.

The numbers of attributes were three and five to reflect numbers commonly found in educational and psychological tests (Kunina-Habenicht et al., 2012). For instance, Chen and Chen (2016) reported a Q-matrix with five attributes by employing the five processes (skills) of reading under the PISA assessment framework. Examples of simulation or real data studies with three or five attributes include de la Torre and Douglas (2004), de la Torre (2009), de la Torre and Lee (2010), Kunina-Habenicht et al. (2012), Templin and Bradshaw (2014), de la Torre and Chiu (2016), Hu et al. (2016), and Sen and Bradshaw (2017). The base rate of mastery for an attribute is the proportion of examinees who have mastered the attribute in the population. This was set to 0.25 and 0.50. A base rate of 0.50 is commonly reported in the literature (e.g., Kunina-Habenicht et al., 2012; Bradshaw and Madison, 2016; Sen and Bradshaw, 2017). As in Sen and Bradshaw (2017), the 0.25 base rate condition was added to investigate item recovery and classification accuracy comparisons under less optimal conditions. Base rate mastery and tetrachoric correlations between each pair of attributes were generated using the SAS macro created by Templin and Hoffman (2013). The SAS code used to generate the 0.25 base rate is presented in the **Supplementary Data**.

Four different data-generating models were simulated, including (a) Reduced LCDM, (b) C-RUM, (c) DINA, and (d) DINO. For the reduced LCDM, the underlying DCM structures were generated to differ across the complex items with the following common sub-models of the LCDM. For a 12-item test, three complex items were generated under the DINA model, three under the DINO model, three under the C-RUM, and three under the saturated LCDM (see **Table 1**). The specific item structures used in LCDMREDUCED model for 12-, 24-, and 36-item tests are presented in **Table 1**. For the DINA, DINO, and C-RUM models, the underlying DCM structure was generated to be the same. That is, all of the items were of a common type. The same patterns were used for the 5-attribute matrix presented in **Table 2**.

Data Generation

Data generation was done using Mplus. First, a full saturated LCDM syntax was created using *MplusDCM_functions.R* function with the MplusAutomation (Hallquist and Wiley, 2018) package in R. The *MplusDCM_functions.R* function created by Andre Rupp and Oliver Wilhelm was used to build, run, and parse Mplus syntax for estimation of the LCDM. After generating the full LCDM syntax using this function, the syntax for specific DCMs was created by modifying the syntax using the descriptions provided by Rupp et al. (2010) and Sen and Terzi (2020). The LCDM syntax for Item 4 from **Table 1**, for example, included an intercept ($\lambda_{4,0}$), two main effects [$\lambda_{4,1(1)}$ and $\lambda_{4,1(2)}$] for Attribute 1 and Attribute 2 and an interaction effect [$\lambda_{4,2(1,2)}$] between these two attributes. In this example, 3 attributes yield 8 attribute patterns or classes. Then, the LCDM kernel (see Equation 2) was specified for each class and each item to assign latent classes to an attribute pattern or profile. Finally, unique item response functions for each item were specified in the Item-to-Profile table (see **Table 3**). The following labels used in the Mplus code as shown in **Table 3** are t4_1, t4_2, t4_3, and t4_4. These were created using the Item-to-Profile table.

The specific part of the Mplus syntax for LCDM is presented in the first column of **Table 4**. As can be seen in the first column, the LCDM syntax to estimate intercept, main effects, and the interaction term is labeled as l4_0, l4_12, l4_11, and l4_212,

TABLE 5 | Partial eta-squared values for manipulated variables in simulation.

	RMSE				Bias			
	Intercept	**Main**	**e parameter**	**Interaction**	**Intercept**	**Main**	**e parameter**	**Interaction**
N	0.976*	0.983*	0.985*	0.960*	0.469*	0.674*	0.356*	0.407
Att	0.540*	0.810*	0.167*	0.788*	0.119*	0.349*	0.001	0.406
k	0.777*	0.903*	0.800*	0.835*	0.236*	0.197*	0.113*	0.739
BR	0.420*	0.460*	0.339*	0.452*	0.007	0.004	0.090*	0.038
Model	0.633*	0.446*	0.404*	-	0.049*	0.001	0.018	-
N×Att	0.127*	0.314*	0.019	0.552*	0.094*	0.247*	0.236*	0.300
N×*k*	0.274*	0.543*	0.217*	0.772*	0.077*	0.492*	0.037	0.222
N×BR	0.137*	0.077	0.039	0.206	0.055*	0.022	0.164*	0.107
N×Model	0.328*	0.158*	0.080	-	0.053	0.014	0.054	-
Att×*k*	0.198*	0.334*	0.148*	0.354*	0.059*	0.092*	0.010	0.673
Att×BR	0.074*	0.005	0.575*	0.060	0.000	0.040*	0.042*	0.036
Att×Model	0.025	0.121*	0.247*	-	0.040*	0.003	0.021	-
k×BR	0.048*	0.001	0.016	0.187*	0.015	0.037	0.025	0.091
k×Model	0.407*	0.224*	0.047*	-	0.140*	0.045*	0.012	-
BR×Model	0.059*	0.001	0.070*	-	0.034*	0.014	0.008	-

*N, sample size; Att, number of attributes; k, test length; BR, base rate of mastery; Model, model type; *p < 0.01.*

respectively. C-RUM, DINA, and DINO models are presented in the next three columns of **Table 4**. The C-RUM does not include any interaction effect. The DINA and DINO models only include intercept and *e* parameters. The Mplus syntax for these latter two models are not the same, however, due to differences in the kernel function.

The MONTECARLO command in Mplus was used to generate 100 data sets for each condition. The true generating values for intercept, main, *e* parameter, and interaction effects were set to be −1.1, 1.3, 3, and 0.24, respectively. These values were selected to produce items with medium-quality with an item discrimination value of 0.60. The item discrimination is specified by taking the difference in the probability of a correct response for two groups of students (0.60 = 0.85–0.25).

Estimation

All of the models were estimated using maximum likelihood estimation (i.e., MLR) as implemented in Mplus. The following Mplus options were used to obtain estimates of model parameters: ANALYSIS: TYPE = MIXTURE; STARTS = 200 20; PROCESSORS = 8;. Eight classes and 32 classes were modeled for the 3-Attribute and 5-Attribute models in the MODEL command using the labels from **Table 3**. The MODEL CONSTRAINT part in Mplus syntax was modified using the model constraints described in **Table 4**. The vectors of attribute classifications were obtained based on expected a posteriori estimation by specifying the FILE = "respondents#.cprob"; option under SAVEDATA command in Mplus. In total, 38,400 Mplus analyses (8 ×3 × 2 × 4 × 2 × 100) were run using a Linux (64-bit Centos 7) high performance computing (HPC) cluster. Item parameter estimates for intercept, main, *e* parameter, and interaction effects and class probability values were extracted from each Mplus files using MplusAutomation package.

Evaluation Criteria

Recovery of item parameters were assessed using the root mean square error (RMSE) and bias across replications. RMSE and bias values for intercept, main, *e* parameter, and interaction terms were calculated using the following formulas:

$$\text{RMSE} = \sqrt{\frac{\sum_{r=1}^{R}\left(\lambda_i - \hat{\lambda}_{ir}\right)^2}{R}}, \tag{7}$$

$$\text{Bias} = \frac{\sum_{r=1}^{R}\left(\lambda_i - \hat{\lambda}_{ir}\right)}{R}, \tag{8}$$

where R is the total number of replications ($r = 1, \ldots R$) and λ_i is the true parameter value of intercept, main, *e* parameter or interaction term for item *i*. $\hat{\lambda}_{ir}$ refers to estimated item parameters for item i under the rth replication. Classification accuracy of attribute profiles under each condition was determined by calculating the percentage of examinees whose estimated attribute profile was the same as the simulated (i.e., true) attribute profile.

RESULTS

Results of item recovery and classification accuracy were obtained for the 384 conditions in the study. Mean RMSE and mean bias values were computed over 100 replications. The percentage of classification accuracy was also calculated for each condition. Mean RMSE and bias values are presented in **Supplementary Tables 1–12**. **Figures 1–5** summarize the mean RMSE and absolute mean bias results for each fitted model. Separate plots are provided for intercept, main, *e* parameter, and interaction effects in each figure. Each plot displays 12 labeled lines representing 12 different conditions for the 2 attributes, 3

test lengths, and 2 base rates. For instance, 3ATT12ITEM25BR label represents a condition with 3 attributes, 12 items, and 0.25 base rate of mastery.

Item Parameter Recovery Results

Recovery of the C-RUM

Figure 1 presents the item parameter recovery results for the C-RUM when model-data fit holds. As can be seen in **Figure 1**, mean RMSE and bias values for the intercept and main parameters of the C-RUM appear to decrease as the sample size increases. Mean RMSE values for the intercept parameter ranged from 0.040 to 0.759 (see **Supplementary Tables 1–3**). Mean bias values for the intercept parameter were between 0.001 and 0.259.

Bias and RMSE values for the intercept parameter in 12-item conditions was higher than in the 24-item conditions, which was than the 36-item conditions. The 5-attribute conditions yielded higher RMSE and bias values than the 3-attribute conditions, and the 0.50 base rate conditions yielded higher RMSE and bias values than 0.25 base rate conditions. Overall, the highest mean RMSE values were for the 5-attribute × 12-item × 0.50 base rate conditions and the lowest mean RMSE and bias values were for the 3-attribute 36-item × 0.25 base rate conditions. Similarly, the highest mean bias values were obtained in the 5-attribute × 12-item for both 0.25 and 0.50 base rate conditions. The lowest mean bias values were in the 3-attribute × 36-item × 0.25 base rate conditions.

The RMSE and bias plots for intercept indicate that sample size, test length, and base rate of mastery had effects on mean RMSE and bias for the intercept parameter. It appears that the number of respondents, test lengths, number of attributes, and mastery base rates had an impact on the recovery of intercept parameter. As the number of respondents increased and test length, number of attributes, and base rate decreased, the recovery of the intercept parameter appeared to increase. Only conditions with 5,000 respondents produced RMSE values <0.10. Some of the 1,000-respondent conditions also yielded RMSE values <0.10. Mean RMSE values for the conditions <200 exceeded 0.20. The conditions with 12 items produced highest bias values. A sharp decline was observed with other conditions after 200 respondents.

In the lower panels in **Figure 1**, the item recovery values are plotted for the main effect of the C-RUM. As can be seen, both RMSE and bias values decreased as the number of respondents increased. This pattern is clearer for the RMSE plot (see the left lower panel) than of the bias plot (see the right lower panel). Mean RMSE values for the main effect ranged from 0.080 to 1.210 (see **Supplementary Tables 1–3**). Mean bias values ranged between 0.001 and 0.297. Item parameter recovery values for the main effect parameter in the 12-item conditions was higher than in the 24-item conditions and both were higher than in the 36-item conditions. The 5-attribute conditions yielded higher RMSE and bias values than the 3-attribute conditions. The pattern is clearer for RMSE as some of the conditions showed reversals for mean bias values. The 0.50 base rate conditions yielded higher RMSE values than the 0.25 base rate conditions. Except for the 36-item conditions, the 0.50 base rate conditions yielded higher bias values than the 0.25 base rate conditions.

Overall, the highest mean RMSE values were obtained with the 5-attribute × 12-item × 0.25 base rate conditions while the lowest mean RMSE values were observed with the 3-attribute × 36-item × 0.50 base rate conditions. Similarly, the highest mean bias values were obtained with 5-attribute × 12-item for both 0.25 and 0.50 base rate conditions. The lowest mean bias values were observed with 3-attribute × 36-item × 0.50 base rate conditions. The RMSE and bias plots for the intercept indicate that sample size, test length, and base rate of mastery had effects on mean RMSE and bias for the main effect parameter. The number of respondents, test length, number of attributes, and base rate of mastery also had effects on the recovery of the intercept parameter. As the number of respondents and base rate increased and test length and number of attributes decreased, the recovery of the main effect parameter increased. Except for the 12-item conditions, only conditions with 5,000 respondents produced RMSE values <0.10. Mean RMSE values for the main effect under conditions with $<1{,}000$ exceeded 0.20. The conditions with 12 items produced the highest bias values. Bias values <0.10 were more likely under conditions with more than 200 respondents. Overall, the recovery of intercept parameter was found to be better than that of the main effect parameter for the C-RUM.

Recovery of the DINA Model

Figure 2 shows item recovery results for the DINA model when model-data fit holds. As can be seen in **Figure 2**, mean RMSE and bias values for intercept and s parameter of the DINA model decreased as sample size increased. Mean RMSE values for the intercept parameter ranged from 0.037 to 0.628 (see **Supplementary Tables 4–6**). Mean bias values for the intercept parameter ranged from 0.001 to 0.218. When the data generating model was the DINA, recovery values for the intercept parameter in the 12-item conditions were higher than in the 24-item conditions and both were higher than in the 36-item conditions. The 5-attribute conditions yielded higher RMSE and bias values than the 3-attribute conditions.

The 0.50 base rate conditions yielded higher RMSE values than the 0.25 base rate conditions for the intercept parameter. The 0.50 base rate conditions produced lower RMSE values than the 0.25 base rate conditions for the 3-attribute conditions. However, the 0.25 base rate conditions produced lower values than for the 0.50 base rate conditions under most of the 5-attribute conditions.

Overall, the highest mean RMSE values were obtained with 5-attribute × 12-item × 0.50 base rate conditions while the lowest mean RMSE values were observed with 3-attribute and 0.25 base rate conditions with 24 and 36 items. Similarly, the highest mean bias values were obtained with 5-attribute × 12-item conditions for both the 0.25 and 0.50 base rate conditions. The lowest mean bias values were observed with the 3 attribute × 0.25 base rate conditions with 24 and 36 items. It appears that the number of respondent, test length, number of attributes, and base rate of mastery had an impact on the recovery of the intercept parameter. As the number of respondents and test length increased, and the number of attributes and base rate decreased, the recovery of the intercept parameter appeared to increase. Only conditions with 5,000 respondents with 24 and 36 items

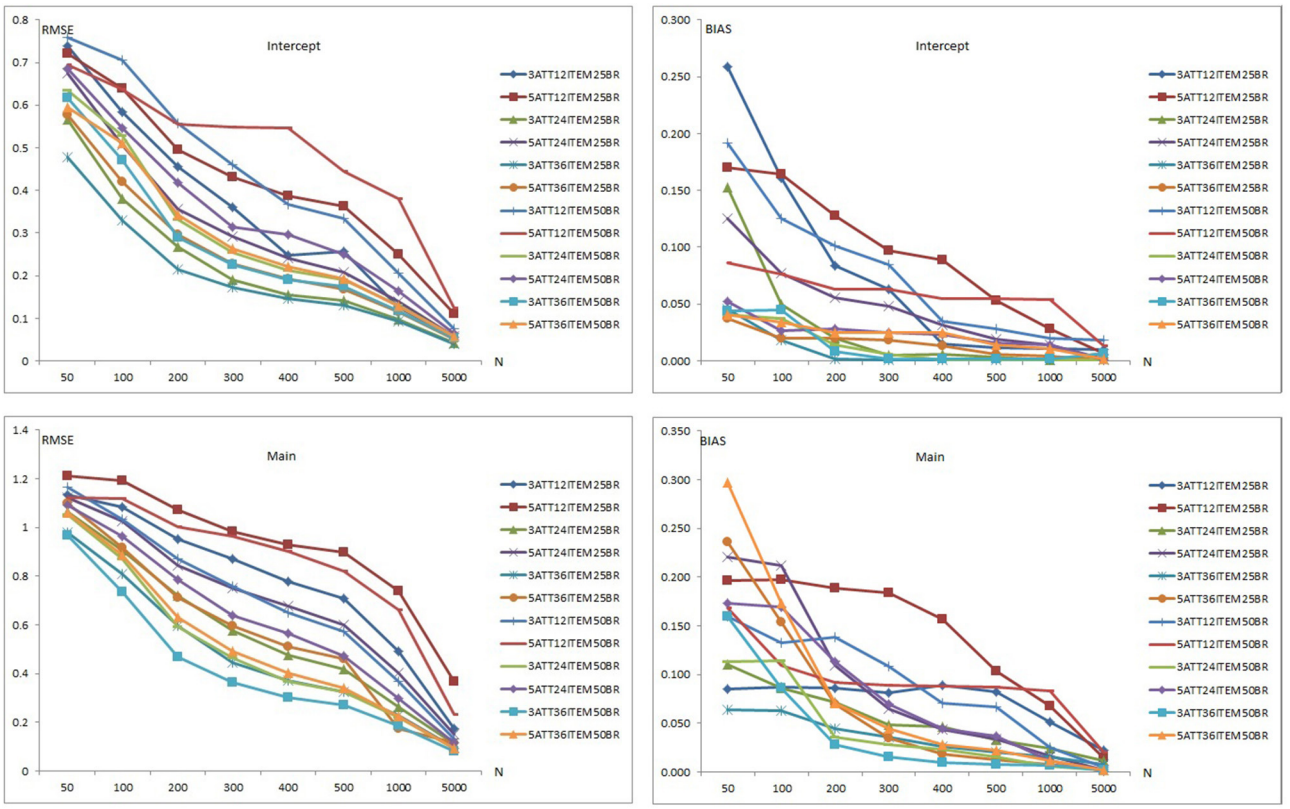


FIGURE 1 | Mean RMSE and bias plots for the C-RUM.

produced RMSE values <0.10. Mean RMSE values, however, were found to be higher than 0.10 for 12-item conditions even with 5,000 respondents. For 24- and 36-item conditions, mean RMSE values exceeded 0.20 with sample sizes <200. This was not the case with the 12-item conditions as conditions with 12 items also produced the highest bias values. Mean bias values for the intercept parameter were found to be between 0.001 and 0.10 under most of the conditions except for two conditions: 5-attribute × 12-item × 0.25 base rate and 5-attribute × 12-item × 0.50 base rate.

As can be seen in the lower part of **Figure 2**, similar patterns were observed for recovery of the *e* parameter. However, both RMSE and bias values of the *e* parameter were higher than for the intercept parameter.

Mean RMSE values for the *e* parameter ranged from 0.080 to 1.067 (see **Supplementary Tables 4–6**). Mean bias values for the *e* parameter were between 0.001 and 0.592. When the data generating model was the DINA model, item recovery values for the *e* parameter in the 12-item conditions was higher than in the 24-item conditions and both were higher than in the 36-item conditions. The 3-attribute conditions yielded higher RMSE values than the 5-attribute conditions for the 0.25 base conditions. However, the 5-attribute conditions yielded higher RMSE values than the 3-attribute conditions for the 0.50 base rate of mastery. Mean bias results for the *e* parameter did not show any clear pattern with respect to the number of attributes and base rate.

Overall, the highest mean RMSE values for *e* parameter were obtained with the 3-attribute × 12-item × 0.25 base rate conditions while the lowest mean RMSE values were observed for the 36-item × 0.50 base rate conditions with 3 and 5 attributes. Similarly, the highest mean bias values for the *e* parameter were obtained with the 5-attribute × 12-item conditions with both the 0.25 and 0.50 base rates. The lowest mean bias values were observed with 3-attribute × 0.25 base rate conditions with both 24 and 36 items. It appears that the number of respondents, test length, number of attributes, and mastery base rates had an impact on the recovery of the intercept parameter. As the number of respondents and test length increased, the recovery of the *e* parameter appeared to increase. The effect of number of attributes and base rates appeared to be less clear. Only the 24- and 36-item conditions with 5 attributes produced RMSE values <0.10 when the sample size was 5,000. As can be seen in **Supplementary Tables 4–6**, a few RMSE values <0.20 were observed, even with 1,000 respondents. Overall, the intercept parameter of the DINA model appeared to be recovered better than the *e* parameter based on mean RMSE and bias values.

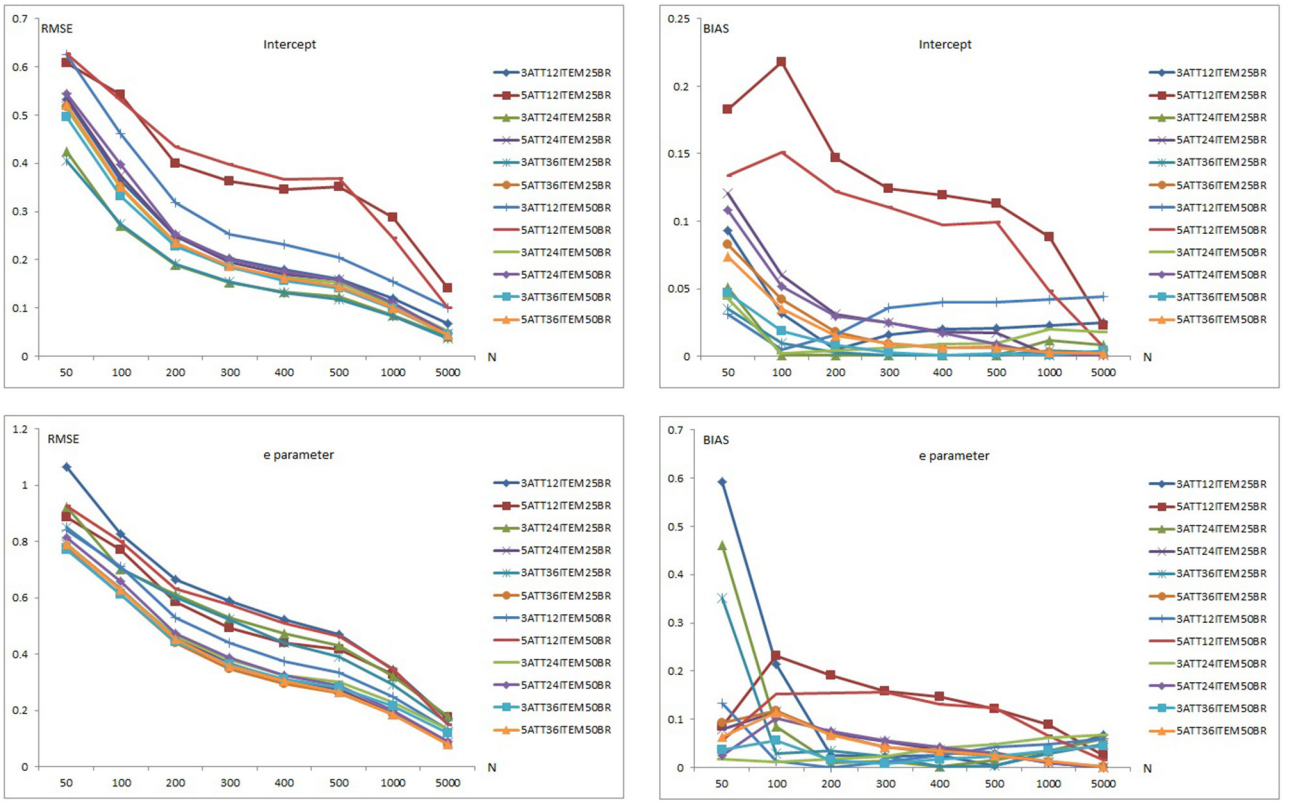


FIGURE 2 | Mean RMSE and bias plots for the DINA Model.

Recovery of the DINO Model

Item recovery results for the DINO model (see **Figure 3**), indicate that mean RMSE and bias values for intercept and *e* parameter of DINO model appear to decrease as the number of respondents increases when model-data fit holds. Mean RMSE values for intercept parameter ranged from 0.038 to 0.688 (see **Supplementary Tables 7–9**). Mean bias values for the intercept parameter ranged between 0.001 and 0.182. Item recovery values for the intercept parameter in the 12-item conditions was higher than in the 24-item conditions and both were higher than in the 36-item conditions. The 5-attribute conditions yielded higher RMSE and bias values than the 3-attribute conditions for the intercept parameter.

The 0.50 base rate conditions yielded higher RMSE values than 0.25 base rate conditions for the intercept parameter. The 0.25 base rate conditions also yielded lower values than the 0.50 base rate conditions under most of the 3-attribute conditions. However, the 0.25 base rate conditions produced lower values than the 0.50 base rate conditions for over half of the 5-attribute conditions.

Overall, the highest mean RMSE values for the intercept parameter were obtained for the 5-attribute × 12-item × 0.50 base rate conditions while the lowest mean RMSE values were observed with the 3-attribute × 0.25 base rate conditions for both 24 and 36 items. Similarly, the highest mean bias values were obtained with the 5-attribute × 12-item × 0.25 base rate conditions in addition to 5-attribute 12-item and 0.50 base rate under the small sample size conditions (i.e., <200). However, the highest mean bias values were obtained with the 3-attribute × 12-item × 0.50 base rate conditions for sample sizes >200. The lowest mean bias values for the intercept parameter of DINO model were observed with the 3-attribute × 0.25 base rate conditions for both 24 and 36 items. It appears that the number of respondents, test length, number of attributes, and mastery base rates had an impact on recovery of the intercept parameter. As the number of respondents and test length increased, and number of attributes and base rate decreased, the recovery of intercept parameter appeared to improve. Conditions with 5,000 simulated respondents produced RMSE values <0.100. The conditions with 12 items also produced the highest bias values for the intercept parameter. Mean bias values for the intercept parameter were between 0.001 and 0.100 except for two conditions: The 5 attribute × 12 item × 0.25 base rate and the 5 attribute × 12 item × 0.50 base rate conditions.

As can be seen in the lower panels of **Figure 3**, similar patterns were observed with the recovery of the *e* parameter, although both RMSE and bias values were higher than for the intercept parameter. Mean RMSE values for the *e* parameter ranged from 0.076 to 0.934 (see **Supplementary Tables 7–9**). Mean bias values for the *e* parameter ranged between 0.002 and 0.269. Recovery

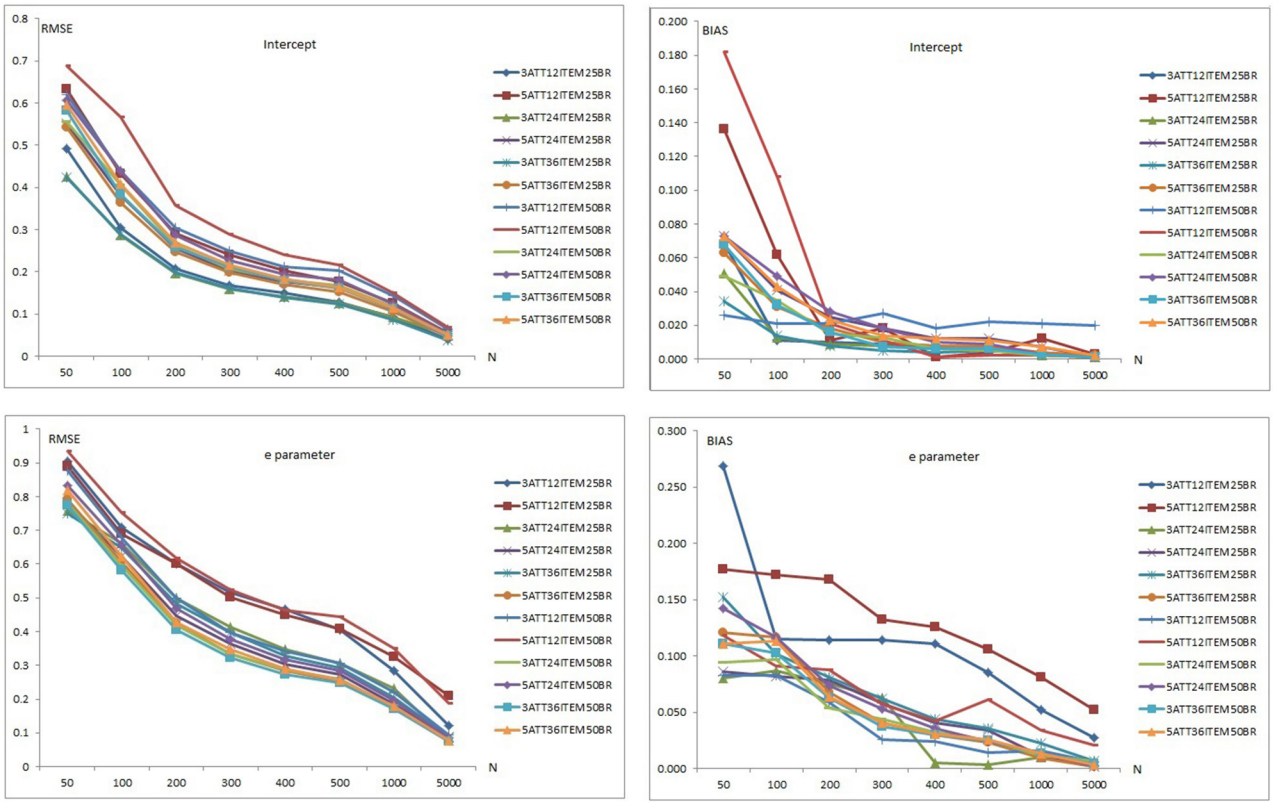


FIGURE 3 | Mean RMSE and bias plots for the DINO Model.

values for the *e* parameter in the 12-item conditions were higher than in the 24-item conditions and both were higher than in the 36-item conditions. The 3-attribute conditions yielded higher RMSE values than 5-attribute conditions for the 0.25 base conditions. However, the 5-attribute conditions yielded higher RMSE values than 3-attribute conditions for the 0.50 base rate conditions. For both 0.25 and 0.50 base rate conditions, mean bias values of the *e* parameter for the 5-attribute conditions were higher than for the 3-attribute conditions.

Overall, the highest mean RMSE values for the *e* parameter were with 5-attribute × 12-item × 0.50 base rate conditions while the lowest mean RMSE values were for the 3-attribute × 36-item × 0.50 base rate conditions. Similarly, the highest mean bias values for the *e* parameter were obtained with the 12-item × 0.25 base rate conditions with both 3 and 5 attributes while the lowest mean bias values were observed with 3-attribute × 12-item × 0.50 base rate conditions.

The number of respondents, test length, number of attributes, and base rates of mastery appeared to affect recovery of the intercept parameter. As the number of respondents and test length increased, recovery of the *e* parameter increased. However, the effects of the number of attributes and base rates were less clear. Only the 5-attribute conditions × 5,000 respondents for the 24 and 36 items yielded RMSE values <0.10. As can be seen in **Supplementary Tables 7–9**, relatively few RMSE values <0.20 were observed even with 1,000 respondents. Overall, the intercept parameter of the DINO model appeared to be recovered better than the *e* parameter.

Recovery of the LCDMREDUCED Model

Figures 4, **5** are plots of the mean RMSE and bias results, respectively, for the LCDMREDUCED model when model-data fit holds. The mean RMSE and bias summaries are presented separately as the number of estimated parameters in the LCDMREDUCED model is higher than C-RUM, DINA, and DINO models. Separate recovery plots are provided for the intercept, main effect, *e* parameter, and interaction effects. As can be seen in **Figure 4**, mean RMSE values for the intercept, main, *e* parameter, and interaction effects of the LCDMREDUCED model decreased as sample size increased. Mean RMSE values for the intercept parameter ranged from 0.041 to 0.741 (see **Supplementary Tables 10–12**). Mean RMSE values for main effects ranged between 0.096 and 1.206. Mean RMSE values for the *e* parameter ranged from 0.078 to 1.266 (see **Supplementary Tables 10–12**). Mean RMSE values for the interaction effects ranged between 0.158 and 1.569.

When the data generating model was the LCDMREDUCED, mean RMSE values for the intercept, main effects, *e* parameter, and interaction effects in the 12-item conditions were higher

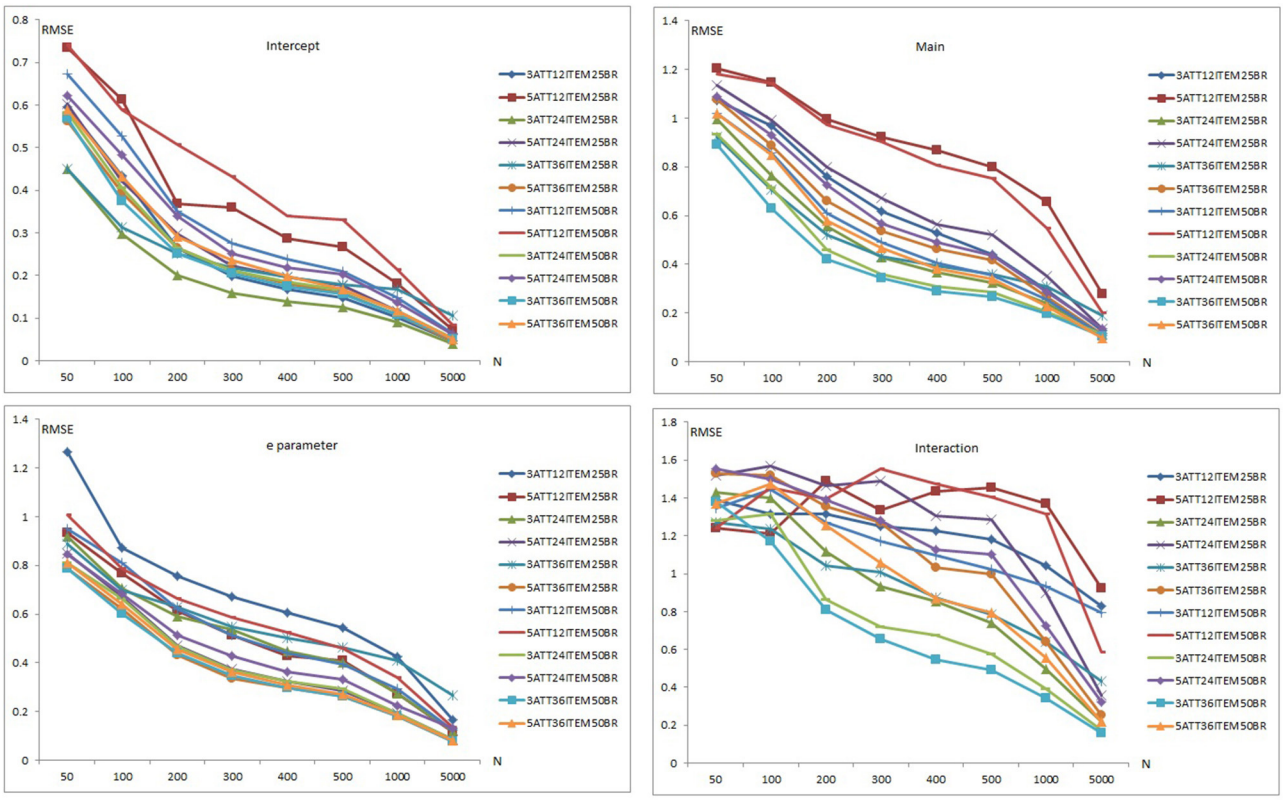


FIGURE 4 | Mean RMSE plots for the LCDMREDUCED Model.

than in the 24- and 36-item conditions. However, the 36-item conditions yielded smaller RMSE values than the 24-item conditions under more than half of the conditions. The 0.25 base rate conditions yielded lower RMSE values than the 0.50 base rate conditions except for the main effect parameters under the 12-item conditions. The base rate did not show any clear pattern of effects for the 24-item and 36-item conditions. Except for the 0.25 base rate for the *e* parameter, the 3-attribute conditions produced lower mean RMSE values than the 5-attribute conditions for intercept, main, *e* parameter, and interaction effects.

Overall, the highest mean RMSE values were obtained with 5-attribute × 12-item conditions for both 0.25 and 0.50 base rate conditions for the intercept, main effects and interaction parameters. The highest RMSE for the *e* parameter was observed with the 3-attribute × 12-item × 0.25 base rate conditions. The lowest mean RMSE values were observed with the 3-attribute × 36-item × 0.50 base rate conditions for the main effects, *e* parameter, and interaction effects. The 3-attribute × 24-item × 0.25 base rate conditions yielded the lowest RMSE values for the intercept parameter.

Separate bias plots are presented in **Figure 5** for the intercept, main effects, *e* parameter, and interaction effects. As can be seen in **Figure 5**, mean bias values for intercept, main effect, *e* parameter, and interaction effects of the LCDMREDUCED model appeared to decrease as the number of respondents increased. Mean bias plots of the intercept and *e* parameter showed clearer patterns than those of the main and interaction effects. Mean bias values for the intercept parameter ranged from 0.001 to 0.194 (see **Supplementary Tables 10–12**). Mean bias values for main effects ranged between 0.001 and 0.232. Mean bias values for the *e* parameter ranged from 0.002 to 0.718 (see **Supplementary Tables 10–12**). Mean bias values for interaction effects were between 0.002 and 0.750.

When the data generating model was the LCDMREDUCED, mean bias values for the intercept, main, *e* parameter, and interaction effects in the 12-item conditions were higher than in the 24- and 36-item conditions. The 36-item conditions yielded smaller bias values than the 24-item conditions for more than half of these conditions. Mean bias values showed less clear patterns with respect to base rate and number of attributes.

Overall, the highest mean bias values were obtained with the 5-attribute × 12-item conditions under both the 0.25 and 0.50 base rate conditions for intercept, main effects and *e* parameters. In addition, the 3-attribute × 36-item × 0.25 base rate conditions yielded higher RMSE values for sample sizes >400. The highest mean bias values for the interaction effect were for the 3-attribute × 12-item conditions under both 0.25 and 0.50 base rate conditions. There was no clear pattern of lowest mean bias values, however, for the intercept, main, *e* parameter, and interaction effects.

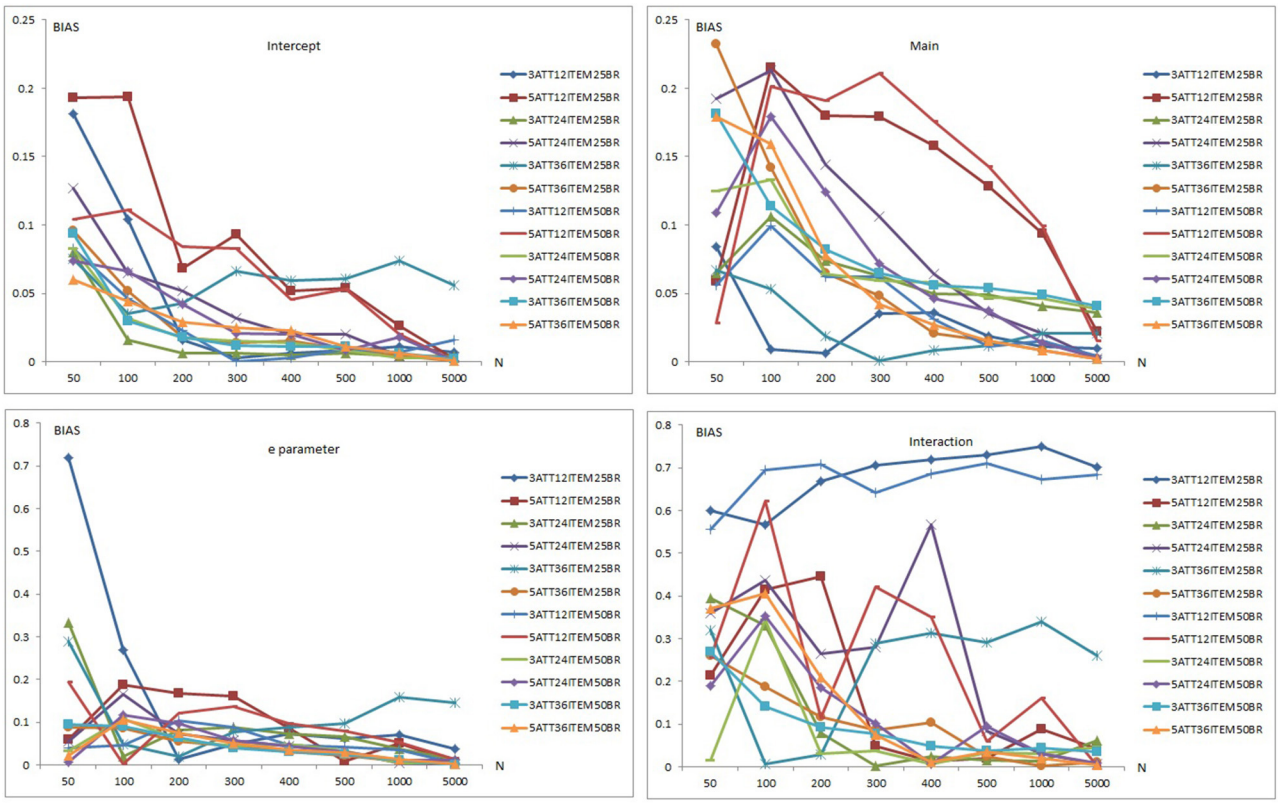


FIGURE 5 | Mean bias plots for the LCDMREDUCED Model.

It appears that the number of respondent, test length, number of attributes, and base rate of mastery had an impact on the recovery of LCDMREDUCED parameters. As the number of respondents and test length increased, the recovery of LCDMREDUCED parameter also improved.

Only the intercept parameter conditions produced RMSE values <0.10 when the sample size was 5,000. Main effect RMSE results were also very close to 0.10 when the sample size was 5,000. As can be seen in **Supplementary Tables 10–12**, RMSE values tended to be higher than 0.20 even with 1,000 respondents. Overall, the intercept parameter of the LCDMREDUCED model appeared to be recovered better than the other parameter based on mean RMSE and bias values, and recovery of the interaction effect appeared to be less accurate than for the other parameters.

A Linear Model Analysis of Item Recovery Statistics

Mean RMSE and bias results were summarized using a linear model. Effects of each of the different conditions for each of the RMSE and bias values were assessed using a factorial ANOVA. **Table 5** presents partial eta-squared values for each of the main effects and the two-way interactions. As can be seen in **Table 5**, sample size (N) was the most influential factor on RMSE and bias for each item parameter except for interaction bias. Test length (k) was the second most influential factor on RMSE and bias values. Base rate of mastery was the least influential factor on RMSE and bias calculated for intercept and the interaction parameters. Model type was the least influential factor on RMSE and bias calculated for the main effect parameters. Number of attributes was the least influential factor on RMSE and bias for the e parameter. Effects of two-way interactions between simulated factors appeared to be less than for main effects (see **Table 5**). Most of the main and interaction effects were found to have significant effects on RMSE and bias values.

Classification Accuracy Results

Figure 6 presents the classification accuracy results for C-RUM, DINA, DINO, and LCDMREDUCED models when model-data fit holds. Classification accuracy percentages appear to increase for the 12-item conditions (i.e., conditions 3ATT12ITEM25BR, 3ATT12ITEM50BR, 5ATT12ITEM25BR, 5ATT12ITEM50BR) as the number of respondents increases. The classification accuracy percentages appear to change only slightly for the 24- and 36-item conditions across the different sample sizes.

For the C-RUM, classification accuracy percentages ranged from 21.640 to 70.344 (see **Supplementary Table 13**). When the data generating model was the C-RUM, classification accuracy results were very close for the 24- and 36-item conditions. Neither the number of attributes nor the base rate, however, appeared to have a significant effect on the classification accuracy for the 24- and 36-item conditions. The sample size increase also appears to have slight effect on the classification

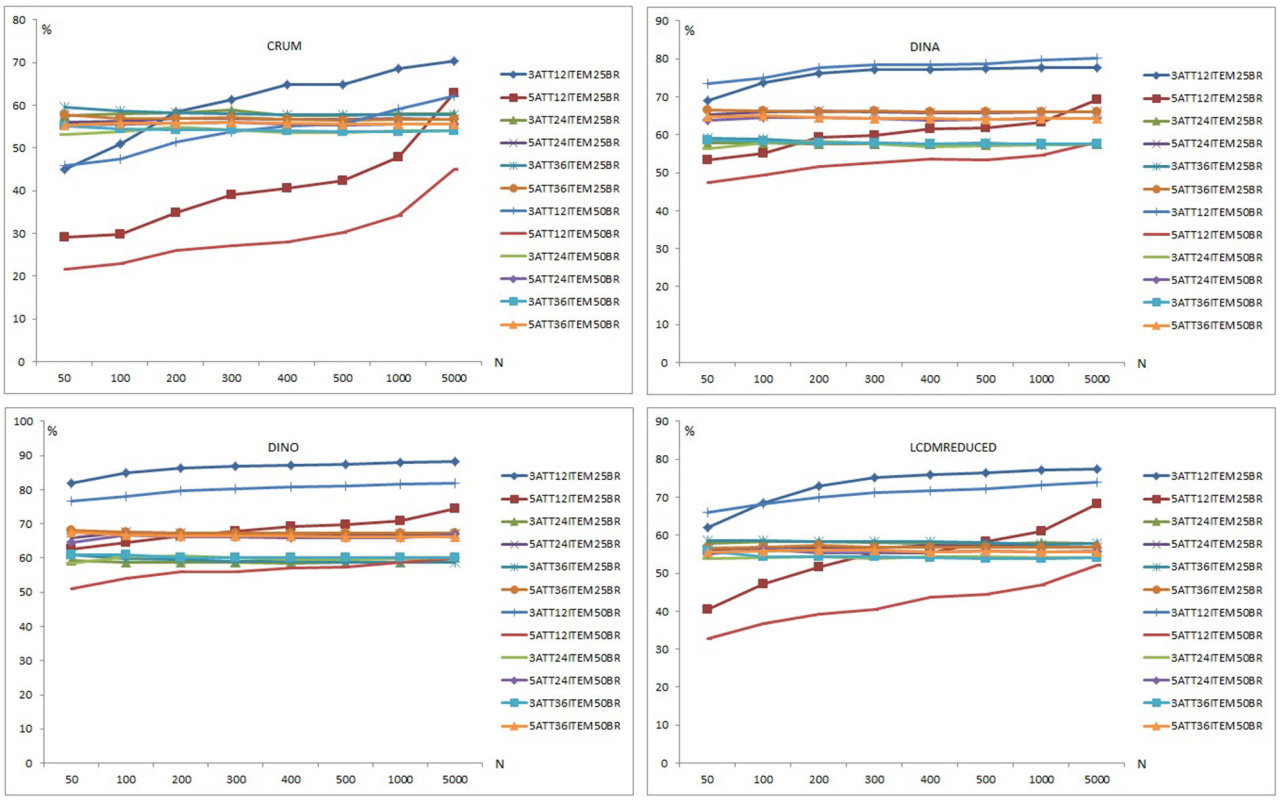


FIGURE 6 | Classification accuracy plots for the C-RUM, DINA, DINO, and LCDMREDUCED Models.

accuracy for the 24- and 36-item conditions. C-RUM results varied for the 12-item conditions. The 5-attribute × 12-item × 0.50 base rate conditions had the lowest percentages (see **Figure 6**). The 3-attribute × 12-item × 0.25 base rate conditions produced the highest percentages for the C-RUM for sample sizes >200.

For the DINA model, classification accuracy percentages ranged from 47.480 to 80.210 (see **Supplementary Table 13**). For the DINA generating model, classification accuracy results for the 24- and 36- item conditions were very close. The 5-attribute conditions had higher percentages than the 3-attribute conditions for the 24- and 36-item conditions. The increase in sample size appeared to have a slight effect on the classification accuracy for the 24- and 36-item conditions. However, DINA model results varied for the 12-item conditions. The 5-attribute × 12-item × 0.50 base rate conditions produced the lowest percentages for the DINA model. The 3-attribute × 12-item conditions for both 0.25 and 0.50 base rates produced the highest percentages. Results for the 12-item conditions appear to be stable for sample sizes of 200 or more.

For the DINO model, classification accuracy percentages ranged from 51.00 to 88.27 (see **Supplementary Table 13**). When the data generating model was the DINO model, classification accuracy results were similar for the 24- and 36-item conditions. The 5-attribute conditions had higher percentages than the 3-attribute conditions for both 24- and 36-item conditions. The sample size increase appears to have a slight effect on the classification accuracy under the 24- and 36-item conditions. The 5-attribute × 12-item × 0.50 base rate conditions had the lowest percentages for the DINO model. The 3-attribute × 12-item × 0.25 base rate conditions produced the highest percentages. DINO model results for 12 items appeared to be more stable for samples of 200 or more.

For the LCDMREDUCED, classification accuracy percentages ranged from 32.820 to 77.456 (see **Supplementary Table 13**). When the data generating model was the LCDMREDUCED, classification accuracy results were similar between the 24- and 36-item conditions. The 5-attribute conditions had higher percentages than the 3-attribute conditions for both 24- and 36-item conditions. The sample size increase appears to have a slight effect on the classification accuracy under the 24- and 36-item conditions. LCDMREDUCED results were variable for the 12-item conditions. The 5-attribute × 12-item × 0.50 base rate conditions had the lowest percentages for the LCDMREDUCED, and the 3-attribute × 12-item × 0.25 base rate conditions had the highest percentages for LCDMREDUCED model.

The number of respondents, number of attributes, and mastery base rate had an impact on the classification accuracy of DCMs in the 12-items conditions. When all models were

compared, the highest classification was observed with the DINO model (the mean across all conditions = 66.35). The models with the next highest classification percentages after the DINO model were the DINA, LCDMREDUCED and C-RUM models, respectively. The average classification percentages across all conditions were 63.16, 57.54, and 52.82 for the DINA, LCDMREDUCED, and C-RUM models, respectively.

DISCUSSION

The present simulation study was designed to investigate the effects of sample size on item parameter recovery and classification accuracy of four DCMs, the C-RUM, DINA, DINO, and LCDMREDUCED. Effects of additional factors including test length, number of attributes, and base rate of mastery were also examined. Bias and RMSE values were computed between true generating parameters and estimated parameters. Effects of simulated factors on attribute assignment were also evaluated using the percentage of classification accuracy.

The present study differed from previous studies (Rojas et al., 2012; Başokcu, 2014) in several respects. Although previous simulations on DCMs showed that classification accuracy and item recovery can be poor with small sample sizes, they tended to focus on a limited number of sample size conditions, making results somewhat difficult to generalize to other practical testing conditions. This study extended the sample size conditions from 50 to 5,000. Results showed that sample size appears to have an impact on recovery of DCM model parameters. Larger sample sizes showed better item parameter recovery. The effect of sample size on item parameter recovery is consistent with previous research (Rojas et al., 2012; Başokcu, 2014). Conditions with sample sizes <200 showed poor results. In general, it appears that sample sizes should be at least 500 for the four DCMs considered in this study in order to obtain precise estimates. This is consistent with previous research in which a sample size of 500 was considered to be a small sample size for DCMs (Bradshaw and Madison, 2016; Madison and Bradshaw, 2018), although small RMSE and bias values were difficult to obtain with samples of 1,000 respondents under some conditions. The results of this simulation study showed that a sample size as small as $N = 1{,}000$ would be sufficient to adequately recover all model parameters, under all the given conditions, adequately for the DINA, DINO and C-RUM models. However, the LCDMREDUCED model does appear to require larger sample sizes for some model parameters such as the interaction effect.

Another important finding from this study is that increase in test length did result in more precise estimates of item parameters. The average RMSE and bias values decreased as test length increased from 12 to 36 items. This finding is consistent with previous research indicating that the CDM framework requires assessments that are at least of moderate test lengths of 15 or 20 items (de la Torre, 2009).

Another important finding obtained is the effect of number of attributes on the recovery of item parameter estimates. The recovery of item parameters worsened when the number of attributes increased from three to five. This is important as the most of the studies in DCM literature use more than three attributes. For example, Sessoms and Henson (2018) has conducted a literature review on the applications of DCMs and found that number of attributes estimated varied from four to 23. The average number of attributes estimated was 8 and almost half of the application studies modeled 8 or more attributes. As shown in this simulation study, higher numbers of attributes clearly required larger sample sizes. Results of this study, in other words, suggest that tests with large numbers of attributes also need larger sample sizes to accurately estimate model parameters.

Another important point to be noted is the distribution of attributes over items. In this study, the distribution of attributes over items was not equal for the five-attribute case. This does have an effect on estimation. For instance, Madison and Bradshaw (2015) compared performance of the LCDM using various Q-matrix designs and found that classification accuracy varied markedly for different Q-matrix designs. For a given number of items an attribute is measured, classification accuracy increased as the number of items measuring the attribute in isolation increased. In contrast, classification accuracy suffered most when a pair of attributes was measured. In this study, the same numbers of attributes over items were used in the LCDMREDUCED model for the four models that were estimated (C-RUM, DINA, DINO, and LCDM). It is important to note that operational tests are constructed to meet the requirements of test blueprints so may not have this type of regular pattern. As is the case for any simulation study, the generalizability of the results of this study is necessarily limited to conditions manipulated in this study. It would be helpful, in this regard, to study the effects of different patterns of attributes in future research. Among the limitations of this study are the lacks of consideration of a general model (e.g., LCDM) and of the number of attributes larger than five.

This simulation study also showed that mastery base rate had a varying effect on item parameter recovery. This effect varied for item parameter types (i.e., intercept, main, e parameter, and interaction effects) and model type. It appears that intercept and *e* parameters were better recovered than main and interaction parameters. Consistent with previous research (Kunina-Habenicht et al., 2012; Bradshaw and Madison, 2016), the recovery of the interaction terms was lower than for the intercepts, main effects, and *e* parameters.

Parameters of DINA and DINO models were more likely to be recovered well than C-RUM and LCDMREDUCED models when model-data fit holds. This was also the case for classification accuracy. The DINO model had better fit with small sample sizes than the other three DCMs. This result is consistent with previous research (Roussos et al., 2007; de la Torre, 2011). Previous simulation studies generated data sets under the assumption of a common underlying model for the whole test. The simulation in the present study also considered different underlying model for each model (i.e., the LCDMREDUCED model). The LCDMREDUCED model consisted of four different model structures including the C-RUM, DINA, DINO, and full LCDM. The item parameter recovery and classification accuracy of the LCDMREDUCED model was worse than for the other DCMs.

Patterns of bias and RMSE values computed between true (i.e., generating) parameters and estimated parameters were consistent for almost all conditions. However, some irregularities were observed in which reversals occurred in the results for bias. Previous research has also reported this finding in which different patterns occurred for bias compared to RMSE (Harwell, 2018). As Harwell has noted, it may be that average bias is masking important patterns in recovery accuracy compared to RMSE.

Consistent with previous research (de la Torre et al., 2010; Rojas et al., 2012), the sample size did not result in any change in classification accuracy percentages for the 24 and 36 item conditions. Higher attribute assignments, however, were observed with larger sample sizes in the 12-item conditions. Overall, the classification accuracy rates were below 90% even for the 5,000 sample size. Consistent with previous research (de la Torre et al., 2010; Kunina-Habenicht et al., 2012; Lei and Li, 2016), results from the present study showed that sample size explained only a small proportion of the variance in classification accuracy. Higher attribute assignments were observed with the 12-item conditions, compared to the 24- and 36-item conditions. This was expected as the longer test lengths provide more information on which the classifications can be based.

Classifications with the DINO and DINA models were more accurate than the C-RUM and LCDMREDUCED models. The simulation study results showed that the DINO and DINA models performed better in this regard for small samples than other two DCMs.

Several practical suggestions may be made from this study for researchers or practitioners who seek to design diagnostic tests from a DCM framework. Results of this study showed that simpler models were recovered better than more complex models. Thus, before drawing any conclusion based on a specific DCM, one might alternatively specify other appropriate DCMs, which can capture potential relationships among the attributes. Another finding of this study was the increasing accuracy of the recovery of the item parameters as sample sizes increased. For instance, to ensure model identifiability and consistent estimation, it is necessary to collect sufficient data (i.e., typically samples of 1,000 or more) that satisfy identifiability, when designing the diagnostic tests. It was also shown that longer tests produced more precise and consistent estimates. Results of this study showed varying effects of mastery base rate on item parameter recovery. It would be useful to explore this issue in future research.

The probability of making a correct classification and accurately recovering item parameters depends at least in part on the fit of the model to the data. In this study, model-data fit was assumed for each condition, as the generating and estimated models were the same. It is difficult to know in practice, however, whether the selected DCM is the best fitting model to real test data. As Ma (2020) has noted, the usefulness of DCMs depends on whether they can adequately fit the data. It is for this reason that fit indices play an important role in selecting the best fitting DCM. Several studies have been conducted to examine the performances of absolute fit (Hu et al., 2016) and relative fit indices (de la Torre and Douglas, 2008; Hu et al., 2016; Sen and Bradshaw, 2017) in the DCM framework. These studies can help guide practitioners to choose appropriate model fit indices to estimate model fit under various conditions.

Results of this study suggest that the precision of the parameter estimates and classification accuracy are a function not only of the sample size but also of test length, number of attributes, base rate of mastery, and model type. Selection of the appropriate DCMs needs to be guided in part by sample size.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

SS and AC contributed equally to the data analyses and reporting parts. Both authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.621251/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

Akbay, L. (2016). Relative efficiency of the nonparametric approach on attribute classification for small sample cases. *J. Eur. Educ.* 6, 19–35. doi: 10.18656/jee.58934

Başokcu, T. O. (2014). Classification accuracy effects of Q-matrix validation and sample size in DINA and G-DINA models. *J. Educ. Pract.* 5, 220–230. doi: 10.7176/JEP

Bozdogan, H. (1987). Model selection and akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/BF02294361

Bradshaw, L. P., and Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *Int. J. Testing* 16, 99–118. doi: 10.1080/15305058.2015.1107076

Chen, H., and Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Lang. Assess. Q.* 13, 218–230. doi: 10.1080/15434303.2016.1210610

Chiu, C. Y., and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classif.* 30, 225–250. doi: 10.1007/s00357-013-9132-9

Choi, H. J., Templin, J. L., Cohen, A. S., and Atwood, C. H. (2010). "The impact of model misspecification on estimation accuracy in diagnostic classification models," in *Paper Presented at the Meeting of the National Council on Measurement in Education* (Denver, CO).

Cui, Y., Gierl, M. J., and Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Educ. Meas.* 49, 19–38. doi: 10.1111/j.1745-3984.2011.00158.x

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Appl. Psychol. Meas.* 33, 163–183. doi: 10.1177/0146621608320523

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 Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika* 81, 253–273. doi: 10.1007/s11336-015-9467-8

de la Torre, J., and Douglas, J. A. (2004). A higher-order latent trait model for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

de la Torre, J., and Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. *Psychometrika* 73, 595–624. doi: 10.1007/s11336-008-9063-2

de la Torre, J., Hong, Y., and Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *J. Educ. Meas.* 47, 227–249. doi: 10.1111/j.1745-3984.2010.00110.x

de la Torre, J., and Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *J. Educ. Meas.* 47, 115–127. doi: 10.1111/j.1745-3984.2009.00102.x

Galeshi, R., and Skaggs, G. (2016). Parameter recovery of a cognitive diagnostic model: evidence from a simulation study. *Int. J. Quant. Res. Educ.* 3, 223–241. doi: 10.1504/IJQRE.2016.082386

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26, 333–352. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hallquist, M. N., and Wiley, J. F. (2018). Mplus automation: an R package for facilitating large-scale latent variable analyses in Mplus. *Struct. Equation Model. A Multidiscipl. J.* 25, 621–638. doi: 10.1080/10705511.2017.1402334

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, IL, United States.

Harwell, M. (2018). A strategy for using bias and RMSE as outcomes in monte carlo studies in statistics. *J. Modern Appl. Stat. Methods* 17:eP2938. doi: 10.22237/jmasm/1551907966

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Hu, J., Miller, M. D., Huggins-Manley, A. C., and Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnosis models. *Int. J. Testing* 16, 119–141. doi: 10.1080/15305058.2015.1133627

Im, S., and Yin, Y. (2009). Diagnosing skills of statistical hypothesis testing using the rule space method. *Stud. Educ. Eval.* 35, 193–209. doi: 10.1016/j.stueduc.2009.12.004

Jang, E. E., Dunlop, M., Park, G., and van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Lang. Testing* 32, 359–383. doi: 10.1177/0265532215570924

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

Kunina-Habenicht, O., Rupp, A. A., and Wilhem, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *J. Educ. Meas.* 49, 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Lei, P. W., and Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Appl. Psychol. Meas.* 40, 405–417. doi: 10.1177/0146621616647954

Ma, W. (2020). Evaluating the fit of sequential G-DINA model using limited-information measures. *Appl. Psychol. Meas.* 44, 167–181. doi: 10.1177/0146621619843829

Madison, M. J., and Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educ. Psychol. Meas.* 75, 491–511. doi: 10.1177/0013164414539162

Madison, M. J., and Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika* 83, 963–990. doi: 10.1007/s11336-018-9638-5

Muthén, L. K., and Muthén, B. O. (1998–2019). *Mplus (Version 8.4) [Computer Software]*. Los Angeles, CA: Muthén and Muthén.

Paulsen, J. (2019). *Examining cognitive diagnostic modeling in small sample contexts* (Unpublished doctoral dissertation). Indiana University, Bloomington, IN, United States.

Rojas, G., de la Torre, J., and Olea, J. (2012). "Choosing between general and specific cognitive diagnosis models when the sample size is small," in *Paper Presented at the Meeting of the National Council on Measurement in Education* (Vancouver).

Roussos, L. A., Templin, J. L., and Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *J. Educ. Meas.* 44, 293–311. doi: 10.1111/j.1745-3984.2007.00040.x

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Rupp, A. A., and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Measurement* 6, 219–262. doi: 10.1080/15366360802490866

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

Sen, S., and Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Appl. Psychol. Meas.* 41, 422–438. doi: 10.1177/0146621617695521

Sen, S., and Terzi, R. (2020). A comparison of software packages available for DINA model estimation. *Appl. Psychol. Meas.* 44, 150–164. doi: 10.1177/0146621619843822

Sessoms, J., and Henson, R. A. (2018). Applications of diagnostic classification models: a literature review and critical commentary. *Meas. Interdiscipl. Res. Perspect.* 16, 1–17. doi: 10.1080/15366367.2018.1435104

Sünbül, S. Ö., and Kan, A. (2016). Bilişsel tani modellerinde parametre kestirimini ve siniflama tutarliligini etkileyen faktörlerin incelenmesi. *H. U. J. Educ.* 31, 778–795. doi: 10.16986/HUJE.2015014663

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

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. Meas.* 32, 37–50. doi: 10.1111/emip.12010

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

Tzou, H., and Yang, Y. H. (2019). Improved performance of model fit indices with small sample sizes in cognitive diagnostic models. *Int. J. Assess. Tools Educ.* 6, 154–169. doi: 10.21449/ijate.482005

von Davier, M. (2005). *A General Diagnostic Model Applied to Language Testing Data* (ETS Technical Report No. RR-05–16). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2014). *The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM) (RR-14-40)*. Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12043

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., et al. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assess. Educ.* 26, 466–488. doi: 10.1080/0969594X.2019.1586642

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 Sen and Cohen. 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: 02 February 2021
doi: 10.3389/fpsyg.2021.623077

Test Assembly for Cognitive Diagnosis Using Mixed-Integer Linear Programming

Wenyi Wang[1], Juanjuan Zheng[1], Lihong Song[2]*, Yukun Tu[1] and Peng Gao[1]*

[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, [2] School of Education, Jiangxi Normal University, Nanchang, China

One purpose of cognitive diagnostic model (CDM) is designed to make inferences about unobserved latent classes based on observed item responses. A heuristic for test construction based on the CDM information index (CDI) proposed by Henson and Douglas (2005) has a far-reaching impact, but there are still many shortcomings. He and other researchers had also proposed new methods to improve or overcome the inherent shortcomings of the CDI test assembly method. In this study, one test assembly method of maximizing the minimum inter-class distance is proposed by using mixed-integer linear programming, which aims to overcome the shortcomings that the CDI method is limited to summarize the discriminating power of each item into a single CDI index while neglecting the discriminating power for each pair of latent classes. The simulation results show that compared with the CDI test assembly and random test assembly, the new test assembly method performs well and has the highest accuracy rate in terms of pattern and attributes correct classification rates. Although the accuracy rate of the new method is not very high under item constraints, it is still higher than the CDI test assembly with the same constraints.

Keywords: cognitive diagnosis, cognitive diagnostic model information index, cluster analysis, mixed-integer linear programming, inter-class distance, correct classification rate

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Chunhua Kang,
Zhejiang Normal University, China
Miguel A. Sorrel,
Autonomous University of Madrid, Spain

****Correspondence:***
Wenyi Wang
wenyiwang@jxnu.edu.cn
Lihong Song
viviansong1981@163.com

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

Received: *29 October 2020*
Accepted: *11 January 2021*
Published: *02 February 2021*

Citation:
Wang W, Zheng J, Song L, Tu Y and Gao P (2021) Test Assembly for Cognitive Diagnosis Using Mixed-Integer Linear Programming. Front. Psychol. 12:623077. doi: 10.3389/fpsyg.2021.623077

INTRODUCTION

The theory of cognitive diagnostic assessment (CDA) is an important part of personalized adaptive learning (Sia and Lim, 2018). Since the cognitive diagnostic model (CDM) was put forward, it has attracted much attention because of its ability to analyze and explain the test results in detail (Hsu et al., 2020). On the other hand, the test is the bridge between the abstract and unobservable ability of the examinees and the real observable item response data, so the quality of the test affects the quality of diagnostic classification directly. A test that meets the test specification needs to be selected from an item bank, then the test assembly will be restricted by many conditions and requirements (Zijlmans et al., 2019; Tang and Zhan, 2020), such as the difficulty and discrimination under the constraints of psychometrics, the maximum number of knowledge points allowed in a test, or the requirements of parallel tests.

How to construct a test with higher quality has always been a research hotspot. In the aspect of test assembly based on cognitive diagnosis, the test assembly method of CDM information index (CDI) proposed by Henson and Douglas (2005) is of great influence. Henson et al. (2008) put forward the attribute level discrimination index (ADI) under uniform and non-uniform

distribution of attributes. However, neither the CDI method nor the ADI method considers the attribute hierarchical structure. When these methods are applied in practice, the performance of CDI and ADI methods will be poor under some conditions if the hierarchical structure exists between attributes (Kuo et al., 2016). In addition, Finkelman et al. (2009) proposed a method of test assembly based on genetic algorithm minimizing the expected posterior error rate for attributes under the framework of CDA. For example, the method of test assembly based on genetic algorithm takes three fitness functions: the average number of classification errors, maximum error rate, and ability to hit attribute-level target error rates. This method can directly optimize the classification errors, but its computational intensity is considerably greater than that of analytic procedures like the CDI. For classroom or formative assessment, we should choose the algorithm with low computational complexity if other algorithms for test assembly are sufficient to meet the needs (Clark, 2013).

In terms of test assembly methods based on cognitive diagnosis, researchers have proposed a large number of methods, but most of these methods are based on a certain CDI, and there are some problems such as lacking of global consideration or requiring large amount of computation. Therefore, it is urgent to consider the global information and the method of test assembly with less calculation in cognitive diagnosis. The method of test assembly based on CDI takes into account the sum of the whole amount of information, but it has been found that this method is not the optimal method of test assembly. In some cases, the total amount of information is the largest, which may due to some of the larger information on a non-trivial subset of the universal set of latent classes (i.e., the set of all possible combinations of attributes). The discriminating power of this strategy with the largest CDI is not necessarily better than the strategy with uniform distribution of information and less overall information. Therefore, the goal of this study is to explore a new method for test construction, and combine the idea of cluster analysis (Guo et al., 2020) and mixed-integer linear programming method (Kantor et al., 2020) to propose a method to maximize the minimum distance (MMD) between latent classes, in order to overcome the shortcomings of the existing methods.

METHODS

Cognitive Diagnostic Model

The purpose of cognitive diagnostic model is to describe the relationship between examinee's item response and his or her potential cognitive attributes (Mao, 2014). It is a psychometric model. The common cognitive diagnostic models are the deterministic input noisy "and" gate (DINA) model, the deterministic input noise "or" gate (DINO) model, and the reduced-reparameterized unified model (R-RUM; Hartz, 2002). The new method proposed in this study mainly focused on these two cognitive diagnosis models. Let K be the number of attributes to be measured by the test. The entry q_{jk} in the Q-matrix indicates whether the attribute k is measured in item j. When $q_{jk} = 1$, the attribute k is measured by item j. And 0 indicates that it has not been measured. α_{ik} indicates the attribute status of examinee i, that is, 1 indicates examinees' mastery of attribute k, and otherwise 0.

The DINA model is a completely non-compensatory model, which requires that the examinees must master all the attributes required by the item for correctly answering. As long as any one of them is not mastered, it will lead to a wrong answer or a very low probability of correct answer. For the value of the ideal response $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q^{jk}}$, a value of 1 indicates that the examinee i has mastered all the attributes measured by the item j, while a value of 0 means that the examinee has not fully mastered the attributes measured by the item j. The corresponding probabilities of correct answer to this item are $(1 - s_j)$ and g_j respectively. The formula of DINA model is as follows (Junker and Sijtsma, 2001)

$$\mathrm{P}\left(X_{ij} = 1|\alpha_i\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{(1-\eta_{ij})}. \tag{1}$$

The DINO model is the compensatory model. As long as the examinees have mastered any of the attributes measured by the item, they can have a higher probability of correctly answering. For the value of the ideal response $\varpi_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}$ is 1, it means that the examinees have mastered at least one attribute measured by item j. A value of 0 indicates that the examinees have not mastered all the attributes of item j. The formula of the DINO model is as follows (Templin and Henson, 2006)

$$\mathrm{P}\left(X_{ij} = 1|\alpha_i\right) = (1 - s_j)^{\varpi_{ij}} g_j^{(1-\varpi_{ij})}. \tag{2}$$

where s_j is the slip probability for the examinees of the ideal response with value 1 on item j, and g_j is the guessing probability for the examinees with value 0 on item j.

Like the DINA model, R-RUM is a non-compensatory model, which is a simplified unified model of reparameterization. The baseline parameter π_j^* indicates the positive response probability for examinees who have mastered all the attributes required by item j. The values are all between 0 and 1. The penalty parameter r_{jk}^* for not possessing the kth attribute is defined at the level of interaction between the item and the attributes and reflects the importance of attribute k on item j. The formula of R-RUM is as follows (Hartz, 2002)

$$P\left(X_{ij} = 1|\alpha_i\right) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{ik})q_{jk}}. \tag{3}$$

For simplify, the correct answer probability $\mathrm{P}\left(X_{ij} = 1|\alpha_i\right)$ is denoted by $P_j(\alpha_i)$, where α_i is the knowledge state of the examinee i.

Kullback-Leibler Information Distance Between Classes

Considering the existing cognitive diagnosis item bank, attribute vectors of all items in the item bank have been specified (Wang et al., 2020), and the parameters of each item have been estimated

by the parameter estimation algorithm of cognitive diagnosis model. The correct answer probability $P_j(\alpha_c)$ of knowledge state α_c on item j can be calculated by item attribute vector q_j and item parameters, where α_c is a knowledge state of the examinees and is an element of the universal set of latent classes. Let M be the size of the item bank. Kullback-Leibler (K-L; Cover and Thomas, 2006; Debeer et al., 2020) information quantity or K-L distance is the most commonly used to measure the distance between any two probability distributions $P_j(\alpha_u)$ and $P_j(\alpha_v)$ for two knowledge states α_u and α_v. Formally, item j is defined as the K-L distance of the item response probability distributions under the knowledge states of α_u and α_v

$$D_{K-L}(\alpha_u, \alpha_v, \mathrm{j}) = P_j(\alpha_u) \log\left[\frac{P_j(\alpha_u)}{P_j(\alpha_v)}\right] + (1 - P_j(\alpha_u)) \log\left[\frac{(1 - P_j(\alpha_u))}{(1 - P_j(\alpha_v))}\right]. \quad (4)$$

In fact, D_{K-L} is the expectation of the function of the logarithmic likelihood ratio of probability distributions $P_j(\alpha_u)$ and $P_j(\alpha_v)$. Although this amount of information is called the distance between the two distributions, and it does have statistical significance for distance measurement, that is, with the increase of D_{K-L}, it is easier to distinguish the two distributions statistically (Rao, 1962). But it is not symmetrical, that is, $D_{K-L}(\alpha_u, \alpha_v) \neq D_{K-L}(\alpha_v, \alpha_u)$.

Kullback-Leibler distance is often used for computer adaptive testing or cognitive diagnostic computer adaptive testing. For instance, Chang and Ying (1996) firstly suggested that K-L distance instead of Fisher information should be used as a more effective item selection index in computer adaptive testing based on one-dimensional IRT model. Madigan and Almond (1995) use K-L distance for test selection strategy of belief networks. Tatsouka and Ferguson (2003) use K-L distance and Shannon entropy for sequential item selection and use it in cognitive diagnostic computer adaptive testing. Different from the amount of Fisher information, the K-L distance does not require that the parameter space must be continuous, so it is suitable for CDM where the attribute pattern is discrete.

Test Assembly Using Mixed-Integer Linear Programming

In cognitive diagnosis, the probability of correct answer or the expected vector of item response of knowledge state α_c on test length of J in a test is $\mathrm{P}(\alpha_c) = (P_1(\alpha_c), P_2(\alpha_c), \ldots, P_J(\alpha_c))$. For knowledge state α_c, the $\mathrm{P}(\alpha_c)$ can be regarded as the center of the class. In pattern recognition or clustering methods, the method of maximum distance between classes can usually be used for classification. If the cognitive diagnostic test can maximize the distance between the class centers of all potential classes $\alpha_c \in Q_s$, where Q_s is the universal set of latent classes, it is easier to classify knowledge states. It is just like in a jigsaw puzzle, if there is a big difference between the sub-images, the difficulty of completing the puzzle will be correspondingly lower.

In order to characterize the distinguishing power of item j to knowledge states α_u and α_v, the following is D_{K-L} as its metric index. For any α_u and α_v, the discrimination power matrix or K-L distance matrix $D_j = (D(\alpha_u, \alpha_v, j))$ is obtained. If the cardinality (i.e., the number of elements) of Q_s is T, we know that the number of rows or columns is T in D_j. In order to use the mixed-integer linear programming for test construction, it is necessary to vectorize the matrix D_j into a single stacked column vector. That is, the sequence of rows in this matrix is composed of a long vector, and then transpose the row vector to get the stacked column vector, which is denoted as $V_j = Vec(D_j)$. When the matrix D_j is vectorized, we remove the main diagonal elements because these values are zeros. For each item in the item bank, V_j can be calculated, and the matrix $V = (V_1, V_2, \ldots, V_M)$ composed of all the items can be obtained, where M is the number of items in the item bank. Based on the mixed-integer linear programming model, we will give a linear programming model which takes into account the mean value of the distance between all classes and maximizes the minimum distance between classes:

$$\min\left(f_1 x + f_2 y\right), \quad (5)$$

Subject to

$$Vx + \mathrm{y} \geq b,$$

$$1^T x = J,$$

$$x_j \in \{0, 1\}, \mathrm{j} = 1, 2, \ldots, \mathrm{M},$$

$$\mathrm{y} \in \mathrm{R}.$$

Among them, $f_1 = (f_{11}, f_{12}, \ldots, f_{1M})^T$, where $f_{1j} = -\sum_{v=1}^{TT-1} V_{vj}/(T(\mathrm{T}-1))$. The negative of f_{1j} is used to convert a maximization problem into a minimization one. Here, $f_2 = J \sum_{u=1}^{T(T-1)} \sum_{v=1}^{M} V_{uv}/(\mathrm{TM}(\mathrm{T}-1))$ is the weight of y, $x = (x_1 x_2 \ldots x_M)^T$, where $x_1 x_2 \cdots x_M$ is the 0-1 vector in the decision vector of linear programming, and the value of the x_j indicated whether the test contains the item j. If $x_j = 1$, it means that the test contains the item j, otherwise it does not include the item j, $\mathrm{b} = (b_1, b_2, \ldots, b_{T(T-1)})^T$ represents the lower limit of K-L distance for all pairs of knowledge states. You can set the bounded distance $b_t = J \sum_{j=1}^{M} \left(V_{tj}\right)/M$, which is the average value of the distance between classes of J items in the item bank. $1^T x = J$ represents the test length constraint, where 1^T is a M-dimensional column vector with all elements 1, and J is the test length. y captures the difference between the t-th pair inter-class distance $V_{(t)}x$ and the target distance b_t, where $V_{(t)}$ is row t in V. Then, adding y to the constraint condition, and adding $f_2 y$ to the objective function, is to maximize the minimum inter-class distance y. For example, if the components in $\boldsymbol{b}$ are equal, and $V_{(t)}x$ is the smallest of all the distances between classes, if $V_{(t)}x < b_t$, then $V_{(t)}x$ can at least add $b_t - V_{(t)}x$ to satisfy the constraint. Because the average distance between other classes is larger than $V_{(t)}x$, $V_{(t)}x$ needs

to add $b_t - V_{(t)}x$ to reach the constraint. And minimizing f_2y in the objective function is minimizing $f_2(b_t - V_{(t)}x)$.Because f_2y is positive and b_t fixed, that is maximizing $V_{(t)}x$ which is the minimum inter-class distance between classes. In the objective function, we also consider the f_1x, linear programming model at the same time, that is, to maximize the distance between all classes, because the model also contains 0-1 vector x and real vector y, so this linear programming model is a typical mixed-integer or mixed 0-1 linear programming model, which can be solved by intlinprog function in Matlab2015a. For the source codes, we provided a user-friendly code in MATLAB into a public repository at the website: https://github.com/JXNU-EduM/MMD-Test-Assembly-for-CD/.

Simplify of K-L Distance Matrix

The distance index D_{K-L} in this study needs to be calculated for mixed-integer linear programming, so it is necessary to process the distance index matrix with vectorizing, transposing and merging. In the case of no hierarchical structure of attributes, there are $T = 2^K$ possible mastery modes for K attributes, and there are M items in the item bank. The size of the distance matrix of M items on the 2^K attribute mastery patterns after vectorizing, transposing and merging is $M*2^K(2^K - 1)$. If M is 300 and K is 4, the size of the distance matrix is 300*240. Although the size of the matrix is within the acceptable range, the amount of calculation for mixed-integer linear programming is a little large, so if possible, the distance matrix should be simplified.

If the u-th row and v-column element in D_j is denoted by D_{juv}, and the corresponding element in V_j is denoted by V_{juv}.

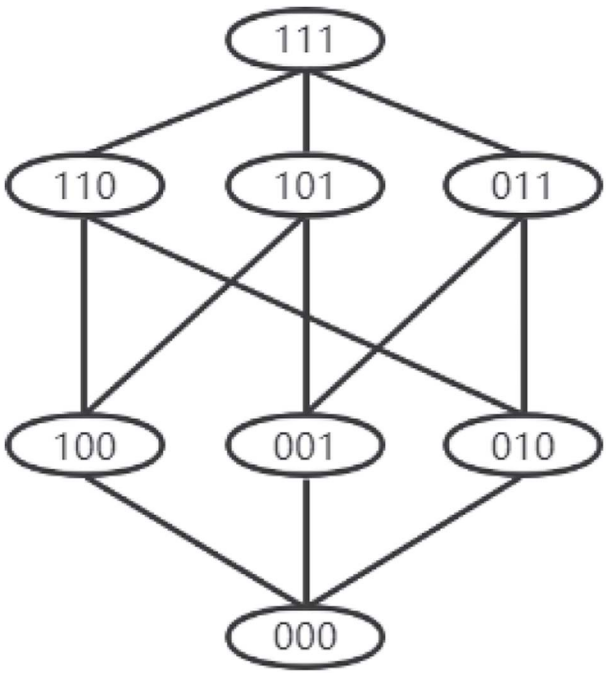


FIGURE 1 | Partial relation for eight possible attributes mastery patterns.

D_{juv} or V_{juv} is the discriminating power for these two different knowledge states of α_u and α_v, and one condition for the smallest difference between the two knowledge states is that there is a k-th attribute in the two attribute mastery patterns, which makes the k-th attribute mastery status of the two patterns different, and all mastery status except k are exactly the same. If only the discriminating power among attribute patterns with the least difference for the item is considered when vectorizing the distance matrix, the V_{juv} can be simplified. In the following, the distance matrix index corresponding to the simplified V_{juv} is recorded as SD_{K-L}. According to the characteristics of attribute patterns, we know that if the number of attributes is K and a certain attribute pattern is given, there are K attribute patterns with the least difference from it. Because of the asymmetry of the distance between α_u and α_v, that is, the D_{K-L} distance from α_u and α_v is different from that from α_v to α_u, both D_{juv} and D_{jvu} should be considered. If the number of attributes is three and the attributes are independent and without hierarchical structure, there are eight possible attribute mastery patterns, as shown in **Figure 1**: the difference of attribute patterns with connections between adjacent levels is the smallest. Thus, only 24 elements needed to be considered in D_j is obviously smaller than the number of non-diagonal elements in D_j, which can greatly save the computational cost.

STUDY DESIGN

Some main factors that may affect the efficiency of constructing test assembly should be considered: cognitive diagnosis model (the DINA model, the DINO model, and the R-RUM), attribute correlation coefficient (0 and 0.5), the number of examinees was fixed at 10000, the size of item bank was fixed at 300, the number of measured attributes was fixed at 4. Attribute correlation coefficient is zero, implying that the attributes were independent of each other, and the knowledge state was distributed evenly. Under each condition, the experiment was repeated for 200 times.

Assuming that the test measured K attributes, there are at most $2^K - 1$ possible item attribute vectors. First of all, all possible item attribute binary vectors were converted to decimal as $1, 2, \ldots, 2^K - 1$, and then 300 random integers in the range $[1, 2^K - 1]$ were randomly generated. Item attribute vectors of 300 items with corresponding numbers were selected to form the Q-matrix for an item bank. Item parameters of each item were randomly generated from specified distributions. The DINA and DINO models have the guessing and slip item parameters, which are randomly generated from a uniform distribution U (0.05, 0.4). Meanwhile the R-RUM also has the baseline and penalty parameters, which are respectively randomly generated from the uniform distribution U (0.75, 0.95) and U (0.2, 0.95). These were the same as the experimental design of Henson and Douglas (2005).

When the examinees are simulated, two aspects need to be considered: one is attribute mastery status α_{ki} at the k-th attribute for the i-th examinee and the other is the correlation coefficient between attributes, denoted by ρ. Multivariate normal

distribution can be used to simulate latent ability, $\tilde{\alpha}_i \sim MVN(0, \Sigma_{K*K})$, where **0** is the zero vector with the length of K and Σ is the correlation matrix

$$\Sigma = \begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix}. \quad (6)$$

In this study, the value of ρ is 0 (independent structure) or 0.5. After getting the value of $\tilde{\alpha}_{ki}$, we need to discretize it. The strategy of discretization of α_{ki} is

$$\alpha_{ki} = \begin{cases} 1 & if\ \tilde{\alpha}_{ki} \geq 0, \\ 0 & otherwise. \end{cases} \quad (7)$$

Two groups of 10000 examinees were simulated. One group of examinees was used to calculate the empirical distribution of knowledge state, which will be applied as the prior distribution for compute the posterior mode in the classification of the other group. We have not changed this condition for the repetition of the study of Henson and Douglas (2005). If a lager sample is available for the calibration of item bank, the empirical distribution of attribute patterns from the large sample will be applied as the prior distribution to computing the posterior mode in the classification of examinees who have taken the tests constructed from the calibrated item bank.

For a set of given attribute mastery pattern, $PX_{ij} = 1|\alpha_i$ depending on the selected model is the probability of correct response to item j for examinee i with attribute mastery pattern α_i. We supposed u was randomly generated from a uniform distribution U (0, 1). The item response of the ith examinee on item j can be obtained by discretizing the probability matrix

$$X_{ij} = \begin{cases} 1\ if\ u \leq P\left(X_{ij} = 1|\alpha_i\right) \\ 0 \qquad otherwise. \end{cases} \quad (8)$$

Since the item parameters were known, the examinees' item responses on the selected items could be simulated, and then the examinees were classified by maximum posterior estimation, and then attribute correct rate (ACR) and pattern correct rate (PCR) could be calculated. The formulas of ACR and PCR are as follows

$$ACR = \frac{1}{NK} \sum_{k=1}^{K} \sum_{i=1}^{N} I\left(\alpha_{ik} = \hat{\alpha}_{ik}\right), \quad (9)$$

and

$$PCR = \frac{1}{N} \sum_{i=1}^{N} I\left(\alpha_i = \hat{\alpha}_i\right). \quad (10)$$

In the above two expressions, N and K represent the number of examinees and the number of attributes, respectively, and $I(x = y)$ is an indicative function, which is defined as follows: when $x = y$, $I(x = y) = 1$, otherwise it is 0. The attribute correct rate (ACR) is the proportion of examinees whose estimated attribute status is equal to the simulated or true attribute status, while the pattern correct rate (PCR) is the proportion of examinees whose estimated attribute patterns is equal to the simulated or true attribute patterns. These two indices are commonly used in the simulation study for evaluating the correct classification rates for attributes or attribute patterns. The higher PCR and ACR for a test construct method implies that it yields considerably higher correct classification rates.

The D_{K-L} distance was used as the inter-class distance, and the mixed-integer linear programming is used to maximize the minimum inter-class distance with additional constraints. The test length is 20 for all test design. The first constraint was no constraint (No Constraints, NC), which directly used the greedy algorithm to construct test, and did not set any constraints based on the CDI or MDD. The second constraint was item-level constraint (Item Constraints, IC), which controls the number of items that measure a specific number of attributes for test assembly. According to the suggestion of Henson and Douglas (2005), among the 20 items that measure a total of 4 attributes, 9 items measured three attributes, 7 items measured two attributes, and the remaining 4 items measured one attribute. The third constraint was the attribute number constraint (Attribute Constraints, AC), which required that each attribute must be measured at least 7 times in a test with four attributes and 20 items.

STUDY 1: COMPARISON BETWEEN THE PROPOSED METHOD AND ITS SIMPLIFICATION

The proposed method uses mixed-integer linear programming to maximize the minimum inter-class distance between classes and comprehensively to consider the overall amount of information in order to achieve better test assembly quality. However, when the number of attributes measured was four, the calculation of the distance matrix D_{K-L} after vectorizing by the new method was a bit large, so when using the new method to construct test assembly, the distance matrix needs to be simplified. The test assembly method using the original and simplified matrices were denoted by D_{K-L} and SD_{K-L}, respectively. In fact, the simplification of the distance matrix will reduce the constraints of mixed-integer linear programming. The simplified matrix aims to discriminate similar attribute patterns, but whether it will lose the amount of information, if it is true, the size of the loss still needs to be verified.

Research Purpose

The purpose of this study is to verify whether the simplified distance matrix will lose information and lead to poor results. Since this study only considered the effect of simplified constraints on the efficiency of the MMD test assembly method, a single factor or one-way analysis of variance (ANOVA) can be performed on the two groups of ACR and PCR before and after the simplification in order to measure the impact of simplified constraints on ACR and PCR. In addition, the mean of ACR or PCR before and after simplification and the index of constructing test assembly time (in seconds) need to be taken into account.

TABLE 1 | Single factor analysis of variance for simplified and non-simplified constraints under the DINA model.

Correlation	Accuracy	Constraints	S_A	S_E	F	*p*-value
0	ACR	NC	5.3222E-06	0.0054	0.3942	0.5304
		IC	3.4223E-09	0.0168	0.0001	0.9928
		AC	6.7185E-06	0.0059	0.4556	0.5001
	PCR	NC	6.9139E-05	0.0604	0.4559	0.4999
		IC	9.8010E-07	0.0966	0.0040	0.9494
		AC	6.4883E-05	0.0654	0.3951	0.5300
0.5	ACR	NC	9.8533E-06	0.0052	0.7510	0.3867
		IC	4.4944E-08	0.0177	0.0010	0.9747
		AC	2.9618E-06	0.0050	0.2346	0.6284
	PCR	NC	1.2589E-04	0.0619	0.8099	0.3687
		IC	4.6923E-07	0.1320	0.0014	0.9700
		AC	3.4047E-05	0.0592	0.2290	0.6326

TABLE 2 | Single factor analysis of variance for simplified and non-simplified constraints under the DINO model.

Correlation	Accuracy	Constraints	S_A	S_E	F	*p*-value
0	ACR	NC	1.1516E-05	0.0060	0.7602	0.3838
		IC	1.2100E-10	0.0163	0.0000	0.9986
		AC	6.0639E-06	0.0060	0.4025	0.5262
	PCR	NC	1.0040E-04	0.0670	0.5961	0.4405
		IC	5.0625E-08	0.0939	0.0002	0.9883
		AC	6.6831E-05	0.0668	0.3985	0.5282
0.5	ACR	NC	5.8443E-06	0.0054	0.4343	0.5103
		IC	3.8813E-07	0.0154	0.0100	0.9202
		AC	2.2801E-06	0.0054	0.1688	0.6814
	PCR	NC	7.1234E-05	0.0658	0.4309	0.5119
		IC	2.6732E-06	0.1209	0.0088	0.9253
		AC	1.8966E-05	0.0654	0.1154	0.7343

TABLE 3 | Single factor analysis of variance for simplified and non-simplified constraints under the R-RUM model.

Correlation	Accuracy	Constraints	S_A	S_E	F	*p*-value
0	ACR	NC	3.0360E-07	0.0082	0.0147	0.9035
		IC	4.1598E-04	0.0114	14.4606	0.0002
		AC	8.2369E-08	0.0085	0.0039	0.9505
	PCR	NC	4.1209E-06	0.0854	0.0192	0.8898
		IC	1.3195E-03	0.0886	5.9273	0.0153
		AC	1.6512E-06	0.0882	0.0075	0.9313
0.5	ACR	NC	1.7222E-09	0.0051	0.0001	0.9908
		IC	1.9847E-04	0.0058	13.6922	0.0002
		AC	2.4602E-07	0.0051	0.0193	0.8896
	PCR	NC	2.3040E-07	0.0584	0.0016	0.9684
		IC	1.1219E-03	0.0507	8.8060	0.0032
		AC	3.2580E-06	0.0586	0.0221	0.8818

Experimental Steps

In order to achieve the purpose of this study, the experiment was designed according to the following steps:

(1) According to the design of Section 3 (four attributes were considered), we simulate two groups of examinees, in which one group was used to calculate the prior distribution, and the other group was used for classification. We simulate the Q matrix and item parameters in the item bank, and simulate the observed complete item response matrix of all examinees on all items in the item bank.
(2) Calculate the D_{K-L} distance and the simplified D_{K-L} distance of all items on all possible attribute mastery patterns.
(3) Choose the items according to the strategies of no restriction, attribute restriction and item restriction;
(4) Take out the response matrix of all the items on the corresponding test according to the test items generated by the test assembly algorithm;

TABLE 4 | Comparison of simplified and non-simplified constraints under the DINA model.

Correlation	Accuracy	Constraints	D_{K-L}	SD_{K-L}	SD_{K-L} Outperforms D_{K-L}	Time for D_{K-L} (seconds)	Time for SD_{K-L} (seconds)
0	ACR	NC	0.9798	0.9795	0.4500	1.9195	0.7373
		IC	0.9463	0.9463	0.9700	0.3661	0.1022
		AC	0.9796	0.9793	0.4250	1.9718	0.7971
	PCR	NC	0.9260	0.9251	0.4500	1.9195	0.7373
		IC	0.8356	0.8357	0.9800	0.3661	0.1022
		AC	0.9253	0.9245	0.4150	1.9718	0.7971
0.5	ACR	NC	0.9810	0.9807	0.4350	1.8395	0.7196
		IC	0.9489	0.9489	0.9800	0.3429	0.0909
		AC	0.9808	0.9806	0.4850	1.9187	0.7774
	PCR	NC	0.9292	0.9281	0.4400	1.8395	0.7196
		IC	0.8349	0.8350	0.9800	0.3429	0.0909
		AC	0.9286	0.9280	0.4800	1.9187	0.7774

TABLE 5 | Comparison of simplified and non-simplified constraints under the DINO model.

Correlation	Accuracy	Constraints	D_{K-L}	SD_{K-L}	SD_{K-L} Outperforms D_{K-L}	Time for D_{K-L} (seconds)	Time for SD_{K-L} (seconds)
0	ACR	NC	0.9794	0.9790	0.4300	1.7908	0.8185
		IC	0.9457	0.9457	0.9800	0.3470	0.1052
		AC	0.9793	0.9791	0.4700	1.8486	0.8369
	PCR	NC	0.9246	0.9236	0.4750	1.7908	0.8185
		IC	0.8346	0.8346	0.9800	0.3470	0.1052
		AC	0.9245	0.9237	0.4800	1.8486	0.8369
0.5	ACR	NC	0.9814	0.9811	0.4450	1.7821	0.8107
		IC	0.9494	0.9495	0.9950	0.3277	0.0955
		AC	0.9811	0.9810	0.5150	1.8327	0.8403
	PCR	NC	0.9305	0.9297	0.4450	1.7821	0.8107
		IC	0.8358	0.8360	0.9950	0.3277	0.0955
		AC	0.9297	0.9292	0.5300	1.8327	0.8403

(5) Estimate the knowledge state of the examinees and calculate the PCR and ACR, according to the selected response matrix, and repeat experiments for a total of 200 times.
(6) A one-way analysis of variance was performed on the data before and after the simplification. The specific steps of the analysis method were as follows:

We conduct a statistical test to compare the means for the PCR and ACR from two methods with the null hypothesis H_0: The simplified constraint has no significant effect on the ACR and PCR of the MMD test assembly method.

In order to express the differences of the means for the PCR or ACR from two methods, the simplified ACR (the same for PCR analysis) is combined into a two-column matrix Y_{ij}, i = 1,2; j = 1,2, …, n. The sum of samples is set to $Y_{i.} = \sum_{j=1}^{n} Y_{ij}$, and the sample mean is $\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$, then the calculation formula for the total mean of the samples is

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n} Y_{ij}. \tag{11}$$

The sum of squares of deviations is an indicator of the degree of dispersion of all data. If the assumption H_0 holds, the simplified constraint will have no significant effect on ACR or PCR, and then the difference of data in Y_{ij} is caused by other random factors. If the assumption is not true, in addition to random factors, the data difference in Y_{ij} also has the influence of simplified constraints. If the influence of simplified constraints is much greater than that of random factors, the simplified constraints should be considered to have a significant impact on ACR or PCR, otherwise it is considered to have no significant impact. Among them, the calculation formulas for the sum of squares between groups S_A and the random error sum of squares (or sum of squares within groups) S_E are

$$S_A = \sum_{i=1}^{2} n(\bar{Y}_i - \bar{Y})^2, \tag{12}$$

and

$$S_E = \sum_{i=1}^{2}\sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2. \tag{13}$$

In this study, only one factor was considered, so the degree of freedom of S_A was 1, and the total observation data was set to 2n, then the degree of freedom of S_E was 2n-2. From this, the formula for the one-way analysis of variance F-test can be calculated

$$F = \frac{S_A}{S_E/(2n-2)}. \tag{14}$$

TABLE 6 | Comparison of simplified and non-simplified constraints under the R-RUM model.

Correlation	Accuracy	Constraints	D_{K-L}	SD_{K-L}	SD_{K-L} Outperforms D_{K-L}	Time for D_{K-L} (seconds)	Time for SD_{K-L} (seconds)
0	ACR	NC	0.9514	0.9514	0.5250	3.3500	0.8915
		IC	0.9332	0.9311	0.2600	0.3594	0.1072
		AC	0.9513	0.9513	0.5150	3.4402	0.9404
	PCR	NC	0.8270	0.8272	0.5500	3.3500	0.8915
		IC	0.7816	0.7780	0.3750	0.3594	0.1072
		AC	0.8267	0.8269	0.5500	3.4403	0.9404
0.5	ACR	NC	0.9590	0.9590	0.4950	3.3598	0.8921
		IC	0.9457	0.9443	0.2750	0.3369	0.1024
		AC	0.9590	0.9589	0.5000	3.4420	0.9420
	PCR	NC	0.8495	0.8495	0.5000	3.3598	0.8921
		IC	0.8115	0.8081	0.3500	0.3369	0.1024
		AC	0.8494	0.8492	0.5050	3.4420	0.9420

TABLE 7 | The accuracy rate of each condition for four attributes under the DINA model.

Correlation	Accuracy	Constraints	Random	CDI	SD_{K-L}
0	ACR	NC	0.8443	0.9692	0.9795
		IC	0.8257	0.9173	0.9463
		AC	0.8439	0.9730	0.9793
	PCR	NC	0.5803	0.8886	0.9251
		IC	0.5507	0.7498	0.8357
		AC	0.5806	0.9054	0.9245
0.5	ACR	NC	0.8801	0.9733	0.9807
		IC	0.8680	0.9401	0.9489
		AC	0.8798	0.9756	0.9806
	PCR	NC	0.6588	0.9013	0.9281
		IC	0.6377	0.8022	0.8350
		AC	0.6587	0.9106	0.9280

After the observed value of F was obtained by analyzing and calculating from the data, we can usually choose a significant level of 0.05 or 0.01 according to the accuracy rate requirements. Then, the p-value was computed based on the observed value of F. Finally, the p-value is compared with 0.05 or 0.01 to decide whether to accept the null hypothesis. In this study, the significance level was set to 0.05.

Experimental Results

Tables 1–3 are results of the one-way analysis of variance of ACR and PCR obtained by the simplified and non-simplified constraint MMD test assembly method under the DINA model, the DINO model and the R-RUM, respectively. It can be seen that the p-value of DINA and DINO models are greater than 0.05 in all relevant cases, indicating that there is no significant difference in ACR or PCR between before and after the simplified constraints. However, the p-value of item constraints on the R-RUM is lower than 0.05, indicating that there is a significant difference in ACR or PCR between before and after the simplified constraints. It shows whether the constraints are simplified or not has little effect on the efficiency of the MMD constructing test assembly, except under the item constraints on the R-RUM.

Tables 4–6 respectively give a detailed comparison of simplified and non-simplified constraints in terms of ACR and PCR under each condition of the DINA model, the DINO model and the R-RUM. The sixth column of the tables indicates that the accuracy rate of simplified constraints higher than the accuracy rate of non-simplified constraints.

TABLE 8 | Comparison of the accuracy rate of each method for four attributes under DINA model.

Correlation	Accuracy	Constraints	CDI Outperforms Random	SD_{K-L} Outperforms Random	SD_{K-L} Outperforms CDI
0	ACR	NC	1.0000	1.0000	0.9150
		IC	0.9950	1.0000	0.9500
		AC	1.0000	1.0000	0.8850
	PCR	NC	1.0000	1.0000	0.9250
		IC	1.0000	1.0000	0.9550
		AC	1.0000	1.0000	0.8600
0.5	ACR	NC	1.0000	1.0000	0.8800
		IC	1.0000	1.0000	0.9050
		AC	1.0000	1.0000	0.8500
	PCR	NC	1.0000	1.0000	0.8850
		IC	1.0000	1.0000	0.9300
		AC	1.0000	1.0000	0.8500

TABLE 9 | Comparison of the accuracy rate of each method for four attributes under DINO model.

Correlation	Accuracy	Constraints	Random	CDI	SD_{K-L}
0	ACR	NC	0.8428	0.9687	0.9790
		IC	0.8256	0.9173	0.9457
		AC	0.8429	0.9728	0.9791
	PCR	NC	0.5779	0.8867	0.9236
		IC	0.5510	0.7490	0.8346
		AC	0.5786	0.9040	0.9237
0.5	ACR	NC	0.8791	0.9739	0.9811
		IC	0.8696	0.9409	0.9495
		AC	0.8792	0.9758	0.9810
	PCR	NC	0.6566	0.9033	0.9297
		IC	0.6406	0.8041	0.8360
		AC	0.6573	0.9114	0.9292

TABLE 10 | Comparison of the accuracy rate of each method for four attributes under the DINO model.

Correlation	Accuracy	Constraints	CDI Outperforms Random	SD_{K-L} Outperforms Random	SD_{K-L} Outperforms CDI
0	ACR	NC	1.0000	1.0000	0.9250
		IC	0.9850	1.0000	0.9800
		AC	1.0000	1.0000	0.8750
	PCR	NC	0.9950	1.0000	0.9350
		IC	0.9900	1.0000	0.9800
		AC	1.0000	1.0000	0.8550
0.5	ACR	NC	1.0000	1.0000	0.8550
		IC	1.0000	1.0000	0.8950
		AC	1.0000	1.0000	0.8500
	PCR	NC	1.0000	1.0000	0.8650
		IC	1.0000	1.0000	0.9100
		AC	1.0000	1.0000	0.8550

It can be seen from **Table 4** that under the DINA model, when the MMD test assembly simplifies the constraints, the overall efficiency is less than 50% although the efficiency of the simplified constraints is higher than that of the non-simplified constraints. Therefore, the simplification of the distance matrix will indeed lose information. From the perspective of the overall mean, the loss of information has a relatively low impact on the efficiency of the test assembly. This conclusion is similar to the results of the one-way analysis of variance. In terms of average time consumption, simplifying the constraints will increase the operating efficiency by 2 to 4 times. Comparing with the information of lost by the simplified constraints, the improvement of the operating efficiency is considerable. Therefore, the simplified constraints on the distance matrix are feasible.

Tables 5, **6** shows that the efficiency of the simplified constraints is higher than that of the non-simplified constraints, the efficiency is more than 50% or close to 50% under the attribute constraints, but the overall situation is still lower than the non-simplified constraints and the difference is still small under the DINO model and R-RUM. In terms of time-consuming, the time-consuming for these two models is similar to that under the DINA model, but simplifying the constraints will still increase the operating efficiency by 2 to 4 times on average, so a similar conclusion can be obtained with the DINA model.

STUDY 2: COMPARISON BETWEEN SIMPLIFIED MMD METHOD AND CDI METHOD

Experimental Purpose

Study 1 has verified that the simplified constraints on the distance matrix is feasible, so how the new method itself compares with the famous method needs to be discussed further. In order to compare the simplified MMD test assembly method and the CDI method (Henson and Douglas, 2005), we performed the second simulation experiments by using the similar condition settings as the study of Henson and Douglas (2005). It should be noted that eight attributes were considered in the second simulation study for exploring the performance of the simplified MMD test assembly method under different conditions.

Experimental Steps

Conduct the simulation experiment as follows:

TABLE 11 | Comparison of the accuracy rate of each method for four attributes under R-RUM model.

Correlation	Accuracy	Constraints	Random	CDI	SD_{K-L}
0	ACR	NC	0.8450	0.9386	0.9514
		IC	0.8447	0.9258	0.9311
		AC	0.8453	0.9408	0.9513
	PCR	NC	0.5418	0.7894	0.8272
		IC	0.5446	0.7594	0.7780
		AC	0.5427	0.7984	0.8269
0.5	ACR	NC	0.8798	0.9504	0.9590
		IC	0.8806	0.9416	0.9443
		AC	0.8801	0.9518	0.9589
	PCR	NC	0.6230	0.8223	0.8495
		IC	0.6272	0.7986	0.8081
		AC	0.6240	0.8275	0.8492

TABLE 12 | Comparison of the accuracy rate of each method for four attributes under the R-RUM model.

Correlation	Accuracy	Constraints	CDI Outperforms Random	SD_{K-L} Outperforms Random	SD_{K-L} Outperforms CDI
0	ACR	NC	1.0000	1.0000	0.9750
		IC	1.0000	1.0000	0.7400
		AC	1.0000	1.0000	0.9700
	PCR	NC	0.9950	1.0000	0.9650
		IC	0.9950	1.0000	0.7850
		AC	1.0000	1.0000	0.9550
0.5	ACR	NC	1.0000	1.0000	0.9650
		IC	1.0000	1.0000	0.7050
		AC	1.0000	1.0000	0.9700
	PCR	NC	1.0000	1.0000	0.9600
		IC	1.0000	1.0000	0.7300
		AC	1.0000	1.0000	0.9650

(1) According to the design of the first study, we simulated two groups of examinees, one of groups was used to calculate the prior distribution and the other was used for classification. The Q matrix and item parameters in the item bank and observed complete item response matrix of all possible attribute mastering patterns on all items in the item bank were simulated;
(2) Calculate the CDI and SD_{K-L} of all items;
(3) Construct cognitive diagnostic test using the random way, the CDI method, or the simplified MMD method, according to the three strategies of no constraints, attribute constraints and item constraints;
(4) Take out the response matrix of all the items on the corresponding test according to the test items generated by the test assembly algorithms;
(5) Estimate the knowledge state of the examinees and calculate the PCR and ACR, according to the selected response matrix, and repeat experiments for a total of 200 times.

Experimental Results

Table 7 shows the average accuracy rate of each condition under measuring four attributes with the DINA model. In the table, CDI represents the CDI test assembly method, SD_{K-L} is the simplified MMD test assembly method, and Random represents random test assembly. Analyzing the data in **Table 7** shows that the new method has a higher improvement compared with the CDI method. In terms of the three constraints, the overall accuracy rate of the attribute constraints is slightly higher than the other two constraints, and the accuracy rate for the item constraints is the worst. Under the condition of item constraints, the ACR and

TABLE 13 | The accuracy rate of each condition for eight attributes under the DINA model.

Correlation	Accuracy	Constraints	Random	CDI	SD_{K-L}
0	ACR	NC	0.6234	0.8294	0.8181
		IC	0.6489	0.7244	0.7289
		AC	0.6234	0.8315	0.8181
	PCR	NC	0.0988	0.3305	0.3678
		IC	0.1304	0.2525	0.2764
		AC	0.0988	0.3438	0.3678
0.5	ACR	NC	0.7474	0.8745	0.8664
		IC	0.7672	0.8267	0.8259
		AC	0.7474	0.8759	0.8664
	PCR	NC	0.3272	0.4683	0.4909
		IC	0.3381	0.4479	0.4547
		AC	0.3272	0.4766	0.4909

TABLE 14 | Comparison of the accuracy rate of each method for eight attributes under the DINA model.

Correlation	Accuracy	Constraints	CDI Outperforms Random	SD_{K-L} Outperforms Random	SD_{K-L} Outperforms CDI
0	ACR	NC	1.0000	1.0000	0.1500
		IC	1.0000	1.0000	0.6950
		AC	1.0000	1.0000	0.0950
	PCR	NC	1.0000	1.0000	0.8850
		IC	1.0000	1.0000	0.9600
		AC	1.0000	1.0000	0.8350
0.5	ACR	NC	1.0000	1.0000	0.1000
		IC	1.0000	1.0000	0.4150
		AC	1.0000	1.0000	0.0650
	PCR	NC	1.0000	1.0000	0.9400
		IC	1.0000	1.0000	0.7800
		AC	1.0000	1.0000	0.8900

TABLE 15 | The accuracy rate of each condition for eight attributes under the DINO model.

Correlation	Accuracy	Constraints	Random	CDI	SD_{K-L}
0	ACR	NC	0.6208	0.8289	0.8171
		IC	0.6479	0.7238	0.7296
		AC	0.6208	0.8302	0.8171
	PCR	NC	0.0963	0.3319	0.3679
		IC	0.1294	0.2525	0.2781
		AC	0.0963	0.3435	0.3679
0.5	ACR	NC	0.7444	0.8754	0.8665
		IC	0.7666	0.8268	0.8264
		AC	0.7444	0.8765	0.8665
	PCR	NC	0.3248	0.4706	0.4910
		IC	0.3368	0.4488	0.4553
		AC	0.3248	0.4776	0.4910

TABLE 16 | Comparison of the accuracy rate of each method for eight attributes under the DINO model.

Correlation	Accuracy	Constraints	CDI Outperforms Random	SD_{K-L} Outperforms Random	SD_{K-L} Outperforms CDI
0	ACR	NC	1.0000	1.0000	0.1900
		IC	1.0000	1.0000	0.7200
		AC	1.0000	1.0000	0.1600
	PCR	NC	1.0000	1.0000	0.8750
		IC	1.0000	1.0000	0.9400
		AC	1.0000	1.0000	0.8450
0.5	ACR	NC	1.0000	1.0000	0.0700
		IC	1.0000	1.0000	0.4350
		AC	1.0000	1.0000	0.0400
	PCR	NC	1.0000	1.0000	0.8950
		IC	1.0000	1.0000	0.8000
		AC	1.0000	1.0000	0.8700

PCR of CDI, MMD, and random test assembly method are lower than the other two constraints.

Table 8 shows the comparison of the accuracy rate of each method when the number of attributes under the DINA model is four in the 200 repeats. Among them, the last column represents the proportion of MMD test assembly method with SD_{K-L} distance as the class distance index more efficient than CDI in the 200 simulation repeats. The fourth or fifth column respectively represents the proportion of CDI test assembly method or MMD test assembly method with SD_{K-L} distance more efficient than the random test assembly method across 200 repetitions.

It can be seen from **Table 8** that the MMD test assembly method with SD_{K-L} distance as an index is stable under various conditions. In the existing conclusions, as the correlation increasing, the accuracy rate of the MMD test assembly method

TABLE 17 | The accuracy rate of each condition for eight attributes under the R-RUM model.

Correlation	Accuracy	Constraints	Random	CDI	SD_{K-L}
0	ACR	NC	0.7427	0.8300	0.8336
		IC	0.7572	0.8142	0.8231
		AC	0.7427	0.8303	0.8336
	PCR	NC	0.1236	0.2648	0.2823
		IC	0.1365	0.2433	0.2651
		AC	0.1236	0.2660	0.2823
0.5	ACR	NC	0.8243	0.8753	0.8787
		IC	0.8300	0.8705	0.8745
		AC	0.8243	0.8757	0.8787
	PCR	NC	0.3261	0.4175	0.4309
		IC	0.3232	0.4183	0.4254
		AC	0.3261	0.4189	0.4309

TABLE 18 | Comparison of the accuracy rate of each method for eight attributes under the R-RUM model.

Correlation	Accuracy	Constraints	CDI Outperforms Random	SD_{K-L} Outperforms Random	SD_{K-L} Outperforms CDI
0	ACR	NC	1.0000	1.0000	0.6350
		IC	1.0000	1.0000	0.8700
		AC	1.0000	1.0000	0.6350
	PCR	NC	1.0000	1.0000	0.8000
		IC	1.0000	1.0000	0.8850
		AC	1.0000	1.0000	0.7950
0.5	ACR	NC	1.0000	1.0000	0.7900
		IC	1.0000	1.0000	0.8450
		AC	1.0000	1.0000	0.7800
	PCR	NC	1.0000	1.0000	0.8300
		IC	1.0000	1.0000	0.6850
		AC	1.0000	1.0000	0.8300

the CDI method based on SD_{K-L} distance increases. Therefore, as the correlation increasing, the gap between the two methods will shrink. Thus, comparing with random test assembly, the average value of each method is greater than the random method. However, in the 200 simulation repeats, the CDI test assembly method is occasionally outperformed by the random test assembly method, which is similar to the simulation results of Henson and Douglas (2005).

On the whole, the result of CDI test assembly method is slightly different from that of Henson and Douglas (2005) in comparison with random test assembly method under measuring four attributes, because the random test assembly method itself is uncertain. In addition, the case that the accuracy rate of CDI test assembly method is lower than that of the random method is concentrated under the item constraints.

Table 9 shows the comparison of several test assembly methods for 200 repetitions under DINO model. From the data in **Table 9**, it can be seen that the MMD method is still superior to CDI method, and the MMD method with SD_{K-L} distance as the distance index does not have the situation that the average accuracy rate is lower than CDI method.

Table 10 shows the comparison of the accuracy rate of each method when the number of attributes is four across replications under the DINO model. It can be seen from **Table 10**, the MMD method with SD_{K-L} distance as the distance index has slightly better accuracy rate than the CDI method under the condition of unconstrained and attribute constraints, respectively. However, the accuracy rate of the MMD test assembly method under the item constraints is better than under the other two constraints. In the existing conclusions, with the increase of correlation, the accuracy rate of MMD test assembly method with SD_{K-L} distance as the index decreased while that of the CDI method increased. Therefore, with the increase of correlation, the gap between the two methods narrowed.

Table 11 shows the comparison of several test assembly methods for the 200 repetitions under the R-RUM. Like the DINA and DINO model, the performance of MMD test assembly method based on SD_{K-L} distance is better than the other two methods, and the performance is almost the same under the condition of both no constraints and attribute constraints.

Table 12 shows the comparison of the accuracy rate of each method when the number of attributes is four across replications under the R-RUM. It can be seen from **Table 12** that in the 200 simulation replications, the MMD test assembly method based on SD_{K-L} distance is better than the CDI test assembly method in every case, and its performance on R-RUM is also better than that of DINO model under the condition of both no constraints and attribute constraints.

Tables 13–18 show accuracy rates and comparison results for eight attributes. The results of eight attributes are similar to that of four attributes. On the whole, the new method is better than the CDI test assembly method and the random assembly method under the DINA model, the DINO model and the R-RUM. Furthermore, the new method has a greater advantage over the CDI method in terms of the PCR. Under the three models, the PCR of the MMD test assembly method based on SD_{K-L} distance is higher than that of the CDI test assembly method, but the ACR of the MMD test assembly method is slightly lower than the CDI test assembly method. It means that the higher the averaged ACR, the PCR is not necessarily higher. For example, the ACRs for two attributes are 0.1 and 0.9 or 0.4 and 0.4. Although the average of ACR for these two cases are 0.5 and 0.4, the former case has the PCR of 0.09, while the latter case has the PCR of 0.16, if the correct classification rates for two attributes are independent.

DISCUSSION

Simulation results show that the MMD test assembly method with the simplified constraints has similar performance to the new method with the full constraints under four attributes, and the new simplified method performs better than the CDI method for four and eight attributes in term of the PCR. The MMD test assembly method with the full constraints suffers a large computational burden due to the optimization problem of complex constraints, but it is fast and performs relatively well when the number of attributes is four. In order to simplify computation, the MMD test assembly method with the simplified constraints can simplify computation effectively and is suitable for a larger number of attributes (i.e., eight attributes). We also found that when the number of measured attributes increases, the advantages in PCR for the MMD method are still obvious, while its performance in ACR tends to be average. This is related to the characteristics of the MMD and CDI-based test assembly methods: the CDI test assembly method pays attention to the local information, while the MMD focuses on the global information. When the ratio of test length to attribute number is large, the MMD test assembly method has enough room to play and select enough high-quality tests to obtain sufficient overall information, in order to make up for the lack of local information. So, the MMD test assembly method has obvious advantages at this condition.

We found that there is a considerably worse performance for item constraints compared to attribute constraints, which is consistent of results of Henson and Douglas (2005). The possible explanations are as follows: First, we think this may be related to the concept of statistical identification that is receiving a lot of attention lately for the case of CDMs. Specifically, for the DINA model, two identity matrices in the Q matrix and an additional third item per attribute would be required (e.g., Chen et al., 2015; Xu and Shang, 2018). This would be never satisfied in the item constraint condition. Second, because the item constraints required only 4 items measured one attribute, the Q-matrix is not complete if all columns of the $K \times K$ identity matrix are not contained in the Q-matrix. A simple example of a complete Q-matrix is the $K \times K$ identity matrix $\boldsymbol{I}$ (Chiu et al., 2009; Cai et al., 2018). Third, item-level expected classification accuracy of attributes for 16 items measured two or three attributes in item constraint condition is often lower than that for items measured only one attribute (Wang et al., 2019).

Test constraints in this study are still rough, since it is only a repetition of Henson and Douglas (2005) experiments. The performance of each method under other constraints needs to be studied. The MMD test assembly method with SD_{K-L} distance as the index is superior to CDI test assembly method in performance, but the combination of this test assembly idea and other distance indexes is worth discussing. This study does not consider the relationship between the number of measured attributes and the length of tests, the influence of the ratio of the test length and number of attributes on the MMD test assembly method, how about the specific relationship between them is, and how to specify item constraints and attribute constraints when the length of tests is different, all of above will need a further investigation.

As in the study of Henson and Douglas (2005), the larger sample in our study was only employed to obtain correct classification rates more stability with simulated item parameters. We have not considered the impact of item banks calibrated by using larger or smaller sample sizes on the performance of test construction methods. As the reviewer motioned, it is true that larger sample sizes are likely to be used to calibrate item banks (e.g., Liu et al., 2013; George and Robitzsch, 2014), while the review of available empirical studies indicates that sample sizes in cognitive diagnosis tend to be much smaller (Sessoms and Henson, 2018). It would be an interesting question to justify whether a difference in performance is expected from the CDI and SD_{K-L} methods for item banks calibrated from different sample sizes. One limitation of this study is that three simple CDMs (the DINA model, the DINO model, and the R-RUM) were considered in the simulation study. If we have a large sample size to calibrate an item bank, we believe that the results can be generalized to a more general model, such as the G-DINA model (de la Torre, 2011), or a combination of reduced models (Ravand, 2016; Sorrel et al., 2017; de la Torre et al., 2018).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

WW and LS designed the study and wrote the manuscript. JZ, YT, and PG drafted and revised the manuscript. WW and PG conducted the simulation study. All authors contributed to the article and approved the submitted version.

FUNDING

This research was partially supported by the National Natural Science Foundation of China (Grant No. 62067005), the Social Science Foundation of Jiangxi (Grant No. 17JY10), the Project of Teaching Reform of Jiangxi Normal University (Grant No. JXSDJG1848), and the grant (Grant No. CTI2019B10) from the Chinese Testing International Co., Ltd.

REFERENCES

Cai, Y., Tu, D., and Ding, S. (2018). Theorems and methods of a complete Q matrix with attribute hierarchies under restricted q-matrix design. *Front. Psychol.* 9:1413. doi: 10.3389/fpsyg.2018.01413

Chang, H.-H., and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Appl. Psychol. Measurem.* 20, 213–229. doi: 10.1177/014662169602000303

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *J. Am. Statist. Assoc.* 110, 850–866. doi: 10.1080/01621459.2014.934827

Chiu, C.-Y., Douglas, J. A., and Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika* 74, 633–665. doi: 10.1007/s11336-009-9125-0

Clark, I. (2013). *Efficacy of Formative Classroom Assessments in Theory and Practice*. Doctoral dissertation University of Washington: Seattle, WA.

Cover, T. M., and Thomas, J. A. (ed.). (2006). *Elements of Information Theory*, 2nd Edn. Hoboken, NJ: John Wiley & Sons, Inc.

de la Torre, J., van der Ark, L. A., and Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement Eval. Counsel. Dev.* 51, 281–296. doi: 10.1080/07481756.2017.1327286

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

Debeer, D., Van Rijn, P. W., and Ali, U. S. (2020). Multidimensional test assembly using mixed-integer linear programming: an application of Kullback-Leibler information. *Appl. Psychol. Measurement* 44, 17–32. doi: 10.1177/0146621619827586

Finkelman, M., Kim, W., and Roussos, L. A. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *J. Educ. Measurement* 46, 273–292. doi: 10.1111/j.1745-3984.2009.00081.x

George, A. C., and Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychol. Test Assessm. Model.* 56, 405–432.

Guo, L., Yang, J., and Song, N. (2020). Spectral clustering algorithm for cognitive diagnostic assessment. *Front. Psychol.* 11:944. doi: 10.3389/fpsyg.2020.00944

Hartz, S. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*, Unpublished doctoral dissertation University of Illinois at Urbana-Champaign: Champaign, IL.

Henson, R., and Douglas, J. (2005). Test construction for cognitive diagnosis. *Appl. Psychol. Measurement* 29, 262–277. doi: 10.1177/0146621604272623

Henson, R., Roussos, L., Douglas, J., and He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Appl. Psychol. Measurement* 32, 275–288. doi: 10.1177/0146621607302478

Hsu, C. L., Jin, K. Y., and Chiu, M. M. (2020). Cognitive diagnostic models for random guessing behaviors. *Front. Psychol.* 11:570365. doi: 10.3389/fpsyg.2020.570365

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Measurement* 25, 258–272. doi: 10.1177/01466210122032064

Kantor, I., Robineau, J. L., Bütün, H., and Maréchal, F. (2020). A mixed-integer linear programming formulation for optimizing multi-scale material and energy integration. *Front. Energy Res.* 8:49. doi: 10.3389/fenrg.2020.00049

Kuo, B.-C., Pai, H.-S., and de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Appl. Psychol. Measurement* 40, 315–330. doi: 10.1177/0146621616638643

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., and Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *J. Classif.* 30, 152–172. doi: 10.1007/s00357-013-9128-5

Madigan, D., and Almond, R. (1995). "On test selection strategies for belief networks," in *Learning From Data: AI and Statistics V*, eds D. Fisher and H.-J. Lenz (New York, NY: Springer-Verlag), 89–98. doi: 10.1007/978-1-4612-2404-4_9

Mao, X. Z. (2014). The attribute mastery probability cognitive diagnostic model. *J. Sichuan Normal University* 37, 373–443.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. R. Statist. Soc. Series B* 24, 46–72. doi: 10.1111/j.2517-6161.1962.tb00436.x

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

Sessoms, J., and Henson, R. A. (2018). Applications of diagnostic classification models: a literature review and critical commentary. *Measurement: Interdiscipl. Res. Perspect.* 16, 1–17. doi: 10.1080/15366367.2018.1435104

Sia, C. J. L., and Lim, C. S. (2018). "Cognitive diagnostic assessment: an alternative mode of assessment for learning," in *Classroom Assessment in Mathematics. ICME-13 Monographs*, eds D. Thompson, M. Burton, A. Cusi, and D. Wright (Cham: Springer).

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., and Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Appl. Psychol. Measurement* 41, 614–631. doi: 10.1177/0146621617707510

Tang, F., and Zhan, P. D. (2020). The development of an instrument for longitudinal learning diagnosis of rational number operations based on parallel tests. 11:2246. doi: 10.3389/fpsyg.2020.02246

Tatsouka, C., and Ferguson, T. (2003). Sequential classification on partially ordered sets. *J. R. Statist. Soc.: Series B* 65, 143–158. doi: 10.1111/1467-9868.00377

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989x.11.3.287

Wang, W. Y., Gao, P., Song, L. H., and Wang, T. (2020). The improved exploratory method of Q-matrix specification with noise preprocessing. *J. Jiangxi Normal University* 44, 136–141.

Wang, W. Y., Song, L. H., Chen, P., and Ding, S. L. (2019). An item-level expected classification accuracy and its applications in cognitive diagnostic assessment. *J. Educ. Measurement* 56, 51–75. doi: 10.1111/jedm.12200

Xu, G., and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *J. Am. Statist. Assoc.* 113, 1284–1295. doi: 10.1080/01621459.2017.1340889

Zijlmans, E. A. O., Tijmstra, J., van der Ark, L. A., and Sijtsma, K. (2019). Item-score reliability as a selection tool in test construction. *Front. Psychol.* 9:2298. doi: 10.3389/fpsyg.2018.02298

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 Wang, Zheng, Song, Tu and Gao. 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: 15 February 2021
doi: 10.3389/fpsyg.2021.614470

Determining the Number of Attributes in Cognitive Diagnosis Modeling

*Pablo Nájera, Francisco José Abad and Miguel A. Sorrel**

Department of Social Psychology and Methodology, Faculty of Psychology, Autonomous University of Madrid, Madrid, Spain

Cognitive diagnosis models (CDMs) allow classifying respondents into a set of discrete attribute profiles. The internal structure of the test is determined in a Q-matrix, whose correct specification is necessary to achieve an accurate attribute profile classification. Several empirical Q-matrix estimation and validation methods have been proposed with the aim of providing well-specified Q-matrices. However, these methods require the number of attributes to be set in advance. No systematic studies about CDMs dimensionality assessment have been conducted, which contrasts with the vast existing literature for the factor analysis framework. To address this gap, the present study evaluates the performance of several dimensionality assessment methods from the factor analysis literature in determining the number of attributes in the context of CDMs. The explored methods were parallel analysis, minimum average partial, very simple structure, DETECT, empirical Kaiser criterion, exploratory graph analysis, and a machine learning factor forest model. Additionally, a model comparison approach was considered, which consists in comparing the model-fit of empirically estimated Q-matrices. The performance of these methods was assessed by means of a comprehensive simulation study that included different generating number of attributes, item qualities, sample sizes, ratios of the number of items to attribute, correlations among the attributes, attributes thresholds, and generating CDM. Results showed that parallel analysis (with Pearson correlations and mean eigenvalue criterion), factor forest model, and model comparison (with AIC) are suitable alternatives to determine the number of attributes in CDM applications, with an overall percentage of correct estimates above 76% of the conditions. The accuracy increased to 97% when these three methods agreed on the number of attributes. In short, the present study supports the use of three methods in assessing the dimensionality of CDMs. This will allow to test the assumption of correct dimensionality present in the Q-matrix estimation and validation methods, as well as to gather evidence of validity to support the use of the scores obtained with these models. The findings of this study are illustrated using real data from an intelligence test to provide guidelines for assessing the dimensionality of CDM data in applied settings.

Keywords: cognitive diagnostic models, dimensionality assessment, parallel analysis, machine learning, model comparison, Q-matrix validation

OPEN ACCESS

Edited by:
Tao Xin,
Beijing Normal University, China

Reviewed by:
Chester Chun Seng Kam,
University of Macau, China
Shiyu Wang,
University System of Georgia, United States

****Correspondence:***
Miguel A. Sorrel
miguel.sorrel@uam.es

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

Received: *06 October 2020*
Accepted: *25 January 2021*
Published: *15 February 2021*

Citation:
Nájera P, Abad FJ and Sorrel MA (2021) Determining the Number of Attributes in Cognitive Diagnosis Modeling. Front. Psychol. 12:614470. doi: 10.3389/fpsyg.2021.614470

INTRODUCTION

The correct specification of the internal structure is arguably the key issue in the formulation process of a measurement model. Hence, it is not surprising that the determination of the number of factors has been regarded as the most crucial decision in the context of exploratory factor analysis (EFA; e.g., Garrido et al., 2013; Preacher et al., 2013). Since the very first proposals to address this issue, such as the eigenvalue-higher-than-one rule or Kaiser-Guttman criterion (Guttman, 1954; Kaiser, 1960), many methods have been developed for assessing the dimensionality in EFA. Despite the longevity of this subject of study, the fact that it is still a current research topic (e.g., Auerswald and Moshagen, 2019; Finch, 2020) is a sign of both its relevance and complexity.

In contrast to the vast research in the EFA framework, dimensionality assessment remains unexplored for other measurement models. This is the case of cognitive diagnosis models (CDMs). CDMs are restricted latent class models in which the latent variables or *attributes* are discrete, usually dichotomous. The popularity of CDMs has increased in the last years, especially in the educational field, because of their ability to provide a fine-grained information about the examinees' mastery or non-mastery of certain skills, cognitive processes, or competences (de la Torre and Minchen, 2014). However, CDM applications are not restricted to educational settings, and they have been employed for the study of psychological disorders (Templin and Henson, 2006; de la Torre et al., 2018) or staff selection processes (García et al., 2014; Sorrel et al., 2016).

A required input for CDMs is the *Q-matrix* (Tatsuoka, 1983). It has dimensions J items $\times$ K attributes, in which each *q-entry* (q_{jk}) can adopt a value of 1 or 0, depending on whether attribute k is relevant to measure item j or not, respectively. Hence, the Q-matrix determines the internal structure of the test, and its correct specification is fundamental to obtain accurate structural parameter estimates and, subsequently, an accurate classification of examinees' latent classes or *attribute profiles* (Rupp and Templin, 2008; Gao et al., 2017). However, the Q-matrix construction process is usually conducted by domain experts (e.g., Sorrel et al., 2016). This process is subjective in nature and susceptible to specification errors (Rupp and Templin, 2008; de la Torre and Chiu, 2016). To address this, several Q-matrix estimation and validation methods have been proposed in the recent years with the aim of providing empirical support to its specification. On the one hand, empirical Q-matrix estimation methods rely solely on the data to specify the Q-matrix. For instance, Xu and Shang (2018) developed a likelihood-based estimation method, which aims to find the Q-matrix that shows the best fit while controlling for model complexity. Additionally, Wang et al. (2018) proposed the *discrete factor loading* (DFL) method, which consists in conducting an EFA and dichotomizing the factor loading matrix up to some criterion (e.g., row or column means). On the other hand, empirical Q-matrix validation methods aim to correct a provisional, potentially misspecified Q-matrix based on both its original specification and the data. For instance, the *stepwise* method (Ma and de la Torre, 2020a) is based on the Wald test to select the q-entries that are statistically necessary for each item, while the *general discrimination index* method (de la Torre and Chiu, 2016) and the *Hull* method (Nájera et al., 2020) aim to find, for each item, the simplest *q-vector* specification that leads to an adequate discrimination between latent classes. These methods serve as a useful tool for applied researchers, who can obtain empirical evidence of the validity of their Q-matrices (e.g., Sorrel et al., 2016).

Despite their usefulness, the Q-matrix estimation and validation methods share an important common drawback, which is assuming that the number of attributes specified by the researcher is correct (Xu and Shang, 2018; Nájera et al., 2020). Few studies have tentatively conducted either a parallel analysis (Robitzsch and George, 2019) or model-fit comparison (Xu and Shang, 2018) to explore the dimensionality of the Q-matrix. However, to the authors' knowledge, there is a lack of systematic studies on the empirical estimation of the number of attributes in CDMs. The main objective of the present research is precisely to compare the performance of a comprehensive set of dimensionality assessment methods in determining the number of attributes. The remaining of the paper is laid out as follows. First, a description of two popular CDMs is provided. Second, a wide selection of EFA dimensionality assessment methods is described. Third, an additional method for assessing the number of attributes in CDMs is presented. Fourth, the design and results from an exhaustive simulation study are provided. Fifth, real CDM data are used for illustrating the functioning of the dimensionality assessment methods. Finally, practical implications and future research lines are discussed.

THE DINA AND G-DINA MODELS

CDMs can be broadly separated into general and reduced, specific models. General CDMs are saturated models that subsume most of the reduced CDMs. They include more parameters and, consequently, provide a better model-data fit in absolute terms. As a counterpoint, their estimation is more challenging. Thus, reduced CDMs are often a handy alternative to applied settings because of their simplicity, which favors both their estimation and interpretation. Let denote by K_j^* the number of required attributes for item j. Under the *deterministic inputs, noisy "and" gate* model (DINA; Junker and Sijtsma, 2001), which is a conjunctive reduced CDM, there are only two parameters per item regardless of K_j^*: the *guessing* parameter (g_j), which is the probability of correctly answering item j for those examinees that do not master, at least, one of the required attributes, and the *slip* parameter (s_j), which is the probability of failing item j for those examinees that master all the attributes involved. The probability of correctly answering item j given latent class l is given by

$$P_j(\boldsymbol{\alpha}_l) = g_j^{1-\eta_{lj}}(1-s_j)^{\eta_{lj}} \tag{1}$$

where η_{lj} equals 1 if examinees in latent class l master all the attributes required by item j, and 0 otherwise.

The *generalized* DINA model (G-DINA; de la Torre, 2011) is a general CDM, in which the probability of correctly answering item j for latent class l is given by the sum of the main effects of

the required attributes and their interaction effects (in addition to the intercept):

$$P_j\left(\boldsymbol{\alpha}_l^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lkk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (2)$$

where $\boldsymbol{\alpha}_l^*$ is the reduced attribute profile whose elements are the $K_j{}^*$ required attributes for item j, δ_{j0} is the intercept for item j, δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$. **Figure 1** depicts the probabilities of success of the four possible latent groups for an item requiring two attributes ($K_j^* = 2$) under the DINA and G-DINA models. For the DINA model, the probability of success for the latent group that masters all attribute is high ($P(11) = 1 - s_j = 1 - 0.2 = 0.8$), while the probability of success for the remaining latent groups is very low ($P(00) = P(10) = P(01) = g_j = 0.1$). For the G-DINA model, the baseline probability (i.e., intercept) is also very low ($P(00) = \delta_{j0} = 0.1$). The increment in the probability of success as a result of mastering the first attribute ($P(10) = \delta_{j0} + \delta_{j1} = 0.1 + 0.25 = 0.35$) is slightly lower than the one due to mastering the second attribute ($P(01) = \delta_{j0} + \delta_{j2} = 0.1 + 0.35 = 0.45$). Finally, although the interaction effect for both attributes is low ($\delta_{j12} = 0.1$), the probability of success for the latent group that masters both attributes is high because the main effects are also considered ($P(11) = \delta_{j0} + \delta_{j1} + \delta_{j2} + \delta_{j12} = 0.1 + 0.25 + 0.35 + 0.1 = 0.80$).

DIMENSIONALITY ASSESSMENT METHODS

In the following, we provide a brief explanation of seven dimensionality assessment methods that were originally developed for determining the number of factors in EFA and will be explored in the present study.

Parallel Analysis

Parallel analysis (PA; Horn, 1965) compares the eigenvalues extracted from the sample correlation matrix (i.e., sample eigenvalues) with the eigenvalues obtained from several randomly generated correlation matrices (i.e., reference eigenvalues). The number of sample eigenvalues that are higher than the average of their corresponding reference eigenvalues is retained as the number of factors. The 95th percentile has also been recommended rather than the mean to prevent from over-factoring (i.e., overestimate the number of factors). However, no differences have been found in recent simulation studies between both cutoff criteria (Crawford et al., 2010; Auerswald and Moshagen, 2019; Lim and Jahng, 2019). Additionally, polychoric correlations have been recommended when working with categorical variables. Although no differences have been found for non-skewed categorical data, polychoric correlations perform better with skewed data (Garrido et al., 2013) as long as the reference eigenvalues are computed considering the univariate category probabilities of the sample variables by, for instance, using random column permutation for generating the random samples (Lubbe, 2019). Finally, different extraction methods have been used to compute the eigenvalues: principal components analysis (Horn, 1965), principal axis factor analysis (Humphreys and Ilgen, 1969), or minimum rank factor analysis (Timmerman and Lorenzo-Seva, 2011). The original proposal by Horn has consistently shown the best performance across a wide range of conditions (Garrido et al., 2013; Auerswald and Moshagen, 2019; Lim and Jahng, 2019). Simulation studies have shown the superiority of PA above other dimensionality assessment methods. Thus, it is usually recommended and considered the gold standard (Garrido et al., 2013; Auerswald and Moshagen, 2019; Lim and Jahng, 2019; Finch, 2020). As a flaw, PA tends to under-factor (i.e., underestimate the number of factors) in conditions with low factor loadings or highly correlated factors (Garrido et al., 2013; Lim and Jahng, 2019).

Minimum Average Partial

The minimum average partial (MAP; Velicer, 1976) method has also been recommended for determining the number of factors with continuous data (Peres-Neto et al., 2005). It is based on principal components analysis and the partial correlation matrix. The MAP method extracts one component at a time and computes the average of the squared partial correlations (MAP index). The MAP method relies on the rationale that extracting the first components, which explain most of the common variance, will result in a decrease of the MAP index. Once the relevant components have been partialled out, extracting the remaining ones (which are formed mainly by unique variance) will make the MAP index to increase again. The optimal number of components corresponds to the lowest MAP index. A variant where the MAP index is computed by averaging the fourth power of the partial correlation was proposed by Velicer et al. (2000). However, Garrido et al. (2011) recommended the use of the original squared partial correlations, in addition to polychoric correlations when categorical variables are involved. They found that MAP method performed poorly under certain unfavorable situations, such as low-quality items or small number of variables per factor, where the method showed a tendency to under-factor.

Very Simple Structure

The very simple structure (VSS; Revelle and Rocklin, 1979) method was developed with the purpose of providing the best interpretable factor solution, understood as the absence of cross-loadings. In this procedure, a loading matrix with K factors is first estimated and rotated. Then, a simplified factor loading matrix ($\mathbf{S'}_{vk}$) is obtained, given a prespecified complexity v. Namely, the v highest loadings for each item are retained and the remaining loadings are fixed to zero. Then, the residual correlation matrix is found by

$$\overline{\mathbf{R}}_{vk} = \mathbf{R} - \mathbf{S}_{vk}\boldsymbol{\Phi}_{\mathbf{k}}\mathbf{S'}_{vk} \quad (3)$$

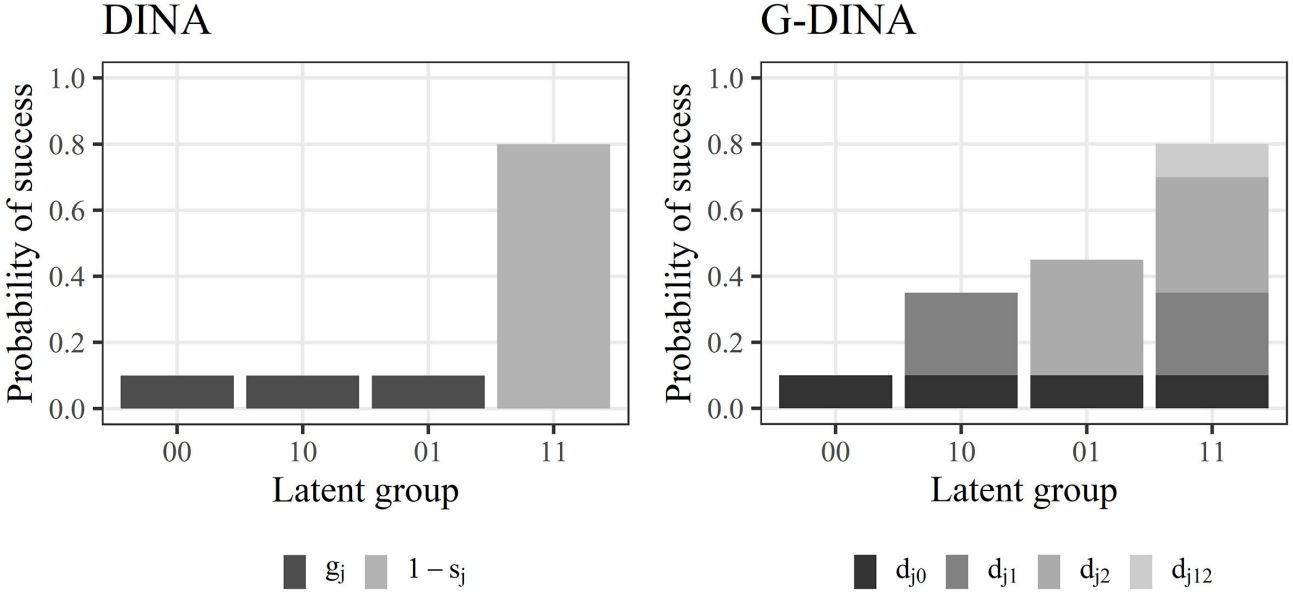


FIGURE 1 | Illustration of item parameters and probabilities of success of the different latent groups for the DINA and G-DINA models involving 2 attributes ($K = 2$).

where $\mathbf{R}$ is the observed correlation matrix and $\mathbf{\Phi_k}$ is the factor correlation matrix. Then, the VSS index is computed as

$$VSS_{vk} = 1 - \frac{MS_{\overline{\mathbf{R}}_{vk}}}{MS_{\mathbf{R}}} \tag{4}$$

where $MS_{\overline{\mathbf{R}}_{vk}}$ and $MS_{\mathbf{R}}$ are the average of the squared residual and observed correlations, respectively. The VSS index is computed for each factor solution, and the highest VSS corresponds to the number of factors to retain. The main drawback of the procedure is that the researcher must prespecify a common expected complexity for all the items, which is usually $v = 1$ (VSS_1; Revelle and Rocklin, 1979). In a recent simulation study, the VSS method obtained a poor performance under most conditions, over-factoring with uncorrelated factors and under-factoring with highly correlated factors (Golino and Epskamp, 2017).

Dimensionality Evaluation to Enumerate Contributing Traits

The dimensionality evaluation to enumerate contributing traits (DETECT; Kim, 1994; Zhang and Stout, 1999; Zhang, 2007) method is a nonparametric procedure that follows two strong assumptions: first, a single "dominant" dimension underlies the item responses, and second, the residual common variance between the items follows a simple structure (i.e., without cross-loadings). The method estimates the covariances of item pairs conditioned to the raw item scores, which are used as a non-parametric approximation to the dominant dimension. If the data are essentially unidimensional, these conditional covariances will be close to zero. Otherwise, items measuring the same secondary dimension will have positive conditional covariances, and items measuring different secondary dimensions will have negative conditional covariances. The DETECT index is computed as

$$D(P) = \frac{1}{J(J-1)/2} \sum_{j<j'} (-1)^{c_{jj'}} (\hat{C}_{jj'} - \overline{C}) \tag{5}$$

where P represents a specific partitioning of items into clusters, $\hat{C}_{jj'}$ is the estimated conditional covariance between items j and j', $\overline{C}$ is the average of the estimated conditional covariances, and $c_{jj'} = 0$ or 1 if items j and j' are part of the same cluster or not, respectively. The method explores different number of dimensions and the one that obtains the highest DETECT index is retained. Furthermore, the method also provides which items measure which dimension. In a recent study, Bonifay et al. (2015) found that the DETECT method had a great performance at the population level, retaining the correct number of dimensions in 97% of the generated datasets with $N = 10{,}000$. As a limitation of the study, the authors only tested a scenario in which the generating model had no cross-loadings (i.e., simple structure).

Empirical Kaiser Criterion

The empirical Kaiser criterion (EKC; Braeken and van Assen, 2017) is similar to PA in that the sample eigenvalues are compared to reference eigenvalues to determine the number of factors to retain. Here, reference eigenvalues are derived from the theoretical sampling distribution of eigenvalues, which is a Marčenko-Pastur distribution (Marčenko and Pastur, 1967) under the null hypothesis (i.e., non-correlated variables). The first reference eigenvalue depends only on the ratio of test length to sample size, while the subsequent reference eigenvalues consider the variance explained by the previous ones. The reference eigenvalues are coerced to be at least equal to 1, and thus it cannot suggest more factors than the Kaiser-Guttman criterion would.

In fact, EKC is equivalent to the Kaiser-Guttman criterion at the population level. EKC has been found to perform similarly to PA with non-correlated variables, unidimensional models, and orthogonal factors models, while it outperformed PA with oblique factors models and short test lengths (Braeken and van Assen, 2017). The performance of EKC with non-continuous data remains unexplored.

Exploratory Graph Analysis

EGA (Golino and Epskamp, 2017) is a recently developed technique that has emerged as a potential alternative for PA. EGA was first developed based on the Gaussian graphical model (GGM; Lauritzen, 1996), which is a network psychometric model in which the joint distribution of the variables is estimated by modeling the inverse of the variance-covariance matrix. The GGM is estimated using the least absolute shrinkage and selector operator (LASSO; Tibshirani, 1996), which is a penalization technique to avoid overfitting. Apart from EGA with GGM, Golino et al. (2020) recently proposed an EGA based on the triangulated maximally filtered graph approach (TMFG; Massara et al., 2016), which is not restricted to multivariate normal data. Regardless of the model (GGM or TMFG), in EGA each item is represented by a node and each edge connecting two nodes represents the association between the two items. Partial correlations are used for EGA with GGM, while any association measure can be used for EGA with TMFG. A strong edge between two nodes is interpreted as both items being caused by the same latent variable. A *walktrap* algorithm is then used to identify the number of clusters emerging from the edges, which will be the number of factors to retain. Furthermore, EGA also provides information about what items are included in what clusters, and clusters can be related to each other if their nodes are correlated. EGA with GGM seems to have an overall better performance than EGA with TMFG (Golino et al., 2020). EGA with GGM has been found to perform similarly to PA in most situations, with slightly worse results with low factor correlations, but better performance with highly correlated factors (Golino and Epskamp, 2017). On the other hand, EGA with TMFG tends to under-factor when there are many variables per factor or highly correlated factors (Golino et al., 2020).

Factor Forest

Factor forest (FF; Goretzko and Bühner, 2020) is an extreme gradient boosting machine learning model that was trained to predict the optimal number of factors in EFA. Specifically, the model estimates the probability associated to different factor solutions and subsequently suggests the number of factors with the highest probability. As opposed to the previously described dimensionality assessment methods, the FF is not based on any particular theoretical psychometric background, and its purpose is to make accurate predictions based on a combination of empirical results obtained from the training datasets. This is commonly referred to as the "black box" character of the machine learning models (Goretzko and Bühner, 2020). In the original paper, the authors trained the FF model using a set of 181 features (e.g., eigenvalues, sample size, number of variables, Gini-coefficient, Kolm measure of inequality) while varying the sample size, primary and secondary loadings, number of factors, variables per factor, and factor correlations, through almost 500,000 datasets. The data were generated assuming multivariate normality. The FF model obtained very promising results, correctly estimating the number of factors in 99.30% of the evaluation datasets. The Kolm measure of inequality and the Gini-coefficient were the most influential features on the predictions of the model. It is remarkable that some evaluation conditions were different from those used in the training stage. Thus, the FF model trained in Goretzko and Bühner (2020) for the EFA framework will be explored in the present paper.

MODEL-FIT INDICES FOR DETERMINING THE NUMBER OF ATTRIBUTES IN CDM

All the aforementioned methods were developed with the purpose of assessing the number of factors in the EFA framework and, thus, their assumptions might not fit the nature of CDM data. **Table 1** shows that some of the most important assumptions required by some of the methods might be usually violated when analyzing CDM data. There is, however, one additional procedure that can be applied to CDMs without any further assumptions: the model comparison approach based on model-fit indices. This approach has also been widely explored in EFA. Previous studies have shown that, even though the traditional cutoff points for some commonly used fit indices (e.g., CFI, RMSEA, SRMR) are not recommended for determining the number of factors (Garrido et al., 2016), the relative difference in fit indices between competing models might even outperform PA under some conditions, such as small loadings, categorical data (Finch, 2020), or orthogonal factors (Lorenzo-Seva et al., 2011). Additionally, Preacher et al. (2013) recommended to use AIC for extracting the number of factors whenever the goal of the research was to find a model with an optimal parsimony-fit balance, while they recommended RMSEA whenever the goal was to retain the true, generating number of factors.

In the CDM framework, relative and absolute fit indices have been used to select the most appropriate Q-matrix specification. Regarding relative model-fit indices, Kunina-Habenicht et al. (2012) and Chen et al. (2013) found that Akaike's information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) perform really well at selecting the correct Q-matrix among competing matrices. In this vein, AIC and BIC always selected the correct Q-matrix when a three-attribute model was estimated for data generated from a five-attribute model, and vice versa (Kunina-Habenicht et al., 2012). Regarding absolute fit indices, Chen et al. (2013) proposed to inspect the residuals between the observed and predicted proportion correct of individual items (p_j), between the observed and predicted Fisher-transformed correlation of item pairs ($r_{jj'}$), and between the observed and predicted log-odds ratios of item pairs ($l_{jj'}$). Specifically, they used the p-value associated to the maximum z-scores of p_j, $r_{jj'}$, and $l_{jj'}$ to evaluate absolute fit. While p_j obtained very bad overall results, $r_{jj'}$ and $l_{jj'}$ performed appropriately at identifying both Q-matrix and CDM misspecification, with a tendency to be conservative.

TABLE 1 | Dimensionality assessment methods assumptions.

Method	Based on	Latent variable	Essential unidim.	Simple structure
PA	Eigenvalues	No	**No**	**No**
MAP	Partial correlations	No	**No**	**No**
VSS[a]	Factor loadings	Continuous	**No**	Yes
DETECT	Conditional covariances	Continuous	Yes	Yes
EKC	Eigenvalues	No	**No**	**No**
EGA	Network psychometrics	No	**No**	**No**
MC	Model-fit indices	**Discrete**	**No**	**No**

The methods assumptions that are aligned with cognitive diagnosis modeling characteristics are highlighted in bold. The factor forest (FF) model has not been included due to its dependence on the conditions employed for training the model. Essential unidim., essential unidimensionality; PA, parallel analysis with principal components extraction; MAP, minimum average partial; VSS, very simple structure; DETECT, dimensionality evaluation to enumerate contributing traits; EKC, empirical Kaiser criterion; EGA, exploratory graph analysis; MC, model comparison based on fit indices.

[a]*Simple structure (understood as a single factor being measured by each item) is technically assumed only by VSS with complexity $v = 1$.*

The aforementioned studies pointed out that these fit indices are promising for identifying the most appropriate Q-matrix. However, further research is required to examine their systematic performance in selecting the most appropriate number of attributes across a wide range of conditions. The use of fit indices to select the most appropriate model among a set of competing models, from 1 to K number of attributes, requires the calibration of K CDMs, each of them requiring a specified Q-matrix. This task demands an unfeasible amount of effort if done by domain experts, but it is viable if done by empirical means. The idea of using an empirical Q-matrix estimation method to generate Q-matrices for different number of attributes and then compare their model-fit has been already suggested by Chen et al. (2015). Furthermore, the `edina` package (Balamuta et al., 2020a) of the R software (R Core Team, 2020) incorporates a function to perform a Bayesian estimation of a DINA model (Chen et al., 2018) with different number of attributes, selecting the best model according to the BIC. In spite of these previous ideas, the performance of fit indices in selecting the best model among different number of attributes has not been evaluated in a systematic fashion, including both reduced (e.g., DINA) and general (e.g., G-DINA) CDMs. More details about the specific procedure used in the present study for assessing the number of attributes using model comparison are provided in the Method section.

GOALS OF THE CURRENT STUDY

The main goal of the present study is to compare the performance of several dimensionality assessment methods in determining the generating number of attributes in CDM. Additionally, following the approach of Auerswald and Moshagen (2019), the combined performance of the methods is also evaluated to explore whether a more accurate combination rule can be obtained and recommended for applied settings. As a secondary goal, the effect of a comprehensive set of independent variables and their interactions over the accuracy of the procedures is systematically evaluated.

Table 1 provides the basis for establishing some hypotheses related to the performance of the methods. First, while CDMs are discrete latent variable models, most methods do not consider the existence of latent variables (PA with principal components extraction, MAP, EKC, EGA) or consider the existence of continuous latent variables (VSS, DETECT). The violation of this assumption might not be too detrimental, given that PA with component analysis violates EFA assumptions and is the current gold standard. On the other hand, both essential unidimensionality and simple structure assumptions are expected to have a great disruptive effect, since CDMs are usually highly multidimensional and often contain multidimensional items. Accordingly, VSS with $v = 1$ (VSS_1) and DETECT are expected to perform poorly. Although VSS with complexity $v > 1$ is not technically assuming a simple structure (understood as a single attribute being measured by each item), its performance is still expected to be poor because of its stiffness and inability to adapt to the usual complex structure (i.e., items measuring a different number of attributes) of CDM items. Even though the remaining methods (i.e., PA, EKC, MAP, and EGA) do not assume a simple structure, their performance under complex structures remains mostly unexplored. Assessing the dimensionality of complex structures is expected to be more challenging compared to simple structures, in a similar fashion as correlated factors are more difficult to extract than orthogonal factors. The extent to which the performance of these methods is robust under complex structures is unknown. All in all, and considering the assumptions of each method, PA, EKC, MAP, and EGA, as well as the CDM model comparison approach based on fit indices (MC), are expected to perform relatively well, except for their idiosyncratic weakness conditions found in the available literature as previously described. Finally, the performance of FF is difficult to predict due to its dependency on the training samples. Even though no training samples were generated based on discrete latent variables in Goretzko and Bühner (2020), the great overall performance and generalizability of the FF model to conditions different from the ones used to train the model might extend to CDM data as well.

METHODS

Dimensionality Estimation Methods

Eight different dimensionality estimation methods, with a total of 18 variants, were used in the present simulation study.

The following text describes the specific implementation of each method.

Parallel Analysis

Four variants of PA were implemented as a function of the correlation matrix type (r = Pearson; ρ = tetrachoric) and the reference eigenvalue criterion (m = mean; 95 = 95th percentile): PA_{rm}, PA_{r95}, $PA_{\rho m}$, and $PA_{\rho 95}$. All variants were implemented with principal components extraction and 100 random samples generated by random column permutation (Garrido et al., 2013; Lubbe, 2019). The `sirt` package (Robitzsch, 2020) was used to estimate the tetrachoric correlations for PA, as well as for the remaining methods that also make use of tetrachoric correlations.

Minimum Average Partial

MAP indices were based on the squared partial tetrachoric correlations computed with the `psych` package (Revelle, 2019). The maximum number of dimensions to extract was set to 9 (same for VSS, DETECT, and MC), so there was room for overestimating the number of attributes (the details of the simulation study are provided in the Design subsection).

Very Simple Structure

VSS was computed using tetrachoric correlations and the `psych` package. In addition to the most common VSS with complexity $v = 1$ (VSS_1), VSS with complexity $v = 2$ (VSS_2) was also explored.

Dimensionality Evaluation to Enumerate Contributing Traits

The DETECT index was computed using the `sirt` package, which uses the hierarchical Ward algorithm (Roussos et al., 1998) for clustering the items.

Empirical Kaiser Criterion

EKC was implemented by using tetrachoric correlations and the `semTools` package (Jorgensen et al., 2019).

Exploratory Graph Analysis

Two variants of EGA were implemented: EGA with GGM (EGA_G) and EGA with TMFG (EGA_T). The `EGAnet` package (Golino and Christensen, 2020) was employed for computing both variants.

Factor Forest

The R code published by Goretzko and Bühner (2020) at *Open Science Framework*[1] was used for the implementation of the FF model trained in their original paper. With this code, FF can recommend between one and eight factors to retain.

Model Comparison Based on Fit Indices

The MC procedure was implemented varying the number of attributes from 1 to 9 as follows. First, the DFL Q-matrix estimation method (Wang et al., 2018) using Oblimin oblique rotation, tetrachoric correlations, and the row dichotomization criterion was used to specify the initial Q-matrix, and the Hull validation method (Nájera et al., 2020) using the PVAF index was then implemented to refine it and provide the final Q-matrix. Second, a CDM was fitted to the data using the final Q-matrix with the `GDINA` package (Ma and de la Torre, 2020b). The CDM employed to fit the data was the same as the generating CDM (i.e., DINA or G-DINA). This resulted in a set of nine competing models varying in K. Third, the models were alternatively compared with the AIC, BIC, and $r_{jj'}$ fit indices. For the AIC and BIC criteria, the model with the lowest value was retained. Regarding $r_{jj'}$, the number of items with some significant pairwise residual (after using Bonferroni correction at the item-level) was counted. Then, the most parsimonious model with the lowest count was retained. The MC procedure with AIC, BIC, or $r_{jj'}$ will be referred to as MC_{AIC}, MC_{BIC}, and MC_r, respectively. Given that the MC procedures rely on empirically specified Q-matrices, their performance will greatly depend on the quality of such Q-matrices. Even though the DFL and Hull methods have provided good results in previous studies, their combined performance should be evaluated to examine the quality of their suggested Q-matrices. For this reason, the proportion of correctly specified q-entries was computed for the estimated (i.e., DFL) and validated (i.e., DFL and Hull) Q-matrices (more details are provided in the Dependent variables subsection). The further the DFL and Hull methods are from a perfect Q-matrix recovery, the greater the room for improvement for the MC procedures. In this vein, the set of nine competing models (using DFL and Hull) were additionally compared to the model using the generating Q-matrix, with the purpose of providing an upper-limit performance for the MC methods when the Q-matrix is perfectly recovered. The results of these comparisons will be referred to as MC_{AIC-G}, MC_{BIC-G}, and MC_{r-G}.

Design

Table 2 shows the factors (i.e., independent variables) used in the simulation study: number of attributes (K), item quality (IQ), sample size (N), ratio of number of items to attribute (JK), underlying correlation among the attributes (AC), and attribute thresholds (AT). The levels of each factor were selected in pursuit of representativeness of varying applied settings. For instance, the most common number of attributes (K) seen in applied studies is 4 (Sessoms and Henson, 2018), while 5 is the most usual value in simulation studies (e.g., de la Torre and Chiu, 2016; Ma and de la Torre, 2020a). The levels selected for item quality (IQ), sample size (N), and ratio of number of items to attribute (JK) are also considered as representative of applied settings

TABLE 2 | Summary of the factors explored in the simulation study.

Factors	Factor levels
Number of attributes (K)	4, 5, 6
Item quality (IQ)	0.40, 0.60, 0.80
Sample size (N)	500, 1,000, 2,000
Ratio of number of items to attribute (JK)	4, 8
Correlation among the attributes (AC)	0, 0.30, 0.60
Attribute thresholds (AT)	0, 0.50, 1
Generating model (M)	DINA, G-DINA

[1] https://osf.io/mvrau/

(Nájera et al., 2019; Ma and de la Torre, 2020a). Regarding the attribute correlations, some applied studies have obtained very high attribute correlation coefficients, up to 0.90 (Sessoms and Henson, 2018). It can be argued that these extremely high correlations may be indeed a consequence of overestimating the number of attributes, where one attribute has been split into two or more undifferentiated attributes. For this reason, we decided to use *AC* levels similar to those used in EFA simulation studies (e.g., Garrido et al., 2013). Additionally, different attribute thresholds (*AT*) levels were included to generate different degrees of skewness in the data, given its importance in the performance of dimensionality assessment methods (Garrido et al., 2013). Finally, both a reduced CDM (i.e., DINA) and a general CDM (i.e., G-DINA) were used to generate data. A total of 972 conditions, resulting from the combination of the factor levels, were explored.

Data Generation

One hundred datasets were generated per condition. Examinees' responses were generated using either the DINA or G-DINA model. Attribute distributions were generated using a multivariate normal distribution with mean equal to 0 for all attributes. All underlying attribute correlations were set to the corresponding *AC* condition level. Attribute thresholds, which are used to dichotomize the multivariate normal distribution to determine the mastery or non-mastery of the attributes, were generated following an equidistance sequence of length K between $-AT$ and AT. This results in approximately half of the attributes being "easier" (i.e., higher probabilities of attribute mastery) and the other half being "more difficult" (i.e., lower probabilities of attribute mastery). For instance, for $AT = 0.50$ and $K = 5$, the generating attributes thresholds were $\{-0.50, -0.25, 0, 0.25, 0.50\}$.

Item quality was generated by varying the highest and lowest probabilities of success, which correspond to the latent classes that master all, $P(\mathbf{1})$, and none, $P(\mathbf{0})$, of the attributes involved in an item, respectively. These probabilities were drawn from uniform distributions as follows: $P(\mathbf{0}) \sim U(0, 0.20)$ and $P(\mathbf{1}) \sim U(0.80, 1)$ for high-quality items, $P(\mathbf{0}) \sim U(0.10, 0.30)$ and $P(\mathbf{1}) \sim U(0.70, 0.90)$ for medium-quality items, and $P(\mathbf{0}) \sim U(0.20, 0.40)$ and $P(\mathbf{1}) \sim U(0.60, 0.80)$ for low-quality items. The expected value for the item quality across the J items is then 0.80, 0.60, and 0.40 for high, medium, and low-quality items, respectively. For the G-DINA model, the probabilities of success for the remaining latent classes were simulated randomly, with two constraints. First, a monotonicity constraint on the number of attributes was applied. Second, the sum of the δ parameters associated to each attribute was constrained to be higher than 0.15 to ensure the relevance of all the attributes (Nájera et al., 2020).

The Q-matrices were generated randomly with the following constraints: (a) each Q-matrix contained, at least, two identity matrices; (b) apart from the identity matrices, each attribute was measured, at least, by another item; (c) the correlation between attributes (i.e., Q-matrix columns) was lower than 0.50; (d) the proportion of one-, two-, and three-attribute items was set to 0.50, 0.40, and 0.10. Constrains (a) and (b) are in line with the identifiability recommendations made by Xu and Shang (2018). Constrain (c) ensures non overlapping attributes. Finally, constrain (d) was based on the proportion of items measuring one-, two-, and three-attributes encountered in previous literature. We examined the 36 applied studies included in the literature revision by Sessoms and Henson (2018) and extracted the complexity of the q-vectors from the 17 studies that reported the Q-matrix (see **Table 3**). The reason why we used a higher proportion of one-attribute items was to preserve constrain a). For instance, in the condition of $JK = 4$, at least 50% of one-attribute items are required to form two identity matrices.

Dependent Variables

Four dependent variables were used to assess the accuracy of the dimensionality assessment methods. The hit rate (*HR*) was the main dependent variable, computed as the proportion of correct estimates:

$$HR = \frac{\sum I\left(\hat{K} = K\right)}{R} \tag{6}$$

where I is the indicator function, $\hat{K}$ is the recommended number of attributes, K is the generating number of attributes, and R is the number of replicates per condition (i.e., 100). A *HR* of 1 indicates a perfect accuracy, while an *HR* of 0 indicates complete lack of accuracy. Additionally, given that a model selection must be done according to both empirical and theoretical criteria, it is a recommended approach to examine alternative models to the one suggested by a dimensionality assessment method (e.g., Fabrigar et al., 1999). The close hit rate (*CHR*) was assessed to explore the proportion of times that a method recommended a number of attributes close to the generating number of attributes:

$$CHR = \frac{\sum I\left[(K-1) \leq \hat{K} \leq (K+1)\right]}{R} \tag{7}$$

Finally, the mean error (*ME*) and root mean squared error (*RMSE*) were explored to assess the bias and inaccuracy of the methods:

$$ME = \frac{\sum\left(\hat{K} - K\right)}{R} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum\left(\hat{K} - K\right)^2}{R}} \tag{9}$$

TABLE 3 | Complexity of q-vectors in applied studies (percentages).

	$q = 1$	$q = 2$	$q = 3$	$q > 3$	$q = K$
Mean	34.9	42.7	14.2	8.2	1.5
Median	30.4	43.5	10.8	0	0

(q = 1, 2, 3) = q-vectors measuring 1, 2, or 3 attributes, respectively; (q > 3) = q-vectors measuring more than 3 attributes; (q = K) = q-vectors measuring all the attributes included in the Q-matrix.

A *ME* of 0 indicates lack of bias, while a negative or positive *ME* indicates a tendency to underestimate or overestimate the number of attributes, respectively. It is important to note that an *ME* close to 0 can be achieved either by an accurate method, or by a compensation of under- and overestimation. On the contrary, *RMSE* can only obtain positive values: the further from 0, the greater the inaccuracy of a method.

Univariate ANOVAs were conducted to explore the effect of the factors on the performance of each method. The dependent variables for the ANOVAs were the hit rate, close hit rate, bias, and absolute error, which correspond to the numerators of Equations (6)–(9) (i.e., *HR*, *CHR*, *ME*, and *RMSE* at the replica-level), respectively. Note that the *RMSE* computed at the replica level is the absolute error (i.e., $\left|\hat{K} - K\right|$). Effects with a partial eta-squared (η_p^2) higher than 0.060 and 0.140 were considered as medium and large effects, respectively (Cohen, 1988).

In order to explore the performance of the combination rules (i.e., two or more methods taken together), the agreement rate (*AR*) was used to measure the proportion of conditions under which a combination rule recommended the same number of attributes, while the agreement hit rate (*AHR*) was used to measure the proportion of correct estimations among those conditions in which an agreement has been achieved:

$$AR = \frac{\sum I\left(\hat{K}_1 = \hat{K}_2\right)}{R} \tag{10}$$

$$AHR = \frac{\sum I\left(\hat{K}_1 = K | \hat{K}_1 = \hat{K}_2\right)}{\sum I\left(\hat{K}_1 = \hat{K}_2\right)} \tag{11}$$

where $\hat{K}_1$ and $\hat{K}_2$ are the recommended number of attributes by any two different methods. Note that these formulas can be easily generalized for more than two methods. Both a high *AR* and *AHR* are required for a combination rule to be satisfactory, since this indicates that it will be accurate and often applicable (Auerswald and Moshagen, 2019).

Finally, for the MC methods, when the model under exploration had the same number of attributes as the generating number of attributes, the Q-matrix recovery rate (*QRR*) was explored to assess the accuracy of the DFL and Hull methods. Specifically, it reflects the proportion of correctly specified q-entries. A *QRR* of 1 indicates perfect recovery. The higher the *QRR*, the closer the methods based on model-fit indices (e.g., MC_{AIC}) should be to their upper-limit performance (e.g., using the generating Q-matrix as in MC_{AIC-G}). All simulations and analyses were conducted using the R software. The data were simulated using the `GDINA` package. The codes are available upon request.

RESULTS

Before describing the main results, the results for the *QRR* are detailed. The overall *QRR* obtained after implementing both the DFL and Hull method was 0.949. The lowest and highest *QRR* among the factor levels were obtained with $IQ = 0.40$ ($QRR = 0.890$) and $IQ = 0.80$ ($QRR = 0.985$), respectively. These results are consistent with Nájera et al. (2020). The DFL method alone (i.e., before validating the Q-matrix with the Hull method) led to a good overall accuracy ($QRR = 0.939$). However, despite this high baseline, the Hull method led to a *QRR* improvement across all factor levels ($\Delta QRR = [0.005, 0.013]$).

Table 4 shows the overall average results, across all conditions, for all the variants and dependent variables considered. The four PA variants, FF, and MC_{AIC} performed reasonably well, with a $HR > 0.700$ and a $CHR > 0.900$. EGA_G also obtained a high *CHR* ($CHR = 0.918$), but a much lower *HR* ($HR = 0.576$). The highest *HR* was obtained by PA_{rm} ($HR = 0.829$), while the highest *CHR* was provided by MC_{AIC} ($CHR = 0.954$). Congruently with these results, the PA variants, FF, MC_{AIC}, and EGA_G showed a low *RMSE* ($RMSE < 1$), being the MC_{AIC} the method with the lowest error ($RMSE = 0.633$). The remaining methods (i.e., MAP, VSS_1, VSS_2, DETECT, EKC, EGA_T, MC_{BIC}, and MC_r) obtained a poorer performance ($HR \leq 0.682$ and $CHR \leq 0.853$). Regarding the bias, most methods showed a tendency to underestimate the

TABLE 4 | Overall performance for all dimensionality estimation methods.

Method	*HR*	*CHR*	*ME*	*RMSE*
PA_{rm}	**0.829**	0.947	−0.144	0.681
PA_{r95}	0.801	0.919	−0.309	0.876
PA_{pm}	0.805	0.938	−0.190	0.734
PA_{p95}	0.770	0.904	−0.369	0.941
MAP	0.518	0.618	−1.337	2.205
VSS_1	0.278	0.378	−1.045	3.135
VSS_2	0.424	0.522	−0.040	2.339
DETECT	0.492	0.691	1.043	1.956
EKC	0.502	0.625	2.200	3.988
EGA_G	0.576	0.918	−0.257	0.870
EGA_T	0.337	0.782	−0.831	1.231
FF	0.824	0.918	−0.191	0.746
MC_{AIC}	0.768	**0.954**	0.010	**0.633**
MC_{BIC}	0.682	0.824	−0.620	1.289
MC_r	0.635	0.853	0.096	1.027
MC_{AIC-G}	**0.886**	**0.975**	−0.022	**0.458**
MC_{BIC-G}	0.713	0.829	−0.593	1.270
MC_{r-G}	0.814	0.922	−0.086	0.774

The dashed line separates the MC methods that are implemented using the generating Q-matrix (i.e., MC_{AIC-G}, MC_{BIC-G}, MC_{r-G}). Best results for HR, CHR, and RMSE are shown in bold, considering the MC methods with the generating Q-matrix separately. HR ≥ 0.700 and CHR ≥ 0.900 results are underlined. HR, hit rate; CHR, close hit rate; ME, mean error; RMSE, root mean squared error; PA_r, parallel analysis with Pearson correlations; PA_p, parallel analysis with tetrachoric correlations; PA_m, parallel analysis with mean eigenvalue criterion; PA_{95}, parallel analysis with 95th percentile eigenvalue criterion; MAP, minimum average partial; VSS_1, very simple structure with complexity v = 1; VSS_2, very simple structure with complexity v = 2; DETECT, dimensionality evaluation to enumerate contributing traits; EKC, empirical Kaiser criterion; EGA_G, exploratory graph analysis with Gaussian graphical model; EGA_T, exploratory graph analysis with triangulated maximally filtered graph; FF, factor forest; MC_{AIC}, model comparison based on AIC; MC_{BIC}, model comparison based on BIC; MC_r, model comparison based on the Fisher-transformed correlations; $MC_{\cdot-G}$, model comparison using the generating Q-matrix when the number of attributes coincides with the generating number of attributes.

number of attributes, especially MAP, VSS_1, and EGA_T ($ME \leq -0.831$). On the contrary, EKC and DETECT showed a tendency to overestimate the number of attributes ($ME \geq 1.043$). Among the methods with low *RMSE*, FF ($ME = -0.191$), PA_{rm} ($ME = -0.144$), and, especially, MC_{AIC} ($ME = 0.010$), showed a very low bias. Finally, the MC methods that rely on the generating Q-matrix (i.e., MC_{AIC-G}, MC_{BIC-G}, MC_{r-G}) generally provided good results, outperforming their corresponding MC method. Specifically, MC_{AIC-G} obtained the highest overall accuracy (*HR* = 0.886; *CHR* = 0.975; *RMSE* = 0.458).

Table 5 shows the results for the methods that obtained the best overall performance as indicated by *CHR* > 0.900 (i.e., PA_{rm}, EGA_G, FF, and MC_{AIC}) across the factor levels. Only PA_{rm} is shown among the PA variants because their results were congruent and PA_{rm} obtained the better overall performance. In addition to these methods, the MC_{AIC-G} is also included to provide a comparison with MC_{AIC}. The results from **Table 5** can be easier interpreted by inspecting the main effect size values obtained in the ANOVAs (see **Table 6**). These main effects offer a proper summary of the results since only one interaction effect, which will be described below, was relevant ($\eta_p^2 > 0.140$).

Regarding the hit rate, *IQ* was the factor that most affected all the methods ($\eta_p^2 \geq 0.161$), except for EGA_G. These methods performed very accurately with $IQ = 0.80$ ($HR \geq 0.916$), but poorly with $IQ = 0.40$ ($HR \leq 0.675$). Other factors obtained a medium effect size for one specific method. EGA_G was affected by *M*, obtaining a higher accuracy with the G-DINA model (*HR*= 0.709) than with the DINA model ($HR = 0.443$). On the other hand, PA_{rm} was affected by *N* and *AC*. PA_{rm} obtained the highest *HR* in most conditions, especially with large sample sizes ($N = 2000$), but was negatively affected by high correlations among the attributes ($AC = 0.60$). In the cases in which PA_{rm} was not the best performing method, the FF obtained the highest accuracy. FF obtained the biggest advantage in comparison to the other methods under $N = 500$ ($\Delta HR = 0.054$). Results for the *RMSE* showed a very similar pattern to those from *HR*. The most notable difference was that MC_{AIC} obtained the lowest *RMSE* under most conditions, especially $AC = 0.60$ ($|\Delta RMSE| = 0.239$). On the contrary, PA_{rm} showed a smaller error with lower attribute correlations, especially $AC = 0.30$ ($|\Delta RMSE| = 0.213$).

The close hit rate of the methods was more robust than the *HR* to the different simulation conditions. Only PA_{rm} and, especially, FF were affected by *IQ*. PA_{rm} was also affected by *AC*, and FF by *JK*. The *CHR* of both EGA_G and MC_{AIC} remained stable across the factor levels. The highest *CHR* was obtained by MC_{AIC} in most conditions, obtaining the biggest advantage under $AC = 0.60$ ($\Delta CHR = 0.059$). PA_{rm} and FF provided the highest *CHR* in those conditions in which MC_{AIC} did not obtain the best result.

With respect to the bias (i.e., *ME*), **Table 6** shows that the only large effect was observed for *AC* on PA_{rm}. However, there was a relevant effect for the interaction between *AC* and *IQ* ($\eta_p^2 = 0.194$). This was the only interaction with a large effect among all the ANOVAs. Namely, the strong tendency to underestimate seen for PA_{rm} under $AC = 0.60$ was mainly due to $IQ = 0.40$. Thus, under $IQ = 0.40$, PA_{rm} showed a strong tendency to underestimate when $AC = 0.60$ ($ME = -1.108$), but a slight tendency to overestimate when $AC = 0$ ($ME = 0.258$). With *IQ* ≥ 0.60 and $AC \leq 0.30$, PA_{rm} obtained a low bias ($ME \leq |0.055|$). Apart from this interaction, other factors with relevant effect sizes were: a) *K*, which had an effect on EGA_G and MC_{AIC}; b) *IQ*, with an effect on PA_{rm}, FF, and MC_{AIC}; and c) *JK*, which had an effect on PA_{rm}, EGA_G, and MC_{AIC}. In general, the most demanding levels of these factors (i.e., $K = 6$, $IQ = 0.40$, $JK = 4$) led to an underestimation tendency for the methods. Finally, while PA_{rm}, EGA_G, and FF showed a negative bias (i.e., $ME < 0$) across almost all conditions, MC_{AIC} showed a positive bias (i.e., $ME > 0$) under several conditions, especially $K = 4$ and the G-DINA model ($ME \geq 0.150$).

Finally, MC_{AIC-G} performed the best under almost all conditions and dependent variables, with the only exception of $AC \leq 0.30$ and G-DINA generated data, where PA_{rm} obtained slightly better results. As expected, MC_{AIC-G} outperformed MC_{AIC} under all conditions. The ANOVA effects were similar for both methods. One of the main differences is that the *HR* of MC_{AIC-G} was more affected by the sample size (a steeper *HR* improvement as *N* increased) and the generating model (performing comparatively better under the DINA model). On the other hand, the *ME* of MC_{AIC-G} was more robust under different levels of *K* and *M*.

Table 7 shows the results for the combination rules split by sample size. VSS_1, VSS_2, DETECT, EKC, and EGA_T are not included because they were not usually consistent with any other method (i.e., $AR < 0.50$). Both the *AR* and the *AHR* tended to increase as the sample size increased. As expected from the results above, the best performing combination rules were mainly formed by PA (especially PA_{rm}), FF, and MC (especially MC_{AIC}). The combination rule formed by PA_{rm} and FF obtained arguably the best balance between agreement and accuracy ($AR \geq 0.70$; $AHR \geq 0.923$), while FF and MC_{AIC} obtained a higher accuracy with a slightly lower agreement ($AR \geq 0.65$; $AHR \geq 0.953$). The best accuracy was obtained by the combination rule formed by MC_{AIC} and MAP ($AHR \geq 0.980$), although at the cost of a lower agreement ($AH \approx 0.46$). In addition to these two-method combination rules, the performance of the three best methods (i.e., PA_{rm}, FF, and MC_{AIC}) taken together was also explored. This combination rule showed a very high overall accuracy while keeping an $AR > 0.50$. Specifically, for $N = 500$, 1000, and 2000, it obtained *AHR (AR)* = 0.976 (0.57), 0.985 (0.65), and 0.992 (0.70), respectively.

REAL DATA EXAMPLE

Real data were analyzed to illustrate the performance of the dimensionality estimation methods explored in the simulation study. This section can be also understood as an illustration of how to approach the problem of determining the number of attributes in applied settings. The data employed for this example was previously analyzed by Chen et al. (2020). The dataset consists of dichotomous responses from 400 participants to 20 items from an intelligence test. Each item consists of nine matrices forming a 3 rows × 3 columns disposition, in which the ninth matrix (i.e., the lower right) is missing. Participants must select the missing matrix out of eight possible options. There

TABLE 5 | Performance of the best methods by factor level.

	K			*IQ*			*N*			*JK*		*AC*			*AT*			*M*	
Method	**4**	**5**	**6**	**0.40**	**0.60**	**0.80**	**500**	**1,000**	**2,000**	**4**	**8**	**0**	**0.30**	**0.60**	**0**	**0.50**	**1**	**D**	**G-D**
Hit rate (HR)																			
PA_{rm}	**0.870**	**0.828**	0.789	**0.606**	**0.911**	0.970	0.726	**0.844**	**0.918**	**0.772**	**0.886**	0.895	**0.905**	0.687	0.857	0.838	**0.793**	**0.810**	**0.848**
EGA_G	0.662	0.571	0.494	0.447	0.625	0.655	0.543	0.590	0.594	0.525	0.627	0.607	0.591	0.529	0.565	0.586	0.576	0.443	0.709
FF	0.858	0.785	**0.827**	0.596	0.894	**0.981**	**0.780**	0.823	0.868	0.768	0.879	**0.897**	0.832	**0.741**	**0.883**	**0.843**	0.745	0.803	0.845
MC_{AIC}	0.721	0.788	0.794	0.552	0.835	0.916	0.719	0.774	0.811	0.754	0.782	0.812	0.774	0.718	0.811	0.772	0.721	0.791	0.745
MC_{AIC-G}	**0.899**	**0.893**	**0.866**	**0.675**	**0.984**	**0.999**	**0.796**	**0.895**	**0.967**	**0.863**	**0.909**	0.895	0.890	**0.872**	**0.907**	**0.897**	**0.853**	**0.966**	0.806
Close hit rate (CHR)																			
PA_{rm}	**0.973**	0.949	0.921	0.860	**0.983**	**0.999**	0.900	0.959	**0.983**	0.913	0.982	**0.979**	**0.983**	0.879	0.942	0.951	**0.949**	0.947	0.948
EGA_G	0.951	0.929	0.875	0.855	0.954	0.946	0.898	0.926	0.931	0.902	0.934	0.944	0.934	0.877	0.879	0.934	0.942	0.891	0.945
FF	0.936	0.961	0.858	0.787	0.969	**0.999**	0.886	0.918	0.950	0.854	**0.983**	0.960	0.926	0.869	0.939	0.924	0.891	0.916	0.921
MC_{AIC}	0.949	**0.970**	**0.942**	**0.891**	0.980	0.989	**0.928**	**0.961**	0.972	**0.948**	0.960	0.966	0.957	**0.938**	**0.964**	**0.956**	0.941	**0.948**	**0.959**
MC_{AIC-G}	**0.979**	**0.981**	**0.963**	**0.924**	**0.999**	**1**	**0.945**	**0.982**	**0.997**	**0.968**	0.981	0.978	0.977	**0.969**	**0.980**	**0.979**	**0.965**	**0.982**	**0.967**
Mean error (ME)																			
PA_{rm}	−0.076	−0.136	−0.219	−0.325	−0.100	−0.006	−0.228	−0.134	−0.068	−0.271	−0.016	0.100	−0.044	−0.487	−0.174	−0.148	−0.109	−0.150	−0.137
EGA_G	−0.007	−0.266	−0.498	−0.412	−0.275	−0.083	−0.240	−0.249	−0.281	−0.507	−0.007	−0.164	−0.217	−0.389	−0.169	−0.291	−0.310	−0.282	−0.232
FF	−0.128	−0.258	−0.188	−0.430	−0.137	−0.008	−0.161	−0.227	−0.186	−0.309	−0.074	−0.126	−0.202	−0.245	−0.104	−0.178	−0.292	−0.217	−0.166
MC_{AIC}	0.232	0.006	−0.207	−0.203	0.140	0.095	−0.107	0.021	0.118	−0.124	0.145	0.038	0.031	−0.037	0.047	0.042	−0.057	−0.129	0.150
MC_{AIC-G}	0.067	−0.003	−0.129	−0.069	0.004	0.001	−0.069	0.002	0.003	−0.127	0.084	0.011	−0.012	−0.064	0.007	−0.009	−0.063	−0.056	0.013
Root mean squared error (RMSE)																			
PA_{rm}	**0.491**	0.661	0.844	1.098	**0.389**	0.184	0.926	0.619	**0.386**	0.858	**0.436**	**0.447**	**0.403**	1.014	0.708	0.656	**0.677**	0.690	0.672
EGA_G	0.749	0.839	1.003	1.076	0.737	0.754	0.940	0.840	0.826	0.928	0.808	0.777	0.817	0.999	0.978	0.817	0.804	1.002	0.713
FF	0.685	0.663	0.871	1.201	0.452	**0.149**	0.884	0.724	0.604	0.921	0.514	0.520	0.693	0.958	0.635	0.708	0.875	0.759	0.733
MC_{AIC}	0.671	**0.555**	**0.668**	**0.925**	0.477	0.348	**0.748**	**0.600**	0.533	**0.669**	0.596	0.555	0.616	**0.719**	**0.562**	**0.624**	0.706	**0.642**	**0.625**
MC_{AIC-G}	**0.416**	**0.405**	**0.539**	**0.781**	**0.134**	**0.024**	**0.646**	**0.411**	**0.205**	**0.513**	**0.395**	**0.425**	0.441	**0.502**	**0.406**	**0.429**	**0.528**	**0.339**	**0.551**

Best results for HR, CHR, and RMSE are shown in bold, considering MC_{AIC-G} separately. Results for HR, CHR, and RMSE that are higher than MC_{AIC-G} are also underlined. K, number of attributes; IQ, item quality; N, sample size; JK, ratio of the number of items to attribute; AC, correlation among the attributes; AT, attribute thresholds; M, generating model; D, DINA model; G-D, G-DINA model; PA_{rm}, parallel analysis with Pearson correlations and mean eigenvalue criterion; EGA_G, exploratory graph analysis with Gaussian graphical model; FF, factor forest; MC_{AIC}, model comparison based on AIC; MC_{AIC-G}, model comparison based on AIC and using the generating Q-matrix when the number of attributes coincides with the generating number of attributes.

TABLE 6 | Univariate ANOVAs main effect size values (η_p^2).

	K	*IQ*	*N*	*JK*	*AC*	*AT*	*M*
Hit rate							
PA_{rm}	0.014	**0.251**	0.076	0.041	0.117	0.009	0.005
EGA_G	0.024	0.042	0.003	0.014	0.006	0.000	0.085
FF	0.009	**0.218**	0.013	0.031	0.041	0.034	0.005
MC_{AIC}	0.009	**0.161**	0.011	0.001	0.012	0.011	0.004
MC_{AIC-G}	0.005	**0.344**	0.103	0.012	0.002	0.013	0.131
Close hit rate							
PA_{rm}	0.014	0.109	0.037	0.036	0.068	0.000	0.000
EGA_G	0.016	0.031	0.003	0.004	0.014	0.012	0.011
FF	0.040	**0.161**	0.015	0.084	0.030	0.009	0.000
MC_{AIC}	0.004	0.051	0.010	0.001	0.004	0.003	0.001
MC_{AIC-G}	0.003	0.059	0.023	0.002	0.001	0.002	0.003
Bias							
PA_{rm}	0.015	0.073	0.019	0.067	**0.216**	0.003	0.000
EGA_G	0.074	0.035	0.001	0.111	0.018	0.008	0.001
FF	0.007	0.069	0.002	0.032	0.006	0.014	0.002
MC_{AIC}	0.126	0.094	0.037	0.075	0.005	0.010	0.080
MC_{AIC-G}	0.048	0.009	0.009	0.078	0.007	0.007	0.009
Absolute error							
PA_{rm}	0.026	**0.260**	0.089	0.064	**0.145**	0.002	0.002
EGA_G	0.029	0.056	0.005	0.013	0.014	0.005	0.071
FF	0.008	**0.223**	0.017	0.054	0.045	0.024	0.002
MC_{AIC}	0.006	**0.167**	0.017	0.002	0.013	0.011	0.001
MC_{AIC-G}	0.007	**0.280**	0.090	0.011	0.003	0.010	0.074

$\eta_p^2 > 0.060$ and $\eta_p^2 > 0.140$ are shown underlined and bolded, respectively. K, number of attributes; IQ, item quality; N, sample size; JK, ratio of the number of items to attribute; AC, correlation among the attributes; AT, attribute thresholds; M, generating model; D, DINA model; G-D, G-DINA model; PA_{rm}, parallel analysis with Pearson correlations and mean eigenvalue criterion; EGA_G, exploratory graph analysis with Gaussian graphical model; FF, factor forest; MC_{AIC}, model comparison based on AIC; MC_{AIC-G}, model comparison based on AIC and using the generating Q-matrix when the number of attributes coincides with the generating number of attributes.

are no missing data. The dataset is available at the `edmdata` package (Balamuta et al., 2020b) and item definitions can be found at *Open Psychometrics*.[2] Chen et al. (2020) defined four attributes involved in the test: (a) learn the pattern from the first two rows and apply it to the third row, (b) infer the best overall pattern from the whole set of matrices, (c) recognize that the missing matrix is different from the given matrices (e.g., applying rotations or stretching), and (d) recognize that the missing matrix is exactly as one of the given matrices. The authors did not explicitly define a Q-matrix for this dataset because they focused on the exploratory estimation of the item parameters. However, they described a procedure to derive a Q-matrix from the item parameter estimates by dichotomizing the standardized coefficients related to each attribute (Chen et al., 2020, pp. 136). This original Q-matrix, which is here referred to as Q_O, is shown in **Figure 2**.

According to the findings from the simulation study, the following steps are recommended to empirically determine the number of attributes in CDM data: (a) if PA_{rm}, FF, and MC_{AIC} agree on their suggestion, retain their recommended number of attributes; (b) if any two of these methods agree, retain their recommended number of attributes; (c) if none of these methods agree, explore the recommended number of attributes by those that suggest a similar (i.e., ± 1) number of attributes; (d) if these methods strongly disagree, explore the recommended number of attributes by each of them. Constructing several Q-matrices is a very challenging and time-consuming process for domain experts; thus, the Q-matrices suggested by the DFL and Hull methods (which are already used to implement the MC methods), can be used as a first approximation. Domain experts should be consulted to contrast the interpretability of these Q-matrices.

All the dimensionality assessment methods included in the simulation study were used to assess the number of attributes of the dataset. Their recommendations were as follows: 1 attribute was retained by MAP and VSS_1; 2 attributes by $PA_{\rho 95}$, $PA_{\rho m}$, VSS_2, and EGA_T; 3 attributes by PA_{rm}, PA_{r95}, and MC_{BIC}; 4 attributes by EGA_G, MC_{AIC}, and MC_r; 5 attributes by EKC and FF; and 8 attributes by DETECT. In accordance with the simulation study results, MAP, VSS_1, and EGA_T, which showed a tendency to underestimate, suggested a low number of attributes, while DETECT, which showed a strong tendency to overestimate, suggested the highest number of attributes.

Following the previously described guidelines, we focused on the recommendations of PA_{rm}, MC_{AIC}, and FF (i.e., 3, 4, and

[2]https://openpsychometrics.org/_rawdata/

TABLE 7 | Performance of the combination rules by sample size.

	PA_{rm}	PA_{r95}	$PA_{\rho m}$	$PA_{\rho 95}$	MAP	EGA_G	FF	MC_{AIC}	MC_{BIC}	MC_r
***N* = 500**										
PA_{r95}	**0.816** (0.80)									
$PA_{\rho m}$	**0.732** (0.93)	**0.789** (0.80)								
$PA_{\rho 95}$	**0.828** (0.74)	**0.697** (0.92)	**0.777** (0.79)							
MAP	0.939 (0.52)	0.876 (0.55)	0.926 (0.50)	0.854 (0.53)						
EGA_G	0.848 (0.54)	0.841 (0.52)	0.837 (0.51)	*0.833* (0.49)	0.904 (0.38)					
FF	**0.923** (0.70)	**0.896** (0.70)	0.913 (0.67)	0.885 (0.67)	0.935 (0.53)	0.903 (0.52)				
MC_{AIC}	0.914 (0.66)	0.912 (0.64)	0.911 (0.62)	0.910 (0.60)	*0.980* (0.47)	*0.862* (0.49)	0.953 (0.65)			
MC_{BIC}	0.885 (0.61)	0.843 (0.62)	0.873 (0.58)	0.834 (0.59)	0.849 (0.57)	*0.837* (0.45)	0.873 (0.62)	0.884 (0.59)		
MC_r	0.893 (0.56)	0.861 (0.56)	0.887 (0.53)	0.854 (0.53)	*0.915* (0.42)	*0.847* (0.40)	0.915 (0.56)	0.849 (0.61)	0.843 (0.52)	
Single	0.726	0.687	0.687	0.646	0.507	0.543	0.780	0.719	0.572	0.584
***N* = 1,000**										
PA_{r95}	**0.891** (0.89)									
$PA_{\rho m}$	**0.849** (0.96)	**0.877** (0.89)								
$PA_{\rho 95}$	**0.897** (0.84)	**0.821** (0.95)	**0.866** (0.88)							
MAP	0.974 (0.53)	0.953 (0.54)	0.974 (0.51)	0.946 (0.52)						
EGA_G	0.900 (0.61)	0.885 (0.60)	0.884 (0.60)	0.871 (0.59)	*0.899* (0.41)					
FF	**0.957** (0.79)	0.929 (0.80)	**0.948** (0.77)	**0.916** (0.78)	0.960 (0.54)	0.911 (0.58)				
MC_{AIC}	**0.951** (0.73)	**0.947** (0.72)	**0.950** (0.71)	0.946 (0.69)	*0.994* (0.46)	0.910 (0.52)	**0.967** (0.70)			
MC_{BIC}	**0.933** (0.70)	**0.904** (0.72)	0.927 (0.69)	0.899 (0.69)	0.934 (0.55)	0.866 (0.52)	**0.900** (0.73)	0.930 (0.65)		
MC_r	0.936 (0.63)	0.926 (0.62)	0.934 (0.61)	0.924 (0.59)	*0.984* (0.38)	*0.879* (0.44)	0.945 (0.60)	0.861 (0.67)	0.908 (0.56)	
Single	0.844	0.815	0.818	0.781	0.522	0.590	0.823	0.774	0.684	0.645
***N* = 2,000**										
PA_{r95}	**0.937** (0.95)									
$PA_{\rho m}$	**0.925** (0.98)	**0.935** (0.95)								
$PA_{\rho 95}$	**0.944** (0.92)	**0.907** (0.97)	**0.929** (0.95)							
MAP	0.985 (0.53)	0.980 (0.53)	0.990 (0.52)	0.982 (0.52)						
EGA_G	0.935 (0.62)	0.922 (0.62)	0.927 (0.62)	0.911 (0.62)	*0.883* (0.41)					
FF	**0.972** (0.86)	**0.956** (0.87)	**0.969** (0.86)	**0.950** (0.87)	0.974 (0.54)	0.911 (0.61)				
MC_{AIC}	**0.978** (0.78)	**0.977** (0.77)	**0.978** (0.77)	**0.978** (0.75)	*0.998* (0.45)	0.953 (0.50)	**0.983** (0.73)			
MC_{BIC}	**0.966** (0.80)	**0.950** (0.81)	**0.964** (0.79)	**0.949** (0.79)	0.985 (0.53)	0.890 (0.57)	**0.928** (0.81)	**0.960** (0.70)		
MC_r	0.970 (0.65)	0.968 (0.64)	0.970 (0.65)	0.968 (0.63)	*0.996* (0.36)	*0.932* (0.41)	0.974 (0.62)	**0.866** (0.72)	0.952 (0.59)	
Single	0.918	0.900	0.909	0.884	0.526	0.594	0.868	0.811	0.791	0.675

Each cell shows the agreement hit rate (AHR) and the agreement rate (AR; within the parentheses) for each combination rule. The AHR of combination rules with AR > 0.70 are shown in bold, and those with AR < 0.50 are shown in italics. PA_r, parallel analysis with Pearson correlations; PA_ρ, parallel analysis with tetrachoric correlations; PA_m, parallel analysis with mean eigenvalue criterion; PA_{95}, parallel analysis with 95th percentile eigenvalue criterion; MAP, minimum average partial; EGA_G, exploratory graph analysis with Gaussian graphical model; FF, factor forest; MC_{AIC}, model comparison based on AIC; MC_{BIC}, model comparison based on BIC; MC_r, model comparison based on the Fisher-transformed correlations; N, sample size.

5 attributes, respectively). Step *c* of the guidelines apply to this case because none of the methods agreed on their suggestion, but PA_{rm} and MC_{AIC}, as well as MC_{AIC} and FF, recommended a close number of attributes. Thus, we explored solutions from 3 to 5 attributes in terms of model fit. A G-DINA model was fitted using each of the Q-matrices from 3 to 5 attributes (Q_3-Q_5) suggested by the DFL and Hull methods (see **Figure 2**). Additionally, Q_O was also used to fit a G-DINA model for comparison purposes. **Table 8** shows the fit indices for each model. Overall, Q_4 obtained the best model fit. This result is in agreement with the number of attributes defined by Chen et al. (2020). Thus, a solution with four attributes was considered the most appropriate. The differences between Q_4 and Q_O were not very pronounced: 81.25% of the q-entries were the same for both matrices. In an applied study in which no original Q-matrix had been prespecified, Q_4 could be used as a starting point for domain experts to achieve a Q-matrix specification that provides both good fit and theoretical interpretability.

DISCUSSION

The correct specification of the Q-matrix is a prerequisite for CDMs to provide accurate attribute profile classifications (Rupp and Templin, 2008; Gao et al., 2017). Because the Q-matrix construction process is usually conducted by domain

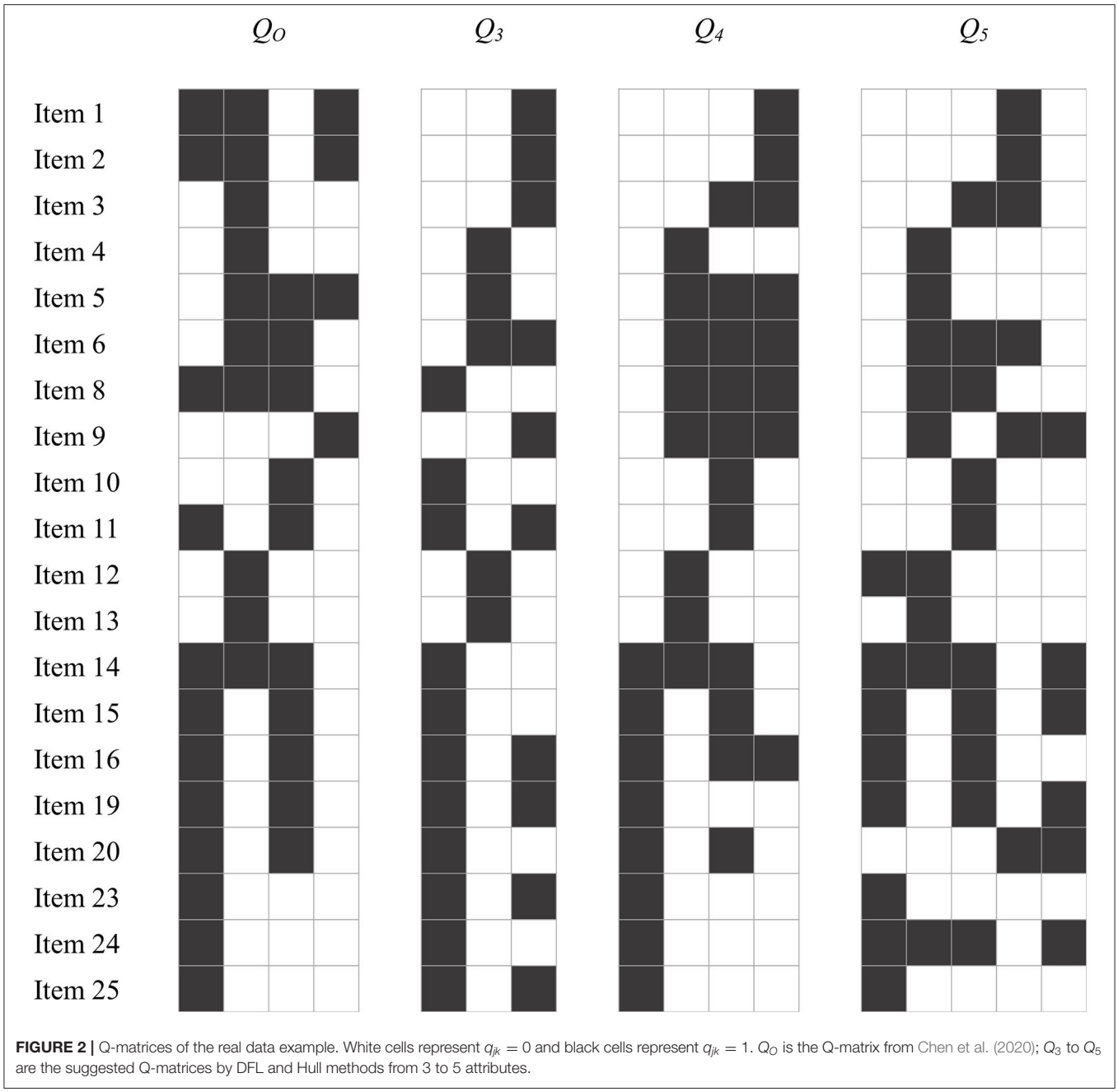


FIGURE 2 | Q-matrices of the real data example. White cells represent $q_{jk} = 0$ and black cells represent $q_{jk} = 1$. Q_O is the Q-matrix from Chen et al. (2020); Q_3 to Q_5 are the suggested Q-matrices by DFL and Hull methods from 3 to 5 attributes.

TABLE 8 | Model-fit for the real data illustration Q-matrices.

	−2LL	*np*	AIC	BIC	min.p(*r*)	items(*r*)
Q_3	8,584	59	8,702	**8,938**	0.006	5
Q_O	8,498	97	8,692	9,079	0.000	8
Q_4	8,441	97	**8,635**	9,022	**0.098**	**0**
Q_5	8,390	133	8,656	9,187	**0.096**	**0**

Best result for AIC, BIC, and items(r), as well as min.p(r) > 0.05, are shown in bold. −2LL, deviance; np, number of parameters; AIC, Akaike's information criterion; BIC, Bayesian information criterion; min.p(r), minimum p-value (adjusted for multiple comparisons) associated to the residual Fisher-transformed correlations; items(r), number of items showing a statistically significant (adjusted for multiple comparisons) Fisher-transformed correlation with at least another item. Q_3, Q_4, Q_5, Q-matrix specified by DFL and Hull methods with 3, 4, and 5 attributes, respectively; Q_O, Q-matrix from Chen et al. (2020).

experts, many Q-matrix validation methods have been recently developed with the purpose of empirically evaluating the decisions made by the experts. Additionally, empirical methods to specify the Q-matrix directly from the data (i.e., Q-matrix estimation methods), without requiring a previously specified one, have been also proposed. The problem with the Q-matrix estimation and validation methods proposed so far is that they do not question the number of attributes specified by the researcher. The assumption of known dimensionality has not been exhaustively explored in the CDM framework. This contrasts with the vast literature on dimensionality assessment methods in the factor analysis framework, where this problem is considered of major importance and has received a high degree of attention (e.g., Garrido et al., 2013; Preacher et al., 2013). All in all, the main goal of the present study was to explore the performance of several dimensionality assessment methods from the available literature in determining the number of attributes in CDMs. A comprehensive simulation study was conducted with that purpose.

Results from the simulation study showed that some methods available can be considered suitable for assessing the dimensionality of CDMs. Namely, parallel analysis with principal components and random column permutation (i.e., PA), the machine learning factor forest model (i.e., FF), and using the AIC fit index to compare CDMs with different number of attributes (i.e., MC_{AIC}) obtained high overall accuracies ($HR \geq 0.768$). PA with Pearson correlations and mean eigenvalue criterion (i.e., PA_{rm}) obtained the highest overall accuracy, while MC_{AIC} obtained the best close accuracy, considering a range of ± 1 attribute around the generating number of attributes. Item quality was found to be the most relevant simulation factor, severely affecting the performance of PA_{rm}, FF, and MC_{AIC}. Thus, the percentage of correct estimates varied from around 60% with low-quality items to more than 90% with high-quality items. Apart from item quality, PA_{rm} was also affected by the sample size and the correlation among the attributes, showing a bad performance with highly correlated attributes. These results are in line with previous studies (e.g., Garrido et al., 2013; Lubbe, 2019). MC_{AIC} and, especially, FF, were more robust to the different explored conditions (other than item quality). However, it should be noted that, unlike PA_{rm} and FF (which consistently tended to underestimate the number of attributes under almost all conditions), MC_{AIC} bias might show a slightly under- or overestimation tendency depending on the number of attributes, item quality, ratio of number of items to attribute, and generating model.

The remaining methods (i.e., MAP, VSS_1, VSS_2, DETECT, EKC, EGA_G, EGA_T, MC_{BIC}, and MC_r) obtained an overall poor performance, and thus their use cannot be recommended for the assessment of CDM data dimensionality. Of these methods, DETECT and EKC showed a heavy tendency to overestimate. Even though EKC was expected to perform better, it was observed that the first reference eigenvalue was usually very high, leaving the remaining ones at low levels. These resulted in the EKC often performing identically to what the Kaiser-Guttman criterion would (which is known for its tendency to overestimate the number of dimensions). On the other hand, MAP, VSS_1, EGA_T, and MC_{BIC} showed a strong tendency to underestimate. Even though a higher performance was expected for MAP and EGA_T, their underestimation tendency is aligned with previous findings (Garrido et al., 2011; Golino et al., 2020). As for the MC methods, while both AIC and BIC have shown good results in selecting the correct Q-matrix among competing misspecified Q-matrices (Kunina-Habenicht et al., 2012; Chen et al., 2013), it is clear that the higher penalization that BIC applies compared to AIC is not appropriate for the dimensionality assessment problem. Finally, EGA_G was the only remaining method that obtained a good performance in terms of close hit rate. However, its overall hit rate was low, especially due to its poor performance when the generating model was the DINA model.

Although the influence of the generating model was most noticeable for EGA_G, most dimensionality assessment methods from the EFA framework performed worse under the DINA model than under the G-DINA model. These results might be due to the non-compensatory nature of the DINA model, in which the relationship between the number of mastered attributes and the probability of correctly answering an item clearly deviates from being linear (in a more pronounced way that under the G-DINA model, as illustrated in **Figure 1**). A greater depart from linearity might produce a greater disruption to the performance of all the methods that are based on correlations (e.g., PA, FF, EGA). On the contrary, the MC methods performed better under the DINA model. Since the MC methods are precisely modeling the response process, they benefit from the parsimony of reduced models. The performance of the dimensionality assessment methods under other commonly used reduced CDMs (e.g., the *deterministic inputs, noisy "or" gate* model or DINO; Templin and Henson, 2006) is expected to follow a similar pattern as the one obtained for the DINA model.

An important finding regarding the MC methods is that the performance of the variants that made use the generating Q-matrix (e.g., $HR = 0.886$ for MC_{AIC-G}) was notably better than that of their corresponding methods (e.g., $HR = 0.768$ for MC_{AIC}). Given that the Q-matrices specified by the DFL and Hull methods obtained a very high overall recovery rate ($QRR = 0.949$), these results imply that a small improvement in the quality of the Q-matrices might have a big impact on the dimensionality assessment performance of the MC procedures. This reiterates the importance of applying empirical Q-matrix validation methods such as the Hull method, even though the improvement over the original Q-matrix (be it empirically estimated or constructed by domain experts) might seem small.

The exploration of combination rules showed that PA_{rm} and FF often agreed on the recommended number of attributes ($AR \geq 0.70$), providing a very high combined accuracy ($AHR \geq 0.923$). FF and MC_{AIC} obtained an even higher accuracy ($AHR \geq 0.953$) with a slightly lower agreement rate ($AR \geq 0.65$). When these three methods agree on their number of attributes, which occurred in more than 60% of the overall conditions, the percentage of correct estimations was, at least, of 97.6%. Given these results, the following guidelines can be followed when aiming to empirically determine the number of attributes in CDM data: (a) if PA_{rm}, FF, and MC_{AIC} agree on their suggestion, retain their recommended number of attributes;

(b) if any two of these methods agree, retain their recommended number of attributes; (c) if none of these methods agree, explore the recommended number of attributes by those that suggest a similar (i.e., ± 1) number of attributes; (d) if these methods strongly disagree, explore the recommended number of attributes by each of them. The number of attributes provided by the dimensionality assessment methods should be understood as suggestions; the final decision should consider theoretical interpretability as well.

These guidelines were used to illustrate the dimensionality assessment procedure using a real dataset. The number of suggested number of attributes greatly varied from 1 attribute (MAP and VSS_1) to 8 attributes (DETECT). The best three methods from the simulation study, PA_{rm}, MC_{AIC}, and FF recommended 3, 4, and 5 attributes, respectively. After inspecting the model fit of the Q-matrices suggested by the DFL and Hull methods from 3 to 5 attributes, it was found that 4 was the most appropriate number of attributes, which was consistent with Chen et al. (2020). The interpretability of the Q-matrices suggested by the DFL and Hull method should be further explored by domain experts, who should make the final decision on the Q-matrix specification.

The present study is not without limitations. First, the CDMs used to generate the data (i.e., DINA and G-DINA) were also used to estimate the models in the MC methods. In applied settings, the saturated G-DINA model should be used for both estimating/validating the Q-matrix and assessing the number of attributes to make sure that there are no model specification errors. After these two steps have been fulfilled, item-level model comparison indices should be applied to check whether more reduced CDMs are suitable for the items (Sorrel et al., 2017). The main reason why the DINA model was used to estimate the models in the MC methods (whenever the generating model was also the DINA model) was to try to reduce the already high computation time of the simulation study. Nevertheless, it is expected that the results of these conditions would have been similar if the G-DINA model were used to estimate these models: it provides similar results as the DINA model given that the sample size is not very small (i.e., $N < 100$; Chiu et al., 2018). Second, the generalization of the results to other conditions not considered in the present simulation study should be done with caution. For instance, the range of the number of attributes was kept around the most common number of attributes encountered in applied settings and simulation studies. Highly dimensional scenarios (e.g., $K = 8$) were not explored because the computation time increases exponentially with the number of attributes and the simulation study was already computationally expensive. Hence, the performance of the dimensionality assessment methods under highly dimensional data should be further evaluated. In this vein, an important discussion might arise when considering highly dimensional CDM data. As Sessoms and Henson (2018) reported, many studies obtained attribute correlations higher than 0.90. These extremely high correlations imply that those attributes are hardly distinguishable, which might indicate that the actual number of attributes underlying the data is lower than what has been specified. It can be argued that CDM attributes are expected to show stronger correlations than EFA factors because attributes are usually defined as fine-grained skills or concepts within a broader construct. However, it is important to note that each attribute should be still distinguishable from the others. Otherwise, the interpretation of the results might be compromised. The proper identification of the number of attributes might be of help in this matter.

Finally, only one of the best three performing methods (i.e., FF) can be directly implemented by the interested researcher in assessing the dimensionality of CDM data, using publicly available functions. With the purpose of facilitating the application of the other two best performing methods, the specific implementations of parallel analysis and model comparison approach used in the present study have been included in the `cdmTools` R package (Nájera et al., 2021). A sample R code to illustrate a dimensionality assessment study of CDM data can be found in **Supplementary Materials**.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://cran.r-project.org/web/packages/edmdata.

AUTHOR CONTRIBUTIONS

PN wrote the R simulation scripts, performed the real data analyses, and wrote the first draft of the manuscript. FA and MS wrote sections of the manuscript. All authors contributed to conception, design of the study, manuscript revision, read, and approved the submitted version.

FUNDING

This research was partially supported by Ministerio de Ciencia, Innovación y Universidades, Spain (Grant PSI2017-85022-P), European Social Fund, and Cátedra de Modelos y Aplicaciones Psicométricas (Instituto de Ingeniería del Conocimiento and Autonomous University of Madrid).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.614470/full#supplementary-material

REFERENCES

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

Auerswald, M., and Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: a comparison of extraction methods under realistic conditions. *Psychol. Methods* 24, 468–491. doi: 10.1037/met0000200

Balamuta, J. J., Culpepper, S. A., and Douglas, J. A. (2020a). *edina: Bayesian Estimation of an Exploratory Deterministic Input, Noisy and Gate Model. R Package Version 0.1.1*. Available online at: https://CRAN.R-project.org/package=edina

Balamuta, J. J., Culpepper, S. A., and Douglas, J. A. (2020b). *edmdata: Data Sets for Psychometric Modeling. R Package Version 1.0.0*. Available online at: https://CRAN.R-project.org/package=edmdata

Bonifay, W. E., Reise, S. P., Scheines, R., and Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Struct. Equ. Model.* 22, 504–516. doi: 10.1080/10705511.2014.938596

Braeken, J., and van Assen, M. A. L. M. (2017). An empirical Kaiser criterion. *Psychol. Methods* 22, 450–466. doi: 10.1037/met0000074

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Meas.* 50, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

Chen, Y., Culpepper, S. A., Chen, Y., and Douglas, J. (2018). Bayesian estimation of the DINA Q-matrix. *Psychometrika* 83, 89–108. doi: 10.1007/s11336-017-9579-4

Chen, Y., Culpepper, S. A., and Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika* 85, 121–153. doi: 10.1007/s11336-019-09693-2

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *J. Am. Stat. Assoc.* 110, 850–866. doi: 10.1080/01621459.2014.934827

Chiu, C.-Y., Sun, Y., and Bian, Y. (2018). Cognitive diagnosis for small educational programs: the general nonparametric classification method. *Psychometrika* 83, 355–375. doi: 10.1007/s11336-017-9595-4

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. Hillsdale, NJ: Erlbaum.

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., et al. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educ. Psychol. Meas.* 70, 885–901. doi: 10.1177/0013164410379332

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 Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika* 81, 253–273. doi: 10.1007/s11336-015-9467-8

de la Torre, J., and Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicol. Educat.* 20, 89–97. doi: 10.1016/j.pse.2014.11.001

de la Torre, J., van der Ark, L. A., and Rossi, G. (2018). Analysis of clinical data from cognitive diagnosis modeling framework. *Measure. Eval. Counsel. Dev.* 51, 281–296. doi: 10.1080/07481756.2017.1327286

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* 4, 272–299. doi: 10.1037/1082-989X.4.3.272

Finch, W. H. (2020). Using fit statistics differences to determine the optimal number of factors to retain in an exploratory factor analysis. *Educ. Psychol. Meas.* 80, 217–241. doi: 10.1177/0013164419865769

Gao, M., Miller, M. D., and Liu, R. (2017). The impact of Q-matrix misspecifications and model misuse on classification accuracy in the generalized DINA model. *J. Meas. Eval. Educ. Psychol.* 8, 391–403. doi: 10.21031/epod.332712

García, P. E., Olea, J., and de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema* 26, 372–377. doi: 10.7334/psicothema2013.322

Garrido, L. E., Abad, F. J., and Ponsoda, V. (2011). Performance of velicer's mínimum average partial factor retention method with categorical variables. *Educ. Psychol. Meas.* 71, 551–570. doi: 10.1177/0013164410389489

Garrido, L. E., Abad, F. J., and Ponsoda, V. (2013). A new look at horn's parallel analysis with ordinal variables. *Psychol. Methods* 4, 454–474. doi: 10.1037/a0030005

Garrido, L. E., Abad, F. J., and Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via monte carlo simulation. *Psychol. Methods* 21, 93–111. doi: 10.1037/met0000064

Golino, H., and Christensen, A. P. (2020). *EGAnet: Exploratory Graph Analysis – A Framework for Estimating the Number of Dimensions in Multivariate Data Using Network Psychometrics. R Package Version 0.9.2*. Available online at: https://CRAN.R-project.org/package=EGAnet

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., et al. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: a simulation and tutorial. *Psychol. Methods* 25, 292–320. doi: 10.1037/met0000255

Golino, H. F., and Epskamp, S. (2017). Exploratory graph analysis: a new approach for estimating the number of dimensions in psychological research. *PLoS ONE* 12:e0174035. doi: 10.1371/journal.pone.0174035

Goretzko, D., and Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychol. Methods* 25, 776–786. doi: 10.1037/met0000262

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika* 19, 149–161. doi: 10.1007/BF02289162

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika* 30, 179–185. doi: 10.1007/BF02289447

Humphreys, L. G., and Ilgen, D. R. (1969). Note on a criterion for the number of common factors. *Educ. Psychol. Meas.* 29, 571–578. doi: 10.1177/001316446902900303

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2019). *semTools: Useful Tools for Structural Equation Modeling. R Package Version 0.5-2*. Available online at: https://CRAN.R-project.org/package=semTools

Junker, B., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric IRT. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20, 141–151. doi: 10.1177/001316446002000116

Kim, H. (1994). *New techniques for the dimensionality assessment of standardized test data* (Doctoral dissertation). University of Illinois at Urbana-Champaign. IDEALS. Available online at: https://www.ideals.illinois.edu/handle/2142/19110

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *J. Educ. Meas.* 49, 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Lauritzen, S. L. (1996). *Graphical Models*, Vol. 17. New York, NY: Clarendon Press.

Lim, S., and Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychol. Methods* 24, 452–467. doi: 10.1037/met0000230

Lorenzo-Seva, U., Timmerman, M. E., and Kiers, H. A. L. (2011). The hull method for selecting the number of common factors. *Multivariate Behav. Res.* 46, 340–364. doi: 10.1080/00273171.2011.564527

Lubbe, D. (2019). Parallel analysis with categorical variables: impact of category probability proportions on dimensionality assessment accuracy. *Psychol. Methods* 24, 339–351. doi: 10.1037/met0000171

Ma, W., and de la Torre, J. (2020a). An empirical Q-matrix validation method for the sequential generalized DINA model. *Br. J. Math. Stat. Psychol.* 73, 142–163. doi: 10.1111/bmsp.12156

Ma, W., and de la Torre, J. (2020b). GDINA: an R package for cognitive diagnosis modeling. *J. Stat. Softw.* 93, 1–26. doi: 10.18637/jss.v093.i14

Marčenko, V. A., and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik* 1, 457–483. doi: 10.1070/SM1967v001n04ABEH001994

Massara, G. P., Di Matteo, T., and Aste, T. (2016). Network filtering for big data: triangulated maximally filtered graph. *J. Complex Netw.* 5, 161–178. doi: 10.1093/comnet/cnw015

Nájera, P., Sorrel, M. A., and Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educ. Psychol. Meas.* 79, 727–753. doi: 10.1177/0013164418822700

Nájera, P., Sorrel, M. A., and Abad, F. J. (2021). *cdmTools: Useful Tools for Cognitive Diagnosis Modeling. R Package Version 0.1.1*. Available online at: https://github.com/Pablo-Najera/cdmTools

Nájera, P., Sorrel, M. A., de la Torre, J., and Abad, F. J. (2020). Balancing fit and parsimony to improve Q-matrix validation. *Br. J. Math. Stat. Psychol.* doi: 10.1111/bmsp.12228. [Epub ahead of print].

Peres-Neto, P. R., Jackson, D. A., and Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* 49, 974–997. doi: 10.1016/j.csda.2004.06.015

Preacher, K. J., Zhang, G., Kim, C., and Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: a model selection perspective. *Multivariate Behav. Res.* 48, 28–56. doi: 10.1080/00273171.2012.710386

R Core Team (2020). *R: A Language and Environment for Statistical Computing (Version 3.6)*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://www.R-project.org/

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research. R Package Version 1.9.12*. Evanston, IL: Northwestern University. Available online at: https://CRAN.R-project.org/package=psych

Revelle, W., and Rocklin, T. (1979). Very simple structure: an alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behav. Res.* 14, 403–414. doi: 10.1207/s15327906mbr1404_2

Robitzsch, A. (2020). *sirt: Supplementary Item Response Theory Models. R Package Version 3.9-4*. Available online at: https://CRAN.R-project.org/package=sirt

Robitzsch, A., and George, A. C. (2019). "The R package CDM for diagnostic modeling," in *Handbook of Diagnostic Classification Models. Methodology of Educational Measurement and Assessment,* eds M. von Davier and Y.-S. Lee (Springer), 549–572.

Roussos, L. A., Stout, W. F., and Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *J. Educ. Meas.* 35, 1–30. doi: 10.1111/j.1745-3984.1998.tb00525.x

Rupp, A. A., and Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educ. Psychol. Meas.* 68, 78–96. doi: 10.1177/0013164407301545

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

Sessoms, J., and Henson, R. A. (2018). Applications of diagnostic classification models: a literature review and critical commentary. *Meas. Interdiscip. Res. Perspect.* 16, 1–17. doi: 10.1080/15366367.2018.1435104

Sorrel, M. A., de la Torre, J., Abad, F. J., and Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology* 13, 39–47. doi: 10.1027/1614-2241/a000131

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., and Lievens, F. (2016). Validity and reliability of situational judgement test scores: a new approach based on cognitive diagnosis models. *Organ. Res. Methods* 19, 506–532. doi: 10.1177/1094428116630065

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconception based on item response theory. *J. Educ. Meas.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Timmerman, M. E., and Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychol. Methods* 16, 209–220. doi: 10.1037/a0023353

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika* 41, 321–327. doi: 10.1007/BF02293557

Velicer, W. F., Eaton, C. A., and Fava, J. L. (2000). "Construct explication through factor or component analysis: a review and evaluation of alternative procedures for determining the number of factors or components," in *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy,* eds R. D. Goffin and E. Helmes (New York, NY: Kluwer Academic/Plenum), 41–71.

Wang, W., Song, L., and Ding, S. (2018). "An exploratory discrete factor loading method for Q-matrix specification in cognitive diagnosis models," in *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics & Statistics,* Vol. 233, eds M. Wilberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar (Cham: Springer), 351–362.

Xu, G., and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *J. Am. Stat. Assoc.* 13, 1284–1295. doi: 10.1080/01621459.2017.1340889

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika* 72, 69–91. doi: 10.1007/s11336-004-1257-7

Zhang, J., and Stout, W. (1999). The theoretical DETECT index of dimensionality and its applications to approximate simple structure. *Psychometrika* 64, 213–249. doi: 10.1007/BF02294536

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 Nájera, Abad and Sorrel. 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: 04 March 2021
doi: 10.3389/fpsyg.2021.564976

The Q-Matrix Anchored Mixture Rasch Model

Ming-Chi Tseng[1] and Wen-Chung Wang[2]**

[1] National University of Tainan, Tainan, Taiwan, [2] The Education University of Hong Kong, Tai Po, Hong Kong

Mixture item response theory (IRT) models include a mixture of latent subpopulations such that there are qualitative differences between subgroups but within each subpopulation the measure model based on a continuous latent variable holds. Under this modeling framework, students can be characterized by both their location on a continuous latent variable and by their latent class membership according to Students' responses. It is important to identify anchor items for constructing a common scale between latent classes beforehand under the mixture IRT framework. Then, all model parameters across latent classes can be estimated on the common scale. In the study, we proposed Q-matrix anchored mixture Rasch model (QAMRM), including a Q-matrix and the traditional mixture Rasch model. The Q-matrix in QAMRM can use class invariant items to place all model parameter estimates from different latent classes on a common scale regardless of the ability distribution. A simulation study was conducted, and it was found that the estimated parameters of the QAMRM recovered fairly well. A real dataset from the Certificate of Proficiency in English was analyzed with the QAMRM, LCDM. It was found the QAMRM outperformed the LCDM in terms of model fit indices.

Keywords: Q-matrix anchored mixture Rasch model, Q-matrix, anchor, mixture Rasch model, Rasch model

OPEN ACCESS

Edited by:
Chun Wang,
University of Washington,
United States

Reviewed by:
Mirka Henninger,
University of Zurich, Switzerland
Sedat Sen,
Harran University, Turkey

****Correspondence:***
Ming-Chi Tseng
mingchi@gms.ndhu.edu.tw
Wen-Chung Wang
wcwang@eduhk.hk

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

Received: *24 June 2020*
Accepted: *12 February 2021*
Published: *04 March 2021*

Citation:
Tseng M-C and Wang W-C (2021) The Q-Matrix Anchored Mixture Rasch Model. Front. Psychol. 12:564976. doi: 10.3389/fpsyg.2021.564976

INTRODUCTION

Measurement invariance is a key assumption that enables score comparison across different groups of respondents (Hambleton et al., 1991). In reality, the assumption may not hold and needs to be checked empirically. In the context of Rasch measurement, different groups of respondents may take different views on items, resulting in measurement non-invariance. Rost (1990) integrated latent class analysis (LCA; Lazarsfeld and Henry, 1968) to the Rasch model (Rasch, 1960) and derived the mixture Rasch model (MRM), which can be viewed as an extension of the Rasch model that allows different groups (latent classes) of respondents to have different item parameters and ability distributions. To the extent that these classes are substantively meaningful, the mixture Rasch model provides a potentially important means to understanding how and why examinees respond in different ways. It is assumed that a Rasch model holds in each class, but each class may have different item and ability parameters. Specifically, the probability of a correct response in the MRM can be given as:

$$P(Y_{ij}=1)=\sum_{g=1}^{G}\pi_g\cdot P(Y_{ijg}=1|g,\theta)=\sum_{g=1}^{G}\pi_g\cdot\frac{\exp(\theta_{ig}-b_{jg})}{1+\exp(\theta_{ig}-b_{jg})}, \quad (1)$$

where g is an index for the latent class, $g = 1,..., G$, $i = 1,..., N$ examinees, π_g is the proportion of examinees for each class, θ_{ig} is the latent ability of examinee i in latent class g, and b_{jg} is the

difficulty parameter of item j for latent class g. The MRM can account for qualitative differences between latent classes and quantitative differences within latent classes (Rost, 1990).

An important feature of the MRM or other mixture models (von Davier and Yamamoto, 2007) is that the number of latent classes must be explored from the data, which is an exploratory approach. Usually, the Akaike information criterion (AIC), Bayesian information criterion (BIC), or deviance information criterion (DIC) are applied to determine the number of latent classes but they do not always provide the same answer. Over- or under-extraction of latent classes may occur, making the interpretation problematic (Alexeev et al., 2011). It is desirable to adopt a constrained approach to the identification of latent classes when there are substantive theories or hypotheses.

Recent developments in the Q-matrix (Tatsuoka, 1983) for diagnostic classification models (DCMs) may help with the identification. Domain-specific assessment experts encode the relationships that they believe exist between the diagnostic assessment items and the latent variables that are used to classify respondents into so-called Q-matrices. The attribute is a latent characteristic of respondents in the Q-matrices.

Choi (2010) develop the diagnostic classification mixture Rasch model (DCMixRM) which combines a Mixture Rasch model with log-linear cognitive diagnostic model (LCDM; Henson et al., 2009). In the DCMixRM, this model includes mastery states of attributes as covariates. To be more specific, in the measurement component, observed item responses are jointly regressed on latent trait and attributes through the Rasch model and the LCDM. Next, in the structural model, ability is regressed on class membership, and class membership is regressed on mastery profile to explain latent class as covariates. Besides, Bradshaw and Templin (2014) develop the Scaling Individuals and Classifying Misconceptions (SICM) model which is presented as a combination of a unidimensional IRT model and LCDM where the categorical latent variables represent misconceptions instead of skills. In the SICM, IRT, and LCDM assumed to be orthogonal as in the original bifactor model (Gibbons and Hedeker, 1992). Theoretically, SICM expected that subjects vary in ability even when they possess the same misconception pattern, meaning a significant correlation between ability and misconception pattern was not expected or modeled.

On the other hand, we developed the Q-matrix anchor mixture Rasch model (QAMRM) by incorporating the Q-matrix into the MRM. The QAMRM is constrained because the number of latent classes is specified by users rather than explored from the data. The latent traits in the QAMRM can be compensatory or non-compensatory. The Q-matrix contains a set of elements q_{jk} indicating whether attribute k is required to answer item j correctly, and $q_{jk} = 1$ if the attribute is required, otherwise it is 0. The total number of attributes and the value of q_{jk} is assigned by content experts. Similar Q-matrices have been adopted in IRT to specify *a priori* which latent traits (components) have been measured by which items, such as the linear logistic test model (LLTM; Fischer, 1973), the multicomponent latent trait model (Whitely, 1980), the loglinear multidimensional IRT model for polytomously scored items (Kelderman and Rijkes, 1994), the multidimensional random coefficients multinomial logic model (Adams et al., 1997), the multidimensional componential IRT model for polytomous items (Hoskens and De Boeck, 2001), and the multicomponent latent trait model for diagnosis (Embretson and Yang, 2013).

The QAMRM uses the Q-matrix to check whether different classes have different measurement characteristics. The utility of this approach lies in the fact that the numbers of latent classes, immediately observable through the Q-matrix, are defined in advance and can be used to help explain item level performance to discover how members in one class differ from another. It is these Q-matrix differences in response propensities that help explain the potential causes of these differential measurement characteristics.

The approach proposed in this study provides the Q-matrix by means of a design matrix describing the composition of the different classes. We begin below by illustrating how the QAMRM can be viewed as incorporating features from the Q-matrix, and then through the Q-matrix to establish class invariant items and allow all model parameter estimates across latent classes to be on a common scale regardless of the ability distribution.

THE Q-MATRIX ANCHOR MIXTURE RASCH MODEL

We adapted **Figure 1** from Wright and Stone (1979) about the Rasch model to express the response difference between the Rasch model, MRM, and QAMRM.

In the Rasch model diagram, person ability θ_i and item difficulty b_j jointly determine response $P(y_{ij})$. In the MRM diagram, conditional on latent class g, person ability θ_{ig} and item difficulty b_{jg} jointly determine response $P(y_{ijg})$. In the QAMRM diagram, conditional on class $g(\alpha)$, person ability $\theta_{ig(\alpha)}$ and item difficulty $b_{jg(\alpha)}$ jointly determine response $P(Y_{ijg(\alpha)})$. When there is no Q-matrix, the number of classes is estimated from data, the QAMRM simplifies to the MRM; when there is only single class, the MRM simplifies to the Rasch model.

The QAMRM has multiple classes that follow the Q-matrix design matrix. Like the MRM, it assumes that there may be heterogeneity in response patterns at different classes which should not be ignored (Mislevy and Verhelst, 1990; Rost, 1990), but should consider the Q-matrix to form the number of classes beforehand, rather than forming the number of classes during the parameter estimation. Viewed in this way, the Q-matrix inside the QAMRM captures the association between the items and classes. The probability of getting a correct response in the QAMRM can be given as follows:

$$P(Y_{ij}=1)=\sum_{g(\alpha)=1}^{G}\pi_{g(\alpha)}\cdot P(Y_{ijg(\alpha)}=1|g_{(\alpha)},\ \theta_{(\alpha)})=$$

$$\sum_{g(\alpha)=1}^{G}\pi_{g(\alpha)}\cdot\frac{\exp(\theta_{ig(\alpha)}-b_{jg(\alpha)})}{1+\exp(\theta_{ig(\alpha)}-b_{jg(\alpha)})}, \tag{2}$$

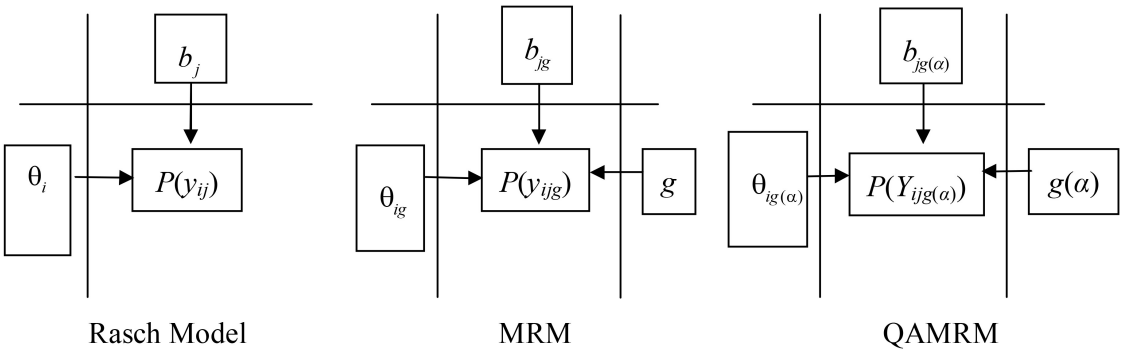


FIGURE 1 | Comparison among the Rasch model, MRM, and QAMRM.

where $Y_{ijg(\alpha)}$ is the score of examinee i ($i = 1, \ldots, N$) on item j ($j = 1, \ldots, J$) in class g conditional on attribute profile α ($\alpha = \alpha_1, \ldots, \alpha_K$)', $\theta_{ig(\alpha)}$ is the latent ability of examinee i within class g conditional on attribute α, and $b_{jg(\alpha)}$ is the difficulty parameter of item j for class g conditional on attribute α. Like the exploratory MRM, there is only one mixing proportion/structural parameter/latent class membership probability in the QAMRM, $\pi_{g(\alpha)}$, which is the probability of being in class g conditional on attribute pattern α.

THE Q-MATRIX SETS CLASS INVARIANT ITEMS *A PRIORI* IN QAMRM

Paek and Cho (2015) posit four scenarios to establish a common scale across latent classes in MRM and suggest proposing the use of class-invariant items, which was also suggested by von Davier and Yamamoto (2004). Those invariant items have the same item difficulties across latent classes. Once a set of class invariant items are available, this ensures a common scale across latent classes.

The challenge in MRM is to identify class invariant items, because selecting the best measurement model regarding the numbers of item parameters, latent groups, and dimensions are mainly decided by statistical procedure *post hoc*. Some studies exist in the MRM literature (e.g., Cho et al., 2016), where a statistical procedure was applied as an attempt to locate class invariant items in real data analyses, but how to find class invariant items in the context of MRM until now remains unclear.

Paek and Cho (2015) suggest the use of class invariant items to recover the parameter differences correctly in both item profiles and ability distributions for latent classes when those differences exist simultaneously in MRM; in QAMRM we do not need to find class invariant items by statistical procedure *post hoc* to let all model parameter estimates across latent classes be on a common scale because the Q-matrix has already done that *a prior*. The invariant items in MRM are data driven, but in QAMRM, even though the latent classes follow different ability distributions in terms of their means, we still can easily set invariant items in QAMRM through the Q-matrix *a priori*. Even though there are no ability distributions in the latent class model, through the Q-matrix being set beforehand, the invariance parameters of the DINA model (de la Torre and Lee, 2010) and LCDM model (Bradshaw and Madison, 2016) still hold when the model fit the data.

LCDM is a flexible model that allows the relationships between categorical variables to be modeled using a latent class model, because most cognitive diagnosis models are typically parameterized to define the probability of a correct response, LCDM is re-expressed in terms of the log-odds of a correct response for each item. In addition, von Davier (2005) discusses the General Diagnostic Model (GDM) as a general approach to log-linear models with latent variables, where the latent variables are both continuous and discrete in addition to focusing on ordered responses for items. As a special case, the GDM general definition easily incorporates LCDM with dichotomous latent variables for dichotomous (von Davier, 2014). Besides, Hong et al., 2015 combines DINA model and non-compensatory item response theory to form DINA-NIRT model, which tries to combine continuous and discrete latent variables in cognitive diagnosis. In contrast, the DINA-NIRT is a special case of LCDM because it does not have compensatory attributes. The QAMRM combines continuous and discrete latent variables in the same framework, which includes compensatory (disjunctive) and non-compensatory (conjunctive) models.

In compensatory QAMRM, a low value on one latent variable can be compensated for by a high value on another latent variable, so it is not necessary to master all attributes that are required by an item to produce a correct response within a class. On the contrary, in non-compensatory QAMRM, a low value on one latent variable cannot be compensated for by a high value on another latent variable, so it is necessary to master all attributes that are required by an item to produce a correct response within a class. In the QAMRM, the relationship among latent variables, either compensatory or non-compensatory, is assumed to be identical across classes.

Substantive theories can help decide the numbers of attributes and classes prior to parameter estimation, as done in the Q-matrix. As an example, let there be 14 items measuring three binary attributes in the QAMRM. In total, there will be eight (2^3) attribute profiles, which are called classes g_1-g_8 as shown in column 1 in **Table 1**.

TABLE 1 | Class, Q-matrix and item difficulty for 14-item 3 attributes QAMRM.

Class				Q-matrix				Non-compensatory									Compensatory								
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
	α_1	α_2	α_3	Item	α_1	α_2	α_3	Item	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	Item	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
g_1	0	0	0	1	1	0	0	1	2[a]	2	2	2	1	1	1	1	1	2	2	2	2	1	1	1	1
g_2	0	0	1	2	0	1	0	2	2	2	1	1	2	2	1	1	2	2	2	1	1	2	2	1	1
g_3	0	1	0	3	0	0	1	3	2	1	2	1	2	1	2	1	3	2	1	2	1	2	1	2	1
g_4	0	1	1	4	1	1	0	4	2	2	2	2	2	2	1	1	4	3	3	2	2	2	2	1	1
g_5	1	0	0	5	1	0	1	5	2	2	2	2	2	1	2	1	5	3	2	3	2	2	1	2	1
g_6	1	0	1	6	0	1	1	6	2	2	2	1	2	2	2	1	6	3	2	2	1	3	2	2	1
g_7	1	1	0	7	1	1	1	7	2	2	2	2	2	2	2	1	7	4	3	3	2	3	2	2	1
g_8	1	1	1	8	1	0	0	8	2	2	2	2	1	1	1	1	8	2	2	2	2	1	1	1	1
				9	0	1	0	9	2	2	1	1	2	2	1	1	9	2	2	1	1	2	2	1	1
				10	0	0	1	10	2	1	2	1	2	1	2	1	10	2	1	2	1	2	1	2	1
				11	1	1	0	11	2	2	2	2	2	2	1	1	11	3	3	2	2	2	2	1	1
				12	1	0	1	12	2	2	2	2	2	1	2	1	12	3	2	3	2	2	1	2	1
				13	0	1	1	13	2	2	2	1	2	2	2	1	13	3	2	2	1	3	2	2	1
				14	1	1	1	14	2	2	2	2	2	2	2	1	14	4	3	3	2	3	2	2	1

[a] *The 2 represents the label we are using for the threshold in Item 1. Therefore, the whole line to label the first threshold of Item 1 is 2. More importantly, each parameter with a label of 2 in Item 1 will be equated, thus providing the same value for the estimated threshold. The label 2 between each item is different; the 2 is the label, not the number.*

Columns 2–4 give the attribute profiles for g_1-g_8. Persons in g_1 (0,0,0) have not mastered any of the three attributes; persons in g_2 (0,0,1) have mastered α_3 but have not mastered α_1 and α_2; and so on for the other classes. Note that we follow substantive theories to set g classes conditional on Q-matrix (2^K) when saturated by QAMRM *a priori*, if the class has almost no examinees in practice (de la Torre and Lee, 2010; Templin and Bradshaw, 2014), and we can thus cancel out the class.

Columns 6–8 show the Q-matrix for the 14 items. For example, item 1 measures α_1, item 4 measures α_1 and α_2, item 14 measures α_1, α_2, and α_3.

Columns 10–17 list hypothetical difficulties of the 14 items for the eight classes when the attributes are non-compensatory. For example, the difficulty of item 1 is 2 for g_1-g_4 but 1 for g_5-g_8. Because persons in g$_1$-g_4 have not mastered the attribute that item 1 measures (α_1), the item difficulty for them would be equally high. On the contrary, persons in g_5-g_8 have mastered α_1, so the item difficulty for them would be equally low. That is, although there are eight classes, item 1 has only two difficulties, one for g_1-g_4 and the other for g_5-g_8. Likewise, item 4 have difficulty of 2 for g_1-g_6 but 1 for g_7-g_8. Persons in g_1-g_6 have not mastered all attributes that are measured by item 4 (α_1 and α_2) so the item difficulty for them would be equally high; on the contrary, persons in g_7-g_8 have mastered both α_1 and α_2 so the item difficulty for them would be equally low. The other items can be interpreted similarity.

Columns 19–26 list hypothetical difficulties of the 14 items for the eight classes when the attributes are compensatory. Item 4 have difficulty of 3 for g_1 and g_2, and 2 for g_3-g_6, and 1 for g_7 and g_8. Persons in g_1 and g_2 have not mastered any of the attributes that item 4 measures (α_1 and α_2) so the item difficulty for them would be equally high; persons in g_3-g_6 have mastered one of α_1 and α_2 so the item difficulty for them would be equally median; persons in g_7 and g_8 have mastered both α_1 and α_2 so the item difficulty for them would be equally low. In other words, although there are eight class, item 4 has three difficulties, one for g_1 and g_2, one for g_3-g_6, and the other for g_7 and g_8, and the three difficulties are expected to be ordered. The other items can be interpreted similarity.

The 14 item parameters were specified according to the Q-matrix. The Q-matrix is like a bridge to connect different latent classes together and sets item parameter constraints across different latent classes *a priori*, like anchor items in different latent classes, hence the Q-matrix in QAMRM can be considered as the "anchor attribute."

For illustrative purposes, the item difficulties in the **Table 1** are set as integers, when in reality, they can be real numbers. However, the ordinal nature of $4 > 3 > 2 > 1$ is expected. If the Q-matrix was not adopted, there are 112 item parameters (assuming there were eight latent classes and 14 item parameters in each latent class), while according to the Q-matrix, in the non-compensatory model, each of the 14 items has two difficult parameters for the eight classes, so the total number of difficulty parameters is 28, in the compensatory model, six, six, and two items have two, three, and four difficult parameters for the eight classes, respectively, so the total number of difficulty parameters is 38.

SIMULATION STUDY

The primary goal of this section is to demonstrate that when the QAMRM fits the item responses, through the Q-matrix setting invariant items, all model parameter estimates across latent

classes are on a common scale and will be invariant regardless of the nature of the latent ability distributions across latent classes.

In the mixture Rasch model for binary data as described by Rost (1990) who used the model constraint $\sum \delta_{ig}=0$ in MRM, where δ_{ig} the item difficulty of the *i*th item is in the *g*th latent class and the summation is over items at a given *g*, indicating the summation of group specific item difficulty parameters over items is 0 within a latent group, and compared item profiles to characterize latent groups. The constraint can be used for scale comparability only when there is no mean difference in a continuous latent variable. Because we do not know the "true" mean difference in real data sets, the constraint cannot be sufficient for all empirical data sets (Cho et al., 2016). An alternative to identifying a common scale is by making the mean of the item parameters on latent ability zero (Wu et al., 1998), and when calculating every item difficulty parameter in latent classes conditional on Q-matrix, the QAMRM uses this setting.

Several aspects of the simulated data were held constant: the number of attributes was fixed to $K = 3$, test length to $J = 14$; the item parameters were generated as follows. For non-compensatory QAMRM (in **Table 1** columns 10–17), each item had two levels of difficulty and the item parameter was set at either −2 or 2. For compensatory QAMRM (in **Table 1** columns 19–26), an item could have 2–4 levels of difficulty. When there were two levels (items 1–3, 8–10), the item parameter was set at −2 or 2; When there were three levels (items 4–6, 11–13), the item parameter was set at −2, 1, or 2. When there were four levels (items 7, 14), the item parameter was set at −2, −1, 1, or 2. The eight classes were uniformly distributed, meaning that the mixing proportion for each class was 12.5%. There were three sample sizes (1,000, 2,000, and 4,000) for the latent ability distributions that involve the QAMRM model, with ability distributions with means $\mu_{\theta}=0.0$ and a common standard deviation $\sigma_{\theta}=1.0$ been used. Note that the mixture Rasch model is employed here for discussion, where the slope parameters of items in the item response function is unity. Therefore, the difference in σ_g^2 does not pose a problem in establishing a common scale between latent classes (Paek and Cho, 2015).

A total of 1,000 replications were generated under each condition. The item parameters were estimated via an EM implementation of the marginal maximum likelihood estimation (MMLE/EM) that was implemented using the computer program Mplus (Muthén and Muthén, 2021). In the EM estimation, maximum likelihood optimization was done in two stages. In the initial stage, 20 random sets of initial values were generated. An optimization was carried out for 10 iterations using each of the 20 random sets of initial values. The final values from the four optimizations with the highest log-likelihoods were used as the starting values in the final stage optimizations (Muthén and Muthén, 2021). The problems with the local maximum in the QAMRM did not occur because the number of classes was

TABLE 3 | ECPE Q-Matrix and the skill profile.

	Skill profile (α_1, α_2, α_3)				Q-matrix		
Class	α_1	α_2	α_3	**Item**	α_1	α_2	α_3
g_1	0	0	0	1	1	1	0
g_2	0	0	1	2	0	1	0
g_3	0	1	0	3	1	0	1
g_4	0	1	1	4	0	0	1
g_5	1	0	0	5	0	0	1
g_6	1	0	1	6	0	0	1
g_7	1	1	0	7	1	0	1
g_8	1	1	1	8	0	1	0
				9	0	0	1
				10	1	0	0
				11	1	0	1
				12	1	0	1
				13	1	0	0
				14	1	0	0
				15	0	0	1
				16	1	0	1
				17	0	1	1
				18	0	0	1
				19	0	0	1
				20	1	0	1
				21	1	0	1
				22	0	0	1
				23	0	1	0
				24	0	1	0
				25	1	0	0
				26	0	0	1
				27	1	0	0
				28	0	0	1

α_1, Morphosyntactic rules; α_2, Cohesive rules; α_3, Lexical rules.

TABLE 2 | Bias and 95% coverage rate for the item parameters in the simulation study.

Sample size	Non-compensatory		Compensatory	
	Bias	**Coverage**	**Bias**	**Coverage**
1,000	0.01	0.95	0.00	0.96
2,000	0.00	0.95	0.00	0.95
4,000	0.00	0.95	0.00	0.95

The detail simulation results about bias and coverage rate for individual parameters can be found through the link: https://www.dropbox.com/sh/b3sb3ju4swk1pmi/AABpbXMRLXFji52PfHmNbi0Za?dl=0.

TABLE 4 | Comparisons of model-data fit among the QAMRM, LCDM.

Model	AIC	BIC	ABIC
LCDM[a]	85641.43	86125.81	85868.44
QAMRM	85131.55	85568.09	85336.14
Hierarchical LCDM[b]	85638.63	86045.08	85829.21
Hierarchical QAMRM	85125.80	85538.42	85319.18

[a] *LCDM = log-linear cognitive diagnostic model; More details on these models can be found in Templin and Hoffman (2013).*
[b] *Hierarchical LCDM; More details on these models can be found in Templin and Bradshaw (2014).*

TABLE 5 | Skill profile distributions obtained for the QAMRM, LCMD.

Skill profile	QAMRM	LCDM[a]	Skill profile	Hierarchical LCDM[b]	Hierarchical QAMRM
(0,0,0)	0.44	0.30	(0,0,0)	0.34	0.44
(0,0,1)	0.11	0.13	(0,0,1)	0.11	0.08
(0,1,0)	0.00	0.01	(0,1,1)	0.18	0.13
(0,1,1)	0.10	0.18	(1,1,1)	0.38	0.34
(1,0,0)	0.00	0.01			
(1,0,1)	0.01	0.02			
(1,1,0)	0.00	0.01			
(1,1,1)	0.33	0.35			

[a] *Templin and Hoffman (2013).*
[b] *Templin and Bradshaw (2014).*

TABLE 6 | ECPE Q-matrix and hierarchical QAMRM item difficulty parameter estimates.

Item	Skill			Skill profile (α_1, α_2, α_3)		
	α_1	α_2	α_3	(0,0,0)	(0,0,1)	(0,1,1)
1	1	1	0	−1.98	−1.98	−2.04
2	0	1	0	−2.30	−2.30	−3.42
3	1	0	1	−0.01	−0.01	−0.01
4	0	0	1	−0.31	−2.54	−2.54
5	0	0	1	−2.27	−5.14	−5.14
6	0	0	1	−1.95	−4.18	−4.18
7	1	0	1	−0.43	−1.58	−1.58
8	0	1	0	−3.09	−3.09	−4.93
9	0	0	1	−0.68	−2.14	−2.14
10	1	0	0	−0.42	−0.42	−0.42
11	1	0	1	−0.42	−1.86	−1.86
12	1	0	1	2.68	0.30	0.30
13	1	0	0	−1.45	−1.45	−1.45
14	1	0	0	−0.68	−0.68	−0.68
15	0	0	1	−2.20	−5.01	−5.01
16	1	0	1	−0.37	−1.62	−1.62
17	0	1	1	−2.94	−2.94	−4.54
18	0	0	1	−2.11	−3.70	−3.70
19	0	0	1	−0.14	−2.83	−2.83
20	1	0	1	1.88	0.49	0.49
21	1	0	1	−0.83	−2.46	−2.46
22	0	0	1	1.06	−2.56	−2.56
23	0	1	0	−1.62	−1.62	−4.31
24	0	1	0	0.46	0.46	−1.12
25	1	0	0	−0.50	−0.50	−0.50
26	0	0	1	−0.86	−1.98	−1.98
27	1	0	0	1.09	1.09	1.09
28	0	0	1	−1.35	−3.97	−3.97

The syntax and detail analysis results can be found in: https://www.dropbox.com/sh/b3sb3ju4swk1pmi/AABpbXMRLXFji52PfHmNbi0Za?dl=0.
For items 1, 3, 11–12, 16–17, and 20–21, by means of the Q-matrix we can see all of these items measuring two attributes, with those inside QAMRM compensatory. Take item 1 for example; item 1 measures attributes α_1 and α_2; if the examiner masters attribute α_1 and α_2, he will have a high probability to answer the item 1 correctly, with the item difficulty in item 1 being−3.99, while if the examiner does not master attributes α_1 and α_2 at the same time, and just masters attribute α_2 compensatorily, he will have a moderate probability to answer the item correctly, with the item difficulty becoming−2.04. However, if the examiner does not master attributes α_1 and α_2, he will have a low probability of answering the item correctly, with the item difficulty becoming−1.98. The other items can be interpreted similarly.
*If the examiner doesn't master any skill, he will fall into latent class (α_1, α_2, α_3) = (0, 0, 0) and will have the lowest probability of correctly answering all the items. If the examiner just masters the α_3 skill, but does not master the α_1 and α_2 skills, the examiner will fall into (α_1, α_2, α_3) = (0, 0, 1), and the examine will have a higher probability of correctly answering items 4–6, 9, 15, 18–19, 22, 26, 28 than the examiner who hasn't mastered the α_3 skill. The details can be seen in **Table 6**. The other items can be interpreted similarly.*

specified by the user rather than explored from the data as in finite mixture models.

The parameters in QAMRM included the mixing proportion, latent class membership conditional on the attribute profile specific item parameters, and the population parameters of latent class conditional on the attribute profile specific continuous latent variable. In the QAMRM, the Q-matrix was adopted to constrain the item parameters to be invariant across latent classes, which could reduce a large number of item parameters and improve parameter estimation (see **Table 1**). Take non-compensatory QAMRM in the simulation as an example, the total parameters to be estimated were 36, including seven mixing proportion parameters (the mixing proportion parameters should add up to 1, so only seven parameters could be estimated when there were three latent attributes and eight latent profiles), one variance parameter, and 28 item parameters for the 14 items which were specified according to the Q-matrix. If the Q-matrix was not adopted, the model became the MRM, which would estimate 120 parameters, including 112 item parameters (assuming there were eight latent classes and 14 item parameters in each latent class), seven mixing proportion parameters, and one variance parameter. Such a large number of parameters would require a large sample size, which would be a practical constraint.

To evaluate the parameter recovery, we computed the bias, the 95% coverage rate for the item parameter estimates in **Table 2**. Due to enhance readability, we do not report bias and coverage rate for individual parameters; rather, we show the mean bias and mean coverage rate across all parameters in **Table 2**.

The EM estimation method yielded very small bias. Besides, the mean coverage was very close to 95%. The results of the simulated data analysis indicate that the invariance property of the QAMRM model is absolute in that the parameter estimates were obtained using different calibration samples. By means of the Q-matrix setting invariant items *a priori* in QAMRM, all model parameter estimates across latent classes to be on a common scale, which does not require any transformation for them to be comparable.

REAL DATA ANALYSIS

We used the Certificate of Proficiency in English (ECPE) data, which is available in the R package CDM (Robitzsch et al., 2011–2014), to demonstrate the advantages of the QAMRM over the LCDM. The ECPE data consist of responses from 2,922 test-takers to 28 items, with each item measuring one or two out of three skills. The data has been analyzed with the LCDM by Templin and Hoffman (2013) and Templin and Bradshaw (2014) and with the GDM by von Davier (2014). As shown previously, when analyze the ECPE data, the LCDM and GDM are mathematically equivalent (von Davier, 2014), hence we fit the QAMRM to the data using Mplus and compared the results with those under the LCDM (Templin and Hoffman, 2013; Templin and Bradshaw, 2014).

The Q-matrix used in ECPE example was the result of psychometric analyses on the ECPE by Buck and Tatsuoka (1998). The analyses showed that items of the test were likely to measure three distinct skills. The left side of **Table 3** shows the skill profile and the right side of **Table 3** shows Q-matrix that maps each item to the three skills. As shown, eight items measure only one skill, seven items measure two skills, and zero items measure three skills. The morphosyntactic (α_1), cohesive (α_2), and lexical (α_3) skills were each measured by 13, 6, and 18 items, respectively.

Therefore, we only report the skill distributions with latent class pattern and model fit results with the values published by Templin and Hoffman (2013) and Templin and Bradshaw (2014) which agree with those obtained from the CDM R-package (Robitzsch et al., 2011–2014).

We use AIC, BIC, and sample-size adjusted BIC (ABIC) to select the best model (Templin and Bradshaw, 2014), the information criteria selected the best model by small value. **Table 4** presents AIC, BIC, and ABIC for the QAMRM, LCDM. It appears that the QAMRM had lower AIC (85131.55–85641.43), BIC (85568.09–86125.81), and ABIC (85336.14–85868.44). **Table 5** shows the distributions of the eight skill profiles (classes) obtained from the QAMRM, LCDM. The distributions were very similar across models and only four skill profiles were substantial: (0,0,0), (1,0,0), (1,1,0), and (1,1,1).

Templin and Hoffman (2013) analyzed a sample of 2,922 examinees who took the ECPE with the non-hierarchical LCDM. But in the ECPE example, the data and results suggest a linear attribute hierarchy: Examinees must master Attribute 3 (lexical rules) before mastering Attribute 2 (cohesive rules) before mastering Attribute 1 (morphosyntactic rules). Gierl et al. (2007) call this structure a linear hierarchy, where mastery of each attribute follows a linear progression. Therefore, Templin and Bradshaw (2014) introduce the hierarchical LCDM where attribute hierarchies are present, the model fit of hierarchical LCDM shown at the bottom of **Table 4**, the hierarchical LCDM is used to test for the presence of a suspected attribute hierarchy in ECPE, through model fit which confirming the data is more adequately represented by hierarchical attribute structure when compared to a crossed, or non-hierarchical structure.

We reanalyzed the data with hierarchical QAMRM, and compare the model fit with hierarchical LCDM. Right column of **Table 5** shows the distributions of the four hierarchical skill profiles (classes) obtained from the hierarchical QAMRM, hierarchical LCDM. It also appears that the hierarchical QAMRM had lower AIC (85125.80–85638.63), BIC (85538.42–86045.08), and ABIC (85319.18–85829.21).

Table 6 presents Q-matrix in ECPE and hierarchical QAMRM item difficulty parameter estimates. In items 4–10,13–15, 18–19, 22–28 only one attribute is measured, with all of these items inside hierarchical QAMRM non-compensatory, having only two different kinds of item difficulty; take item 4 for example; if the examiner masters attribute α_3, he will have a high probability to answer item 4 correctly, with the item difficulty being -2.54, while, on the other hand, if the examiner does not master attribute α_3, he will have a low probability to answer the item correct, with the item difficulty becoming -0.31, and the other items can be interpreted similarity.

If we only use the LCDM or Hierarchical LCDM for analysis, a second calculation is still needed to find the probability

of a correct response for each item, which may not easy for practitioners.

CONCLUSION AND DISCUSSION

The QAMRM was used to describe for modeling the Q-matrix at the mixture Rasch model. The model developed in this study used features of a Rasch model, a restricted latent class model, and a Q-matrix. The Q-matrix of the model provides an opportunity to determine the number of latent class in advance through substantive theory and not through model fitness or parameter estimation *post hoc*. Information in the Q-matrix can be used to reveal possible differences that might be due to differences among latent classes.

A simulation study through the EM algorithm estimation was presented to investigate the performance of the model. Generated parameters were well recovered for the conditions considered. The QAMRM makes it possible to describe the differential item performance of target attributes using descriptions of Q-matrix characteristics associated with the items compared with characteristics associated with other items not in the same latent classes. This description can then be used to provide the Q-matrix with a framework within which to compare the results in their latent classes and in the other latent classes. Examiners in each of the latent class can be characterized by differences in attribute, as well as by differences in response strategies, particularly at the end of the test.

The real data comparison performed between the QAMRM, LCDM, by means of the ECPE data, shows that when a Rasch model is included inside the diagnostic classification models, QAMRM achieves a more desirable result and has better fit indices than the LCDM variants.

If we want to provide a single, continuous estimate of overall ability and classify the subjects at the same time, we should consider the QAMRM rather than mixture Rasch model or DCM. In QAMRM, the Q-matrix sets class invariant items a priori, if the Q-matrix design is not correct, then the analysis in the QAMRM will be wrong. Kopf et al. (2015) who have discussed anchoring strategies in details which can help to correct the Q-matrix design. Future research can focus on the misspecify Q-matrix design with the model unfit of the QAMRM.

On the other hand, future research can focus on the estimation limitations of the QAMRM. Specifically, as the number of attributes included in the Q-matrix increases and as its complexity increases, the number of parameters estimated by this model will also increase. In these cases, expectations of its performance in estimated attribute mastery and item parameters must be explored. In addition, model comparisons using common indices such as the AIC, BIC, and ABIC must continue to be explored, which could result to clear guidelines for model identification. Finally, possible expansions of this model such as the addition of a continuous ability measure to imply an incomplete Q-matrix (much like what is used in the testlet IRT) will be explored.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://cran.r-project.org/web/packages/CDM/index.html.

AUTHOR CONTRIBUTIONS

M-CT: writing and analyzing, 80%. W-CW: correction and feedback, 20%. Both authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.564976/full#supplementary-material

REFERENCES

Adams, R., Wilson, M., and Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Appl. Psychol. Meas.* 21, 1–23.

Alexeev, N., Templin, J., and Cohen, A. S. (2011). Spurious latent classes in the mixture rasch model. *J. Edc. Meas.* 48, 313–332. doi: 10.1111/j.1745-3984.2011.00146.x

Bradshaw, L., and Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *Int. J. Testing* 16, 99–118. doi: 10.1080/15305058.2015.1107076

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

Buck, G., and Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing* 15, 119–157. doi: 10.1177/026553229801500201

Cho, S.-J., Suh, Y., and Lee, W.-Y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Edc. Meas. Issues Practice* 35, 48–61. doi: 10.1111/emip.12093

Choi, H. J. (2010). *A model that combines diagnostic classification assessment with mixture item response theory models*. Georgia: University of Georgia.

de la Torre, J., and Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *J. Edc. Meas.* 47, 115–127. doi: 10.1111/j.1745-3984.2009.00102.x

Embretson, S. E., and Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika* 78, 14–36.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychol.* 37, 359–374. doi: 10.1016/0001-6918(73)90003-6

Gibbons, R. D., and Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika* 57, 423–436. doi: 10.1007/BF02295430

Gierl, M. J., Leighton, J. P., and Hunka, S. M. (2007). "Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills," in *Cognitive diagnostic assessment for education:theory and applications*, eds J. P. Leighton and M. J. Gierl (Cambridge: Cambridge University Press), 242–274. doi: 10.1017/CBO9780511611186.009

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

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

Hong, H., Wang, C., Lim, Y., and Douglas, J. (2015). Efficient models for cognitive diagnosis with continuous and mixed-type latent variables. *Appl. Psychol. Meas.* 39, 31–43.

Hoskens, M., and De Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Appl. Psychol. Meas.* 25, 19–37. doi: 10.1177/01466216010251002

Kelderman, H., and Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika* 59, 149–176. doi: 10.1007/BF02295181

Kopf, J., Zeileis, A., and Strobl, C. (2015). Anchor selection strategies for DIF analysis: review, assessment, and new approaches. *Edc. Psychol. Meas.* 75, 22–56. doi: 10.1177/0013164414529792

Lazarsfeld, P. F., and Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Mislevy, R. J., and Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika* 55, 195–215.

Muthén, L., and Muthén, B. (2021). *Mplus [Computer software]*. Los Angeles, CA: Author.

Paek, I., and Cho, S. J. (2015). A note on parameter estimate comparability across latent classes in mixture IRT modeling. *Appl. Psychol. Meas.* 39, 135–143. doi: 10.1177/0146621614549651

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydiche.

Robitzsch, A., Kiefer, T., George, A. C., and Ünlü, A. (2011–2014). *R-package documentation for CDM: Cognitive diagnosis modeling*. Available Online at: http://cran.r-project.org/web/packages/CDM/index.html

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

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *J. Edc. Meas.* 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. *Edc. Meas. Issues Prac.* 32, 37–50. doi: 10.1111/emip.12010

von Davier, M. (2005). *A general diagnostic model applied to language testing data (ETS Research Rep. No. RR-05-16)*. Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM) (ETS Research Rep. No. RR-14-40)*. Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12043

von Davier, M., and Yamamoto, K. (2004). Partially observed mixtures of IRT models: an extension of the generalized partial-credit model. *Appl. Psychol. Meas.* 28, 389–406. doi: 10.1177/0146621604268734

von Davier, M., and Yamamoto, K. (2007). "Mixture-distribution and HYBRID Rasch models," in *Multivariate and mixture distribution Rasch models*, eds M. von Davier and C. H. Carstensen (New York, NY: Springer), 99–115. doi: 10.1007/978-0-387-49839-3_6

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika* 45, 479–494. doi: 10.1007/BF02293610

Wright, B. D., and Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Wu, M. L., Adams, R. J., and Wilson, M. R. (1998). *Conquest: Generalized item response modeling software*. Camberwell, Victoria: Australian Council for Educational Research.

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 Tseng 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: 24 March 2021
doi: 10.3389/fpsyg.2021.628607

Remedial Teaching and Learning From a Cognitive Diagnostic Model Perspective: Taking the Data Distribution Characteristics as an Example

He Ren[1], *Ningning Xu*[2], *Yuxiang Lin*[3], *Shumei Zhang*[2] *and Tao Yang*[1]*

[1] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China, [2] School of Statistics, Beijing Normal University, Beijing, China, [3] Department of Statistics, University of California, Davis, Davis, CA, United States

In response to the big data era trend, statistics has become an indispensable part of mathematics education in junior high school. In this study, a pre-test and a post-test were developed for the six attributes (sort, median, average, variance, weighted average, and mode) of the data distribution characteristic. This research then used the cognitive diagnosis model to learn about the poorly mastered attributes and to verify whether cognitive diagnosis can be used for targeted intervention to improve students' abilities effectively. One hundred two eighth graders participated in the experiment and were divided into two groups. Among them, the intervention materials read by the experimental group students only contained attributes that they could not grasp well. In contrast, the reading materials of the control group were non-targeted. The results of the study showed the following: (1) The variance and the weighted average were poorly mastered by students in the pre-test; (2) compared with the control group, the average test score of the experimental group was significantly improved; (3) in terms of attributes, the experimental group students' mastery of variance and the weighted average was significantly improved than the pre-test, while the control group's mastery was not. Based on this, some teaching suggestions were put forward.

Keywords: cognitive diagnostic models, DINA, mathematics teaching, data distribution characteristics, formative assessment

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Miguel A. Sorrel,
Autonomous University of Madrid, Spain
Xiangzi Ouyang,
The University of Hong Kong, Hong Kong

***Correspondence:**
Tao Yang
yangtao@bnu.edu.cn

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

Received: *12 November 2020*
Accepted: *18 February 2021*
Published: *24 March 2021*

Citation:
Ren H, Xu N, Lin Y, Zhang S and Yang T (2021) Remedial Teaching and Learning From a Cognitive Diagnostic Model Perspective: Taking the Data Distribution Characteristics as an Example. Front. Psychol. 12:628607. doi: 10.3389/fpsyg.2021.628607

INTRODUCTION

For a long time, statistics has been the concern of a few researchers in the field of middle school mathematics education (Burrill, 1990). The research of Garfield and Ahlgren (1988) showed that until the 1980s, students still had few opportunities to learn statistics before entering the university. However, since the beginning of the 21st century, the explosive growth of data accessibility has made understanding and application of statistical literacy essential in all walks of life (Galesic and Garcia-Retamero, 2010; Schield, 2010; Ridgway et al., 2011; Watson, 2014). Indeed, citizens who lack statistical knowledge may not be able to distinguish between credible and unreliable information, and it is difficult for them to make decisions based on data rather than feelings (English and Watson, 2016). For the first time, people see the value of statistical literacy and widely

regard it as an essential life skill for fully functional citizens (Ridgway et al., 2011). In response to this trend, statistics has become a focus of concern for many countries and has become an indispensable part of the middle school math curriculum (Lee and Lee, 2008; Arican and Kuzu, 2019). In the United States, the National Council of Teachers of Mathematics [National Council of Teachers of Mathematics (NCTM), 2000] explicitly regarded data analysis and probability as one of the five content standards for the mathematics curriculum. In China, the mathematics curriculum standards included the content of "probability and statistics" as an independent learning module for the first time in 2001 (Ministry of Education of the People's Republic of China, 2001). Furthermore, studies have shown that an essential part of middle schools' statistics education is to teach students to choose appropriate statistical methods to analyze data and extract the information in the data (Franklin et al., 2007). In the *Curriculum Focal Points* [National Council of Teachers of Mathematics (NCTM), 2010], "analyzing and summarizing data sets" is also regarded as one of the key points of middle school math learning. In China's mathematics curriculum standards, it is clearly stated that students should master commonly used methods of measuring the center and variation of data sets in order to carry out an analysis. Therefore, this study will select the data distribution characteristics: the center and variation of the data as the research content.

Although the above statistical knowledge is highly valued, studies have shown that there are often some problems in the learning and teaching of this knowledge (Franklin et al., 2007; Batanero and Díaz, 2012). Suppose we can accurately obtain students' mastery of relevant knowledge and conduct analysis to provide teachers with teaching guidance. In that case, we can help improve these learning and teaching problems and enhance students' ability to solve statistical problems. Nevertheless, unfortunately, so far, most schools' common reporting practices in math tests are still based on classical test theories, which means that each student can only be provided with a total score. Although this can serve the purpose of ranking students, selecting candidates for projects, etc., it should be emphasized that the design of these assessments cannot naturally provide students with more refined information about their mastery (de la Torre and Minchen, 2014). Therefore, after the test, many teachers can only lead all students to review almost all the knowledge involved in the test and cannot give targeted intervention according to each student's knowledge state. This not only limits the progress of students but also limits teachers to reflect on and adjust their own teaching methods and content. In recent decades, researchers have proposed the cognitive diagnostic theory to solve the above problems (Gentile et al., 1969; Embretson, 1998; Tatsuoka, 2009; Wu, 2019).

As a product of the combination of psychometrics and cognitive psychology, the cognitive diagnostic theory is considered the new generation of measurement theory. Specifically, the cognitive diagnosis is a modeling approach aimed at providing examinees' fine-grained information on unobservable (i.e., latent) attributes required to solve specific items (Templin and Bradshaw, 2013). These attributes refer to the knowledge, skills, strategies, etc., which describe mental processing when solving the problem (Nichols et al., 1995; Chen et al., 2012). We can analyze the students' strengths and weaknesses in specific learning fields to remedy students' learning and improve teaching quality based on such fine-grained information.

Researchers have carried out some research in statistics or mathematics tests in middle schools based on cognitive diagnostic theory. For example, Chen (2012) produced a diagnostic description of urban and rural students' cognitive knowledge, abilities, and skills related to TIMSS 1999 mathematics items in Taiwan; Arican and Sen (2015) analyzed the differences between Turkish and South Korean eighth-grade students in attributes involved in the TIMSS 2011 mathematics test and analyzed Turkish students' strengths and weaknesses on these attributes; Lee et al. (2011) used the data of TIMSS 2007 fourth grade mathematics test to show that when using a specific cognitive diagnostic model, there is an incredible wealth of fine-grained information that can be translated directly for classroom application at the attribute level.

Although researchers have conducted some studies with cognitive diagnosis in mathematical tests, unfortunately, most of these studies (1) are based on large-scale assessments with a large number of participants and many items covering a wide range of content. Therefore, these studies may be used to provide references for the macropolicy formulation, such as modifying the focus of the curriculum in subsequent years, but they may not be able to promote immediate changes in the teaching of specific knowledge in a particular classroom; (2) use existing tests in the analysis, which are not guided by cognitive diagnosis when they are developed; (3) usually only use cognitive diagnosis to analyze the test data but do not use the analysis results to carry out targeted interventions on students, nor do they investigate the effects of the interventions (Lee et al., 2011; Chen, 2012).

To sum up, based on the class's actual teaching process, this study selected a specific topic of junior high school mathematics curriculum in China, specifically, the characteristics of the data distribution as the research content, and used the cognitive diagnosis to analyze the students' knowledge mastery. Then, guided by the analysis results, this study carried out targeted interventions and verified the intervention effects. This research's primary purpose is to answer the following two questions in the context of actual class tests: (1) how do students master each attribute involved in the data distribution characteristics and (2) whether the results of cognitive diagnosis can be used to intervene with students so as to improve their mastery of attributes effectively. By answering these two questions, this study not only provides an example for in-service educators to conduct cognitive diagnosis but also provides some suggestions for classroom teaching on the data distribution characteristics at both the classroom and individual levels. Besides, following the cognitive diagnosis process, this study developed two sets of instruments for measuring the knowledge about the data distribution characteristics, which were used for pre- and post-tests (the details will be mentioned later).

MATERIALS AND METHODS

Participants

In this study, 105 eighth-grade students in two parallel classes taught by the same mathematics teacher participated with no payment offered. Among them, 52 students in the first class were in the experimental condition, and 55 students in the other class were in the control condition. In addition, three students in the experimental group and two students in the control group failed to participate on the test day. Therefore, in the end, there were 49 people in the experimental condition and 53 people in the control condition. Gender distribution was as follows: 43 boys (42%) and 59 girls (58%) in total; 18 boys (37%) and 31 girls (63%) in the experimental group; and 25 boys (47%) and 28 girls (53%) in the control group.

It should be noted that although the students in the experimental and control groups were directly grouped by class, these students still had similar ability levels because of the school's class placement policy. Moreover, all participants were taught by the same mathematics teacher, making it more likely that students in these two classes have the same or similar attribute mastery.

The participants were from an ordinary junior high school in a central province of China. The educational situation in this region is a good representative of the overall educational situation in China.

Q-Matrix and Model

In general, in order to conduct cognitive diagnosis analysis, the preliminary work is to determine the Q-matrix and select the appropriate cognitive diagnosis model (CDM).

The Q-matrix is a binary matrix whose elements only take 0 or 1. In cognitive diagnosis, it represents the relations between the test items and attributes; that is, it indicates which attributes are required to correctly answer each item (Tatsuoka, 1983, 1985; Henson et al., 2007; Ravand, 2016). Therefore, the Q-matrix can guide the development of instruments from the perspective of cognitive psychology, making it play an essential role in cognitive diagnosis (Leighton and Gierl, 2007). The columns in the Q-matrix represent attributes, and the rows represent items. Given J items and K attributes, if the attribute k ($k = 1, 2, \ldots, K$) is measured in item j ($j = 1, 2, \ldots, J$), then the element $q_{jk} = 1$, otherwise $q_{jk} = 0$. It should be noted that when constructing the Q-matrix, the hierarchical structure between attributes should be considered (Gierl et al., 2007). In the following instrument development, the Q-matrix determined by this research will be presented in detail.

In terms of CDMs, researchers have proposed a large number of models in the past few decades. According to different assumptions about the relationship between attributes, these CDMs can be roughly divided into three categories: (1) compensatory models, in which participants' lack of specific knowledge required for correct answers can be compensated by other knowledge they mastered, such as linear logistic test model (LLTM; Fischer, 1973) and deterministic inputs, noisy "or" gate model (DINO; Templin and Henson, 2006); (2) non-compensatory models, in which participants' lack of knowledge required for correct answers cannot be compensated by other knowledge, such as rule space method (RSM; Tatsuoka, 1983, 2009) and deterministic inputs, noisy "and" gate model (DINA; Haertel, 1989; Junker and Sijtsma, 2001); and (3) general models, that is, models that can be converted into compensated or non-compensated models after adding constraints, such as general diagnostic model (GDM; von Davier, 2008) and log-linear CDM (LCDM; Henson et al., 2009).

Among these models, the non-compensatory models may be more appropriate for mathematical tests because all steps must be successfully answered when solving a mathematical problem, which is consistent with the assumption of the non-compensatory models (Chen, 2012). Although theoretically, the non-compensatory models are more suitable for the content of this study, in the analysis process, we still considered both the classic non-compensatory models and the compensatory models, specifically DINA, DINO, and generalized DINA (G-DINA; de la Torre, 2011), and compared their fitness.

Design and Procedure

This study followed an experimental design to verify the effectiveness of the DINA model's analysis results and the analysis-based interventions in the context of the actual class test, that is, a small number of participants in specific classes and a small number of items with specific content. The study was divided into three stages: pre-test, intervention, and post-test. The pre- and post-tests were the same for the two groups, but the intervention was different.

During the intervention, the experimental group was asked to read the one-page targeted intervention materials, which included explanations of their poorly mastered knowledge and corresponding exercises. The control group was asked to read the non-targeted intervention materials that contained all the knowledge involved in the tests. In all reading materials, the exercises were accompanied by ideas and answers. Students in both groups had 2 days to read their materials on their own.

Before and after the intervention (also called experimental treatment), pre- and post-tests were applied, respectively. Each test had a proximate duration of 25 min to respond to 17 items. Moreover, the two tests were exactly the same in terms of item format, and the items in the corresponding positions measured the same attributes. In addition, since the composition of the experimental group's intervention materials was determined by the analysis results of the DINA model, the interventions of the experimental and the control groups started on the day after the pre-test.

According to the Helsinki Declaration (World Medical Association, 2013), we strictly followed the ethical principles for psychological research. We informed all the participants of this study's purpose and ensured that they all understood our purpose and the possible benefits of proper participation. It was possible to drop out of the study, but no participant dropped out.

Instruments Development

In this study, following the development process of cognitive diagnosis tests described below, two sets of instruments for

measuring the knowledge about data distribution characteristics were developed, which were used for pre- and post-tests.

The Attributes of Data Distribution Characteristics and Their Hierarchical Structure

The quality of the diagnostic assessment is affected by how correctly the attributes underlying the items of any given test have been specified (Ravand, 2016). A variety of sources can be used to define attributes involved in a test, such as test specifications, analysis of item content, think-aloud protocol, and the results obtained from related research (Leighton et al., 2004; Leighton and Gierl, 2007).

Through studying the Chinese Mathematics Curriculum Standards for Full-Time Compulsory Education, the think-aloud protocol, and consulting experts, researchers determined the six main attributes in the characteristics of the data distribution and the hierarchies of these attributes. The six attributes were sort (A1), median (A2), average (A3), variance (A4), weighted average (A5), and mode (A6).

Then, the hierarchical structure between the above attributes was preliminarily analyzed as follows.

First, consider the relationship between sort, average, and median. For a set of data with odd numbers, as long as the students master the concepts of sort and median, they can find the median. However, for a set of data with even numbers, students also need to calculate the average of the two numbers in the middle. That is, they need to master the concept of average. Therefore, the sort and average are the direct prerequisites for the median.

Second, the average is a direct prerequisite for the variance. Since the average is involved in the calculation of variance, the average is a prerequisite for the variance. Moreover, there is no such attribute. While it is a prerequisite for variance, the average is its prerequisite. Therefore, the average is a direct prerequisite for the variance.

Next, the average is also a direct prerequisite for the weighted average. Only by mastering the average can students master the weighted average, so the average is a prerequisite for the weighted average. Furthermore, there is no such attribute. While it is a prerequisite for the weighted average, the average is its prerequisite. Thus, the average is said to be the direct prerequisite for the weighted average.

Lastly, the mode is not related to other attributes. For a particular set of data, students can find the mode of this set by simple counting. On the other hand, if students do not master the concept of mode, they may still master other attributes. For this reason, the mode is independent of other attributes.

Then, we selected six students who were also in the eighth grade to conduct think aloud. Then, the relationship between these attributes was further analyzed through this method. This process included three steps: training, formal experiment, and analysis. First, the researcher explained the specific requirements for thinking aloud and took three specific items as examples to describe the thinking process in solving the problems. After the students fully understood this process, we asked each student to use the think-aloud method to report on the process of solving 10 items about the data distribution characteristics. Finally, the students' reports were summarized and analyzed. Based on the above analysis and the opinions provided by the think-aloud protocol and experts, it can be considered that the hierarchical structure of these attributes is as follows.

Q-Matrix and Instrument Development

Based on the hierarchies, we can obtain the reachability matrix, R, which reflects the direct and indirect connections between attributes. If the attribute k is reachable from attribute $k^{'}$, that is, attribute $k^{'}$ is a prerequisite of attribute k, then $r_{k^{'}k} = 1$, otherwise, $r_{k^{'}k} = 0$. Therefore, the reachability matrix, R, corresponding to **Figure 1** is,

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The next step was to construct a set of potential items. In the case of the six attributes in this study, if the hierarchical structure is not considered, then the size of this set is related to the number of attributes K (i.e., $2^6 - 1 = 63$). As mentioned above, the Q-matrix can be used to display such a set of potential items. However, if the attribute hierarchy is taken into consideration, how to obtain a set of potential problems that can satisfy these hierarchies? Ding et al. (2009) proposed that starting from the R matrix and using the expansion algorithm, a set of potential items that conforms to the hierarchical structure can be efficiently obtained. Based on their method, this research obtained a set of all the potential items that satisfy the hierarchies, denoted as initial Q-matrix, as shown in **Table 1** (if a row with all 0s was added, it can also represent all possible knowledge states of students).

It should be pointed out that 10 items were removed from the initial Q-matrix for the following three reasons. First, this study's sample size is relatively small, so if 27 items are all used, it may bring challenges to model estimation. Second, in practice, the attribute A1 (sort) is straightforward to be mastered by middle school students, and it is often examined together with the median in tests. Simultaneously, there are many (66.667%) items related to the sort in the initial Q-matrix. Therefore, we removed a part of those items that examined the sort without examining the median at the same time. Third, for a daily class test, it is unlikely to contain so many items. Therefore, based on experts' and teachers' opinions, we selected 17 typical items from the initial Q-matrix to form the final Q-matrix (**Table 2**).

Finally, this research developed the tests based on the final Q-matrix. For example, according to the third row in the final Q-matrix, the third item should only measure the attribute A3 (i.e., the average). Therefore, researchers compiled an item that only measured the attribute A3, as shown in **Table 3**.

It is easy to know that this item only measures the attribute A3, which is consistent with the Q-matrix. According to this process, the pre- and post-test papers, each containing 17 items, were compiled. All items are multiple-choice items with four response options, and all items are 0–1 scored, with a full score of 17.

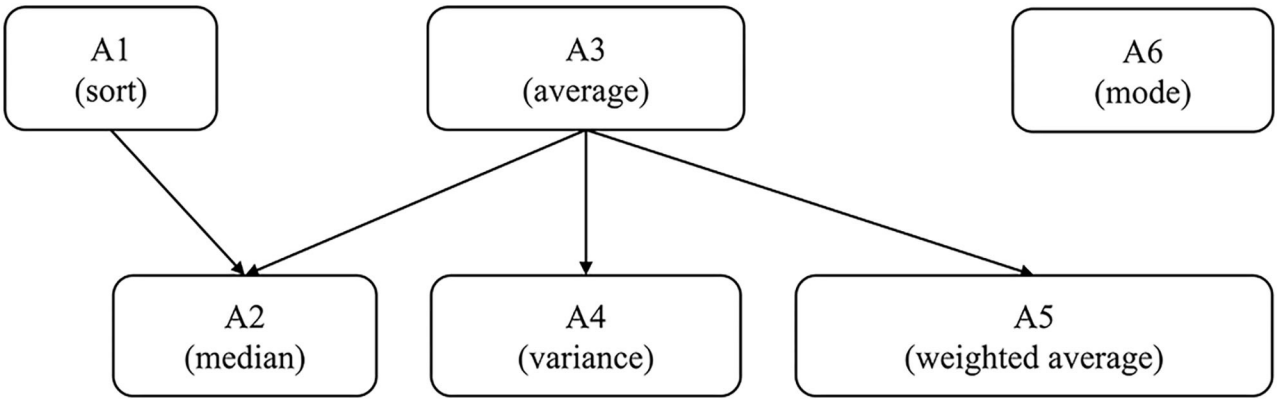


FIGURE 1 | The hierarchical structure of the six attributes of this study.

TABLE 1 | Initial Q-matrix.

Items	A1	A2	A3	A4	A5	A6
1	1	0	0	0	0	0
2	0	0	1	0	0	0
3	1	0	1	0	0	0
4	1	1	1	0	0	0
5	0	0	1	1	0	0
6	1	0	1	1	0	0
7	1	1	1	1	0	0
8	0	0	1	0	1	0
9	1	0	1	0	1	0
10	1	1	1	0	1	0
11	0	0	1	1	1	0
12	1	0	1	1	1	0
13	1	1	1	1	1	0
14	0	0	0	0	0	1
15	1	0	0	0	0	1
16	0	0	1	0	0	1
17	1	0	1	0	0	1
18	1	1	1	0	0	1
19	0	0	1	1	0	1
20	1	0	1	1	0	1
21	1	1	1	1	0	1
22	0	0	1	0	1	1
23	1	0	1	0	1	1
24	1	1	1	0	1	1
25	0	0	1	1	1	1
26	1	0	1	1	1	1
27	1	1	1	1	1	1

TABLE 2 | Final Q-matrix.

Items	A1	A2	A3	A4	A5	A6
1	1	0	0	0	0	0
2	1	1	1	0	0	0
3	0	0	1	0	0	0
4	0	0	1	1	0	0
5	0	0	1	0	1	0
6	0	0	0	0	0	1
7	1	1	1	0	0	1
8	1	1	1	1	0	0
9	1	1	1	0	1	0
10	0	0	1	1	0	1
11	1	0	1	0	0	1
12	0	0	1	1	1	0
13	1	1	1	1	0	1
14	0	0	1	0	1	1
15	1	1	1	0	1	1
16	1	1	1	1	1	1
17	1	0	1	0	0	0

TABLE 3 | Item example.

Item	Item content
Item 3	If the average of a set of data a, b, c, d is M, then the average of another set of data 2a + 2, 2b + 2, 2c + 2, 2d + 2 is (). A. 2M B. 2M + 1 C. 2M + 1.5 D. 2M + 2

Data Analysis

We conducted descriptive and inferential analyses of the students' responses. Specifically, the analysis of variance (ANOVA) was used to examine the difference in average scores between both the groups (experimental and control group) and the moments (pre- and post-tests). We also used three CDMs mentioned in *Q-Matrix and Model* to analyze the students' responses and chose the most suitable CDM. It is then used for item analysis and used to obtain the attributes of poor mastery of students and students' probability of belonging to each knowledge state. In addition, the reliability of the model was examined. In addition, it should be pointed out that the

modeling with the selected CDM was carried out according to the moment. In other words, for the selected model, a total of two modeling was performed, and then, the person parameter estimates were segmented by the condition. There are no missing values. All the data analysis processes were completed in R, and the CDM package was used.

RESULTS

Difference Analyses of the Tests Before and After the Intervention

Table 4 shows descriptive analyses of the tests in experimental and control conditions before and after the intervention and the difference between these two moments.

TABLE 4 | Means and standard deviations of pre-test, post-test, and the difference between tests in both groups.

	Pre-test		Post-test		Pre–post	
Group	**M**	**SD**	**M**	**SD**	**M**	**SD**
Experimental group	12.878	2.395	11.918	3.347	0.960	3.409
Control group	12.906	2.452	9.717	3.559	3.189	3.258

In order to further clarify whether the difference between average scores in **Table 4** is statistically significant, we took an ANOVA with group (experimental and control) as an intersubject variable and time (pre- and post-tests) as a within-subject variable. In addition, **Figure 2** shows the interaction between time and groups in the ANOVA.

The results of the ANOVA show that the time × group interaction is significant ($F = 11.401, p = 0.001, partial\ \eta^2 = 0.102$). This shows that from the pre- to post-test, the difference between the experimental group and the control group has changed. Combining with **Figure 2**, we can know that (1) the experimental and control groups did not show significant differences in the average scores of the pre-test ($b = -0.028, t = -0.059, p = 0.953$). To some extent, this confirms our previous assumption that the knowledge mastery level of the two groups of students should be basically the same (because the two groups of students came from two parallel classes). (2) The experimental and control groups showed significant differences in the average scores of the post-test ($b = 2.201, t = 3.211, p = 0.001$). In addition, it should be pointed out that students' average score in the post-test is lower than that in the pre-test, mainly because the items in the post-test are more difficult. This study took into account that the practice effect may cause the overall students' post-test scores to be too high, making it impossible to distinguish the difference between the experimental group and the control group. Therefore, although the pre- and post-test

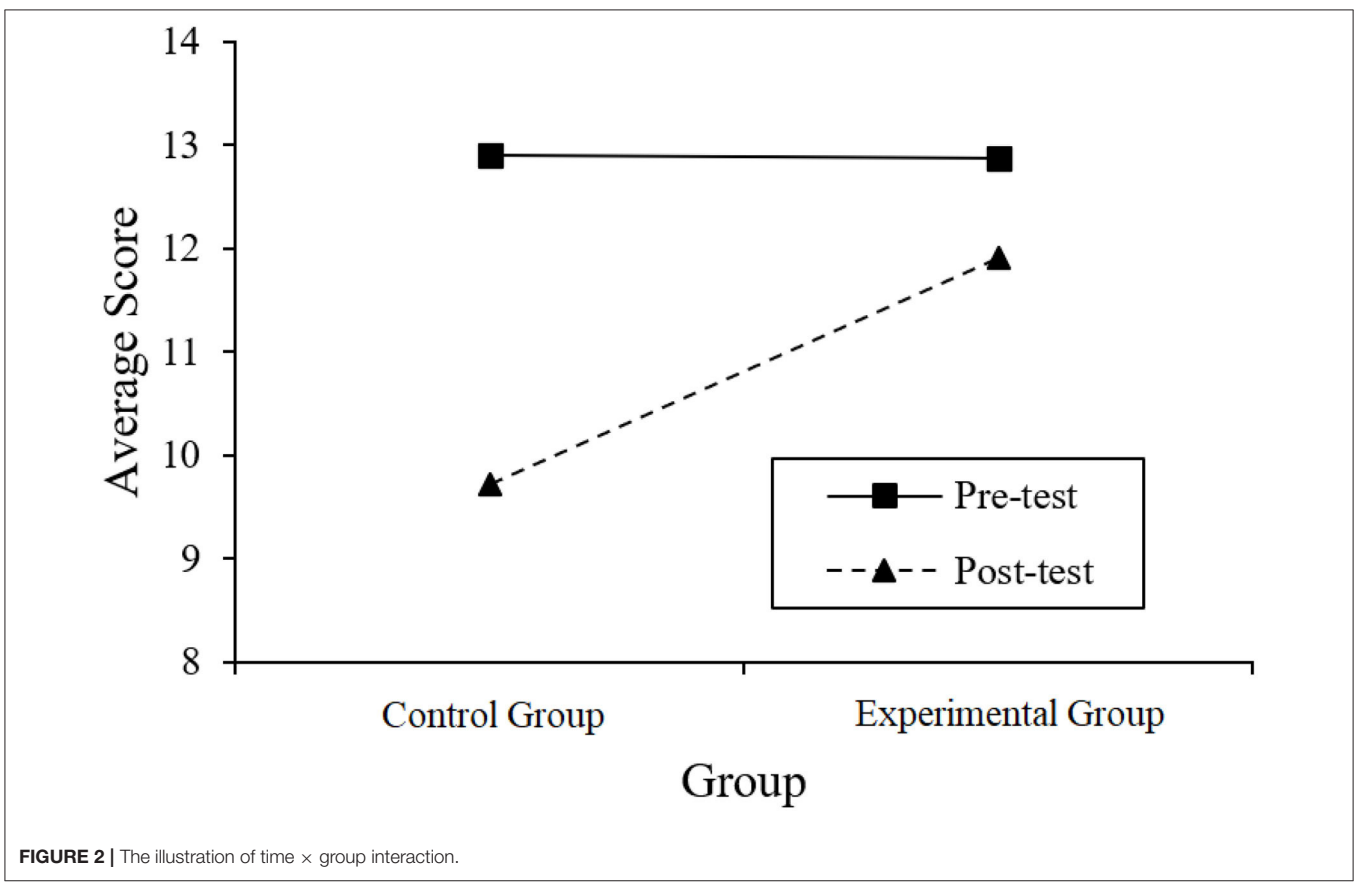


FIGURE 2 | The illustration of time × group interaction.

papers of this study were developed according to the same Q-matrix, the post-test was more difficult.

Diagnosis Results of the Pre-test and Post-test of the Two Groups

The following shows the fit of the three CDMs to the pre-test data, and the most suitable model is selected for subsequent analysis. Then, some reliability indicators of the model are shown. Last but not the least, it presents the pre-test and post-test diagnosis results of the students in two groups, which is an attribute-aspect analysis of the pre- and post-test of the two groups.

Model Fit and Item Analysis

As mentioned earlier, we considered three CDMs (DINA, DINO, and G-DINA). DINA and DINO are conjunctive and disjunctive models, respectively, while G-DINA is a general model that combines DINA and DINO (Sorrel et al., 2016). By evaluating each model's absolute and relative fit, the most suitable CDM can be selected (Sessoms and Henson, 2018). Considering that the DINA and DINO models are nested in the G-DINA model, the DINA and DINO models will always have a lower log likelihood (de la Torre, 2011). Therefore, the likelihood ratio (LR) test can be used to assess whether the observed difference in model fit is statistically significant. If LR is significantly different from 0, the general model fits the data significantly better than the simplified model. In addition, we present the Bayesian information criterion (BIC) of the model, which can also measure the model fit. The smaller the BIC, the better the model fit. Regarding the absolute fit, we used the method proposed by Chen et al. (2012), which is to evaluate the absolute fit. If the evaluated model fits the data well, the maximum χ^2 statistics should not be zero significantly different.

Table 5 shows the relative fit and absolute fit indices calculated for the G-DINA, DINA, and DINO models. Among them, the BIC of the DINA model is the smallest. The two LR tests, respectively corresponding to the comparison of the G-DINA model with the DINA ($LR = 124.572$) and DINO ($LR = 165.026$) models, are not significant ($p > 0.05$), which shows that the more parsimonious models (DINA and DINO) do not result in significant loss of fitting. Absolute item fit statistics also indicated that the DINA model has a better fit than the other models. For all three models, the maximum χ^2 statistics were not significant at a-level of 0.05 after applying the Holm–Bonferroni correction (Holm, 1979). In short, the DINA model fits the pre-test best, which is also consistent with the theoretical analysis mentioned in *Q-Matrix and Model*. Thus, the DINO and G-DINA models are discarded, and the DINA model is further examined for its adequacy to model the post-test.

In addition, we also need to confirm whether the hierarchical structure defined in **Figure 1** and the final Q-matrix defined in **Table 2** are correct. To test the correctness of the hierarchy, we modeled the pre-test using a saturated DINA model without considering the hierarchical structure. The results show that the BIC of the saturated model is 1915.794, which is slightly larger than the simple model. According to the study of Akbay and de la Torre (2020), this means that the structure is correct because the simpler model considering the hierarchical structure has a similar fit to the saturated model where the probability for all the possible latent classes is estimated (i.e., 2^6). For the Q-matrix, according to the research of Sorrel et al. (2016), the modification of the Q-matrix should make theoretical sense. For the test instruments developed by this research, it is very clear whether each item has measured specific attributes. Therefore, although a few rows of the Q-matrix are suggested to be modified according to the methods proposed by Chiu (2013) and de la Torre (2008), we did not adjust the final Q-matrix.

TABLE 5 | Model fit indices for different cognitive diagnosis models.

				LR test			Absolute fit		
Model	**Log-like**	**BIC**	**Np**	**LR**	***df***	***p***	**abs(fcor)**	**max(χ^2)**	***p***
G-DINA	−671.300	2527	256				0.251	8.148	0.586
DINA	−733.586	1749.295	61	124.572	195	0.999	0.271	5.413	1.000
DINO	−753.813	1789.749	61	165.026	195	0.942	0.374	12.109	0.068

Np is the number of parameters.

TABLE 6 | Classification accuracy and consistency.

			Attributes					
Time	**Indicator**	**Pattern**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
Pre-test	Accuracy	0.765	0.961	0.917	0.990	0.931	0.833	0.980
	Consistency	0.784	0.980	0.931	0.980	0.902	0.824	0.961
Post-test	Accuracy	0.817	0.931	0.906	0.960	0.936	0.941	0.941
	Consistency	0.723	0.881	0.832	0.941	0.891	0.901	0.901

Finally, the reliability of CDM scores was tested. According to the pre- and post-test data, the classification accuracy and consistency proposed by Cui et al. (2012) are calculated based on simulation, as shown in **Table 6**. It can be seen that the classification accuracy and consistency of the student pattern (i.e., the student's knowledge states) are >0.75, and the classification accuracy and consistency of the attributes are >0.9, except for a few cases.

So far, it can be considered that by modeling the pre- and post-test through the DINA model, we can obtain accurate diagnosis results. The following are analyses of the items and the attributes. **Table 7** shows some information about the pre- and post-test at the item level. It can be seen that although some guess parameters are relatively large, the guessing parameters and slipping parameters for most items are within a reasonable range. As mentioned earlier, the tests are 0–1 scored, and no points are deducted for wrong answers, which can partly explain the larger guessing parameters. Simultaneously, due to the small sample size, these estimates' standard deviation may be large, which brings some challenges to interpreting those extreme item parameter estimates. These explanations need to be given in conjunction with the specific item content, and interested readers can contact us for the item content. In addition, according to the standard of Kunina-Habenicht et al. (2009, 2012), 76.471% of the items in the pre-test show moderate or good fit [root mean square

TABLE 7 | Item information.

	Pre-test					Post-test			
Item	**P_c**	**g**	**s**	**RMSEA**	**Item**	**P_c**	**g**	**s**	**RMSEA**
1	0.990	1.000	0.010	0.036	1	0.961	0.878	0.000	0.087
2	0.882	0.769	0.094	0.066	2	0.814	0.746	0.143	0.083
3	0.843	0.000	0.104	0.038	3	0.765	0.214	0.125	0.072
4	0.569	0.185	0.182	0.060	4	0.627	0.002	0.188	0.060
5	0.941	0.856	0.000	0.046	5	0.588	0.252	0.263	0.076
6	0.971	0.751	0.011	0.061	6	0.686	0.344	0.169	0.077
7	0.824	0.269	0.063	0.054	7	0.725	0.409	0.000	0.067
8	0.637	0.374	0.169	0.087	8	0.608	0.315	0.166	0.085
9	0.912	0.830	0.000	0.068	9	0.333	0.124	0.485	0.080
10	0.676	0.218	0.000	0.053	10	0.696	0.452	0.172	0.102
11	0.745	0.000	0.191	0.054	11	0.735	0.401	0.045	0.099
12	0.500	0.259	0.260	0.126	12	0.627	0.406	0.253	0.099
13	0.696	0.730	0.329	0.126	13	0.529	0.382	0.318	0.207
14	0.980	0.954	0.000	0.041	14	0.627	0.108	0.142	0.068
15	0.363	0.267	0.534	0.111	15	0.490	0.272	0.321	0.107
16	0.539	0.463	0.376	0.187	16	0.343	0.115	0.421	0.070
17	0.824	0.509	0.150	0.053	17	0.618	0.289	0.226	0.094

P_c is the proportion of correct; g and s are the item parameters involved in the DINA model, representing the probabilities of slipping and guessing, respectively; RMSEA is the root mean square error approximation.

TABLE 8 | The probabilities of students' mastery of the attributes in both conditions and moments.

	Pre-test		Post-test	
Attributes	**Experimental group**	**Control group**	**Experimental group**	**Control group**
A1	1.000	0.962	0.837	0.566
A2	0.857	0.830	0.776	0.566
A3	0.959	0.925	0.918	0.717
A4	0.551	0.717	0.878	0.679
A5	0.653	0.679	0.837	0.585
A6	0.959	0.887	0.837	0.585

error approximation (RMSEA) < 0.1], and the proportion in the post-test is 82.353%.

The Probabilities of Students' Mastery of the Attributes

Table 8 shows the probabilities of students' mastery of the attributes. In the pre-test, the students' mastery of the attributes is generally good, and the probabilities of students' mastery of the attributes show consistent characteristics in both groups. Specifically, students had a good grasp of A1, A2, A3, and A6 (the probabilities are above 0.7) but had a poor grasp of A4 and A5 (the probabilities are around 0.6).

The intervention materials mentioned above were developed based on the results of this analysis. For the experimental group, the one-page targeted intervention materials they read only include the explanation of A4 and A5 and the corresponding exercises, while the materials read by the control group involve all the six attributes, from A1 to A6.

In the post-test, compared with the control group, the probabilities of experimental group show that the targeted intervention was successful. Specifically, in the experimental group, the probability of students' mastery of A4 was increased from 0.551 to 0.878 ($\chi^2 = 12.800, p = 0.000$), and the probability of students' mastery of A5 was raised from 0.653 to 0.837 ($\chi^2 = 4.350, p = 0.037$). However, in the control group, the probability of students' mastery of A4 was changed from 0.717 and 0.679 to 0.679 and 0.585 ($\chi^2 = 0.179, p = 0.672$; $\chi^2 = 1.014, p = 0.314$), respectively. It indicates that from the pre-test to the post-test, the probabilities of the experimental group students' mastery of the two attributes, A4 and A5, were increased obviously. However, for the control group, the changes in the probabilities were fluctuated and relatively smaller. What is more, the changes in the control group were not significant. Therefore, it shows that the post-test result of the control group, especially the students' mastery of A4 and A5, is not as good as the experimental group.

The Probability of Students Belonging to Each Possible Knowledge State

Table 9 shows the probability of students belonging to each possible knowledge state, which can be seen as a more specific expansion of **Table 8**. In the pre-test, although there are some differences in the probabilities between the experimental group and the control group, the knowledge states with higher probability all correspond to not mastering A4 or A5 or both of them. The results of the post-test show that except for the knowledge state of 111111, the probability of the experimental group students belonging to any other knowledge state is <0.1, which indicates that the students in the experimental group do not have a particularly poor grasp of any attribute, but the students in the control group still have poor grasp of certain attributes.

DISCUSSION

Through cognitive diagnosis, this study analyzed the junior high school students' mastery of the six attributes involved in the data distribution characteristics and used the analysis results to conduct targeted interventions on the students in the experimental group. The results show that among the attributes, students had a relatively poor grasp of A4 (variance) and A5 (weighted average) before the intervention. After the intervention, compared with those who read the non-targeted material, those who read the targeted material significantly improved their mastery of the variance and the weighted average. These results answer the two questions raised by this research very well.

On the one hand, the pre-test results in **Table 5** directly indicate that the probability of students mastering the attributes A4 and A5 is low. At the same time, the analysis of the students' knowledge states in **Table 6** shows that before the intervention, students have a higher probability of belonging to the following knowledge states: 111111, 111001, and 101011. It is consistent

TABLE 9 | The probability of students belonging to each possible knowledge state in both conditions and moments.

	Pre-test		Post-test	
Knowledge states	**Experimental group**	**Control group**	**Experimental group**	**Control group**
000000	–	–	0.082	**0.283**
100000	0.041	0.075	–	–
001000	–	–	–	–
101000	–	–	–	–
111000	–	–	–	–
001100	–	–	0.020	0.038
101100	–	–	–	–
111100	–	–	0.061	0.094
001010	–	–	–	–
101010	–	–	–	–
111010	–	–	–	–
001110	–	0.038	–	–
101110	–	–	–	–
111110	–	–	–	–
000001	–	–	–	–
100001	–	–	–	–
001001	–	–	–	–
101001	–	–	–	–
111001	**0.306**	**0.170**	–	–
001101	–	–	–	–
101101	–	–	–	–
111101	–	0.075	–	–
001011	–	–	–	–
101011	**0.102**	0.038	–	–
111011	–	–	0.041	0.038
001111	–	–	0.061	**0.113**
101111	–	0.019	0.061	–
111111	**0.551**	**0.585**	**0.673**	**0.434**

In bold are the probabilities >0.1. The dashes indicate that these probabilities are <0.001.

with the analysis of attributes in **Table 5** because most of these states indicate that attribute variance or weighted average is not mastered. In many previous studies, the weighted average was also regarded as a difficulty (Pollatsek et al., 1981; Day et al., 2014). However, except for a few studies, the variance is rarely mentioned. In fact, as a measure of the data variation, the variance is also difficult to grasp by students (Koparan, 2015). This is mainly due to the following two reasons: (1) The calculation of variance involves multiple steps such as calculating the average, square, and the sum of polynomials, which makes it easy to make mistakes. (2) The concept of variance is abstract (Sinitsky and Ilany, 2009). As for the weighted average, it is calculated by averaging after assigning different weights to each data. After communicating with the class's mathematics teacher, it was found that because she believed that the weighted average was an extension of the average, she did not spend much time explaining to the students in detail, which may lead to students' poor understanding of the concept of weight.

On the other hand, the results show that those participants who read the targeted material improved their mastery of the variance and the weighted average. In contrast, those who read the non-targeted material did not experience any significant improvement, which reveals that the targeted reading material is effective. In addition, since the targeted reading material was developed based on the pre-test diagnosis results, the diagnosis results are also valid.

According to the results of this research, some teaching suggestions on data distribution characteristics can also be provided. Here are some suggestions for classroom teaching through the diagnosis results, the intervention, and the communication with teachers. In the teaching of weighted average, it is recommended that teachers should not only teach students its calculation formula but also pay more attention to the explanation of the concept of weight in the weighted average. Considering the weight's abstractness, it is suggested that teachers should actively use examples from life to help students understand the weight. On the one hand, in the teaching of variance, it is necessary to help students understand the concept as intuitively as possible. On the other hand, since the calculation process of variance is relatively complicated, more exercises should be given to students on the variance calculation.

Furthermore, the following are some learning suggestions on the individual level for certain knowledge states. These knowledge states are those whose probabilities are >0.1 in the pre-test or those who have mastered other attributes but not A4 or A5.

(1) The students with a knowledge state of 111011 only have a poor grasp of variance, a difficult-to-understand attribute, indicating that they have sufficient learning ability. Therefore, it is recommended to conduct a targeted review, grasp the basic concept, and calculation method of variance.
(2) The students with a knowledge state of 111001 have a poor grasp of the two attributes of weighted average and variance. These two attributes are the attributes that the entire classes have a poor grasp of. It is recommended that such students review the definitions of weighted average and variance, find examples in life to help understand, and do a certain number of exercises.
(3) The students with a knowledge state of 111101 have a poor grasp of the weighted average. It is recommended that such students start with the concept of weight, grasp the basic concept of weighted average, and do exercises to help themselves master it.
(4) Students with a knowledge status of 101011 have a poor grasp of the median and variance. Compared with other attributes, especially the weighted average, it is not that difficult to grasp the attribute of the median. It should also be noted that the calculation of the median involves judging whether the number of data is odd or even. Therefore, this kind of students may not study hard and may think that they have mastered certain attributes, when in fact they only know a rough idea. It is recommended that they actively adjust their

mentality, review the textbooks, and do a certain number of exercises for the median and variance.

This study is not exempt from limitations. A major limitation is that the post-testing items are too difficult. Consequently, the diagnosis results show that in the post-test, the probabilities of students' mastery of certain attributes are lower than in the pre-test. This brings certain difficulties to the interpretation of the results and makes the results unintuitive. In addition, the sample size of this study is indeed small, which brings certain challenges to parameter estimation. Finally, the intervention of this study can be further refined, such as conducting the individual-based intervention.

In general, this study provides an example to show that in the actual class tests that usually have few participants and few items, the cognitive diagnosis can be used to obtain a relatively accurate students' knowledge state. Then, remedial teaching can be developed based on these results. In other words, this study guides in-service educators to use cognitive diagnosis to reflect on their teaching methods, adjust teaching content, and carry out remedial teaching in the teaching process. Finally, based on the CDM, some suggestions for classroom teaching and individual learning on the topic of data distribution characteristics are given.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of the Faculty of Psychology, BNU. Written informed consent to participate in this study was provided by the participants' legal guardian/ next of kin.

AUTHOR CONTRIBUTIONS

HR provided original thoughts and completed the writing of this article. NX and YL provided key technical support. SZ and TY participated in all research procedures and gave important guidance. All authors contributed to the article and approved the submitted version.

FUNDING

Funding was provided by the Key Project of the Beijing Educational Science Thirteenth Five-Year Plan Foundation in China (Grant Number: CADA17079).

ACKNOWLEDGMENTS

Thanks to Yingshi Huang and Sijian Wang from Beijing Normal University for their assistance in revising this article.

REFERENCES

Akbay, L., and de la Torre, J. (2020). Estimation approaches in cognitive diagnosis modeling when attributes are hierarchically structured. *Psicothema* 32, 122–129. doi: 10.7334/psicothema2019.182

Arican, M., and Kuzu, O. (2019). Diagnosing preservice teachers' understanding of statistics and probability: developing a test for cognitive assessment. *Int. J. Sci. Math. Educ.* 18, 771–790. doi: 10.1007/s10763-019-09985-0

Arican, M., and Sen, H. (2015). "A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment," in *Paper Presented at the 2015 Annual Meeting of the American Educational Research Association* (Chicago, IL).

Batanero, C., and Díaz, C. (2012). Training school teachers to teach probability: reflections and challenges. *Chil. J. Stat.* 3, 3–13.

Burrill, G. (1990). Implementing the standards: statistics and probability. *Math. Teacher* 83, 113–118. doi: 10.5951/MT.83.2.0113

Chen, P., Xin, T., Wang, C., and Chang, H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika* 77, 201–222. doi: 10.1007/s11336-012-9255-7

Chen, Y. H. (2012). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan. *Assess. Educ.* 19, 193–209. doi: 10.1080/0969594X.2011.560562

Chiu, C. Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Appl. Psychol. Meas.* 37, 598–618. doi: 10.1177/0146621613488436

Cui, Y., Gierl, M. J., and Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Educ. Meas.* 49, 19–38. doi: 10.1111/j.1745-3984.2011.00158.x

Day, J., Holmes, N., Roll, I., and Bonn, D. (2014). "Finding evidence of transfer with invention activities: teaching the concept of weighted average," in *Paper presented at the 2013 Physics Education Research Conference* (Portland, OR). doi: 10.1119/perc.2013.pr.017

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: development and applications. *J. Educ. Meas.* 45, 343–362. doi: 10.1111/j.1745-3984.2008.00069.x

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 Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educ.* 20, 89–97. doi: 10.1016/j.pse.2014.11.001

Ding, S. L., Zhu, Y. F., Lin, H. J., and Cai, Y. (2009). Modification of Tatsuoka's Q matrix theory. *Acta Psychol. Sin.* 41, 175–181. doi: 10.3724/SP.J.1041.2009.00175

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychol. Methods* 3, 380–396. doi: 10.1037/1082-989X.3.3.380

English, L. D., and Watson, J. M. (2016). Development of probabilistic understanding in fourth grade. *J. Res. Math. Educ.* 47, 28–62. doi: 10.5951/jresematheduc.47.1.0028

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychol. (Amst)* 37, 359–374. doi: 10.1016/0001-6918(73)90003-6

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., et al. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A PreK-12 Curriculum Framework.* Alexandria, VA: American Statistical Association.

Galesic, M., and Garcia-Retamero, R. (2010). Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. *Arch. Intern. Med.* 170, 462–468. doi: 10.1001/archinternmed.2009.481

Garfield, J., and Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *J. Res. Math. Educ.* 19, 44–63. doi: 10.2307/749110

Gentile, J., Kessler, K., and Gentil, K. (1969). Process of solving analogies items. *J. Educ. Psychol.* 60, 494–502. doi: 10.1037/h0028379

Gierl, M. J., Leighton, J. P., and Hunka, S. M. (2007). "Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills," in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. P. Leighton and M. J. Gierl (Cambridge: Cambridge University Press), 242–274. doi: 10.1017/CBO9780511611186.009

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26, 301–321. doi: 10.1111/j.1745-3984.1989.tb00336.x

Henson, R., Templin, J., and Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *J. Educ. Meas.* 44, 361–376. doi: 10.1111/j.1745-3984.2007.00044.x

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

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

Koparan, T. (2015). Difficulties in learning and teaching statistics: teacher views. *Int. J. Math. Educ. Sci. Technol.* 46, 94–104. doi: 10.1080/0020739X.2014.941425

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: comparing results from confirmatory factor analysis and diagnostic classification models. *Stud. Educ. Eval.* 35, 64–70. doi: 10.1016/j.stueduc.2009.10.003

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *J. Educ. Meas.* 49, 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Lee, H. S., and Lee, J. T. (2008). "The use of intervals to help coordinate understandings of center and spread: a preliminary report," in Paper presented at the 11th Conference on Research in Undergraduate Mathematics Education (San Diego, CA).

Lee, Y. S., Park, Y. S., and Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. National sample using the TIMSS 2007. *Int. J. Testing* 11, 144–177. doi: 10.1080/15305058.2010.534571

Leighton, J. P., and Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educ. Meas.* 26, 3–16. doi: 10.1111/j.1745-3992.2007.00090.x

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's Rule-Space Approach. *J. Educ. Meas.* 41, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Ministry of Education of the People's Republic of China (2001). *Mathematics Curriculum Standards for Full-Time Compulsory Education (Experimental Draft)*. Beijing: Beijing Normal University Press.

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics, NCTM.

National Council of Teachers of Mathematics (NCTM) (2010). *Focus in Grades 6–8: Teaching with Curriculum Focal Points*. Reston, VA: National Council of Teachers of Mathematics, NCTM.

Nichols, P. D., Chipman, S. F., and Brennan, R. L. (1995). *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pollatsek, A., Lima, S., and Well, A. D. (1981). Concept or computation: students' understanding of the mean. *Educ. Stud. Math.* 12, 191–204. doi: 10.1007/BF00305621

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

Ridgway, J., Nicholson, J., and McCusker, S. (2011). "Developing statistical literacy in students and teachers," in *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education*, eds C. Batanero, G. Burrill, and C. Reading (Dordrecht: Springer), 311–322. doi: 10.1007/978-94-007-1131-0_30

Schield, M. (2010). "Assessing statistical literacy: take care," in *Assessment Methods in Statistical Education: An International Perspective*, eds P. Bidgood, N. Hunt, and F. Jolliffe (Milton: John Wiley and Sons), 133–152. doi: 10.1002/9780470710470.ch11

Sessoms, J., and Henson, R. A. (2018). Applications of diagnostic classification models: a literature review and critical commentary. *Measurement* 16, 1–17. doi: 10.1080/15366367.2018.1435104

Sinitsky, I., and Ilany, B. S. (2009). "Conception of variance and invariance as a possible passage from early school mathematics to algebra," in *Proceedings of the Sixth Congress of the European Society for Research in Mathematics Education*, eds D. G. Viviane, S. L. Sophie, and A. Ferdinando (Lyon: Institut National De Recherche Pédagogique), 639–648.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., and Lievens, F. (2016). Validity and reliability of situational judgement test scores: a new approach based on cognitive diagnosis models. *Organ. Res. Methods* 19, 506–532. doi: 10.1177/1094428116630065

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

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception in the pattern classification approach. *J. Educ. Behav. Stat.* 12, 55–73. doi: 10.3102/10769986010001055

Tatsuoka, K. K. (2009). *Cognitive Assessment.* An Introduction to the Rule Space Method. New York: Routledge/Taylor and Francis. doi: 10.4324/9780203883372

Templin, J., and Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *J. Class.* 30, 251–275. doi: 10.1007/s00357-013-9129-4

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

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

Watson, J. (2014). "Curriculum expectations for teaching science and statistics," in *Proceedings of the Ninth International Conference on Teaching Statistics*, eds K. Makar, B. deSousa, and R. Gould (Voorburg: International Statistical Institute). Available online at: http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_1A1_WATSON.pdf?1405041553

World Medical Association (2013). *WMA Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*. Ferney-Voltaire: World Medical Association.

Wu, H. (2019). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. *Educ. Psychol.* 39, 1218–1232. doi: 10.1080/01443410.2018.1494819

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 Ren, Xu, Lin, Zhang and Yang. 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: 16 April 2021
doi: 10.3389/fpsyg.2021.608320

An Optimal Choice of Cognitive Diagnostic Model for Second Language Listening Comprehension Test

Yanyun Dong[1]*, *Xiaomei Ma*[1], *Chuang Wang*[2] and *Xuliang Gao*[3]

[1] School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China, [2] Faculty of Education, University of Macau, Taipa, China, [3] School of Psychology, Guizhou Normal University, Guiyang, China

Cognitive diagnostic models (CDMs) show great promise in language assessment for providing rich diagnostic information. The lack of a full understanding of second language (L2) listening subskills made model selection difficult. In search of optimal CDM(s) that could provide a better understanding of L2 listening subskills and facilitate accurate classification, this study carried a two-layer model selection. At the test level, A-CDM, LLM, and R-RUM had an acceptable and comparable model fit, suggesting mixed inter-attribute relationships of L2 listening subskills. At the item level, Mixed-CDMs were selected and confirmed the existence of mixed relationships. Mixed-CDMs had better model and person fit than G-DNIA. In addition to statistical approaches, the content analysis provided theoretical evidence to confirm and amend the item-level CDMs. It was found that semantic completeness pertaining to the attributes and item features may influence the attribute relationships. Inexplicable attribute conflicts could be a signal of suboptimal model choice. Sample size and the number of multi-attribute items should be taken into account in L2 listening cognitive diagnostic modeling studies. This study provides useful insights into the model selection and the underlying cognitive process for L2 listening tests.

Keywords: cognitive diagnostic model, L2 listening subskills, model selection, mixed-CDMs, inter-attribute relationship

OPEN ACCESS

Edited by:
Ping Chen,
Beijing Normal University, China

Reviewed by:
Lan Shuai,
Haskins Laboratories, United States
Zhehan Jiang,
University of Alabama, United States

***Correspondence:**
Yanyun Dong
maggiedong@mail.xjtu.edu.cn

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

Received: *20 September 2020*
Accepted: *15 March 2021*
Published: *16 April 2021*

Citation:
Dong Y, Ma X, Wang C and Gao X (2021) An Optimal Choice of Cognitive Diagnostic Model for Second Language Listening Comprehension Test. Front. Psychol. 12:608320. doi: 10.3389/fpsyg.2021.608320

INTRODUCTION

Cognitive diagnosis models (CDMs), also known as diagnostic classification models (Rupp et al., 2010), show great promise for producing rich diagnostic information about students' strengths and weaknesses on a set of finer-grained attributes (Rupp and Templin, 2008). Although a wide array of CDMs have been developed and widely used in language assessment, listening comprehension receives little attention compared with other language skills in the second language (L2) research.

Previous studies, though sparse, have shown the feasibility of applying CDMs to L2 listening comprehension tests and demonstrated the potential usefulness of cognitive diagnostic approaches (CDAs) to understanding various subskills (Buck and Tatsuoka, 1998; Lee and Sawaki, 2009a; Sawaki et al., 2009; Aryadoust, 2011; Meng, 2013; Yi, 2017). However, it is not very clear which CDM should be used for L2 listening comprehension tests. Some studies used non-compensatory models (e.g., Buck and Tatsuoka, 1998; Sawaki et al., 2009), whereas others concluded that

compensatory and non-compensatory models produced striking similar diagnostic results for listening comprehension tests (Lee and Sawaki, 2009a). Still, others used G-DINA, which allows compensatory and non-compensatory inter-attribute relationships within the same test (Meng, 2013).

Selecting the right CDM(s) for a given test is of critical importance because model selection affects the diagnostic classification of examinees (Lee and Sawaki, 2009a) and thus influences the accuracy of diagnostic feedback. Providing learners with accurate diagnostic feedback to guide their remedial learning is the ultimate goal of CDA (Lee and Sawaki, 2009b). Wrong CDM(s) for a given dataset will lead to the wrong classification of examinees and misleading feedback. In addition, CDMs can provide information on the underlying inter-attribute relationships (Yi, 2017), that is, whether a compensatory or non-compensatory inter-attribute relationship can produce correct answers. Wrong CDM(s) for given data will generate a wrong interpretation of inter-attribute relationships and skill mastery status. The sample size is also one concern. Reduced models require a smaller sample size to be estimated accurately, although the saturated model can provide better model-data fit at test level compared with other reduced models; appropriate reduced models can provide better classification rates than saturated models, particularly when the sample size is small (Rojas et al., 2012). Non-parametric models can accommodate small samples, but they could not commonly use model-data fit indices of parametric methods to compare with parametric models. This is because it will be difficult to tell whether parametric or non-parametric models are better for the given data without parameter estimation. Therefore, it is inconvenient to make a model selection with other commonly used CDMs (Kang et al., 2019). Non-parametric models need prior knowledge of the inter-attribute relationships (compensatory or non-compensatory) of a given skill to decide whether a compensatory or non-compensatory model should be selected (Chiu and Douglas, 2013). In sum, inappropriate CDMs lead to inaccurate diagnostic classification, wrong interpretation of skill mastery status, and misunderstanding of inter-skill relationships. Few empirical studies, however, have examined the model comparison and selection for L2 listening comprehension tests, and little is known about L2 listening inter-skill relationships.

THEORETICAL FRAMEWORK AND LITERATURE REVIEW

Listening Comprehension Skills

Listening comprehension is the least-researched skill among the four skills of reading, listening, writing, and speaking (Bae and Bachman, 1998; Field, 2013). Although some researchers assert that listening comprehension is a more integrated skill (Levine and Revers, 1988) and not empirically multi-divisible (Oller, 1983; Wagner, 2004), most researchers agree that listening comprehension involves multiple subskills (Rivers, 1966; Carroll, 1972; Clark and Clark, 1977; Bae and Bachman, 1998; Buck and Tatsuoka, 1998; Song, 2008; Rost, 2011; Field, 2013). Munby (1978) and Richard (1983) presented a complete taxonomy of subskills, but the inter-subskill relationships are hard to explain. In contrast, Aitken (1978) provided a succinct taxonomy from the communicative approach by recognizing major listening subskills. Weir (1993) is along a similar line of defining important listening subskills only. The idea of major listening subskills benefits listening comprehension test development and studies in terms of the construct validity of these tests, especially studies with CDA (Buck and Tatsuoka, 1998; Lee and Sawaki, 2009a; Sawaki et al., 2009; Yi, 2017).

In addition to the widely discussed issue of the divisibility of listening comprehension (e.g., Bae and Bachman, 1998), listening subskill relationships were manifested by factor analysis of test-takers' responses (Liao, 2007; Shin, 2008; Song, 2008). Goh and Aryadoust (2015), however, argued that the structure of interactive and interdependent listening subskills was much more complicated than what factor analysis could represent. This argument echoed the view of Buck (2001) that "various types of knowledge involved in understanding language are not applied in any fixed order—they can be used in any order, or even simultaneously, and they are all capable of interacting and influencing each other" (p. 3). It is worth noting that the assertion of Buck (2001) shows the interactive and varied subskill relationships of listening, which also implies that the relationships among listening subskills are not yet clear.

Model Selection for L2 Listening Comprehension Tests

Each CDM has unique assumptions about the latent attribute relationships (e.g., compensatory or non-compensatory). Under a compensatory CDM, successfully mastering only one or some of the required attributes may compensate for the non-mastery of others. In contrast, under a non-compensatory CDM, an item can be correctly answered only if all the required attributes have been mastered. If the assumption of a CDM does not match the latent attribute relationships of given data, the CDM is improper for the test and cannot offer accurate classification and diagnostic feedback. Test-level model selection is based on clear inter-attribute relationships. If the relationships are not clear, selecting the most appropriate CDM(s) will be a challenge.

In the literature, both compensatory and non-compensatory CDMs were applied to the L2 listening comprehension test. Buck and Tatsuoka (1998) applied the rule-space model to an L2 listening comprehension test. Aryadoust (2011) used the fusion model (FM) to a version of the International English Language Testing System listening comprehension test, and Sawaki et al. (2009) also used FM to the Test of English as a Foreign Language iBT listening comprehension items. The rule-space model and FM are both non-compensatory models. Meng (2013) applied G-DINA to an L2 listening comprehension test. G-DINA accommodates both compensatory and non-compensatory inter-attribute relationships. In addition, Lee and Sawaki (2009a) concluded that compensatory and non-compensatory models produced strikingly comparable diagnostic results. Yi (2017), however, argued that a compensatory model (C-RUM) was the best to interpret the listening subskill relationships. As both compensatory and non-compensatory CDMs were applied to L2

listening comprehension tests in the literature, CDM selection for L2 listening comprehension tests is still inconclusive and deserves further exploration.

Cognitive Diagnostic Models

G-DINA model (de la Torre, 2011) is a saturated model and considers all possible interaction effects among required subskills for an item. It classifies examinees into $2^{K_j^*}$ latent groups based on mastery of required skills for each item. K_j^* is the number of attributes required for item j. α_{lj}^* is the reduced attribute vector whose elements are the required attributes for item j. If one item needs two attributes, these two attributes lead to four latent groups: those who mastered both attributes, one of the attributes, or none of the attributes. Its item response function based on $P(a_{lj}^*)$ is as follows:

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}$$

δ_{j0} is the intercept for item j, representing the baseline probability of a correct response when none of the required subskills is present. δ_{jk} is the main effect of mastering a single-skill α_k, representing the change in the probability of a right answer (PRA) as a result of mastering a single skill. $\delta_{jkk'}$ is the first-order interaction effect due to mastering both α_k and $\alpha_{k'}$. $\delta_{j12\ldots K_j^*}$ is the highest order interaction effect due to mastering all the required subskills up to K_j^* (de la Torre, 2011). G-DINA is often used as the benchmark model when the true model is not known (Chen et al., 2013; Li et al., 2016).

Under the framework of G-DINA, there are some special cases called the reduced models: DINA, DINO, A-CDM, R-RUM, and LLM, which are used in this study and introduced as follows:

The *DINA* (deterministic inputs, noisy, "and" gate; Juncker and Sijtsma, 2001) model is a special case by setting all the parameters of G-DINA, except δ_{j0} and $\delta_{j12\ldots K_j^*}$, to zero. Thus, DINA is a non-compensatory model, assuming that examinees have to master all the required skills simultaneously to choose the correct answer.

If setting the values of all the main and interaction effect parameters of G-DINA to be the same or, in other words, the main and interaction effects are identical to each other, then *DINO* (deterministic input, noisy, "or" gate; Templin and Heson, 2006) is obtained. DINO is the compensatory counterpart of DINA, assuming that examinees can have the same PRA whether they master one required subskill or all.

A-CDM (additive CDM; de la Torre, 2011) can be obtained when all the interaction effects in G-DINA are set to zero while keeping the compensatory property. This model indicates that mastering one subskill increases the PRA on an item, and its contribution is independent of the other subskills.

R-RUM (Reduced Reparameterized Unified Model; Hartz, 2002) is a non-compensatory model with a log link, setting the interaction terms to zero. It is considered a non-compensatory counterpart of A-CDM (Hartz, 2002).

If using a logit link, setting the interaction terms to zero and keeping the compensatory property, *LLM* (linear logistic model; Hagenaars, 1990, 1993; Maris, 1999) is obtained. Similar to A-CDM, it also assumes that the mastery of one subskill will increase the PRA to the item.

There is another group of models, named non-parametric models, which do not require a sample size. Because there is no parameter estimation in non-parametric models, it is not possible to make model comparisons with parametric models based on common fit indices (Kang et al., 2019). Prior knowledge of attribute relationships is required for non-parametric model selection, but unknown attribute relationships in this study make it more difficult to compare with parametric CDMs. As a result, non-parametric models were not be used in this study.

The following research questions guided this study.

1. Which model is the best for the second language listening comprehension test at the test level when the sample size is small?

2. Which model is the best for the second language listening comprehension test at the item level when the sample size is small?

3. What are the inter-attribute/subskill relationships of second language listening comprehension?

METHODS

Participants

Participants were 500 freshmen (149 females and 351 males) conveniently sampled from four universities in the northwest region of China. They all majored in science and technology and aged between 17 and 20 years old. This sample size is of practical importance, although it was considered small in previous simulation studies of CDMs (de la Torre and Lee, 2013; Ma et al., 2016).

Instrument

L2 Listening Diagnostic Assessment (L2LDA)

L2LDA is part of the *English as Foreign Language Listening Diagnostic Test* in the *PELDiaG system (Personalized English Learning Diagnosis and Guidance system)* designed for the diagnostic purpose (Meng, 2013; Ma and Meng, 2014; Du and Ma, 2018). The original test in the *PELDiaG system* has two types of items: multiple-choice items and sentence-dictation items. Only the 19 multiple-choice items, which are dichotomously scored, were used in this study. Sentence-dictation items were not dichotomous and scored holistically within a score range from 0 to 3.5 points. For the reason of convenience, these items were excluded in this study. *L2LDA* has four sections of short conversations, a long conversation, short passages, and a video clip. The topics of it cover campus life, social life, and common scientific knowledge, which largely reduced the possibility of bias caused by topic preference. The participants' total scores of the *L2LDA* followed a normal distribution with a mean score of 11.50 (out of a total score of 19) and a standard deviation of 4.05.

Q-Matrix

A Q-matrix, a critical input of CDMs, specifies the relationship between attributes and test items. Because sentence-dictations (tapping into the attribute of Short-term Memory and Note Taking) were excluded from this study, the attributes of Short-term Memory and Note Taking were accordingly canceled from the original Q-matrix (Dong et al., 2020), and then, six attributes were retained in the Q-matrix (**Table 1**) for this study.

The six subskills/attributes for this study in relation to the existing listening skill taxonomies are presented in **Table 2**. Their definitions in accord with the ones identified by Meng (2013) are the following:

A1: Sound Discrimination: Recognizing special phonological and prosodic information, such as liaison and assimilation, stress and weak forms, intonation.
A2: Less Frequent Vocabulary and Expressions: Understanding less frequent words, oral expressions, and slangs.
A3: Difficult Structures: Difficult sentence structure and grammatical functions such as subjunctive mood, inversion, and negation.
A4: Facts and Details: Understanding detailed expressions of time, places, and relationships.
A5: Main Idea: Recognizing and summarizing main ideas and major points.
A6: Situational Context and Cultural Background Inferences: Obtaining motivations, purposes, reasons, and interactive functions by inferring from the context, implied expressions, and cultural background.

Data Analytical Procedure

Three major procedures were followed: (a) Model selection; (b) Empirical comparisons between G-DINA, the most comparable CDM with G-DINA at the test level and the selected CDM(s) at the item level in terms of psychometric characteristics; and (c) Content analysis that is required to confirm or amend the selected item-level models.

CDM Selection

The R GDINA package (Ma and de la Torre, 2016) was used for model estimation and selection. G-DINA was used as the baseline model and was compared with the other reduced models: DINA, DINO, R-RUM, A-CDM, and LLM. The absolute model fit and the relative model fit were used to compare the models. The absolute fit indices were calculated based on the residuals between the observed and predicted Fisher-transformed correlations of item pairs [Max.z(r)] and between the observed and predicted log-odds ratios of item pairs [Max.z(l)]. The least critical p-value was 1% (Chen et al., 2013). The second absolute fit index is M_2 (Maydeu-Olivares and Joe, 2006), which is a limited-information fit statistic, and 0.05 is the critical p-value. The third is the root mean squared error approximation (RMSEA), which reflects the discrepancy between the predicted and the observed tetrachoric correlation for all pairs of items. RMSEA value of 0.05 was used to assess model fit (Henson and Templin, 2007). The fourth is the standardized root mean square residual (SRMSR), which is the square root of the sum of the squared differences of the observed correlation and the model implied correlation of all item pairs. SRMSR below 0.05 indicates a good absolute fit (Maydeu-Olivares, 2013). All the above are the absolute fit indices provided by the G-DINA package, and they serve as initial screening tools (Yi, 2017). Subsequently, the relative fit indices play a more critical role in narrowing down the scope of CDMs.−2Log-likelihood(-2LL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values are relative fit indices. Because−2LL always selects the saturated model and BIC imposes the biggest penalty (Lei and Li, 2014), AIC was used to compare G-DINA with the other models in this study.

In addition to holistically selecting models at the test level, CDMs were also selected at the item level. Language tests commonly include two kinds of items in terms of how many attributes are measured in each item: single-attribute item and multi-attribute item. The Wald test was used to select the most appropriate reduced CDMs for multi-attribute items (de la Torre and Lee, 2013; Ma et al., 2016). For single-attribute items, no distinction can be made between general and reduced CDMs, and then G-DINA was used. In *L2LDA*, Items 3, 4, 10, 11, 13, 16, and 19 are two-attribute items, but all other items were single-attribute. The reduced CDM whose p-value of Wald statistics was > 0.05 was accepted. When more than one reduced model was acceptable, the model with the largest p-value stayed. If DINA or DINO was one of the retained models, the DINA or DINO models were preferred over the other three models because of their simplicity. If all reduced models were rejected for an item, G-DINA was chosen. Thus, Mixed-CDMs were formed for this test.

CDM Comparison

Accordingly, after the selection of Mixed-CDMs, psychometric properties (i.e., absolute fit, absolute item fit, relative fit, person fit, and attributes classification reliability) under G-DINA as a baseline, the most comparable CDM with G-DINA at the test level, Mixed-CDMs, and G-DINA were compared. The absolute fit and relative fit used the same indices as in the test-level model selection. The $l_z$index was used for person fit. The $l_z$index is standardized, so a value of 0.0 reflects an ideally perfect typical response string (Drasgow et al., 1985). $l_z > 2.0$ indicates over-fitness, whereas $l_z$below -2.0 indicates poor fit. Test-level and attribute-level classification accuracy indices estimated from the GDINA function followed the approaches in Iaconangelo (2017) and Wang et al. (2015).

Then item parameters under G-DINA and the selected models were examined. Item parameter statistics can inform the inter-attribute relationships, which give evidence to inference interpretability analysis. Next, content analysis of the two-attribute items was carried out to justify or modify the results of model selection. Inter-attribute relationships can be informed by item parameter estimations, which are different under different CDMs. Content analysis was to examine whether the inter-attribute relationships manifested by the selected model were reasonable.

TABLE 1 | Configuration of Q-matrix.

	Attributes							Attributes					
	A1	**A2**	**A3**	**A4**	**A5**	**A6**		**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
Item 1	1	0	0	0	0	0	Item 11	0	1	0	0	0	1
Item 2	1	0	0	0	0	0	Item 12	0	0	0	0	0	1
Item 3	0	1	0	1	0	0	Item 13	0	0	0	1	1	0
Item 4	0	1	0	1	0	0	Item 14	0	0	0	0	0	1
Item 5	0	1	0	0	0	0	Item 15	0	0	0	0	0	1
Item 6	0	0	1	0	0	0	Item 16	0	0	1	0	1	0
Item 7	0	0	0	1	0	0	Item 17	0	0	0	0	0	1
Item 8	0	0	1	0	0	0	Item 18	0	0	0	0	1	0
Item 9	0	0	1	0	0	0	Item 19	0	0	0	1	1	0
Item 10	0	0	0	1	1	0							

TABLE 2 | Listening attributes/subskills and relationship with existing listening skill taxonomies.

Listening attributes/subskills and definitions	**Listening Cognitive Ability of China Standards of English (He and Chen, 2017)**	**Field's (2009)**	**Aitken's (1978)**	**Sawaki et al.'s (2009)**
A1: Sound discrimination	Identify/retrieve	Input decoding	Understand prosodic patterns	
A2: Less frequent vocabulary and expressions	Identify/retrieve	Lexical search	Understand vocabulary	Understand vocabulary
A3: Difficult structures	Identify/retrieve/analyze	Parsing	Understand syntactic patterns	
A4: Facts and details	Identify/retrieve/analyze	Meaning construction		Understand important information
A5: Main idea	Analyze/summarize/create	Meaning construction; Discourse construction		Understand overall topic/gist;
A6: Situational context and cultural background inference	Analyze/summarize/ create/evaluate	Meaning construction; Discourse construction	Identify speaker's purpose, attitudes, views, and intentions; Making inferences;	Making inferences
			Identify rhetorical devices.	Understand the structure (rhetorical, discourse).

The attributes and definitions stem from the study of Meng (2013, p. 78, p. 95).

RESULTS

Model Selection at Test Level

Table 3 summarizes the model fit results and the numbers of parameters of the six models. G-DINA, A-CDM, LLM, and R-RUM could be accepted with the significant levels of both Max.z(r) and Max.z(l) being much higher than 1% (the least critical p-value), whereas DINA and DINO could not be accepted with the significant levels of Max.z(r) being lower than 1%, and the p-values of Max.z(l) for them (0.0097, 0.01) were not good enough either. DINA was also rejected by M_2 with a p-value below 0.05. DINO was narrowly accepted by M_2 (p-value is 0.0799, close to 0.05). RMSEA and SRMSR gave favorable acceptance to all the CDMs, but it also indicated that the two indices might not be very sensitive about the model selection. G-DINA is a saturated model that can accommodate both compensatory and non-compensatory inter-attribute relationships, whereas A-CDM and LLM are compensatory models, and R-RUM is a non-compensatory one. A-CDM and R-RUM had the same absolute model fit at RMSEA (0 for both). These all indicated that both compensatory and non-compensatory models fit the data and the attributes of L2 listening, therefore, manifested both (non-)compensatory relationships. As DINA and DINO were rejected by Max.z(r) and Max.z(l), the inter-attribute relationships could not be interpreted simply as what either of the two models could accommodate.

Comparing among the three accepted reduced models (A-CDM, LLM, and R-RUM), LLM performed the best with the smallest−2LL (11268.92), AIC (11484.93), and BIC (11940.11). LLM had the smaller AIC (11484.93) and BIC (11940.11) than G-DINA (11489.23 and 11973.91, respectively). It does not mean that LLM is better than G-DINA, just that G-DINA invites larger penalties than LLM because AIC and BIC both introduce a

TABLE 3 | Absolute fit.

CDMs	#par	Max.z(*r*)	*p*	Max.z(*l*)	*p*	M_2	*p*	RMSEA	SRMSR	-2LL	AIC	BIC
G-DINA	115	3.12	0.3046	2.89	0.6521	89.0495	0.128	0.0194	0.0431	11,259.24	11,489.23	11,973.91
DINA	101	4.68	0.0005	4.03	0.0097	118.336	0.0205	0.0257	0.0452	11,308.10	11,510.09	11,935.77
DINO	101	4.62	0.0006	4.02	0.0100	108.348	0.0799	0.0209	0.0209	11,300.98	11,502.98	11,928.65
A-CDM	108	3.34	0.1437	3.10	0.3272	75.9016	0.6686	0	0.0437	11,273.82	11,489.81	11,944.99
LLM	108	3.15	0.2831	2.96	0.5182	93.2377	0.1862	0.0166	0.0433	11,268.92	11,484.93	11,940.11
R-RUM	108	3.45	0.0947	3.21	0.2237	81.1406	0.5061	0	0.0441	11,278.34	11,494.35	11,949.52

(a) #par., number of parameters; (b) Max.z(r) & Max.z(l), maximum z-score for transformed correlation and log odds ratio; (c) M_2, a limited-information fit statistic for dichotomous response; (d) RMSEA, root mean square error of approximation; (e) SRMSR, standardized root mean square residual.

TABLE 4 | Wald statistics for multi-attribute items.

	Selected CDM	Wald statistics	*p*-values
Item 3	DINO	5.51	0.06
Item 4	DINA	3.57	0.17
Item 10	LLM	0.00	0.99
Item 11	DINO	1.16	0.56
Item 13	DINA	2.69	0.26
Item 16	DINA	0.17	0.92
Item 19	DINO	2.73	0.26

penalty for model complexity. In this sense, LLM is the most comparable model with G-DINA at the test level.

Model Selection at Item Level

As shown in **Table 4**, three models (DINO, DINA, and LLM) were selected; four items manifested compensatory inter-attribute relationships under DINO and LLM, whereas the other three items illustrated non-compensatory relationships under DINA. These results showed that the inter-attribute relationships tapped into by *L2LDA* were compensatory in some multi-attribute items and non-compensatory in others. Thus, the three models plus G-DINA formed Mixed-CDMs.

Table 5 shows that Mixed-CDMs had a high level of absolute fit, the significant levels of Max.z(r) and Max.z(l) were much higher than 1%, the p-value of M_2 (0.3609) was much larger than 0.05, and RMSEA (0.0097) and SRMSR (0.0435) were below 0.05.

Comparisons Among G-DINA, LLM, and Mixed-CDMs

According to the results of the previous section, LLM was the most comparable model to G-DINA at the test level. Then Zoom-in comparisons were made among G-DINA, LLM, and Mixed-CDMs, and relative fit, absolute item fit, person fit, and the classification accuracy were concerned psychometric characteristics. As for the absolute fit at the test level, the three models all met the higher critical requirement ($p >$ 0.10), but Mixed-CDMs were better than the other two with an increase from 0.3046 for G-DINA and 0.2831 for LLM to 0.9120 on Max.z(r) statistics and from 0.6521 and 0.5182 to 1 on Max.z(l) statistics (see **Tables 3**, **5**). Moreover, the RMSEA of Mixed-CDMs was the smallest among the three. As for AIC, Mixed-CDMs performed the best on the test relative fit with the smallest value (11478.36), compared with G-DINA (11,489.23) and LLM (11,484.93). As shown in **Table 6**, the absolute item-level fit statistics for two items (Items 3 and 9) got improvement under Mixed-CDMs, whereas no statistically significant differences were noticed for other items. Mixed-CDMs improved the significant levels of Max.z(r) and Max.z(l) statistics from 3% for G-DINA and LLM to 10% and from 7% for G-DINA and 6% for LLM to 18%, respectively. Thus, Mixed-CDMs had better absolute item-level fit than G-DINA and LLM, especially on Items 3 and 9.

Regarding person fit (**Table 7**), only one examinee (ID331) was over-fit ($l_z > 2.0$) under the three models. l_z for ID331 under Mixed-CDMs was the smallest, and the mean absolute value of $l_z(|l_z|)$ was also the smallest under Mixed-CDMs. Therefore, Mixed-CDMs were slightly better than G-DINA and LLM on person fit. In addition, the classification accuracy at the test level and attribute level were very close under Mixed-CDMs and G-DINA, which means that Mixed-CDMs were as reliable as G-DINA on the classification for the data.

Item parameters of multi-attribute items are presented in **Table 8**. The standard errors were reduced under Mixed-CDMs in comparison with G-DINA and LLM. The estimates of item parameters under Mixed-CDMs were, therefore, statistically more accurate than those under G-DINA and LLM. G-DINA had the highest SE among the three CDMs.

The item parameters also displayed the inter-attribute relationships *via* different PRAs (**Table 8**). If the PRA of an attribute was higher than 0.50, this attribute could compensate the other attribute with a more than 50% probability of giving the right answer. If the PRA of an attribute was below 0.50, this attribute could not compensate the other to give the right answer. Some common features were found under the three models. For Items 3 and 19, one attribute could compensate for the other under the three models. For Item 10, A4 could compensate A5 but not *vice versa*. So, the inter-attribute relationships exhibited compensatory traits in Items 3, 10, and 19 under the three models. In addition, compared with LLM, G-DINA shared more common features with Mixed-CDMs in terms of inter-attribute relationships. For Items 11, A2, and A6 could mutually compensate with each other under G-DINA and Mixed-CDMs, whereas A6 could compensate A2, not *vice versa*, under LLM. For

TABLE 5 | Model fit of mixed-CDMs.

#par	Max.z(*r*)	*p*	Max.z(*l*)	*p*	M_2	*p*	RMSEA	SRMSR	-2LL	AIC	BIC
102	2.78	0.9120	2.72	1.0000	92.1224	0.3609	0.0097	0.0435	11,274.36	11,478.36	11,908.25

TABLE 6 | Absolute item-level fit.

	Item 3				Item 9			
	Max.z(*r*)	*p*	Max.z(*l*)	*p*	Max.z(*r*)	*p*	Max.z(*l*)	*p*
G-DINA	3.12	0.03	2.89	0.07	3.12	0.03	2.89	0.07
LLM	3.14	0.03	2.92	0.06	3.14	0.03	2.92	0.06
Mixed-CDMs	2.79	0.10	2.58	0.18	2.79	0.10	2.58	0.18

Items 13 and 16, under both G-DINA and Mixed-CDMs, PRA for each attribute was below 0.50, which means that attributes could not compensate mutually for a right answer. Under LLM, however, compensatory traits were exhibited.

In addition to the inter-attribute relationships mentioned earlier, conflicting relationships were also detected in **Table 8**. For instance, under G-DINA, mastering both required subskills lowered the PRA (0.66) for Item 10 compared with mastering A4 only (PRA = 1). This sort of attribute conflict (mastering both attributes has lower PRA than mastering one attribute only) also existed in Item 11 under G-DINA and in Items 4 and 13 under LLM. However, this conflict was not exhibited under Mixed-CDMs.

Content Analysis

The outcome of selecting models at the item level should be examined, whether theoretically valid or not, through content analysis. It deserved more attention when DINA and DINO were rejected at the test level but accepted at the item level. For better illustration, a special pair of attributes A2-A4 were focused because they were measured in two items: DINO was selected for one item, and DINA was for the other. Looking closely into the content of these items (**Table 9**), it was found that A2 (Vocabulary and Expressions) could compensate for the lack of A4 (Facts and Details) in Item 3, but A4 could not provide an equal probability of right answer if there were a lack of A2. Whether A4 could give a 50% probability of the right answer was partly because it was a multiple-choice item, and only two choices were logically pertinent. Therefore, DINO that allowed attribute to compensate for the lack of the other was not appropriate for this item, whereas G-DINA, which provided the real picture of the compensability: A2 could compensate for the lack of A4, but A4 could not compensate equally, should be retained. Item 4, in contrast with Item 3, required mastery of both A2 and A4. The expression "run out of" was one attribute (A2) measured in this item, but it did not provide a complete meaning unless it was combined with the detailed information "milk" (A4), which means when a test taker mastered both attributes, he/she could find the right answer. Thus, DINA was justified to be chosen for this item. In addition, mastery of the expression "run out of" (A2) offered lower PRA than mastery of none under G-DINA. This is unreasonable and uninterpretable, so G-DINA is not acceptable for this item. Therefore, item-level model selection still needs content analysis to detect flaws. The same pair of attributes may be able to exhibit different relationships: compensatory or non-compensatory, although DINO seemed to be improper for Item 3.

Following the same way, the rest two-attribute items were analyzed one by one. Interpretability is an important concern: if the inter-attribute relationship under a model is uninterpretable, then this model will be inferior, and the model that gives reasonable relationships will be accepted. In this way, it was found that the inter-attribute relationships under Mixed-CDMs are more reasonable and interpretable without conflicts, and then Mixed- CDMs were accepted for the rest two-attribute items.

DISCUSSIONS

The current study examined the selection of CDMs for an L2 listening comprehension test. Two-layer model selections mutually justified the mixed inter-attribute relationships of L2 listening subskills. The significance of this study is that statistically fit models at the item level require theoretical evidence informed by content analysis. The procedure used in this study can also serve as guidance for other studies aiming at choosing optimal CDM(s).

At the test level, A-CDM, LLM, and R-RUM were accepted and had a comparable relative fit with that of G-DINA. Based on the features of these CDMs, compensatory and non-compensatory inter-attribute relationships coexisted in the test, rather than a monotonous compensatory or non-compensatory one, which is different from previous findings (e.g., Yi, 2017). However, LLM and A-CDM have a smaller relative fit than R-RUM, which seemingly indicates that the inter-attribute relationships exhibit more compensatory than non-compensatory. However, how exactly the attributes interact with each other cannot be informed so far at the test level and would be overgeneralized under the assumed framework of a single reduced CDM if it is imposed on all items. Hence, multiple CDM selection at the item level within the same assessment is tenable and warranted.

TABLE 7 | Psychometric characteristics under both models.

	Mean of $\lvert l_z \rvert$ for person fit	l_z Index for ID331	Classification accuracy							
			Test-level	A1	A2	A3	A4	A5	A6	Means
G-DINA	0.76	2.13	0.7176	0.8731	0.9119	0.9107	0.9253	0.9140	0.9152	0.9084
LLM	0.74	2.09	0.7954	0.8790	0.9931	0.9168	0.9274	0.9956	0.9197	0.9386
Mixed- CDMs	0.73	2.05	0.7157	0.8713	0.9158	0.9096	0.9083	0.9093	0.9138	0.9047

TABLE 8 | Item parameters estimates (EST) and standard errors (SE) of multi-attribute items.

Items	Attributes		G-DINA				Mixed-CDMs				LLM			
			P(00)	P(10)	P(01)	P(11)	P(00)	P(10)	P(01)	P(11)	P(00)	P(10)	P(01)	P(11)
Item 3	A2 + A4	EST	0.48	0.82	0.57	0.89	0.47	0.88	0.88	0.88	0.45	0.61	0.82	0.89
		SE	0.05	0.13	0.14	0.02	0.04	0.02	0.02	0.02	0.04	0.06	0.05	0.02
Item 4	A2 + A4	EST	0.34	0.17	0.57	0.71	0.35	0.35	0.35	0.74	0.33	0.31	0.71	0.69
		SE	0.04	0.15	0.14	0.03	0.03	0.03	0.03	0.03	0.04	0.05	0.05	0.03
Item 10	A4 + A5	EST	0.22	1.00	0.37	0.66	0.29	0.63	0.36	0.71	0.32	0.66	0.32	0.67
		SE	0.05	0.23	0.07	0.03	0.05	0.09	0.06	0.03	0.05	0.07	0.05	0.03
Item 11	A2 + A6	EST	0.23	0.79	1.00	0.89	0.26	0.90	0.90	0.90	0.31	0.38	0.89	0.92
		SE	0.04	0.10	0.26	0.03	0.04	0.02	0.02	0.02	0.04	0.06	0.04	0.02
Item 13	A4 + A5	EST	0.43	0.44	0.26	0.87	0.42	0.42	0.42	0.89	0.40	0.87	0.35	0.85
		SE	0.06	0.17	0.08	0.03	0.03	0.03	0.03	0.03	0.05	0.04	0.05	0.03
Item 16	A3 + A5	EST	0.34	0.29	0.35	0.82	0.34	0.34	0.34	0.84	0.28	0.73	0.37	0.80
		SE	0.06	0.13	0.06	0.03	0.03	0.03	0.03	0.03	0.04	0.06	0.04	0.03
Item 19	A4 + A5	EST	0.59	0.74	0.95	0.97	0.58	0.97	0.97	0.97	0.61	0.88	0.86	0.97
		SE	0.06	0.14	0.05	0.01	0.05	0.01	0.01	0.01	0.05	0.05	0.03	0.01

EST, estimates; SE, standard error; P(11) refers to the probability of the right answer (PRA) to the item when two attributes are mastered; P(10) stands for the PRA when the first attribute is mastered; P(01) is the PRA when the second attribute is mastered.

At the item level, based on the selection criteria, DINA, DINO, and LLM were selected by Wald test for the seven multi-attribute items, G-DINA remained for 12 single-attribute items, and Mixed-CDMs were hence formed for the whole assessment by auto GDINA function. When the absolute fit, the relative fit, and the person fit were considered, Mixed-CDMs performed better over G-DINA and LLM, the most comparable with G-DINA. It is easy to understand that LLM is not optimal because it fails to postulate "both-and" relationships in its framework, even it is the most comparable with G-DINA. However, it is against the intuition that G-DINA is not superior to Mixed-CDMs for this dataset, although saturated G-DINA accommodates all the interactions among subskills and should have fit the data better. Sample size likely contributed to the result because saturated G-DINA needs a large sample size, and 500 was considered a small sample in previous studies (e.g., Chen et al., 2013; Ma et al., 2016). The small sample ($n = 500$) of this study might have constrained G-DINA's performance, which echoes the opinion that saturated models are not always the best choice when the sample size is small (Rojas et al., 2012; Ma et al., 2016). Based on the estimation of item parameters, standard errors of the estimates were the smallest under Mixed-CDMs, which suggests that the item parameter estimation under Mixed-CDMs was more accurate. These results largely agree with the claim of DiBello et al. (2007) that specific (or reduced) CDMs could reduce the standard errors when the sample size and the total number of items are small. It also renders empirical evidence to the simulation study of Ma et al. (2016) that "reduced CDMs usually require smaller sample sizes for accurate parameter estimation" (p. 201).

Although Mixed-CDMs were reported to be able to perform well-statistically, slightly better than G-DINA in some aspects, theoretical supports informed by the content analysis were needed. In this study, it was found that DINO selected by auto-GDINA function was not appropriate for Item 3 because the inter-attribute relationship manifested under the model could not reflect the underlying cognitive process of that item. G-DINA, however, could reflect this process and was thus retained for that item. Inexplicable conflicting inter-attribute relationships were found for some items under LLM and G-DINA (**Table 8**). No literature could interpret the conflicts. One plausible reason could be that LLM and G-DINA were not optimal models for the data and could not exhibit the real inter-attribute relationship. Content analysis of Item 4 showed that A2 (Vocabulary and expressions) and A4 (Facts and details) were non-compensatory with each other. DINA was accepted for this item, so inexplicable conflicting inter-attribute relationships occur under LLM and

TABLE 9 | Items tapping into attributes A2 and A4.

Item	Attributes utilized
3. M: Nancy, why are you late today? W: I overslept and missed the bus. Q: Why is Nancy late? a. The bus was late. b. Her clock was slow. c. She got up late. d. She forgot her class.	The item requires two attributes, A2 (Vocabulary and Expressions: overslept) and A4 (Facts and Details: missed the bus). A test taker would find answer key (c) if he/she only knows A2 (overslept), as it means the same as the answer key. However, if he/she only knows A4 (missed the bus), he/she would find (b) or (c), which means he/she has around 50% probability to find the right answer (c).
4. M: Where is Cindy? W: She ran out of milk and went to get some. Q: Why did Cindy go out? a. She went out jogging. b. She had no more milk. c. She went out for a walk. d. She was delivering milk.	A test taker would find answer key (b) only if he/she knows both the attributes, A2 (Vocabulary and Expressions: ran out of…) and A4 (Facts and Details: milk).

G-DINA. These sorts of conflicts under G-DINA were also mentioned in other studies (Meng, 2013; Chen and Chen, 2016; Ravand, 2016), and no linguistic explanation was provided because the interpretation of the conflicts was difficult based on the available language acquisition theories (Ravand, 2016), and they were considered as inherent inter-attribute relationships (Chen and Chen, 2016). Given that G-DINA and LLM provided inexplicable conflicts between subskills for some items, amended Mixed-CDMs are preferred because they better capture the inherent inter-attribute relationships of *L2LDA* and better reflect the processing of L2 listening without attribute conflicts.

After the amendment of the Mixed-CDMs by content analysis, four items selected compensatory models, whereas three items chose non-compensatory models. This provides evidence that compensatory and non-compensatory models coexist in the literature of listening tests (Buck and Tatsuoka, 1998; Sawaki et al., 2009; Yi, 2017). The items that chose compensatory models are slightly more than those that chose non-compensatory models. This could explain why, at the test level, compensatory LLM and A-CDM gained slight preference compared with non-compensatory R-RUM. This could also imply that test-level model acceptance could roughly predict inter-attribute relationships, and the accepted model having a better relative fit index could predict the dominant inter-attribute relationship. The small differences can also be accounted for by the small number of multi-attribute items (only seven items in this study), which has also been detected in the study of Lee and Sawaki (2009a) and the study of Yi (2017).

It is worth noting that two pairs of attributes were measured repetitively: A2–A4 and A4–A5. A2 and A4 were either compensatory or non-compensatory in two items of short conversations. In Item 4, only when "run out of" (A2) was combined with "milk" (A4), a complete semantic meaning could be understood. Therefore, A2–A4 was non-compensatory for this item. However, A2–A4 was compensatory in Item 3 because "overslept" (A2) and "missed the bus" (A4) each provided complete semantic meaning for understanding. As we can see, the semantic completeness of the target attributes may influence the attribute relationships for short conversations. The A4–A5 pair also showed flexible relationships, compensatory or non-compensatory. They were measured in a short conversation, a long conversation, and a video clip, respectively. In the short conversation, A4 (Facts and Details) was more important than A5 (Main Idea) and compensated A5. In the long conversation, A4 and A5 were not compensatory and had to work together to give the right answer. It seems that understanding facts and details (A4) is more important in processing less information (a short conversation) than understanding the main idea (A5), whereas understanding facts and details has to cooperate with understanding the main idea in processing more information (a long conversation). In the video item, A4 and A5 could compensate each other. It was not clear whether visual aid interacted with the two attributes in this item. The inconsistent inter-attribute relationships of L2 listening subskills are also congruent with the claim of Buck (2001) about varied patterns of relations and interactions of listening subskills. It depicted more detailed and varied relationships, which is in line with the finding of Yi (2017) on the aspect of rendering indirect evidence against the ability to define a hierarchy of contribution among listening subskills. The relationships vary from one item to another, and different features of items are likely to interact with subskills and influence their relationships. This interaction was referred to as "item–level interaction," i.e., the same set of attributes may or may not exhibit interaction depending on the items that measure them (de la Torre et al., 2018). These findings imply that more consideration must be taken in test construction and validation. Compensatory relationships indicate that a correct answer could not guarantee every attribute in an item is actually used by the test taker.

Based on the discussion earlier, this study is significant in the following aspects:

First, it is found that test-level model acceptance by absolute fit indices can roughly predict inter-attribute relationships. This can also suggest whether item-level model selection should be needed or not. The model with a better relative fit index and an acceptable absolute fit index can predict the dominant inter-attribute relationship.

Second, the content analysis showed that the inexplicable attribute conflicts could be a signal of suboptimal model choice, and thus, item-level models are justified for better interpretations. The conflicts were also found in other studies (Meng, 2013; Chen and Chen, 2016; Ravand, 2016) but were not discussed and considered inherent inter-attribute relationships (Chen and Chen, 2016). Along this line, this study makes a step forward.

Third, the Mixed-CDMs are comparable with G-DINA for the L2 listening comprehension test and even better in some aspects. The amended Mixed-CDMs are optimal for *L2LDA*. Previous studies that involved the comparison between

Mixed-CDMs and G-DINA were mainly based on simulations (Ma et al., 2016). This study provides a piece of useful empirical evidence for this topic and renders evidence to the study of L2 listening inter-subskill relationships.

Fourth, it is found that both compensatory and non-compensatory inter-subskill relationships exist in *L2LDA*, and even the relationships between the same pair of attributes are also non-fixed at different items. This is a new finding because the previous research reported only compensatory L2 listening inter-subskill relationships based on model selection (Yi, 2017). Semantic completeness of the attributes and item features are likely to interact with the subskills and influence the relationships. This was also rarely reported in previous studies, even less in L2 listening CDA research.

Last but not least, this study is significant in the procedure of model selection. Model selection is often seen in other assessment studies (Li et al., 2016; Ravand, 2016; Yi, 2017), but few on L2 listening assessment. In this study, the item-level model selection was initiated or ignited by the results of test-level model comparison and then was justified and amended by content analysis. The logic of the procedure provides useful insight into CDM studies.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

Although the findings of this research help provide a process of selecting the right CDM for L2 listening diagnostic assessment and shed some light on the inter-attribute relationships of L2 listening, part of the findings are possibly limited to the particular dataset and test used in this study.

L2LDA consisted of only 19 items. Although individual attributes were measured 4.15 times on average, one attribute was measured twice. According to Rupp et al. (2010), each attribute should be measured at least three times for accurate measurement with CDA.

Restricted by the test, only a few attribute pairs (i.e., A2–A4, A4–A5, A3–A5, and A2–A6) were measured. It may be inaccurate to judge inter-attribute relationships when the number of multi-attribute items is small. A3–A5 and A2–A6 were measured only once; this study did not opine what inter-attribute relationships they are. Thus, further research is needed to examine more pairs of attributes and more multi-attribute items.

Only one sample size ($n = 500$) was used in this study. Other small sample sizes (e.g., $n = 300$ or $n = 600$) should also be examined in future studies so that a feasible threshold of a small sample can be found to decide whether G-DINA or Mixed-CDMs should be used.

This study only focused on dichotomous items, so future studies are recommended to consider changing polychotomous items into binary coding or using some CDMs that accommodate polychotomous scales.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by School of Foreign Studies of Xi'an Jiaotong University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YD: design of the study, data collection, data analysis, writing original draft, and revision. XM: supervision and writing-review. CW: supervision, writing-review, and editing. XG: data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Social Science Fund of China (Fund No. 17BYY015).

ACKNOWLEDGMENTS

We would like to thank all the team members of PELDiaG for the great efforts they have made on constructing the diagnostic listening test prior to this study.

REFERENCES

Aitken, K. G. (1978). "Measuring listening comprehension in english as a second language," *TEAL Occasional Papers, Volume 2* (Vancouver, BC: British Columbia Association of Teachers of English as an Additional Language).

Aryadoust, V. (2011). Application of the fusion model to while-listening performance tests. *Shiken* 15, 2–9. Available online at: https://www.researchgate.net/publication/265249148

Bae, J., and Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Lang. Test.* 15, 380–414. doi: 10.1177/026553229801500304

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511732959

Buck, G., and Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Lang. Test.* 15, 119–157. doi: 10.1177/026553229801500201

Carroll, J. B. (1972). "Defining language comprehension," in *Language Comprehension and the Acquisition of Knowledge*, eds R. O. Freedle and J. B. Carroll (New York, NY: John Wiley and Sons), 1–29.

Chen, H., and Chen, J. (2016). Exploring reading comprehension skill relationships through the G-DINA model. *Educ. Psychol.* 36, 1049–1064. doi: 10.1080/01443410.2015.1076764

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Measure.* 50, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

Chiu, C.-Y., and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classifi.* 30, 225–250. doi: 10.1007/s00357-013-9132-9

Clark, H. H., and Clark, E. V. (1977). *Psychology and Language: An Introduction to Psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.

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 Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *J. Educ. Measure.* 50, 355–373. doi: 10.1111/jedm.12022

de la Torre, J., van der Ark, L. A., and Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measure. Evaluat. Counsel. Dev.* 51, 281–296. doi: 10.1080/07481756.2017.1327286

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. Volume 26: Psychometrics,* eds C. R. Rao and S. Sinharay (Amsterdam: Elsevier), 979–1030.

Dong, Y., Ma, X., and Meng, Y. (2020). Optimizing the Q-Matrix for a cognitive diagnostic EFL listening test. *Mod. Foreign Lang.* 43, 389–401. Available online at: https://mall.cnki.net/magazine/Article/XDWY202003009.htm

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *Br. J. Mathemati. Statist. Psychol.* 38, 67–86. doi: 10.1111/j.2044-8317.1985.tb00817.x

Du, W., and Ma, X. (2018). CDA-based diagnostic model construction for EFL reading tests. *Foreign Lang. Teach. Res.* 50, 74–88. Available online at: https://mall.cnki.net/magazine/article/WJYY201801007.htm

Field, J. (2009). *Listening in the Language Classroom*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511575945

Field, J. (2013). "Cognitive validity," in *Examining Listening,* eds A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 77–151.

Goh, C. C. M., and Aryadoust, V. (2015). Examining the notion of listening subskill divisibility and its implications for second language listening. *Int. J. List.* 29, 109–133. doi: 10.1080/10904018.2014.936119

Hagenaars, J. A. (1990). *Categorical Longitudinal Data: Loglinear Panel, Trend, and Cohort Analysis*. Thousand Oaks, CA: Sage.

Hagenaars, J. A. (1993). *Loglinear Models With Latent Variables*. Thousand Oaks, CA: Sage. doi: 10.4135/9781412984850

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality (Unpublished doctoral dissertation)*. University of Illinois at Urbana–Champaign, Urbana, IL.

He, L., and Chen, D. (2017). Developing common listening ability scales for Chinese learners of English. *LTA.* 7:4. doi: 10.1186/s40468-017-0033-4

Henson, R., and Templin, J. (2007). "Large-scale language assessment using cognitive diagnosis models," *Paper Presented at the Annual Meeting of the National Council for Measurement in Education* (Chicago, IL).

Iaconangelo, C. (2017). *Uses of Classification Error Probabilities in the Three-Step Approach to Estimating Cognitive Diagnosis Models (Unpublished doctoral dissertation)*. New Brunswick, NJ: Rutgers University.

Kang, C., Yang, Y., and Zeng, P. (2019). Approach to cognitive diagnosis: the manhattan distance discriminating method. *J. Psychol. Sci.* 42, 455–462. doi: 10.16719/j.cnki.1671-6981.20190228

Lee, Y.-W., and Sawaki, Y. (2009a). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Lang. Assessment Q.* 6, 239–263. doi: 10.1080/15434300903079562

Lee, Y.-W., and Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: an overview. *Lang. Assessment Q.* 6, 172–189 doi: 10.1080/15434300902985108

Lei, P.-W., and Li, H. (2014). "Fit indices' performance in choosing cognitive diagnostic models and Q-matrices," *Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME)* (Philadelphia, PA).

Levine, A., and Revers, T. (1988). The FL receptive skills: same or different? *System* 16, 327–336. doi: 10.1016/0346-251X(88)90075-9

Li, H., Hunter, C. V., and Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Lang. Test.* 33, 391–409. doi: 10.1177/0265532215590848

Liao, Y. (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-grammatical ability as a predictor of L2 listening ability. *Spaan Fellow Working Papers in Second or Foreign Language Assessment* (Ann Arbor, MI: University of Michigan English Language Institute).

Ma, W., and de la Torre, J. (2016). GDINA: *The Generalized DINA Model Framework. R Package version 0.13.0.* Available online at: http://CRAN.R-project.org/package=GDINA (accessed February 26, 2021).

Ma, W., Iaconangelo, C., and de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Appl. Psychol. Measure.* 40, 200–217. doi: 10.1177/0146621615621717

Ma, X., and Meng, Y. (2014). Towards personalized english learning diagnosis: Cognitive diagnostic modelling for EFL listening. *AJEEL.* 2, 336–348. Available online at: https://www.ajouronline.com/index.php/AJEEL/article/view/1669

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measur. Interdiscipl. Res. Perspect.* 11, 71–137. doi: 10.1080/15366367.2013.831680

Maydeu-Olivares, A., and Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika* 71, 714–732. doi: 10.1007/s11336-005-1295-9

Meng, Y. (2013). *Developing a Model of Cognitive dIagnostic Assessment for College EFL Listening.* Unpublished doctoral dissertation, Shanghai International Studies University, China.

Munby, J. (1978). *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Oller, J. W. (1983). "Evidence for a general proficiency factor: an expectancy grammar," in *Issues in Language Testing Research,* ed J. W. Oller (Rowley, MA: Newbery House Publishers), 3–10.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assessment* 34, 782–799. doi: 10.1177/0734282915623053

Richard, J. C. (1983). Listening comprehension: approach, design, procedure. *TESOL Q.* 17, 219–240. doi: 10.2307/3586651

Rivers, W. M. (1966). Listening comprehension. *Modern Lang. J.* 50, 196–202. doi: 10.1111/j.1540-4781.1966.tb01817.x

Rojas, G., de la Torre, J., and Olea, J. (2012). "Choosing between general and specific cognitive diagnosis models when the sample size is small," *Paper Presented at the Annual Meeting of the National Council of Measurement in Education* (Vancouver, BC).

Rost, M. (2011). *Teaching and Researching Listening.* Edinburgh Gate: Pearson Education.

Rupp, A. A., and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Measure. Interdisc. Res. Perspect.* 6, 219–262. doi: 10.1080/15366360802490866

Rupp, A. A., Templin, J. L., and Henson, R. A. (2010). *Diagnostic Assessment: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Sawaki, Y., Kim, H.-J., and Gentile, C. (2009). Q-Matrix construction: defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Lang. Assessment Q.* 6, 190–209. doi: 10.1080/15434300902801917

Shin, S. (2008). "Examining the construct validity of a web-based academic listening test: an investigation of the effects of response formats," *Spaan Fellow Working Papers in Second or Foreign Language Assessment* (Ann Arbor, MI: University of Michigan English Language Institute).

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modelling approach. *Lang. Test.* 25, 435–464. doi: 10.1177/0265532208094272

Templin, J., and Heson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–307. doi: 10.1037/1082-989X.11.3.287

Wagner, E. (2004). A construct validation study of the extended listening sections of the ECPE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment* (Ann Arbor, MI: University of Michigan English Language Institute).

Wang, W., Song, L., Chen, P., Meng, Y., and Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *J. Educ. Measure.* 52, 457–476. doi: 10.1111/jedm.12096

Weir, C. J. (1993). *Understanding and Developing Language Tests.* New York, NY: Prentice Hall.

Yi, Y.-S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Lang. Test.* 34, 337–355. doi: 10.1177/0265532216646141

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 Dong, Ma, Wang and Gao. 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 May 2021
doi: 10.3389/fpsyg.2021.619771

New Item Selection Method Accommodating Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing: Maximum Deviation and Maximum Limitation Global Discrimination Indexes

*Junjie Li, Lihua Ma, Pingfei Zeng and Chunhua Kang**

Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China

OPEN ACCESS

Edited by:
Chun Wang,
University of Washington,
United States

Reviewed by:
Yong Luo,
Educational Testing Service,
United States
Wenchao Ma,
University of Alabama, United States

***Correspondence:**
Chunhua Kang
akang@zjnu.cn

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

Received: *21 October 2020*
Accepted: *26 March 2021*
Published: *17 May 2021*

Citation:
Li J, Ma L, Zeng P and Kang C (2021) New Item Selection Method Accommodating Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing: Maximum Deviation and Maximum Limitation Global Discrimination Indexes. Front. Psychol. 12:619771. doi: 10.3389/fpsyg.2021.619771

Maximum deviation global discrimination index (MDGDI) is a new item selection method for cognitive diagnostic computerized adaptive testing that allows for attribute coverage balance. We developed the maximum limitation global discrimination index (MLGDI) from MDGDI, which allows for both attribute coverage balance and item exposure control. MLGDI can realize the attribute coverage balance and exposure control of the item. Our simulation study aimed to evaluate the performance of our new method against maximum global discrimination index (GDI), modified maximum GDI (MMGDI), standardized weighted deviation GDI (SWDGDI), and constraint progressive with SWDGDI (CP_SWDGDI). The results indicated that (1a) under the condition of realizing the attribute coverage balance, MDGDI had the highest attribute classification accuracy; (1b) when the selection strategy accommodated the practical constraints of the attribute coverage balance and item exposure control, MLGDI had the highest attribute classification accuracy; (2) adding the item exposure control mechanism to the item selection method reduces the classification accuracy of the attributes of the item selection method; and (3) compared with GDI, MMGDI, SWDGDI, CP_SWDGDI, and MDGDI, MLGDI can better achieve the attribute-coverage requirement, control item exposure rate, and attribute correct classification rate.

Keywords: balance attribute coverage, cognitive diagnostic computerized adaptive testing, attribute discrimination index, item exposure control, mastery pattern correct classification rate

INTRODUCTION

Cognitive diagnostic assessment (CDA) is a recently popular assessment method in theoretical studies on psychological testing. CDA was developed to measure cognitive skills (Leighton and Gierl, 2007; Gierl et al., 2008). When based on the classical test theory (CTT), CDA provides examinee scores. When based on the multidimensional item response theory, CDA provides

multidimensional ability scores, which details the advantages and disadvantages of the examinee in a given content domain, aiding the assessment of the examinees by administrators (Yao and Boughton, 2007; Lee et al., 2012).

Interest in cognitive diagnosis is largely motivated by the need for formative assessments. Computerized adaptive testing (CAT) combines test theory with computer technology to improve testing efficiency (Weiss, 1982), which has become a promising method in psychological and educational measurement. In addition, items in CAT have been executed in examinations for items that have matched the estimating ability for candidates (Mao and Xin, 2013; Chang, 2015). Recently, to maximize the benefits of both CDA and CAT, researchers have attempted to combine CDA with CAT and named it cognitive diagnostic CAT (CD-CAT) (Xu et al., 2003; McGlohen and Chang, 2008; Cheng, 2009a,b). CD-CAT, which has the characteristics of a tailor-made test, is promising and will be influential in future educational practices. CD-CAT has received an increasing scholarly attention worldwide (Kang et al., 2017; Huebner et al., 2018).

The goal of CAT is to conduct individualized item selection tests based on the most currently estimated ability of the participant; thus, the determination of an optimal item selection method is key in CAT. Although many item selection strategies have been constructed in the item response theory–based CAT, few applicable item selection strategies are currently available in CD-CAT. Therefore, this study aimed to construct a selection strategy that is suitable for CD-CAT. Based on the difference distribution criteria of the potential attribute-mastery pattern at the item level, researchers have proposed a selection criteria, such as the Kullback–Leibler (KL)-based global discrimination index (GDI), Shannon entropy procedure (Xu et al., 2003), and the posterior-weighted KL information (PWKL; Cheng, 2009a,b). However, the aforementioned item selection methods focus on the maximum information of the item without considering the attribute coverage balance of the test and exposure control of the item. Therefore, the aforementioned item selection methods face the following two problems. First, attribute coverage imbalance may cause the test results to be unreliable. Cheng (2010) also pointed out that it is of great importance to ensure that each attribute in the test has been measured adequately or the reliability of the test will be reduced. Second, an unevenly applied item bank will result in the following two situations: (1) some items increase their exposure rate in a different test, which endangers the security of the item bank, and (2) if some items are applied adequately, the item bank is poorly utilized and resources (including labor) are wasted. Although CD-CAT is increasingly used in the classroom, test security is not fundamental to the CD-CAT practices, whereas security and property balance are critical to CD-CAT developers. Specifically, the item bank must be secured because CD-CAT is a complex and expensive project. As for every item written for CD-CAT, it must be based on a complex blueprint of cognitive requirements. In addition, when specific items are used for each test, practice or memory effects may produce invalid diagnostic information for candidates who has taken the test repeatedly. Improving the utilization rate of an item bank also constitutes a research problem for the practical application of CD-CAT.

To balance the attribute coverage in CD-CAT, Cheng (2010) developed the modified maximum global discrimination index (MMGDI) to build the item selection method using the number of items that measure each attribute as the MMGDI did. The MMGDI method is based on the global discrimination index (GDI) developed by Xu et al. (2003). Although the MMGDI method achieves the balance in attribute coverage and improves the accuracy of the attribute-mastery pattern, MMGDI does not consider the exposure rate of items, which leads the MMGDI to repeatedly select some items in different tests. Lin and Chang (2018) proposed a method, the constraint progressive with standardized weighted deviation GDI (CP_SWDGDI), which allows for attribute coverage balance and exposure control (named considering the attribute balancing and exposure control). Although CP_SWDGDI considers both the attribute coverage balance and exposure control, the CP_SWDGDI selection method fails under some conditions, such as when the attribute coverage balance is satisfied.

The objective of this article is to propose a maximum deviation (MD) index and a maximum limit (ML) index, and combine them with GDI for use in CD-CAT. We first developed an item selection method MDGDI, which can achieve attribute coverage balance. Subsequently, we added an exposure control mechanism based on MDGDI and developed a CD-CAT item selection method MLGDI, that can achieve attribute coverage balance and items exposure control. The rest of this paper is organized as follows.

First, we discuss the CDM used in this study and introduce the four existing item selection algorithms for CD-CAT. Next, we introduce the MDGDI and MLGDI methods. We then evaluate MDGDI and MLGDI against the existing item selection algorithms via two simulation studies. Finally, we discuss the consequences of the simulation results and provide suggestions for further research.

REDUCED REPARAMETERIZED UNIFIED MODEL

The reduced reparameterized unified model (RRUM) is used in the current study (Hartz, 2002; Hartz and Roussos, 2008), because previous studies have demonstrated that the RRUM is very useful for formative assessment in practice (Wang et al., 2011). The item response function of the RRUM is defined as,

$$p(Y_{ij} = 1|a_i) = \pi_j^* \prod_{k \in A_j} r_{jk}^{*(1-a_{ik})q_{jk}}$$

where, $0 < \pi_j^* < 1$ is the probability of a correct answer for an examinee who has mastered all the attributes required for item j, and $0 < r_{jk}^* < 1$ is a penalty parameter that reduces the probability of a correct response by a factor of r_{jk}^* for examinees who do not possess attribute k.

EXISTING ITEM SELECTION METHODS

Global Discrimination Index (GDI)

The KL information was first introduced to CAT research in Chang and Ying's (1996) groundbreaking paper on global information. The KL information has since been applied to various studies on CAT. For example, CAT was established based on a non-parametric item response theoretical model (Xu and Douglas, 2006), and CAT has been applied to classification (Weissman, 2007) and cognitive diagnostic applications (McGlohen and Chang, 2008; Cheng, 2009a,b). The KL information, which measures the distance or divergence between two probability distributions $f(x)$ and $g(x)$ (Cover and Thomas, 1991; Kaplan et al., 2015), is defined as follows:

$$d[f,g] = E_f\left[\log\left[\frac{f(x)}{g(x)}\right]\right]$$

In CD-CAT, information refers to the ability of an item to distinguish between a pair of attribute patterns. In this sense, KL information in diagnostic classification reflects the distance between two conditional distributions, that is, $f(X_{ij}|\hat{a}_i)$ is the distribution on the currently estimated attribute under condition X_{ij}, and $f(X_{ij}|a_t)$ is the distribution in the real state under condition X_{ij}. This logic gives the KL equation of CD-CAT:

$$KL_j(\hat{a}_i||a_t) = \sum_{x=0}^{1}\log\left(\frac{p(X_{ij}=x|\hat{a}_i)}{p(X_{ij}=x|a_t)}\right)p(X_{ij}=x|\hat{a}_i)$$

Xu et al. (2003), who considered that the true potential is unknown and that 2^k possible states exist, proposed the GDI with the following formula:

$$GDI_j(\hat{a}_i) = \sum_{c=1}^{2^k}\left[\sum_{x=0}^{1}\log\left(\frac{p(X_{ij}=x|\hat{a}_i)}{p(X_{ij}=x|\hat{a}_c)}\right)p(X_{ij}=x|\hat{a}_i)\right]$$

This index is the sum of the KL distances between $p(X_{ij}|\hat{a}_i)$ and all possible potential states $p(X_{ij}|\hat{a}_c)$. Items with large GDI values have a correspondingly high recognition between the estimating attribute patterns and all other possible cognitive profiles. An item with a maximum GDI (MGDI) will be administered as the next item for a specific examinee. In Xu et al. (2003), the MGDI method exhibited a good performance in restoring the pattern of student attribute mastery.

The Maximum Modified Global Discrimination Index (MMGDI)

The disadvantage of the GDI approach is that it does not consider property balancing or exposure control. Cheng and Chang (2009) introduced the maximum priority index (MPI) method for the selection of items that satisfy the constraints in the IRT-based CAT. In a subsequent study, Cheng (2010) extended the MPI method to CD-CAT to achieve balance attribute override. The attribute coverage balance index (ABI) is defined as follows:

$$ABI_j = \prod_{k=1}^{K}\left(\frac{B_k - b_k}{B_k}\right)^{q_{jk}}$$

where, B_k is the lower bound of the number of items required to measure attribute k, b_k is the number of items measuring attribute k that has already been selected, and q_{jk} is the element of the Q matrix. Cheng (2010) added the ABI to GDI and constructed the MMGDI item selection method, which is defined as follows:

$$MMGDI_j(\hat{a}_i) = \prod_{k=1}^{K}\left(\frac{B_k - b_k}{B_k}\right)^{q_{jk}}.GDI(\hat{a}_i)$$

Modified maximum GDI makes a GDI-based strategy more precise. Specifically, MMGDI attributes in the balance tends only toward the choice of measurement index in the selected item of a single attribute (Mao and Xin, 2013), and, in ABI, there may be situations where negative and negative multiply to be positive, which affects the efficiency of the item selection method.

The Standardized Weighted Deviation GDI Method (SWDGDI)

Lin and Chang (2018) proposed a new attribute-balancing item selection criterion, namely the Weighted Deviation GDI (WDGDI), which multiplies GDI by the Weighted Deviation Index (WD). To place the WD and the GDI metrics on an equal footing, they standardized the WD and GDI values and named it the standardized WDGDI (SWDGDI). The SWDGDI method is defined as follows:

$$SWDGDI_j(\hat{a}_i) = \left(\frac{Max(WD_j) - WD_j}{Max(WD_j) - Min(WD_j)}\right)\times \left(\frac{GDI_j(\hat{a}_i) - Min(GDI_j(\hat{a}_i))}{Max(GDI_j(\hat{a}_i)) - Min(GDI_j(\hat{a}_i))}\right)$$

$$WD_j = \sum_{k=1}^{K}(W_k D_{jL_k}) + \sum_{k=1}^{K}(W_k D_{jU_k})$$

$$D_{jL_k} = \begin{cases} L_k - q_k, q_k < L_k \\ 0, q_k \geq L_k \end{cases} \qquad D_{jU_k} = \begin{cases} q_k - U_k, q_k \geq U_k \\ 0, q_k < U_k \end{cases}$$

where, W_k is the weight for the k^{th} attribute, and D_{jLk} and D_{jUk} correspond to the positive deviations from the minimal (i.e., lower boundary) and maximal (i.e., upper boundary) numbers, respectively, of the items required to assess the k^{th} attribute when item j is included in the test. For each constraint k, D_{jLk} is defined as $(L_k - q_k)$and D_{jUk} is defined as $(U_k - q_k)$, where L_k and U_k, respectively, denote the lower and upper bounds for the k^{th} attribute constraint. The term q_k represents the expected number of items measuring the k^{th} attribute that would have been obtained if item candidate j was included in the test.

With the attribute balancing considered, the largest SWDGDI item is selected first in the test rather than the GDI's largest project.

The Constraint Progressive With SWDGDI (CP_SWDGDI)

In order to balance the attribute coverage and control the item exposure rate, Lin and Chang (2018) adopted a progressive

exposure control algorithm in SWDGDI. The Constrained Progressive Algorithm is described as follows:

$$P_INFO_j = \left(1 - \frac{X}{L}\right) R_j + \frac{X}{L} \times R_{jI}$$

$$R_j = U(0, \max(\text{information})) \qquad R_{jI} = U(LB_j, UB_j)$$

$$LB_j = Info_j - (Info_j - Min)/s \qquad UB_j = Info_j + (Max - Info_j)/s$$

In the progressive exposure control algorithm constructed by Lin and Chang (2018), the adjustment information interval parameter s was added. However, with regard to practical applications, Lin and Chang (2018) offered no specific suggestions for determining the value of s. Therefore, the appropriate value of s may differ between the conditions and number of attributes, which makes determining the value of s difficult in practical applications.

When replacing $Info_j$ with $SWDGDI_j$, the CP_SWDGDI became:

$$CP_SWDGDI_j(\hat{a}_i) = \frac{r_\max}{r_j} \times \left[\left(1 - \frac{X}{L}\right) R_j + \frac{X}{L} \times R_{jI}\right] R_{jI} = U(LB_j, UB_j)$$

$$LB_j = SWDGDI_j(\hat{a}_i) - (SWDGDI_j(\hat{a}_i) - Min)/s$$

$$UB_j = SWDGDI_j(\hat{a}_i) + (Max - SWDGDI_j(\hat{a}_i))/s$$

where, r_max is the maximum exposure rate for the title and r_j is the current exposure rate for the item.

PROPOSED ITEM SELECTION METHODS

Maximum Deviation Index With GDI (MDGDI)

In order to make all the attributes relatively balanced throughout the test and to reduce the tendency of the selection strategy to choose certain types of items more often due to the index added, the maximum deviation index (MD) was developed. MD limits the difference between the maximum and minimum measurement times of an attribute within a certain range. The definition of MD is as follows:

$$MD_j = \begin{cases} 1, \max(q_{jk}) - \min(q_{jk}) \le LB \\ 0, \max(q_{jk}) - \min(q_{jk}) > LB \end{cases}$$

where, LB is the lowest number of attributes, q_{jk} is the number of attributes to be investigated if the next item is j, and MD is the deviation index.

Now the maximum deviation global discrimination index (MDGDI) becomes:

$$MDGDI_j = MD_j \times GDI_j$$

The item that yields the largest MDGDI is offered for a specific examinee as the next item.

Combining MD Index and Limited Exposure Control Index With GDI (MLGDI)

Although CP_SWDGDI considers both the attribute coverage balance and exposure control in the selection strategy, for the exposure control part of CP_SWDGDI, the variablesmust be established by the manager themself, and the appropriate s value may differ under different conditions, which makes it difficult to determine the value of s.

Limiting Exposure Index

In this study, we proposed a limiting exposure index to control the exposure rate of items. The idea of limiting exposure index was built upon with the aims of (1) eliminating the need to determine the crucial parameter s in the random part and (2) making the exposure index more concise. The limit exposure index is comprised of two parts: the random part and the limit maximum exposure. The random part is based on the idea of asymptotic behavior, and the amount of information in item J after increasing the random part is expressed as RI_j, where $RI_j = U(LI_j, UI_j)$ and RI_j is generated randomly from the uniform distribution $U(LI_j, UI_j)$.

$$LI_j = GDI_j - (x/L) \times (GDI_j - \min(GDI))$$

$$UI_j = GDI_j + (x/L) \times (\max(GDI) - GDI_j)$$

where, GDI is the GDI information of the remaining items in the item bank; x is the current test length; L is the maximum test length; and LI_j and UI_j are the lower bound and upper line of $U(LI_j, UI_j)$, respectively. As the length of the test increases, LI_j and UI_j approach the original GDI_j, and the random RI_j approaches GDI_j. Therefore, the information of the items becomes more accurate.

In addition, the component that limits the maximum exposure rate is as follows,

$$Lr_j \begin{cases} 1, r_j < r \\ 0, r_j \ge r \end{cases}$$

where, r is the maximum exposure rate and r_j is the current exposure rate of item j. If the exposure rate r_j of the next problem j is greater than or equal to the maximum exposure rate r of the problem, then $Lr_j = 0$; if r_j is less than r, then $Lr_j = 1$.

To maximize the participant's exposure rate restrictions, the GDI item selection method was applied with the limited exposure index as follows:

$$LGDI_j = Lr_j \times RI_j$$

Combining MD Index and Limited Exposure Control Index

According to the aforementioned MD index and limited exposure index, this study proposed the maximum limitation global

discrimination indexes that considers both the attribute coverage balance and exposure control as follows:

$$MLGDI_j = MD_j \times LGDI_j$$

During the MLGDI procedure, an item with the maximum MLGDI value will be selected for administration.

SIMULATION STUDY

Study I

Study I is a simulation conducted to investigate the performance of MDGDI against GDI, MMGDI, and SWDGDI.

Item Pools

Item pools were constructed based on the study of Wang et al. (2020). Three item pools were designed in this study, denoted as the low discrimination (LD), high discrimination (HD), and hybrid discrimination (HyD) item pools, respectively. Each item pool contained 775 items and measured five attributes in total (Wang et al., 2011; Huebner et al., 2018). In the LD item pool, item parameters π_j^* and r_{jk}^* were generated from uniform distributions U(0.75, 0.95) and U(0.15, 0.50), respectively. In the HD item pool, π_j^* and r_{jk}^* were generated from uniform distributions U(0.75, 0.95) and U(0.05, 0.40), respectively. In the HyD item pool, π_j^* and r_{jk}^* were generated from uniform distributions U(0.75, 0.95) and U(0.05, 0.50), respectively. **Table 1** represents the descriptive statistics of item parameters of LD item pool, HD item pool, and HyD item pool.

Examinee Populations

Three examinee populations were generated, each containing 3,200 examinees. The first population (denote as Unif) assumed that the Attribute Mastery Pattern (AMP) of each examinee was generated from the uniform distribution of 32 possible pattern profiles with a probability of 1/32. In this way, each AMP had 100 examinees; meanwhile, each examinee had a 0.5 chance to master each attribute. Considering that correlations among attributes is common in practice, a multivariate normal distribution was used to describe the relationship among attributes for the second and third populations (denote as Norm). In these two groups, the mastery probabilities for the five attributes are defined as 0.45, 0.50, 0.55, 0.60, and 0.65, respectively. The correlations between attributes were set at 0.5 for the second population (low correlation) and 0.8 for the third population (high correlation). **Table 2** represents the frequencies of examinees who possess each possible number of attributes.

Constraints of Attribute-Balance Coverage

The minimum measurement time of each attribute was B_k = 3. The s parameter in CP_SWDGDI was 1.6, r_max = 0.2 and LB = 3.

We generated a total of 27 conditions in this study (3 item pools × 3 examinee populations × 3 item selection methods). We fixed the number of items in the test to have 20 in all conditions. The first item was selected randomly from the item pool, with a maximum *a posteriori* (MAP) method used to estimate the examinee's AMP, and the prior information of AMP assumed to follow a uniform distribution. The study procedures were implemented by the R software.

Evaluation Criteria

We evaluated the methods with respect to six criteria: attribute correct classification rate (ACCR), average marginal match rate (AAMR), mastery pattern correct classification rate (PCCR), item-bank exposure rate χ^2, test overlap rate (TOR), and maximum exposure rate. The computation of the first five criteria is as follows:

$$ACCR_k = \sum_{i=1}^{N} I(\widehat{a}_{ik} = a_{ik})/N$$

$$AAMR = \sum_{k=1}^{K}\sum_{i=1}^{N} I(\widehat{a}_{ik} = a_{ik})/(N^*K)$$

$$PCCR = \sum_{i=1}^{N} I(\widehat{a}_i = a_i)/N$$

$$TOR = \frac{N \times \sum_{j=1}^{N_{item}} er_j^2}{(N-1)\times L} - \frac{1}{N-1}$$

$$\chi^2 = \sum_{j=1}^{N_{item}} (er_j - J/N_{item})^2/(J/N_{item})$$

where, $\hat{\alpha}_i$ and a_i are the real and estimated values, respectively, of the attribute of participant i mastering the pattern, I(...)

TABLE 1 | Descriptive statistics of item parameters of the LD, HD, and HyD item pools.

		π_j^*	r_1^*	r_2^*	r_3^*	r_4^*	r_5^*
LD item pool	Min	0.750	0.151	0.152	0.152	0.151	0.151
	Mean	0.850	0.327	0.323	0.325	0.326	0.324
	Max	0.950	0.500	0.499	0.499	0.499	0.497
	SD	0.058	0.100	0.101	0.103	0.100	0.100
HD item pool	Min	0.750	0.051	0.050	0.052	0.051	0.052
	Mean	0.850	0.229	0.224	0.230	0.226	0.225
	Max	0.950	0.399	0.399	0.398	0.399	0.397
	SD	0.058	0.101	0.102	0.100	0.102	0.101
HyD item pool	Min	0.750	0.051	0.052	0.052	0.053	0.050
	Mean	0.849	0.273	0.268	0.279	0.275	0.275
	Max	0.950	0.499	0.499	0.497	0.499	0.499
	SD	0.057	0.127	0.133	0.129	0.129	0.135

LD item pool, low discrimination item pool; HD item pool, high discrimination item pool; HyD item pool, hybrid discrimination item pool.

TABLE 2 | Frequencies of examinees exhibiting each possible number of attributes in each population.

Number of attributes		0	1	2	3	4	5
Number of examinees	Unif	100	495	999	1,005	494	107
	Norm-0.5	208	338	400	541	691	1,022
	Norm-0.8	486	270	265	301	431	1,447

is an indicator function. A higher ACCR and AMMR value indicate a more accurate estimate of each participant attribute. A higher PCCR value indicates a more accurate estimate of the participant's overall knowledge status; er_j is the exposure rate of item j, N_{item} is the size of the item bank, χ^2 is the exposure rate index of an item, and TOR is the overlapping rate index of the test. The smaller the values of χ^2 and TOR are, the more fully and uniformly the item strategy utilizes the item bank.

Results

Table 3 compares the recovery rate of each attribute and of the entire profile obtained from the four item selection methods (GDI, MMGDI, SWDGDI, and MDGDI). Clearly, the MMGDI, SWDGDI, and MDGDI methods outperformed the GDI method especially in the entire pattern recovery rate. This was because recovering the entire profile requires correctly recovering every attribute and gain the attribute level aggregates. This is in line with Cheng (2010) and Lin and Chang (2018). Among the four methods, the MDGDI method was superior. Besides, all of the methods performed best in the HD item pool, followed by the HyD item pool, and the LD item pool was the worst.

Study II

Study II evaluated the performance of MLGDI, which had the item exposure control mechanism and was based on MDGDI, against competing item selection strategies. The results of Study 1 indicated that when the test termination rule is reached, MDGDI has the highest classification accuracy

TABLE 3 | Accuracy of the attribute classification for five attributes.

Item selection method		Attribute (ACCR)					AAMR	PCCR
		A1	A2	A3	A4	A5		
HD-unif	GDI	0.910	0.814	0.975	0.949	0.956	0.921	0.642
	MMGDI	0.968	0.964	0.965	0.976	0.968	0.968	0.853
	SWDGDI	0.937	0.867	0.969	0.945	0.959	0.935	0.688
	MDGDI	0.990	0.984	0.992	0.991	0.990	0.989	0.953
HD-norm-0.5	GDI	0.946	0.854	0.856	0.901	0.942	0.900	0.598
	MMGDI	0.978	0.969	0.976	0.973	0.982	0.976	0.887
	SWDGDI	0.962	0.944	0.948	0.944	0.967	0.953	0.778
	MDGDI	0.991	0.988	0.991	0.991	0.995	0.991	0.959
HD-norm-0.8	GDI	0.833	0.871	0.882	0.939	0.907	0.886	0.592
	MMGDI	0.967	0.971	0.964	0.980	0.970	0.971	0.864
	SWDGDI	0.947	0.957	0.952	0.967	0.965	0.958	0.800
	MDGDI	0.989	0.994	0.991	0.993	0.990	0.991	0.962
LD-unif	GDI	0.973	0.852	0.982	0.913	0.821	0.908	0.599
	MMGDI	0.955	0.946	0.964	0.937	0.939	0.948	0.769
	SWDGDI	0.964	0.921	0.972	0.922	0.901	0.936	0.728
	MDGDI	0.972	0.963	0.978	0.960	0.960	0.967	0.868
LD-norm-0.5	GDI	0.866	0.914	0.914	0.950	0.793	0.887	0.541
	MMGDI	0.964	0.959	0.959	0.960	0.950	0.958	0.815
	SWDGDI	0.928	0.944	0.949	0.953	0.908	0.936	0.715
	MDGDI	0.977	0.975	0.981	0.985	0.967	0.977	0.900
LD-norm-0.8	GDI	0.912	0.866	0.885	0.913	0.935	0.902	0.595
	MMGDI	0.958	0.960	0.958	0.954	0.973	0.961	0.819
	SWDGDI	0.941	0.933	0.932	0.929	0.954	0.938	0.709
	MDGDI	0.965	0.970	0.974	0.964	0.985	0.972	0.877
HyD-unif	GDI	0.937	0.908	0.983	0.813	0.970	0.922	0.648
	MMGDI	0.969	0.967	0.966	0.953	0.963	0.964	0.837
	SWDGDI	0.946	0.953	0.975	0.865	0.970	0.942	0.728
	MDGDI	0.985	0.990	0.989	0.977	0.988	0.986	0.938
HyD-norm-0.5	GDI	0.893	0.872	0.948	0.849	0.926	0.897	0.593
	MMGDI	0.958	0.970	0.975	0.971	0.973	0.970	0.859
	SWDGDI	0.935	0.947	0.964	0.940	0.963	0.950	0.764
	MDGDI	0.988	0.993	0.990	0.989	0.991	0.990	0.953
HyD-norm-0.8	GDI	0.846	0.873	0.950	0.938	0.912	0.904	0.642
	MMGDI	0.968	0.963	0.976	0.978	0.972	0.971	0.864
	SWDGDI	0.949	0.930	0.975	0.971	0.965	0.958	0.803
	MDGDI	0.984	0.993	0.995	0.997	0.994	0.993	0.964

TABLE 4 | Accuracy of the attribute classification for five attributes.

Item selection method		Attribute (ACCR)					AAMR	PCCR
		A1	A2	A3	A4	A5		
HD-unif	GDI	0.910	0.814	0.975	0.949	0.956	0.921	0.642
	CP_SWDGDI	0.931	0.924	0.937	0.943	0.926	0.932	0.709
	MLGDI	0.981	0.973	0.986	0.987	0.971	0.979	0.911
HD-norm-0.5	GDI	0.946	0.854	0.856	0.901	0.942	0.900	0.598
	CP_SWDGDI	0.910	0.931	0.930	0.931	0.929	0.926	0.676
	MLGDI	0.985	0.984	0.988	0.988	0.993	0.988	0.945
HD-norm-0.8	GDI	0.833	0.871	0.882	0.939	0.907	0.886	0.592
	CP_SWDGDI	0.924	0.918	0.938	0.932	0.931	0.929	0.681
	MLGDI	0.987	0.988	0.986	0.989	0.988	0.988	0.944
LD-unif	GDI	0.973	0.852	0.982	0.913	0.821	0.908	0.599
	CP_SWDGDI	0.903	0.914	0.919	0.906	0.903	0.909	0.643
	MLGDI	0.956	0.949	0.967	0.939	0.944	0.951	0.811
LD-norm-0.5	GDI	0.866	0.914	0.914	0.950	0.793	0.887	0.541
	CP_SWDGDI	0.895	0.903	0.904	0.900	0.886	0.898	0.580
	MLGDI	0.968	0.976	0.963	0.973	0.952	0.966	0.859
LD-norm-0.8	GDI	0.912	0.866	0.885	0.913	0.935	0.902	0.595
	CP_SWDGDI	0.890	0.879	0.897	0.902	0.908	0.895	0.555
	MLGDI	0.963	0.960	0.969	0.961	0.974	0.965	0.850
HyD-unif	GDI	0.937	0.908	0.983	0.813	0.970	0.922	0.648
	CP_SWDGDI	0.923	0.918	0.927	0.913	0.912	0.919	0.658
	MLGDI	0.973	0.978	0.980	0.964	0.971	0.973	0.885
HyD-norm-0.5	GDI	0.893	0.872	0.948	0.849	0.926	0.897	0.593
	CP_SWDGDI	0.914	0.911	0.927	0.926	0.924	0.921	0.653
	MLGDI	0.982	0.977	0.991	0.984	0.986	0.984	0.929
HyD-norm-0.8	GDI	0.846	0.873	0.950	0.938	0.912	0.904	0.642
	CP_SWDGDI	0.903	0.912	0.930	0.919	0.919	0.917	0.642
	MLGDI	0.978	0.986	0.990	0.988	0.988	0.986	0.934

compared with MGDI, MMGDI, and CP_SWGDI. MLGDI is a new item selection method based on MDGDI with an additional exposure control mechanism, whereas CP_SWGDI is a new item selection method based on SWGDI with an additional exposure control mechanism. We expect MLGDI to have the highest attribute classification accuracy in MGDI, CP_SWGDI, and MLGDI when the test satisfies the test termination rule.

Study II was conducted to investigate the performance of MLGDI against CP_SWDGDI and GDI. The data generation and evaluation criteria are the same as study I.

Results

Table 4 lists the estimates of ACCR, AAMR, and PCCR in each condition. The MLGDI stands out in both the recovery rate of each attribute and the entire profile, followed by CP_SWDGDI. As evident in **Table 4**, compared with the PCCR of CP_SWDGDI (which also includes an exposure control mechanism), the PCCR of MLGDI increased by approximately 0.15–0.30. **Table 4** also indicates that when the test sample reached the test termination condition, MLGDI exhibited the highest accuracy in attribute classification.

MLGDI Can Reduce the Participant's Exposure Rate of MDGDI and Yield a High Accuracy in the Attribute Classification

Table 5 presents the exposure indicators of each item under the different examinee populations, item pools, and six item strategies (MGDI, MMGDI, SWDGDI, CP_SWDGDI, MDGDI, and MLGDI). It is worth noting that, as the exposure control index was added to the MLGDI, the decrease in PCCR was relatively small compared to MDGDI which has the highest PCCR comparing to the other selection item methods, but result in a better item bank usage. As detailed in **Table 5**, the chi-square value of the item-bank exposure rate of the four item selection strategies without exposure restriction exceeded 110, the TOR exceeded 0.15, and the maximum item exposure rate reached >0.50. Although the accuracy of the MDGDI's attribute classification was the highest among the six item strategies, the exposure rate of the relevant item bank was also higher than those of the other five strategies. For example, LD-norm-0.5, the chi-square value of MDGDI's exposure rate was as high as 250, the TOR was as high as 0.349, and the maximum exposure rate of the title was

TABLE 5 | Item exposure of each item selection method for five attributes.

	Item selection method	Index			
		χ^2	TOR	max_expose	PCCR
HD-unif	GDI	125.444	0.187	0.764	0.642
	MMGDI	302.007	0.415	1.000	0.853
	SWDGDI	121.937	0.183	0.566	0.688
	MDGDI	326.991	0.448	0.936	0.953
	CP_SWDGDI	0.025	0.026	0.030	0.709
	MLGDI	75.726	0.123	0.190	0.911
HD-norm-0.5	GDI	112.473	0.171	0.574	0.598
	MMGDI	263.488	0.366	1.000	0.887
	SWDGDI	139.819	0.206	0.433	0.778
	MDGDI	250.257	0.349	0.753	0.959
	CP_SWDGDI	0.031	0.026	0.032	0.676
	MLGDI	75.008	0.122	0.190	0.945
HD-norm-0.8	GDI	122.464	0.184	0.458	0.592
	MMGDI	253.052	0.352	1.000	0.864
	SWDGDI	181.618	0.260	0.488	0.800
	MDGDI	264.183	0.366	0.660	0.962
	CP_SWDGDI	0.038	0.026	0.032	0.681
	MLGDI	78.197	0.126	0.190	0.944
LD-unif	GDI	159.785	0.232	0.786	0.599
	MMGDI	297.878	0.410	1.000	0.769
	SWDGDI	136.547	0.202	0.595	0.728
	MDGDI	291.227	0.401	0.893	0.868
	CP_SWDGDI	0.024	0.026	0.030	0.643
	MLGDI	67.731	0.113	0.190	0.811
LD-norm-0.5	GDI	111.453	0.169	0.533	0.541
	MMGDI	284.401	0.393	1.000	0.815
	SWDGDI	141.231	0.208	0.480	0.715
	MDGDI	235.690	0.330	0.693	0.900
	CP_SWDGDI	0.032	0.026	0.031	0.580
	MLGDI	72.465	0.119	0.190	0.859
LD-norm-0.8	GDI	146.151	0.214	0.488	0.595
	MMGDI	270.456	0.375	1.000	0.819
	SWDGDI	179.267	0.257	0.491	0.709
	MDGDI	253.893	0.353	0.642	0.877
	CP_SWDGDI	0.036	0.026	0.033	0.555
	MLGDI	79.315	0.128	0.190	0.850
HyD-unif	GDI	178.761	0.256	0.843	0.648
	MMGDI	250.533	0.349	1.000	0.837
	SWDGDI	148.291	0.217	0.729	0.728
	MDGDI	301.685	0.415	0.918	0.938
	CP_SWDGDI	0.029	0.026	0.030	0.658
	MLGDI	72.495	0.119	0.190	0.885
HyD-norm-0.5	GDI	128.105	0.191	0.567	0.593
	MMGDI	260.627	0.362	1.000	0.859
	SWDGDI	151.950	0.222	0.429	0.764
	MDGDI	263.093	0.365	0.691	0.953
	CP_SWDGDI	0.038	0.026	0.032	0.653
	MLGDI	80.168	0.129	0.190	0.929
HyD-norm-0.8	GDI	151.469	0.221	0.525	0.642
	MMGDI	286.640	0.395	1.000	0.864
	SWDGDI	197.969	0.281	0.480	0.803
	MDGDI	274.848	0.380	0.682	0.964
	CP_SWDGDI	0.037	0.026	0.032	0.642
	MLGDI	79.190	0.128	0.190	0.934

as high as 0.753. MLGDI was the selected item strategy that integrated exposure inhibition based on MDGDI. As indicated in **Table 5**, MLGDI could effectively reduce the exposure index of each item while considering the high accuracy of attribute classification. Similarly, the chi-square value of MLGDI's item exposure rate was 75, which was 175 less than that of MDGDI. The TOR of MLGDI was 0.122, which was less than that of MDGDI by 0.227. The maximum exposure rate of the MLGDI's item was 0.190, which was 0.563 lower than that of MDGDI. With respect to the mastery pattern correct classification rate, the PCCR value of MDGDI was 0.959, and that of MLGDI was 0.945. The mastery pattern correct classification rate decreased by 0.014, and the attribute classification accuracy remained high. In addition, the MLGDI's mastery pattern correct classification rate remained higher than those of MGDI, MMGDI, SWDGDI, and CP_SWDGDI, and its mastery pattern correct classification rate was second only to that of MDGDI. **Table 5** also indicates that although CP_SWDGDI had the highest performance in each index of item exposure rate, the excessive exposure inhibition component added by CP_SWDGDI resulted in a low item selection efficiency and a low accuracy in attribute classification. In the case of LD-norm-0.5, the MLGDI's PCCR value was 0.945, but that of CP_SWDGDI's was only 0.676, which was less than that of MLGDI by 0.269. Therefore, although CP_SWDGDI can reduce the items exposure rate, it has a low item selection efficiency and a low accuracy in attribute classification. Therefore, CP_SWDGDI cannot execute a desirable exposure control while maintaining a relatively high classification accuracy in the item selection test.

Table 6 shows the percentage of tests that met the attribute-coverage requirement, both at the attribute and overall test levels. For instance, the first entry in the table is 0.675, meaning 67.5% of the tests under the GDI method met the coverage constraint of the first attribute, or that 67.5% of the tests had at least three items measuring the first attribute. Compared with the uncontrolled method, MMGDI, SWDGDI, MDGDI, MLGDI, and CP_SWDGDI produced noticeably better results in balancing the attribute coverage: 100% of the tests met all the attribute coverage

TABLE 6 | Attribute coverage balance of each item selection method under five attributes.

	Item selection method	Attribute coverage balance					Total balance
		A1	A2	A3	A4	A5	
HD-unif	GDI	0.675	0.395	0.761	0.699	0.708	0.118
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	1.000	1.000	0.999
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
HD-norm-0.5	GDI	0.883	0.607	0.697	0.739	0.747	0.276
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	0.999	1.000	0.999
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
HD-norm-0.8	GDI	0.776	0.810	0.833	0.890	0.691	0.403
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
LD-unif	GDI	0.948	0.638	0.952	0.585	0.411	0.103
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
LD-norm-0.5	GDI	0.763	0.844	0.880	0.918	0.466	0.241
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	1.000	1.000	0.999
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
LD-norm-0.8	GDI	0.903	0.820	0.830	0.828	0.898	0.481
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	1.000	0.999	0.999
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
HyD-unif	GDI	0.699	0.671	0.928	0.382	0.860	0.125
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	1.000	1.000	0.999
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
HyD-norm-0.5	GDI	0.785	0.781	0.926	0.586	0.718	0.275
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000
HyD-norm-0.8	GDI	0.818	0.789	0.898	0.900	0.815	0.542
	MMGDI	1.000	1.000	1.000	1.000	1.000	1.000
	SWDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	MDGDI	1.000	1.000	1.000	1.000	1.000	1.000
	CP_SWDGDI	1.000	1.000	1.000	0.999	1.000	0.999
	MLGDI	1.000	1.000	1.000	1.000	1.000	1.000

requirements. This was more pronounced at the overall test level: with the GDI method, only approximately 10–54% of the tests had an adequate attribute coverage among the conditions, whereas the other three methods ensured that every test is so.

As shown in **Table 6**, both the MDGDI and MLGDI methods yielded a perfect attribute balancing, with 100% of the tests under all the conditions fulfilling all attribute coverage, or 100% of these tests having three or more items measuring each of the five attributes.

In addition, the ABI of MDGDI and MLGDI incorporates the dynamic balance of test attributes. Consequently, in the entire test process, the measurement frequency of all attributes is relatively balanced; that is, the difference between the maximum and minimum number of attribute measurements are kept within a given range, which increases the item selection efficiency. Therefore, MDGDI and MLGDI have a higher attribute classification accuracy than do MGDI, MMGDI, SWDGDI, and CP_SWDGDI.

DISCUSSION AND CONCLUSION

Cognitive diagnostic CAT captures the advantages of both CDA and CAT, allowing the individualized diagnostic feedback with fewer items. In this article, two new item selection methods, the MLGDI and MDGDI, were introduced, and their efficiency were compared with the existing methods. The results indicated that the MDGDI method successfully balanced the attribute coverage in CD-CAT and the MLGDI method simultaneously achieved balance over the attribute coverage and ensured the test security.

Both the MDGDI and the MLGDI outperformed the GDI, MMGDI, SWDGDI, and CP_SWDGDI in terms of the classification accuracy. Compared with MDGDI, MLGDI

provides a better item exposure control. The studies also showed that items with HD or high correlations among attributes provided better classification rates.

MDGDI and MLGDI Have Higher Pattern Determination Rates

The study demonstrated that MDGDI and MLGDI had a higher attribute correct classification rate than GDI, MMGDI, SWDGDI, and CP_SWDGDI under the different conditions. The PCCRs of MMGDI, SWDGDI, and CP_SWDGDI were worse than those of MDGDI and MLGDI. This was attributable to the multiplicative form of the attribute balance indicator in the MMGDI (Cheng, 2010). In such a form, negative–negative–positive cases can occur, which reduces the item selection efficiency of MMGDI. In addition, in the process of the prophase research item, because of the ABI, SWDGDI and CP_SWDGDI (Lin and Chang, 2018) attribute the propensity to choose more items. Specifically, Lin and Chang (2018) found that compared with the simple q vector (i.e., a vector with less or a single measurement attribute), an excessively complex q vector (i.e., one with multiple measured attributes) reduces the classification accuracy of the measurement (Madison and Bradshaw, 2015; Huebner et al., 2018). The MD index was adopted in MDGDI and MLGDI to achieve attribute coverage balance. Consequently, the measurement drying of all attributes are relatively balanced in the whole test process, that is, the deviation between the minimum and maximum number of attributes measured is within a given range. Therefore, the attributes of MDGDI and MLGDI are more balanced in the test process, which reduces the interference of the original selection strategy of the index and disallows the selection strategy from being more inclined to select some types of items due to the addition of the ABI. Therefore, MDGDI and MLGDI have a higher attribute classification accuracy.

MLGDI Provides a Better Exposure Control and High Attribute Classification Accuracy

Studies have shown that MDGDI has the highest attribute correct classification rate among MGDI, MMGDI, SWDGDI, CP_SWDGDI, MDGDI, and MLGDI. However, MDGDI also has problems such as the item overexposure, high TOR, and overuse of some items. Therefore, we added the restricted exposure index to MDGDI to construct the MLGDI item selection method. We found that MLGDI (1) greatly reduced the exposure rate and TOR but still had a high attribute classification accuracy and (2) had a pattern determination rate that was second only to MDGDI. In addition, the selected item strategy of CP_SWDGDI considers both the attribute coverage balance and exposure control. The exposure control method of CP_SWDGDI contains two key parameters, namely the adjustment information interval parameter s and the exposure parameter r, which is the maximum exposure required for a specific test purpose. Therefore, CP_SWDGDI uses the exposure parameter$r$and the adjustment information interval parameters$s$to control the item exposure. However, determining the value of the information interval parameters$s$that is appropriate for a given test length and number of attributes is difficult, which makes the CP_SWDGDI difficult to apply in practice (Zheng and Wang, 2017). Compared with CP_SWDGDI, MLGDI only realizes the exposure control of the participant through the exposure parameter$r$. The absence of the information interval parameters$s$and the need to determine the appropriates$s$value under the different conditions makes MLGDI more practicable. In addition, MLGDI has a higher attribute classification accuracy than does CP_SWDGDI. In conclusion, MLGDI can better meet the requirements of exposure control and has a high attribute classification accuracy, making MLGDI more suitable for practical applications.

The simulation studies yielded the following conclusions:

(1) When only the accuracy of attribute classification and attribute-coverage requirement are considered, MDGDI had the best attribute classification accuracy among GDI, MMGDI, SWDGDI, and MDGDI.
(2) When the accuracy of attribute classification, attribute-coverage requirement, and control item exposure rate are considered, MLGDI had the best attribute classification accuracy among GDI, CP_SWDGDI, and MLGDI.
(3) Adding a restricted item exposure mechanism to the item selection method will reduce the classification accuracy of the attributes of the item selection method.
(4) Compared with GDI, MMGDI, SWDGDI, CP_SWDGDI, and MDGDI, MLGDI can better achieve the attribute-coverage requirement, control item exposure rate, and attribute correct classification rate.

Directions for Future Research

Future studies can build upon our analysis of the performance of the six item selection strategies (GDI, MMGDI, SWDGDI, CP_SWDGDI, MDGDI, and MLGDI) under different conditions. (1) In the simulations, we found that MDGDI and MLGDI methods can be well used in the selection of CD-CAT projects. However, simulation results are limited to the given simulation conditions. Therefore, to further demonstrate the effectiveness of our method, future research should involve the practical application of the two proposed methods in the use of CD-CAT item banks. (2) For simplicity, we assumed that the correlations between the attributes were set at 0.5 and 0.8 in our simulations. Future studies can test the effectiveness of our thematic strategies (MDGDI and MLGDI) under more realistic conditions. (3) Future studies can extend the MDI and ML indexes to the method based on the expected Shannon entropy and the method based on the *a posteriori*, weighted KL information.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JL and LM proposed the original concept, designed the fundamental study of the manuscript, wrote the simulation study code, and organized the article. All authors contributed to the manuscript revision.

FUNDING

This research was funded by Open Research Fund of College of Teacher Education, Zhejiang Normal University (No. jykf20050).

REFERENCES

Chang, H., and Ying, Z. (1996). A global information approach to computerized adaptive testing. *Appl. Psychol. Meas.* 20, 213–229. doi: 10.1177/014662169602000303

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika* 80, 1–20. doi: 10.1007/s11336-014-9401-5

Cheng, Y. (2009a). *Computerized Adaptive Testing for Cognitive Diagnosis*. ed. D. J. Weiss, *Presented at the 2009 GMAC Conference*.

Cheng, Y. (2009b). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method. *Educ. Psychol. Meas.* 70, 902–913. doi: 10.1177/0013164410366693

Cheng, Y., and Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br. J. Math. Stat. Psychol.* 62, 369–383. doi: 10.1348/000711008X304376

Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: John Wiley.

Gierl, M. J., Wang, C., and Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT© . *J. Technol. Learn. Assess.* 6, 1–53.

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*. Urbana-Champaign, IL: University of Illinois. unpublished doctoral dissertation.

Hartz, S. M., and Roussos, L. A. (2008). *The Fusion Model for Skill Diagnosis: Blending Theory with Practicality (Research Report No. RR-08-71)*. Princeton, NJ: Educational Testing Service.

Huebner, A., Finkelman, M. D., and Weissman, A. (2018). Factors affecting the classification accuracy and average length of a variable-length cognitive diagnostic computerized test. *J. Comput. Adapt. Test.* 6, 1–14. doi: 10.7333/1802-060101

Kang, H. A., Zhang, S. S., and Chang, H. H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *J. Educ. Meas.* 54, 165–183. doi: 10.1111/jedm.12139

Kaplan, M., de la Torre, J., and Ramón Barrada, J. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Appl. Psychol. Measurement* 39, 167–188. doi: 10.1177/0146621614554650

Lee, Y. S., de la Torre, J., and Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. *Asia Pac. Educ. Rev.* 13, 333–345. doi: 10.1007/s12564-011-9196-3

Leighton, J., and Gierl, M. (eds) (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press.

Lin, C. J., and Chang, H. H. (2018). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educ. Psychol. Meas.* 79:001316441879063.

Madison, M. J., and Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educ. Psychol. Meas.* 75, 491–511. doi: 10.1177/0013164414539162

Mao, X., and Xin, T. (2013). The application of the monte carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Appl. Psychol. Meas.* 37, 482–496. doi: 10.1177/0146621613486015

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

Wang, C., Chang, H. H., and Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *J. Educ. Meas.* 48, 255–273. doi: 10.1111/j.1745-3984.2011.00145.x

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Appl. Psychol. Meas.* 6, 473–492. doi: 10.1177/014662168200600408

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educ. Psychol. Meas.* 67, 41–58. doi: 10.1177/0013164406288164

Xu, X., and Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika* 71, 121–137. doi: 10.1007/s11336-003-1154-5

Xu, X. L., Chang, H. H., and Douglas, J. (2003). *A Simulation Study to Compare CAT Strategies for Cognitive Diagnosis*. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL.

Yao, L., and Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale profificiency estimation and classifification. *Appl. Psychol. Meas.* 31, 83–105. doi: 10.1177/0146621606291559

Wang, Y., Sun, X., Chong, W., and Xin, T. (2020). Attribute Discrimination index-based method to balance attribute coverage for short-length cognitive diagnostic computerized adaptive testing. *Front. Psychol.* 11:224. doi: 10.3389/fpsyg.2020.00224

Zheng, C. J., 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: 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 Li, Ma, Zeng and Kang. 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 May 2021
doi: 10.3389/fpsyg.2021.665441

An Improved Parameter-Estimating Method in Bayesian Networks Applied for Cognitive Diagnosis Assessment

Ling Ling Wang[1,2]*, *Tao Xin*[3]* and *Liu Yanlou*[4]

[1] School of Educational Science, Shenyang Normal University, Shenyang, China, [2] Faculty of Psychology, Beijing Normal University, Beijing, China, [3] Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, China, [4] China Academy of Big Data for Education, Qufu Normal University, Qufu, China

Bayesian networks (BNs) can be employed to cognitive diagnostic assessment (CDA). Most of the existing researches on the BNs for CDA utilized the MCMC algorithm to estimate parameters of BNs. When EM algorithm and gradient descending (GD) learning method are adopted to estimate the parameters of BNs, some challenges may emerge in educational assessment due to the monotonic constraints (greater skill should lead to better item performance) cannot be satisfied in the above two methods. This paper proposed to train the BN first based on the ideal response pattern data contained in every CDA and continue to estimate the parameters of BN based on the EM or the GD algorithm regarding the parameters based on the IRP training method as informative priors. Both the simulation study and realistic data analysis demonstrated the validity and feasibility of the new method. The BN based on the new parameter estimating method exhibits promising statistical classification performance and even outperforms the G-DINA model in some conditions.

Keywords: Bayesian Networks, ideal response pattern, cognitive diagnostic model, parameter estimating method, cognitive diagnostic assessment

OPEN ACCESS

Edited by:
Alexander Robitzsch,
IPN - Leibniz Institute for Science and Mathematics Education, Germany

Reviewed by:
Kang Xue,
NWEA, United States
Lihong Song,
Jiangxi Normal University, China

***Correspondence:**
Ling Ling Wang
wling-007@163.com
Tao Xin
xintao@bnu.edu.cn

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

Received: *08 February 2021*
Accepted: *26 April 2021*
Published: *24 May 2021*

Citation:
Wang LL, Xin T and Yanlou L (2021) An Improved Parameter-Estimating Method in Bayesian Networks Applied for Cognitive Diagnosis Assessment. Front. Psychol. 12:665441. doi: 10.3389/fpsyg.2021.665441

INTRODUCTION

Cognitive diagnosis models (CDMs) are psychometric models developed mainly to assess students' specific strengths and weaknesses on a set of finer-grained skills or attributes within a domain. Researchers have developed many kinds of CDMs for cognitive diagnostic assessment (CDA), including the rule space model (Tatsuoka, 1985), the attribute hierarchy model (AHM, Leighton et al., 2004), the deterministic inputs noisy and gate model (DINA, Junker and Sijtsma, 2001), the Deterministic Input, Noisy Output "Or" gate model (DINO; Templin and Henson, 2006), the Log-Linear Cognitive Diagnosis Model (LCDM, Henson et al., 2009), the General Diagnostic Model (GDM, von Davier, 2005), the G-DINA model (de la Torre, 2011), and so on. In addition to these traditional CDMs, some researchers also proposed to use Bayesian networks (BNs, Pearl, 1988) in CDA (Mislevy, 1995; Mislevy et al., 1999; Almond et al., 2007, 2015; Wu, 2013; Levy and Mislevy, 2017). These literatures have documented the efficiency of BN used for diagnostic assessment, especially in the construction of large-scale evaluation system (e.g., Steinberg et al., 2003; Levy and Mislevy, 2004; Almond et al., 2015).

In BNs for CDA, the observable variables are the item responses and the latent variables are the knowledge states that can be treated as missing data. In Bayesian Network, like the traditional CDMs, the parameter estimation is also needed. The parameter estimation algorithms include the Markov Chain Monte Carlo (MCMC) method, the expectation maximization (EM) method, and the gradient descent (GD) learning method. In educational assessment, the MCMC method was used extensively (Wu, 2013; Almond et al., 2015; Levy and Mislevy, 2017), whereas the EM and the GD algorithms are rarely used. Some researches (Lauritzen, 1995; Russell et al., 1995; Korb and Nicholson, 2010) have described the algorithms. The EM algorithm is an iterative method to find (local) maximum likelihood estimates of parameters in BNs, in which the model depends on unobserved latent variables. EM iteration alternates between performing an expectation (E) step, which uses regular BN inference with the existing BN to compute the expected value of all the latent variables (missing data), and a maximization (M) step, which computes parameters maximizing the log-likelihood of BN given the resulting extended data (i.e., original data plus the expected value of missing data). These parameter estimates are used to determine the distribution of the latent variables in the next E step. The GD learning method tries to minimize the objective function of the negative log likelihood to determine the BN parameters. This method can find a better BN by using BN inference to calculate the direction of the steepest gradient and to change the parameters to follow the steepest direction of the gradient (i.e., maximum improvement). Actually, it uses a much more efficient approach than always taking the steepest path, by taking into account its previous path, which is why it is called the conjugate gradient descent (Norsys Inc., 2004). However, both the EM and the GD methods are sensitive to the initial condition and liable to be trapped in a local optimization. Moreover, there is a fundamental issue with using the EM and the GD algorithms in BN. In educational applications, the variables (both proficiency variables and observable outcomes) are usually ordinal, and the parameters are monotonic (greater skill should lead to better item performance). However, most parameter estimation algorithms do not force monotonicity constraints on the parameters. In the case when informative priors are available, they might support to keep the estimates from violating the monotonicity assumptions (von Davier and Lee, 2019). Almond (2015) describes a flexible parameterization for conditional probability tables based on item response theory (IRT) that preserves monotonicity and extends the EM algorithm to a generalized EM algorithm. However, this method needs to map each configuration of parent variables to an effective theta, a point on a latent normal scale, and calculate the conditional probability tables using the IRT model. That is to say, this method needs to introduce other models to assist the parameterization.

Besides the above CDMs, researchers also proposed to use machine learning algorithms for CDA, such as neural networks (NNs, Gierl et al., 2008; Shu et al., 2013; Cui et al., 2016), and support vector machines (SVMs, Liu and Cheng, 2018). Gierl et al. (2008) were the first to propose using the ideal response patterns (IRPs) and the corresponding attribute profiles constructed in the cognitive model of the CDA as the training data of NNs. The ideal response patterns are the students with the knowledge states (attribute profiles) to answer the questions with no slipping and guessing errors. The attribute profiles are the entire possible attribute combinations in the cognitive model. If there are K attributes in the test and the attributes are independent, then the number of the attribute profiles is 2^K, while in the AHM, the number is $<2^K$ due to the hierarchical relationship of the attributes. Shu et al. (2013) also used this approach to train the NN and investigated its performance in small samples. They pointed out that, "In doing this all items are assumed to behave in the same way and so no differentiation is made between items with respect to item quality. A limitation of the use of a NN analysis with IRP is that all 'calibration' of the model is based on connections between response patterns and attribute patterns calibrated in the training process, no empirical information (i.e., students' response patterns) is used to influence the procedure." Thus, they proposed to combine the NN approach using IRP and a MCMC estimation algorithm or a Joint Maximum Likelihood Estimation (JMLE) algorithm. Inspired by this research, we propose to use the IRP data to estimate the BN parameters first, then continue to estimate the BN parameters by EM or GD methods, taking the previously estimated parameters as informative priors. The introduction of IRP into the EM or GD method can overcome the inaccurate estimation limitation caused by the violations of the monotonicity constraints. This combination also provides suitable starting values to continue the EM or GD estimation to overcome the local optimality problem. On the other hand, this method can also improve the performance of the IRP method without empirical information. And comparing to the solution in Almond (2015), we propose to apply the information contained in each CDA (IRP data) for the initial parameterization and do not need to introduce other models to assist the parameterization.

To demonstrate the effectiveness of our proposed approach, we performed a simulation study comparing the parameter estimating algorithms in BN, including the EM method, the GD method, the IRP training method, and the combinations of IRP and EM or GD methods. We considered the classification rates as the performance indicators of the algorithms and the G-DINA model's diagnostic classification performance as the evaluation criterion. Also, we carried out a real data analysis to demonstrate the validity of the proposed method.

Regarding the simulation method, usually various CDMs are adopted as the data-generating models. The simulation data is of great value in verifying the properties of the models. However, there are also some limitations in the CDM-based data. Real data do not necessarily conform to the assumptions of the CDMs. Using a CDM to process the data conforming to the same model assumption may overrate the properties of the model. Thus, other data-generating methods are needed. Wu (2013) proposed to generate data based on BNs, but the parameters of BNs were determined based on traditional CDMs, and the patterns of the simulated data by the BNs were essentially the same as the CDM-generated data. This research introduced a new simulation data generating method by using BNs to generate data with the empirical data's pattern and these simulated data do not necessarily satisfy the G-DINA assumption. To evaluate

the BN and G-DINA methods more objectively, we designed the simulation study using both the G-DINA model and the BN model to generate data.

OVERVIEW OF THE COGNITIVE DIAGNOSTIC MODELS AND BAYESIAN NETWORKS

The Cognitive Diagnostic Model

The G-DINA model is a general CDM and can be converted to the constrained CDMs by setting appropriate constraints (de la Torre, 2011). According to de la Torre (2011) notation, for a test with J items and K attributes, let K_j^*represent the number of required attributes for item j, and α_{lj}^* be the reduced attribute vector whose elements are the required attributes for item j. Then, the item response function of the G-DINA model is expressed as follows:

$$P(\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}$$

Where δ_{j0} is the intercept for item j; δ_{jk} is the main effect due to α_k; $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$; and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1 \ldots \alpha_{K_j^*}$. This function is the identity link function and the other two link functions can also be used to express the G-DINA model according to de la Torre (2011), namely, the logit link and the log link.

The Bayesian Networks

Bayesian networks (BN, Pearl, 1988) are a notation for expressing the joint distribution of probabilities over both observed and latent variables. They are used to represent knowledge in an uncertain domain and have been successfully applied in computer science, especially in artificial intelligence. A BN consists of a directed acyclic graph (DAG) to represent their structure and a corresponding set of conditional probability distributions to represent the parameters (Culbertson, 2016; Hu and Templin, 2019). In this graph, each node represents a random variable and each edge constitutes the probabilistic dependence relationship among the variables represented by the two nodes that are joined. Each pair of connected nodes has a directed edge flowing from a "parent" node to a "child" node. A conditional probability distribution is specified for each node, given its parent nodes. For discrete random variables, this conditional probability is described with conditional probability tables (CPTs). The structure of BNs effectively reflects the conditional independent relationship between the variables. According to the conditional independent relationship, a joint probability distribution is decomposed into a product of a series of conditional distributions, which reduces the number of parameters required to define the joint probability distribution of the variables, and to compute the posterior probabilities efficiently, given the data.

A BN applied for cognitive diagnosis is shown in **Figure 1**. In **Figure 1**, the item nodes (observed variables) can be connected to the attribute nodes (hidden variables), similar to the Q-matrix that represented relationships between attributes and items in traditional CDMs (Tatsuoka, 1983; Almond, 2010). The attribute nodes can also be connected to each other, denoting the attribute hierarchical relationships. According to the specific structure of the BN, the joint probability distribution of all the variables can be factorized into a product of a series of conditional probabilities. Once the structure of the BN has been determined, we need to specify these parameters in the BN first in order to make an inference, i.e., specify the conditional probabilities and the marginal probabilities. If all the nodes of a discrete BN are fully observed in a sample, the CPTs can be updated via a simple counting algorithm. This is the case when we train the BN using the IRP and the corresponding attribute profile data. If there are missing values or latent variables, then the CPTs can be calculated using the EM algorithm, the GD method, or the MCMC method.

After obtaining the structure and parameters of the BN, we can use the BN to predict the students' knowledge state by probability inference. According to the Bayesian Theorem, the probability inference is when the posterior probability of the hidden variables (attributes) is calculated using the values of the observed variables (e.g., item response) as input. This process is also called network propagation or belief updating, which can be realized by a number of algorithms, such as the message passing algorithm (Pearl, 1988), or trees of cliques (Lauritzen and Spiegelhalter, 1988; Jensen, 1996).

METHODS

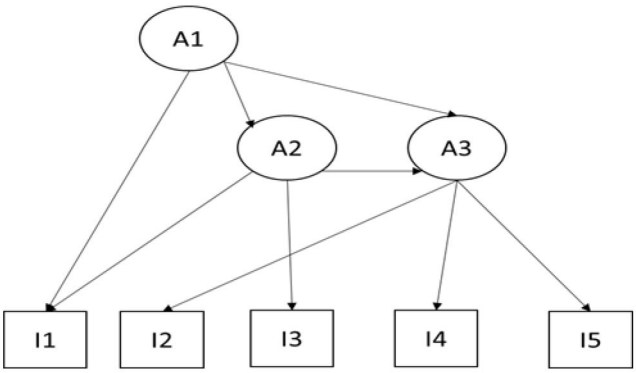
Simulation Design

The study presented in this paper compared several parameters estimating methods in Bayesian networks for cognitive diagnosis. We conducted a simulation study to evaluate the performance of different parameter estimating methods. The Q-matrix (see **Table 1**) with 30 items and 5 attributes in de la Torre (2011) was adopted. The interconnections between attributes and test items in the Q-matrix could also be represented in a BN graph in which the arcs connecting two nodes denoted the associations between the corresponding items and the attributes.

The attribute patterns were generated in two different methods. The first method sampled attribute patterns from a uniform distribution with probability $1/2^K$ of every possible value. In the second method, the discrete attribute pattern α was linked to an underlying multivariate normal distribution, $MVN(0_K, \Sigma)$, with covariance matrix Σ, structured as

$$\sum = \begin{pmatrix} 1 & & \rho \\ & \cdots & \\ \rho & & 1 \end{pmatrix}$$

and $\rho = 0.5$. Let $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})$ denote the K-dimensional vector of latent continuous scores for examinee i. The attribute



$$P(I_1,\cdots,I_5,A_1,A_2,A_3)=P(I_1\mid A_1,A_2)P(I_2\mid A_3)P(I_3\mid A_2)\times$$
$$P(I_4\mid A_3)P(I_5\mid A_3)P(A_1)P(A_2\mid A_1)P(A_3\mid A_1,A_2)$$

FIGURE 1 | A Bayesian network applied for cognitive diagnosis.

TABLE 1 | The Q-Matrix of the simulation study.

	A1	A2	A3	A4	A5		A1	A2	A3	A4	A5
I1	1	0	0	0	0	I16	0	1	0	1	0
I2	0	1	0	0	0	I17	0	1	0	0	1
I3	0	0	1	0	0	I18	0	0	1	1	0
I4	0	0	0	1	0	I19	0	0	1	0	1
I5	0	0	0	0	1	I20	0	0	0	1	1
I6	1	0	0	0	0	I21	1	1	1	0	0
I7	0	1	0	0	0	I22	1	1	0	1	0
I8	0	0	1	0	0	I23	1	1	0	0	1
I9	0	0	0	1	0	I24	1	0	1	1	0
I10	0	0	0	0	1	I25	1	0	1	0	1
I11	1	1	0	0	0	I26	1	0	0	1	1
I12	1	0	1	0	0	I27	0	1	1	1	0
I13	1	0	0	1	0	I28	0	1	1	0	1
I14	1	0	0	0	1	I29	0	1	0	1	1
I15	0	1	1	0	0	I30	0	0	1	1	1

pattern $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$ was determined by

$$\alpha_{ik} = \begin{cases} 1, if \theta_{ik} \geq \Phi^{-1}(\frac{k}{K+1}) \\ 0, \text{otherwise} \end{cases}$$

The item response data that were used in our experiments were generated using the G-DINA model.

The parameter setting include two conditions, one is that both the g and s parameters in each item are fixed at 0.2, the other is a mixed test with high-discriminating items and low-discriminating items (10 items with $g = 0.3, s = 0.3$; 10 items with $g = 0.2, s = 0.2$; and 10 items with $g = 0.1, s = 0.1$), which is more realistic in practice. The success probabilities of the examinees who master none and all attributes [i.e., P(0) and P(1)] were fixed as g and $1 - s$, and the success probabilities of examinees with other attribute patterns were randomly generated from the distribution of Unif [P(0), P(1)] with monotonic constraint. All the above simulated data were generated by CDMs. We also used BNs to generate the data. When using BNs to generate the simulated data, the BN parameters need to be determined first because these parameters reflect the patterns behind the generated simulation data. Two datasets were used to estimate the BN parameters (i.e., the CPTs), including the simulated data previously generated by G-DINA, and the fraction subtraction dataset. The fraction subtraction dataset contains 536 students, 15 items, and 5 attributes, and the Q-matrix was defined as given by de la Torre (2009). Two BN models were constructed to process the two datasets and the parameters of the BN models were estimated based on the two datasets using IRP–EM method. Then, we used the two BN models to generate two types of simulated datasets. The patterns behind the two sets of simulated data reflect the patterns of the two sets of data previously used to

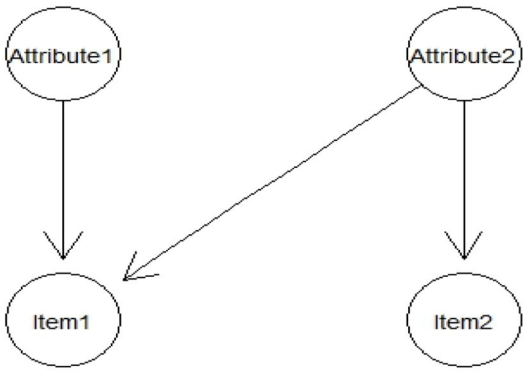


FIGURE 2 | A simple Bayesian network applied for cognitive diagnosis.

TABLE 2 | The ideal response pattern and the corresponding attribute profiles.

	Attribute 1	Attribute 2	Item 1	Item 2
Case 1	1	1	1	1
Case 2	1	0	0	0
Case 3	0	1	0	1
Case 4	0	0	0	0

TABLE 3 | The conditional probability table of the Bayesian network.

P (Attribute1)		**P (Attribute2)**	
Attribute 1 = 1	**Attribute 1 = 0**	**Attribute 2 = 1**	**Attribute 2 = 0**
2/4	2/4	2/4	2/4
P (Item 1|Attribute 1, Attribute 2)			
Attribute 1	**Attribute 2**	**Item 1 = 1**	**Item 1 = 0**
1	1	1	0
1	0	0	1
0	1	0	1
0	0	0	1
P (Item 2|Attribute 1, Attribute 2)			
Attribute 1	**Attribute 2**	**Item 2 = 1**	**Item 2 = 0**
1	1	1	0
1	0	0	1
0	1	1	0
0	0	0	1

estimate the parameters. In other words, the BN-generated data not only reflect the pattern of the G-DINA simulated data, but also the pattern of the real test data.

Accordingly, four types of datasets were simulated by CDMs (two level parameters × two different attribute profile sampling method) and two types of datasets were simulated by BNs. Each condition had two sample sizes of 500 and 1,000. And all the simulations are repeated 30 times to compute the average classification rate and the standard error.

The Realization of the IRP–EM and IRP–GD Methods

The BN and G-DINA models were applied to analyze the data. The parameter estimating method of BNs includes the EM method, the GD method, the IRP training method, and the combinations of IRP with EM or GD methods. The classification rates are considered as the performance indicators of the algorithms and the diagnostic classification performance of the G-DINA model is considered as the evaluation criterion. When using IRP and the corresponding attribute profiles to train the BN, the CPT of the BN can be obtained through counting the observed frequencies in the IRP training dataset using algorithms (Neapolitan, 2004; Lee and Corter, 2011). A sample BN based on the problem of diagnosing attributes is shown in **Figure 2**. **Table 2** presents the ideal response pattern and the corresponding attribute profiles; **Table 3** shows the respective CPTs through counting the observed frequencies in the training dataset.

The example with all the results presented above illustrated the BN training process using the IRP information. Two different types of IRP data were computed based on the DINA model and the DINO model representing the non-compensatory model and the compensatory model. After the BN was trained based on the IRP data, the GD or EM algorithm were used to estimate the parameters of BN treating the IRP parameters as priors. Generally, the DINA–IRP data were used first and when the GD or EM algorithms were trapped in the local optimality, the algorithms transfer to use the DINO–IRP data to train the BN model and continue to estimate the final parameters. No matter which IRP was used to train the BN, they provided the suitable starting values for the EM or the GD method and this avoided the algorithms to be trapped in the local optimality. The implementation of the parameter estimation algorithm is realized by Netica (www.norsys.com), a professional Bayesian network software. In this software, the IRP and the corresponding attribute patterns information is first entered into the constructed BN structure and the prior of the parameters in BN is obtained. After that, it continues to incorporate empirical data to update the BN posterior parameters based on the EM or GD algorithm.

RESULTS

We evaluated the performance of each algorithm under the conditions mentioned in the simulation design, using the average attribute classification rate (AACR) and the pattern classification rate (PCR) as the performance index for each condition. The standard error of each classification rate indicator is computed to evaluate the consistency of the methods. From **Table 4**, we could observe that in each condition, the PCRs of BNs based on EM or GD method were very low because the monotonic

TABLE 4 | The PCR of BNs and G-DINA from the data generated by G-DINA.

gs Level	Distribution	Sample size	EM_BN	GD_BN	IRP_BN	IRP_EM_BN	IRP_GD_BN	G-DINA
g=*s*=0.2	Uniform	1,000	0.196	0.162	0.320	0.611	0.627	0.622
		500	0.138	0.091	0.348	0.577	0.598	0.593
	Mvnorm	1,000	0.064	0.013	0.374	0.647	0.654	0.698
		500	0.112	0.084	0.344	0.627	0.641	0.681
Mixed *gs* (0.3, 0.2, 0.1)	Uniform	1,000	0.076	0.024	0.197	0.536	0.574	0.567
		500	0.002	0.000	0.190	0.471	0.536	0.533
	Mvnorm	1,000	0.036	0.038	0.263	0.582	0.644	0.677
		500	0.006	0.000	0.184	0.579	0.590	0.629

TABLE 5 | The AACR of BNs and G-DINA from the data generated by G-DINA.

gs Level	Distribution	Sample size	EM_BN	GD_BN	IRP_BN	IRP_EM_BN	IRP_GD_BN	G-DINA
g=*s*=0.2	Uniform	1,000	0.741	0.740	0.816	0.902	0.908	0.906
		500	0.350	0.307	0.829	0.896	0.898	0.898
	Mvnorm	1,000	0.736	0.660	0.831	0.916	0.921	0.933
		500	0.358	0.421	0.831	0.910	0.916	0.924
Mixed *gs* (0.3, 0.2, 0.1)	Uniform	1,000	0.740	0.714	0.757	0.880	0.887	0.885
		500	0.412	0.275	0.763	0.862	0.872	0.871
	Mvnorm	1,000	0.738	0.431	0.782	0.898	0.916	0.922
		500	0.371	0.287	0.696	0.894	0.898	0.907

constraint was not satisfied. When the IRP training method was applied to estimate the parameters of BNs, the PCR was improved to some extent but still very low. In all these datasets, the effect of sample size on accuracy was not straightforward. When we used the IRP data to estimate the BN parameters first and continue to estimate the BN parameters by EM or GD methods by regarding the previously estimated parameters as informative priors, the PCRs of the BNs based on these two methods were significantly improved compared to the former three methods. The PCR of BN based on the combination of IRP and GD methods (denoted as BN–IRP–GD method) was higher than the BN based on the combination of IRP and EM method (denoted as BN–IRP–EM method) in these G-DINA datasets. The sample size had a certain influence on the accuracy rate, especially in the four G-DINA datasets processed by the IRP–EM method. The performance gap between the 1,000 and 500 samples of the other conditions was not so high. When comparing with the G-DINA analysis, in the four datasets generated by G-DINA, the PCR of BN–IRP–GD method was a little higher than the PCR of G-DINA if the attribute patterns of the students conformed to the uniform distribution, and the PCR of G-DINA was higher than the BN–IRP–GD method if the attribute patterns of the students conformed to the multivariate normal distribution.

From **Table 5**, the results of AACR were similar to the PCR results. The AACRs of BNs based on EM or GD methods were very low and it was improved by training the BNs with the IRP data. The AACRs of BN–IRP–EM and BN–IRP–GD methods were improved further and the best result was achieved through the BN–IRP–GD method. When comparing with the G-DINA analysis, in the four datasets generated by G-DINA, the AACR of BN–IRP–GD method was a little higher than the AACR of G-DINA if the attribute patterns students conformed to the uniform distribution, and the AACR of G-DINA was higher than that of the BN-IRP-GD method if the attribute patterns of the students conformed to the multivariate normal distribution. We have also evaluated the consistency of the BN–IRP–EM method, the BN-IRP-GD method and the G-DINA method through standard errors of the PCR and AACR in **Tables 6**, **7**. The standard errors of the AACR were lower than that of the PCR. Most conditions are lower than 0.01. The BN–IRP–GD method and the BN–IRP–EM method have the similar level of standard errors as the G-DINA method in all conditions.

The PCR and AACR by BN and G-DINA models from the data generated by G-DINA and the data generated by BN based on the G-DINA data and fraction data are shown in **Table 8**. And the standard errors of each PCR and AACR are given in parentheses. In the initial BN construction, the BN parameters were estimated by IRP–EM method, then the BN models with determined parameters were used to simulate two types of datasets for analyzing. It can be expected that the results of the simulated datasets analyzed by the IRP–EM method to be higher than that analyzed by the IRP–GD method. Thus, these data were not analyzed by IRP–EM method to avoid overrating the IRP–EM method. The BN–IRP–GD method and the G-DINA model were used to process these data. In the dataset generated by BN based on the G-DINA data, the PCR and AACR of BN-IRP-GD method was lower than the G-DINA model, and this result was similar to the dataset directly generated by G-DINA as shown in **Table 6**.

TABLE 6 | The standard error of PCR by BNs and G-DINA from the data generated by G-DINA.

gs Level	Distribution	Sample size	IRP_EM_BN	IRP_GD_BN	G-DINA
s=*g*=0.2	Uniform	1,000	0.025	0.025	0.019
		500	0.032	0.024	0.030
	Mvnorm	1,000	0.014	0.027	0.019
		500	0.038	0.036	0.044
Mixed *gs* (0.3, 0.2, 0.1)	Uniform	1,000	0.027	0.021	0.024
		500	0.061	0.069	0.037
	Mvnorm	1,000	0.050	0.032	0.020
		500	0.030	0.036	0.050

TABLE 7 | The standard error of AACR by BNs and G-DINA from the data generated by G-DINA.

gs Level	Distribution	Sample size	IRP_EM_BN	IRP_GD_BN	G-DINA
s=*g*=0.2	Uniform	1,000	0.006	0.007	0.005
		500	0.002	0.008	0.009
	Mvnorm	1,000	0.006	0.006	0.005
		500	0.011	0.009	0.011
Mixed *gs* (0.3, 0.2, 0.1)	Uniform	1,000	0.008	0.007	0.007
		500	0.020	0.023	0.011
	Mvnorm	1,000	0.025	0.011	0.026
		500	0.008	0.060	0.067

These two datasets were essentially conforming to the ideal G-DINA model assumption. However, when empirical data is used to estimate the BN parameters and generate data based on this BN model, the simulated data pattern might violate the G-DINA assumption or at least does not conform to the assumption of the G-DINA model as ideally as the G-DINA generated data. The BN-IRP-GD method provides higher classification rate than the G-DINA model.

REAL DATA EXAMPLE

To demonstrate the real-world applicability of the BN method, we used the dataset on the buoyancy concept developed by Gao et al. (2020) for cognitive diagnosis. These data have seven attributes, namely, (A1) know that the buoyancy direction is vertically upward, (A2) identify not only the gravity but also the buoyancy exerted on an object that is afloat, suspended, or immersed in liquid, (A3) know that the density is an object property, whose value is the mass divided by the volume, but still invariant to mass or volume changes, (A4) understand the meaning of a displaced liquid volume, (A5) calculate the buoyancy magnitude by analyzing the forces on objects, (A6) understand Archimedes' Principle, and (A7) decide whether objects will float or sink by comparing the object and liquid densities. These seven attributes have the hierarchy relationship as displayed in **Figure 3**. The Q-matrix with 14 items is shown in **Table 9**. A total of 1,089 eighth-grade students were chosen as subjects from 10 schools located in five east-coast cities in China. After excluding the subjects with blank test answers, 1,036 subjects remained. Two physics experts were invited to label the 50 randomly selected students on their mastery of the above attributes.

According to the attribute hierarchy relationships and the Q-matrix, we constructed a BN structure and trained the BN based on the IRP data first, and continued to estimate the BN parameters by GD method by regarding the previously estimated parameters as informative priors. The attribute patterns of the students could be predicted by BN based on the item response data. These data were also analyzed by the G-DINA model. From the previously selected 50 subjects, the estimated attribute patterns by the BN and G-DINA methods were compared with the labeling results obtained by the experts, and the PCR and AACR were calculated. **Table 10** showed the agreement percentage of the BN and G-DINA analysis with the experts' labeling attribute patterns in the randomly selected 50 samples. We can see that the BN–IRP–GD method can achieve promising classification performance, and even higher than the G-DINA model in PCR. The agreement percentage between the BN and experts' labeling in each attribute was also a little higher than that observed between the G-DINA and experts' labeling. These results demonstrated the validity and feasibility of the BN–IRP–GD method.

DISCUSSION

In this research, we conducted a simulation study comparing the parameter estimating algorithms in BN, including the EM method, the GD learning method, the IRP training method, and the combinations of IRP and EM (IRP–EM) or GD (IRP–GD) methods. The classification rates are considered as the performance indicators of the algorithms and the performance of G-DINA was adopted as a criterion to evaluate the performance of the improved parameter estimating method in BN. Real data analysis is followed to demonstrate the validity of the proposed method. The results show that the classification performances of the BN–IRP–EM and BN–IRP–GD methods are promising and even higher than the G-DINA model in certain conditions. The classification rates of EM and GD method are very low due to the reason that monotonic constraints may not be upheld. Moreover, both the EM and the GD methods are sensitive to the initial condition and liable to be trapped in a local optimality. We

TABLE 8 | The PCR and AACR by BNs and G-DINA from the data generated by G-DINA and the data generated by BNs based on the G-DINA data and Fraction data.

DATA/model	PCR		AACR	
	GDINA	IRP_GD_BN	GDINA	IRP_GD_BN
G-DINA-gen	0.701 (0.020)	0.658 (0.029)	0.931 (0.005)	0.919 (0.007)
BN-gen-based on-GDINA data	0.643 (0.015)	0.618 (0.008)	0.910 (0.046)	0.901 (0.008)
BN-gen-based on-Fraction data	0.502 (0.082)	0.670 (0.004)	0.850 (0.033)	0.908 (0.002)

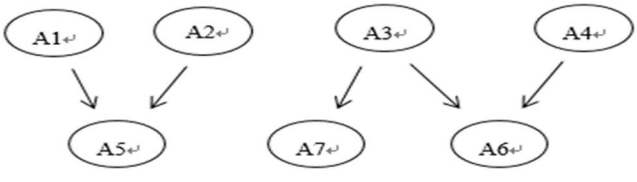


FIGURE 3 | The attribute hierarchy relationship of the buoyancy.

introduce the IRP training method to estimate the parameters of BN first, then continue to update the posterior values of the parameters by taking the previously estimated parameters as informative priors. The initial parameters estimated by IRP training are better starting values for EM and GD algorithms and can also solve the local optimality problems. The IRP data are constructed theoretically in every CDA and reflect a fair amount of test information. However, the IRP training method has no empirical information and the IRP–EM and IRP–GD methods can also solve these limitations and improve the classification rate of the IRP training method if be provided enough empirical samples.

This study demonstrated that, in the data generated by CDMs condition, the analysis by G-DINA was a little higher than the BN analysis based on the IRP–EM or IRP–GD method, which was as expected because the pattern of the data conformed to the assumption of G-DINA. However, in the data generated by BN based on the empirical data condition, the BN analysis based on the IRP-GD method outperformed the G-DINA analysis. But we cannot conclude that the BN methods are superior than the G-DINA model, just as we cannot claim that the G-DINA model are better than the BN methods in the first condition of the G-DINA simulated data. These two simulation conditions jointly demonstrated that the BN and G-DINA models each has their own advantages in the simulated data. From the above two conditions, we can see that the BN-data-generating method is more flexible. Different from the traditional CDMs, there are no explicit assumptions in BN. The parameters of BN reflect the patterns of the generated data and the parameters are determined by the data used to estimate the BN model. When the G-DINA data were used to estimate the parameters of the BN model, the generated data based on this BN model are conforming to the G-DINA assumption. Similarly, when the empirical data were used to estimate the parameters of the BN model, the generated data based on this BN model reflect the pattern of the empirical data, and these data do not necessarily satisfy the G-DINA assumption.

In the empirical study, the consistency of classification between BN and experts' judgment are higher than that between G-DINA and experts' judgment. Although both the average attribute classification rate and the pattern classification rate in BN are higher than the G-DINA model in this case, we are not intended to advocate that the BN model is more beneficial than the G-DINA model. These results are merely implying that the effectiveness of improved parameter-estimating method in BN is validated. In fact, the G-DINA model is a powerful framework that covers a wide variety of psychometric models, and through the comparison with the G-DINA, it demonstrates that if there are appropriate parameter-estimating methods, BN is also another great modeling framework that is compatible with many different types of data patterns in cognitive diagnostic assessment.

Shu et al. (2013) proposed to combine the NN trained by IRP with the traditional CDM to improve the performance of the traditional CDM. In their study, the IRP information cannot directly be used in traditional CDM, and the NN cannot incorporate the empirical information of samples. However, in BN, the parameter estimation can be accumulated according to the Bayesian theorem. First, the IRP and the corresponding attribute profile data are used to train the BN as a supervised learning method. Then the parameters obtained by the IRP training are regarded as the prior of the next stage of parameter estimation. When using EM or GD to continue the estimation, only the practical item responses are available, and the values of attributes are unknown and can be regarded as latent variable, thus this estimation stage is an unsupervised learning method. And the final parameters are the posterior probability according to the Bayesian theorem. An independent and complete model can accommodate the functions of both the NN and the traditional CDM (Shu et al., 2013). In other words, the supervised learning and unsupervised learning can be realized in the same model. From this perspective, the proposed method is a new progress.

BNs have been used extensively in the artificial intelligence community as student models for intelligent tutoring systems. The review article by Culbertson (2016) outlined many existing research studies on BN in educational assessment and it pointed that "BNs readily lend themselves to modular model building in which fragments of BN may be developed separately and combined freely based on a wide variety of types of

TABLE 9 | The Q-matrix for buoyancy concept learning.

Item/attribute	A1	A2	A3	A4	A5	A6	A7
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	0	0	0	1	0	0	0
5	0	0	1	0	0	0	1
6	0	0	1	0	0	0	1
7	1	1	0	0	1	0	0
8	1	1	0	0	1	0	0
9	0	0	1	1	0	0	0
10	0	0	1	1	0	1	0
11	1	1	1	0	1	0	1
12	0	0	1	1	0	1	1
13	1	1	1	1	0	1	0
14	1	1	1	1	1	1	1

TABLE 10 | The agreement of BN or G-DINA analysis with the experts' labeling attribute patterns in randomly selected 50 samples.

	A1	A2	A3	A4	A5	A6	A7	AACR	PCR
G-DINA	0.96	0.96	0.92	0.94	0.88	0.76	0.88	0.90	0.50
BN	0.98	0.98	0.96	0.94	0.86	0.82	0.96	0.93	0.62

relationships." This aspect of a BN renders it a great potential for constructing the intelligent diagnostic assessment system to realize personalized learning. And this goal of realizing personalized learning is shared by CDA. But the BN can accommodate more node variables (e.g., Levy and Mislevy, 2004; Shute et al., 2008), which is more attractive to the educational assessment practice. In fact, in CDA, the MCMC method was used extensively for estimating the parameters of BNs (Wu, 2013; Almond et al., 2015; Levy and Mislevy, 2017). However, the MCMC method has some limitations. The first concern is the computational complexity and the efficiency, especially in cognitive diagnosis that needs immediate feedback of diagnostic information for instructions and learnings. Also, this method depends on the starting values to obtain stable and reliable estimates. The improved parameter-estimating method proposed in this article provides another approach to realize CDA based on BNs. And this method can be embedded into the diagnostic assessment system to achieve immediate feedback for personalized instruction. From the practical view, the optimization of the EM or GD method can provide more efficient computation compared to the MCMC method and can promote the CDA to be used in classroom assessment in which the instant feedback is needed. Moreover, due to the feature of modular model building and freely combining, BN can accommodate more attributes than traditional CDMs, which is needed for a practical diagnostic assessment system.

Naturally, we are aware of certain limitations in this study. The evaluation indicator of the BNs is relatively single and only the classification rate is adopted as the criteria of the BN's performance in this article, similar to what Sinharay (2006) described, "Model checking for BNs is not straightforward. For even moderately large number of items, the possible number of response patterns is huge, and the standard χ^2 test of goodness of fit does not apply directly." Sinharay (2006) applied the posterior predictive model checking (PPMC, Rubin, 1984) method to assess several aspects of fit of BNs applied in educational assessment. The PPMC method is a popular Bayesian model checking tool and combines well with the MCMC algorithms (e.g., Gelman et al., 2003). From the model comparison perspective, the most popular model fit indices are the AIC (Akaike, 1973) and BIC (Schwarz, 1978) criteria, which require knowing the number of free parameters to be estimated in the model. It is not so straightforward in complex BN model to count the number of parameters. Spiegelhalter et al. (2002) introduce a measure called DIC, which includes a Bayesian notion of dimensionality. But the DIC measure is also based on the MCMC algorithms. Thus, developing the new model fit indices of BN based on the IRP–EM or IRP–GD parameter-estimating method is a desired investigation area for future studies.

Additionally, Shu et al. (2013) combined the NN analysis based on the IRP data with the MCMC method to estimate the DINA model parameters in a small sample size. In traditional CDMs, the item parameters can be estimated to reflect the test quality. However, BNs have no straightforward item parameters to evaluate test quality. In future studies, we plan to apply the sensitivity analysis of BNs to determine the key attributes that influence each item response performance and the key items that determine each attribute diagnosis. Sensitivity analysis refers to an uncertain analysis technique to quantify the contribution of an observation node towards the uncertainty of the node of interest in BN. The sensitivity analysis results might provide insights into

the knowledge structure and cognitive process of the students and assist in instruction design and personalized learning.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Academic Ethics Committee in Shenyang Normal University. Written informed consent to participate in this study was provided by the participants legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

LW has made substantial contributions to propose the new algorithm, developed the R code, and wrote the first draft of the manuscript. TX and LW designed the research. YL collected the empirical data. YL and TX contributed to the final manuscript revision. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the National Social Science Foundation of Chinahe Major Project of 2019 (No. 19ZDA359) and National Natural Science Foundation of China (Nos. U1911201 and 32071093).

REFERENCES

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory,* eds B. N. Petrov and F. Caki (Budapest: Akademiai Kiado), 267–281.

Almond, R. G. (2010). I can name that Bayesian network in two matrixes!. *Int. J. Approx. Reas.* 51, 167–178. doi: 10.1016/j.ijar.2009.04.005

Almond, R. G. (2015). "An IRT-based parameterization for conditional probability tables," in *Bayesian Modelling Application Workshop at the Uncertainty in Artificial Intelligence (UAI) Conference*, eds J. M. Agosta and R. N. Carvalho (Amsterdam). Available online at: http://pluto.coe.fsu.edu/RNetica/

Almond, R. G., DiBello, L. V., Moulder, B., and Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *J. Educ. Meas.* 44, 341–359. doi: 10.1111/j.1745-3984.2007.00043.x

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., and Williamson, D. (2015). *Bayesian Networks in Educational Assessment*. New York, NY: Springer.

Cui, Y., Gierl, M., and Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: an artificial neural network approach. *Educ. Psychol.* 36, 1065–1082. doi: 10.1080/01443410.2015.1062078

Culbertson, M. J. (2016). Bayesian networks in educational assessment: the state of the field. *Appl. Psychol. Meas.* 40, 3–21. doi: 10.1177/0146621615590401

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *J. Educ. Behav. Stat.* 34, 115–130. doi: 10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

Gao, Y. Z., Zhai, X. M., Andersson, B., Xin, T., and Zeng, P. F. (2020). Developing a learning progression of buoyancy to model conceptual change: a latent class and rule space model analysis. *Res. Sci. Educ.* 50, 1369–88. doi: 10.1007/s11165-018-9736-5

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

Gierl, M. J., Zheng, Y., and Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *J. Educ. Meas.* 45, 65–89. doi: 10.1111/j.1745-3984.2007.00052.x

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Hu, B., and Templin, J. (2019). Using diagnostic classification models to validate attribute hierarchies and evaluate model fit in Bayesian networks. *Multivar. Behav. Res.* 55, 300–311. doi: 10.1080/00273171.2019.1632165

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. New York, NY: Springer.

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

Korb, K., and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence*, 2nd Edn. London: Chapman and Hall.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Comput. Stat. Data Anal.* 19, 191–201. doi: 10.1016/0167-9473E0056-A

Lauritzen, S. L., and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Ser. B* 50, 157–224. doi: 10.1111/j.2517-6161.1988.tb01721.x

Lee, J., and Corter, J. E. (2011). Diagnosis of subtraction bugs using Bayesian networks. *Appl. Psychol. Measur.* 35, 27–47. doi: 10.1177/0146621610377079

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* 41, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Levy, R., and Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer- based interactive assessment. *Int. J. Test.* 4:333369. doi: 10.1207/s15327574ijt0404_3

Levy, R., and Mislevy, R. J. (2017). *Bayesian Psychometric Modeling*. Boca Raton, FL: Chapman and Hall/CRC.

Liu,. C., and Cheng,. Y. (2018). An application of the support vector machine for attribute-by-attribute classification in cognitive diagnosis. *Appl. Psychol. Meas.* 42, 58–72. doi: 10.1177/0146621617712246

Mislevy, R. J. (1995). "Probability-based inference in cognitive diagnosis," in *Cognitively Diagnostic Assessment*, eds P. Nichols, S. Chipman, and R. Brennan (Mahwah, NJ: Erlbaum), 43–71.

Mislevy, R. J., Almond, R. G., Yan, D., and Steinberg, L. S. (1999). "Bayes nets in educational assessment: where do the numbers come from?," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, eds K. B. Laskey and H. Prade (San Francisco, CA: Morgan Kaufmann), 437–446.

Neapolitan, R. E. (2004). *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

Norsys Inc. (2004). *Netica [Computer Software Manual]*. Retrieved from: http://www.norsys.com

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12, 1151–1172. doi: 10.1214/aos/1176346785

Russell, S., Binder, J., Koller, D., and Kanazawa, K. (1995). "Local learning in probabilistic networks with hidden variables," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, Vol. 2, ed C. S. (San Mateo, CA: Morgan Kaufmann) 1146–1152.

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

Shu, Z., Henson, R., and Willse, J. (2013). Using neural network analysis to define methods of DINA model estimation for small sample sizes. *J. Classif.* 30, 173–194. doi: 10.1007/s00357-013-9134-7

Shute, V. J., Hansen, E. G., and Almond, R. G. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. *Int. J. Artif. Intellig. Educ.* 18, 289–316. doi: 10.5555/1497124.1497126

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *J. Educ. Behav. Stat.* 31, 1–33. doi: 10.3102/10769986031001001

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Steinberg, L. S., Almond, R. G., Baird, A. B., Cahallan, C., Chernick, H., DiBello, L. V., et al. (2003). *Introduction to the Biomass Project: An Illustration of Evidence-Centered Assessment Design and Delivery Capability* (CSE Report No. 609). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from: http://www.cse.ucla.edu/reports/R609.pdf

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *J. Educ. Behav. Stat*. 12, 55–73. doi: 10.3102/10769986010001055

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

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

von Davier, M. (2005). *A General Diagnostic Model Applied to Language Testing Data (Research Report No. RR-05-16)*. Princeton, NJ: Educational Testing Service.

von Davier, M., and Lee,Y. S. (eds.). (2019). *Handbook of Diagnostic Classification Models*. Gewerbestrasse: Springer Nature Switzerland AG.

Wu, H. Y. (2013). *Comparison of general diagnostic models (GDM) and Bayesian networks*. using a middle school mathematics test [Unpublished doctoral dissertation]. Florida State University, Tallahassee, FL, United States.

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 Wang, Xin and Yanlou. 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: 13 August 2021
doi: 10.3389/fpsyg.2021.644764

Model Selection for Cogitative Diagnostic Analysis of the Reading Comprehension Test

Hui Liu[1,2] *and Yufang Bian*[2]*

[1] Faculty of Linguistic Sciences, Beijing Language and Culture University, Beijing, China, [2] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

Reading subskills are generally regarded as continuous variables, while most models used in the previous reading diagnoses have the hypothesis that the latent variables are dichotomous. Considering that the multidimensional item response theory (MIRT) model has continuous latent variables and can be used for diagnostic purposes, this study compared the performances of MIRT with two representatives of traditionally widely used models in reading diagnoses [reduced reparametrized unified model (R-RUM) and generalized deterministic, noisy, and gate (G-DINA)]. The comparison was carried out with both empirical and simulated data. First, model-data fit indices were used to evaluate whether MIRT was more appropriate than R-RUM and G-DINA with real data. Then, with the simulated data, relations between the estimated scores from MIRT, R-RUM, and G-DINA and the true scores were compared to examine whether the true abilities were well-represented, correct classification rates under different research conditions for MIRT, R-RUM, and G-DINA were calculated to examine the person parameter recovery, and the frequency distributions of subskill mastery probability were also compared to show the deviation of the estimated subskill mastery probabilities from the true values in the general value distribution. The MIRT obtained better model-data fit, gained estimated scores being a more reasonable representation for the true abilities, had an advantage on correct classification rates, and showed less deviation from the true values in frequency distributions of subskill mastery probabilities, which means it can produce more accurate diagnostic information about the reading abilities of the test-takers. Considering that more accurate diagnostic information has greater guiding value for the remedial teaching and learning, and in reading diagnoses, the score interpretation will be more reasonable with the MIRT model, this study recommended MIRT as a new methodology for future reading diagnostic analyses.

Keywords: continuous variable, diagnostic study, multidimensional item response theory, model selection, reading comprehension test

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Wenchao Ma,
University of Alabama, United States
Yu Bao,
James Madison University, United States

***Correspondence:**
Yufang Bian
bianyufang66@126.com

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

Received: *21 December 2020*
Accepted: *12 July 2021*
Published: *13 August 2021*

Citation:
Liu H and Bian Y (2021) Model Selection for Cogitative Diagnostic Analysis of the Reading Comprehension Test. Front. Psychol. 12:644764. doi: 10.3389/fpsyg.2021.644764

INTRODUCTION

In the area of language testing, it is reasonable to expect diagnostic information because any language assessment actually has the potential to provide some diagnostic information (Bachman, 1990; Mousavi, 2002), and indeed, there is a series of reading diagnostic studies that have successfully been conducted. However, in the previous reading diagnoses, little discussion on the

continuous nature of reading subskills has been provided, though reading subskills are generally regarded as continuous variables (Griffin and Nix, 1991; Lumley, 1993; Grosjean, 2001; Smith, 2004). Although Buck and Tatsuoka (1998) noted that dichotomizing continuous variables are a problem of their diagnostic study, diagnostic studies focusing on the continuous nature of reading comprehension test data have remained elusive. Indeed, when previous reading diagnostic studies have been performed, scarce diagnostic models could have addressed the continuous latent variables, and this lack of diagnostic models for continuous latent variables has been partially responsible for dichotomizing continuous variables under the methodological requirements in previous studies. Model selection is highly important for data analyses (Burnham and Anderson, 2002), and reading diagnostic analysis is no exception. The potential to provide abundant diagnostic information about the reading ability of each test-taker is what makes reading diagnoses so attractive to teachers, administrators, and other language educators who are concerned with teaching and learning of reading, whereas, without a proper model, it is very difficult to make accurate inferences about weaknesses and strengths of test-takers, let alone to make the reading diagnosis truly become the interface between learning and assessment (Alderson, 2005; Kunnan and Jang, 2009). Reading comprehension test data consisting of continuous latent variables are the target data of this present study, and we investigate the performances of different models when they are used in reading diagnoses, aiming to determine whether a diagnostic model with continuous latent variables will gain an advantage when used in reading diagnostic analyses.

Reading Diagnostic Studies: Current Status and Issues

The existence of distinguishable reading subskills (Grabe, 2009; Bernhardt, 2011) implies the multidimensional nature of reading ability, which renders the diagnostic score report possible. The first batch of reading diagnostic studies is based on the rule-space model (RSM), and the study by Buck et al. (1997) is representative of this line of research. Buck et al. (1997) analyzed the reading scores of 5,000 Japanese students on the Test of English for International Communication (TOEIC), classified 91% of the test-takers into their latent knowledge states, and provided them with diagnostic information based on a set of attributes. The study by Jang (2009) is typical of more recent research. In the context of *LanguEdge*™ test items, Jang suggested that the fusion model with all C parameters being set to 10 [i.e., the reduced reparametrized unified model, R-RUM, which was discussed in detail by Jang (2005)] is appropriate and can provide test-takers with more fine-grained diagnostic information regarding their reading abilities. Readers interested in additional details about previous reading diagnostic studies are referred to by Lee and Sawaki (2009b) and Li et al. (2016). Meanwhile, we also observe limitations in these reading diagnostic studies: the chief models contain only dichotomous (mastery vs. non-mastery) subskills.

Regarding the subskill scale, the cognitively diagnostic psychometric model has two categories, which are the continuous scale and the discrete (i.e., dichotomous or polytomous) scale (Fu and Li, 2007). The major distinction between them is the type of the latent variables used: continuous latent variables are used in the former, whereas discrete latent variables are used in the latter (Rupp et al., 2010). Researchers have emphasized that whether the variable in question is continuous or discrete determines which type of model should be chosen for the actual application (Stout, 2007). Specific to reading comprehension tests, the variable in question is the reading subskill.

A continuous variable is a variable that can take on any value in its possible range, which is in opposition to a discrete variable (Mackey and Gass, 2005). Although a consensus is lacking, many linguists believe that reading develops gradually and that the reading subskills of individuals are at different points on a continuous scale (Griffin and Nix, 1991; Lumley, 1993; Grosjean, 2001; Smith, 2004). Moreover, on reviewing the subskills specified in previous reading diagnostic studies, it is found that subskills are continuous in opposition to discrete latent variables; regardless, the grain size of a subskill is larger (Alderson, 2005; Lee and Sawaki, 2009b) or smaller (Jang, 2009). The continuous nature of reading subskills implies that the continuous scale model is theoretically more appropriate in the context of reading comprehension test data.

The literature review, however, reveals that dichotomous scale models are the traditional choice in reading diagnoses and that virtually all previous reading diagnostic studies have used dichotomous scale models (Lee and Sawaki, 2009b; Li et al., 2016). Some problems may exist with this traditional choice because, in general, statistical models should be used only for data that meet their theoretical assumptions; moreover, it is acknowledged that difficulty will arise in the score interpretation of a reading diagnosis under the framework of a dichotomous subskill scale (Jang, 2009). At present, continuous scale models that are able to conduct diagnostic analysis are available. The multidimensional item response theory (MIRT) model is one of the most popular diagnostic models with a continuous scale (Reckase, 2009), and it has already been used to report subskill scores (Yao and Boughton, 2007; Haberman and Sinharay, 2010). Considering the continuous nature of reading subskills, the MIRT is expected to have an advantage in extracting diagnostic information from reading comprehension tests.

Representative Models

Comparison studies have been conducted on the popular models in reading diagnoses. Lee and Sawaki (2009a) compared the functioning of three diagnostic models, i.e., the general diagnostic model (GDM), the reparametrized unified model (RUM), and the latent class analysis, when used to analyze the reading and listening sections of the TOEFL iBT. Li et al. (2016) to compare the performances of five models, namely, the generalized deterministic inputs, noisy "and" gate (G-DINA), additive cognitive diagnostic model (ACDM), the R-RUM, and the DINA and DINO models, according to their applicability to the Michigan English Language Assessment Battery (MELAB) reading test.

Overall, the R-RUM has received a fairly intensive study and shown good performances in previous research (Jang, 2009;

Lee and Sawaki, 2009b), and Li et al. (2016) showed that G-DINA is better fitted for use in a reading diagnostic study. Consequently, in this study, the R-RUM and G-DINA are selected as representatives of the traditionally widely used models for comparison with the MIRT model. For the sake of brevity, the introduction of the three models is relegated to the Appendix (**Supplementary Material**).

In addition, it is necessary to indicate that, in this study, the MIRT model considered is the compensatory MIRT model. The two basic types of MIRT include the compensatory and non-compensatory models. Bolt and Lall (2003) suggested that the compensatory MIRT model fits the real data from an English usage test better than the non-compensatory MIRT and fits the data generated from the non-compensatory model nearly as well as the non-compensatory model itself. Given the comparison results of Bolt and Lall (2003) and the suggestion of many current language researchers that reading subskills are generally compensatory (Stanovich, 1980; Grabe, 2009), the MIRT model considered in this study is the compensatory MIRT model. In addition, MIRT analyses can be either exploratory or confirmatory, and when used in diagnostic analyses, the model should be confirmatory (Reckase, 2009), because it requires a hypothesis for the structure of the response data, which is identical to the Q-matrix in dichotomous subskill scale models.

As noted above, there have been continuous scale diagnostic models. Given the diagnostic potential of the MIRT model (Stout, 2007) and its capacity to be fitted to continuous latent variables, we expect to determine whether the MIRT model is more appropriate than the traditionally widely used models in reading diagnoses. To that end, the performances of MIRT and representatives of the traditionally widely used models (R-RUM and G-DINA) are examined with both real and simulated data based on a reading comprehension test. The first aim of this study is to examine whether the MIRT model has a better model-data fit. Second, we emphasize assessing the extent to which the estimated subskill scores represent the true abilities of the test-takers. Third, we put emphasis on assessing the person parameter recovery with the correct classification rates. Finally, this study is undertaken to compare the three candidate models deviations from the true values on the frequency distribution of subskill mastery probabilities. We believe that this comparison will provide insight into model selection for cogitative diagnostic analyses of reading comprehension tests.

TABLE 1 | Q-matrix for this study.

Item	Subskill 1 (retrieving information)	Subskill 2 (making straightforward inference)	Subskill 3 (making integrated inference)
1	1	0	0
2	1	1	0
3	0	1	0
4	1	1	0
5	1	1	0
6	0	0	1
7	1	0	0
8	1	0	0
9	1	1	0
10	0	1	0
11	0	1	0
12	1	1	0
13	1	0	0
14	1	0	0
15	1	1	0
16	0	1	0
17	1	1	0
18	0	0	1
19	1	1	0
20	1	0	0
21	0	0	1
22	0	0	1
23	0	0	1
24	0	0	1
25	0	0	1
26	0	0	1
27	0	0	1
28	0	1	1
29	0	1	0
30	0	0	1

MATERIALS AND METHODS

The reduced reparametrized unified model, G-DINA, and MIRT were used to calibrate real and simulated data based on a reading comprehension test, and their performances were compared in this study. Different indices, such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be used to evaluate the model-data fit in a model comparison (Ma et al., 2016); thus, in this study, first, these relative model-data fit indices were compared across the three models with real data and the absolute model-data fit indices were also provided to assess whether the chosen model really fits the real data well. Then, further analyses were conducted with simulated data. Because the "true" values are known in simulation studies, the deviation of the estimated values from "true" values is comparable. With simulated data, the functioning of the three models was examined in terms of the extent to which the true abilities were represented and the correct classification rates under different research conditions, and the corresponding frequency distributions of the subskill mastery probabilities were also compared.

Real Data

Response data from 3,077 students on 30 reading comprehension items were collected from a large-scale Grade 5 and 6 reading comprehension test in Beijing, P.R. China. The test is an existing reading comprehension test that was not designed to be diagnostic, and the Q-matrix in **Table 1**, which was built and validated through the literature review and think-aloud protocols, displays the correspondence between the three subskills and 30 items.

TABLE 2 | True item parameters.

Item	Discrimination			d	c
	a_1	a_2	a_3		
1	1.10	0.00	0.00	2.51	0.23
2	0.61	0.72	0.00	0.54	0.17
3	0.00	0.98	0.00	1.12	0.15
4	0.85	0.77	0.00	0.49	0.29
5	0.70	0.58	0.00	−0.42	0.10
6	0.00	0.00	1.30	−0.53	0.56
7	1.38	0.00	0.00	0.48	0.53
8	1.32	0.00	0.00	−0.15	0.44
9	1.05	1.02	0.00	2.09	0.34
10	0.00	1.34	0.00	0.93	0.28
11	0.00	0.90	0.00	0.87	0.21
12	0.68	0.85	0.00	2.54	0.38
13	1.33	0.00	0.00	2.73	0.13
14	1.32	0.00	0.00	1.44	0.14
15	1.40	1.08	0.00	1.10	0.32
16	0.00	2.03	0.00	4.89	0.17
17	1.08	1.55	0.00	5.01	0.15
18	0.00	0.00	0.70	1.95	0.23
19	0.56	0.56	0.00	2.58	0.22
20	1.40	0.00	0.00	5.01	0.23
21	0.00	0.00	0.98	1.79	0.20
22	0.00	0.00	0.87	1.44	0.12
23	0.00	0.00	1.38	2.33	0.19
24	0.00	0.00	1.52	2.43	0.18
25	0.00	0.00	0.50	−0.30	0.16
26	0.00	0.00	1.60	3.17	0.17
27	0.00	0.00	1.07	0.94	0.07
28	0.00	0.93	1.04	1.10	0.23
29	0.00	0.53	0.00	0.23	0.23
30	0.00	0.00	0.69	0.58	0.13

Item d is a scalar parameter related to the difficulty of the corresponding item.

Simulation Study

We simulated response data based on the real data calibration with the MIRT model, which is a typical choice when multidimensional continuous latent variables are desired.

Item Parameters

The item parameters used to generate the simulated data were obtained from the real data with a three-dimensional MIRT calibration estimated by the flexMIRT software program (Cai, 2015). **Table 2** shows the true item parameters. The first column lists the item number, the second to fourth columns present the three discrimination parameters for subskill 1, subskill 2, and subskill 3, respectively, and the last two columns show the intercept and guessing parameters.

Simulation Conditions

The abilities for 3,000 test-takers were generated from multinormal distributions with the mean of (0, 0, 0) and σ as the variance-covariance matrix, and the simulated response datasets were produced with these abilities, in which:

$$\sigma = \begin{pmatrix} a_1 & r_1 & r_2 \\ r_1 & a_2 & r_3 \\ r_2 & r_3 & a_3 \end{pmatrix}$$

where, $a_1 = a_2 = a_3 = 1$ and $r_1 = r_2 = r_3 = 0.1$, 0.3, 0.5, 0.7, and 0.9.

The manipulated design factor for the simulated data generation was the subskill correlations (SCs). Considering that their correlations can vary over a relatively wide range, from a negligibly small value (Guthrie and Kirsch, 1987) to a moderately and a relatively high relationship between reading subskills (Droop and Verhoeven, 2003; Alderson, 2005), the SCs were found to be 0.1, 0.3, 0.5, 0.7, and 0.9, as noted above, and the correlations between the subskills were set equal to reduce the simulation conditions to be studied (Yao and Boughton, 2007).

Therefore, we had five simulation conditions. Twenty-five different seeds were used to obtain 25 replications across all simulation conditions; thus, there were 125 simulated datasets in total.

Subskillability Estimates and Classification Procedures

For each simulated dataset, three calibrations using R-RUM, G-DINA, and MIRT were performed. The R-package CDM was used to estimate the R-RUM and G-DINA model, and flexMIRT was used to estimate the MIRT model. R-RUM and G-DINA report the MLE and MAP estimates of subskill mastery patterns to each test-taker, and MLE estimates were used in this study; they also provide the posterior mastery probabilities of test-takers on each subskill. The MIRT model provides every test-taker a $\boldsymbol{\theta}$-vector that indicates their proficiency on each subskill, which can easily be transformed into vectors of mastery probabilities, given the multinormal nature of the simulated datasets.

In this simulation part of the present study, there should be four subskill scores per test-taker per subskill, i.e., one true subskill score representing the true ability of the test-taker from the simulated true value, and three estimated subskill scores representing the estimated abilities of the test-taker from R-RUM, G-DINA, and MIRT. The true subskill score and the estimated subskill score from MIRT are the θ abilities on each subskill, which would fall along a continuous subskill scale, and the estimated subskill scores from R-RUM and G-DINA indicate the classification result of "mastery" or "non-mastery," termed the mastery state, which would fall along a dichotomous subskill scale. In this study, the extent to which the estimated subskill scores represent the true abilities of the test-takers were evaluated through comparison of relations between the estimated scores from MIRT, R-RUM, and G-DINA and the true scores.

Furthermore, in the diagnostic study context, the classification results are expected. Classifications of test-takers are the bases for providing remedial strategies to facilitate learning, so whether the remedial strategies provided for improvement are effective depend mainly on the accuracy of the classification results. Therefore, it is also very important for the diagnostic model

selection to evaluate the extent to which the approximating models classify the test-takers into their intended groups. As mentioned above, we classify the performances of the test-takers on a subskill into different groups, such as the "mastery" and "non-mastery" group, and the classification result is termed the mastery state. In this simulation study, there should be four classification results per test-taker per subskill, i.e., one true mastery state from the simulated true value, and three estimated mastery states from R-RUM, G-DINA, and MIRT. We know that the estimated mastery states from R-RUM and G-DINA can be obtained directly, while true mastery states of the test-takers and their estimated mastery states from MIRT cannot be obtained unless we dichotomize each continuous θ scale with a cut-off point (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014). There is no natural cut-off points for a continuous variable; in addition, the establishment of a cut-off point should be based on a labor-intensive and time-consuming process (Hambleton and Pitoniak, 2006), and the implementation of this process is beyond the scope of this present study. Therefore, in this simulation study, different cut-off points were used to cover different practical needs: we set the cut-off point to increase from 0.1 to 0.9 with a step of 0.01 on the subskill mastery probability scale transformed from the θ scale, considering that a cut-off point is often defined in the form of a value established on the subskill mastery probability scale in diagnostic analyses (DiBello and Stout, 2008). In this way, we obtained the true subskill mastery state and the estimated subskill mastery state from the MIRT with each of the 81 cut-off points. Consequently, the comparison of classification results can be conducted. For example, if on subskill 1, the true mastery probability of test-taker A is 0.11, the estimated mastery states from R-RUM and G-DINA are "non-mastery" and "mastery," and the estimated mastery probability from MIRT is 0.15, then with a cut-off point of 0.1, the true mastery state of test-taker A on subskill 1 is "mastery," and the estimated mastery states of test-taker A to this subskill from R-RUM, G-DINA, and MIRT are "non-mastery," "mastery," and "mastery," which indicate the successful estimations of G-DINA and MIRT and of the mastery of subskill 1 of test-taker A; while with a cut-off point of 0.12, only R-RUM would correctly estimate the non-mastery of subskill 1 of test-taker A.

We have five simulation conditions, i.e., the five subskill correlations used in generating simulated data, and 81 cut-off points are used to dichotomize each θ scale in each simulated dataset, producing 405 (5×81) research conditions in total.

Evaluation Criteria

In the simulation section, the deviation of the estimated values from true values was compared in three ways. First, we compared the relations between the estimated subskill scores and the true subskill score, with the former containing both the estimated mastery states on each subskill reported in the subskill mastery patterns from the two dichotomous scale models and the estimated θ values on each subskill from MIRT. Second, the subskill classification accuracies were compared, and the indices are described in detail in the Results section. Finally, the estimated frequency distributions of the subskill mastery probabilities were compared to the true distribution too.

When comparing the subskill classification accuracies, previous studies always use the correct classification rates to evaluate the classification consistency between the true values and the estimated values (Ma et al., 2016), which involves the pattern correct classification rate (PCCR) and the subskill correct classification rate (SCCR) in this study. PCCR and SCCR are defined as,

$$PCCR = \frac{\sum_{r=1}^{Rep}\sum_{i=1}^{N} t_i}{N \times Rep}, \tag{1}$$

where *PCCR* is the index used to evaluate the classification recovery accuracy of the subskill mastery pattern, which refers to a vector that involves the mastery states of a test-taker on all subskills; N is the total number of test-takers; Rep is the number of replications; and t_i indicates whether the estimated mastery pattern of a test-taker i is the same as the true pattern, with a value of 1 if the two are identical and a value of 0, if not.

$$SCCR = \frac{\sum_{r=1}^{Rep}\sum_{i=1}^{N}\sum_{k=1}^{K} g_{ik}}{N \times K \times Rep}, \tag{2}$$

Where, $SCCR_k$ is the index used to evaluate the classification recovery accuracy of subskill k, and g_{ik} indicates whether the estimated mastery state of test-taker, i on subskill, k is the same as the true state, with a value of 1 if the two are identical, and a value of 0, if not.

In the present study, PCCRs and SCCRs were compared among R-RUM, G-DINA, and MIRT across all research conditions.

RESULTS

Model-Data Fit With Real Data

The real data are calibrated using the three models, i.e., R-RUM and G-DINA with R-package CDM, and MIRT with flexMIRT, and the estimations are based on the EM algorithm with the default data (de la Torre, 2011; Cai, 2015; Robitzsch et al., 2015). Indices, such as the AIC and BIC are used to evaluate the model-data fit and to aid in the model comparison (Henson et al., 2009), which is also used in the model selection for reading diagnoses (Li et al., 2016). These goodness-of-fit (GOF) indices are reported by both R-package CDM and flexMIRT and are used for model selection. A summary of these model-data fit indices is presented in **Table 3**, which contains the $-2 \times$ log-likelihood, AIC, and BIC. The smaller values of these indices tend to accept the null hypothesis, which indicates a good model-data fit.

As shown in **Table 3**, the GOF statistics are strikingly different between the MIRT model and the two representatives of the traditionally widely used models, where MIRT has the smallest value in all three indices. Although G-DINA and R-RUM have similar performances on these GOF statistics, the G-DINA model has a smaller $-2 \times$ log-likelihood and AIC values and R-RUM has a smaller BIC value. Given that the AIC performed better in choosing diagnostic models (Li et al., 2016) and the BIC

TABLE 3 | Summary of goodness-of-fit statistics of the three models.

Index	R-RUM	G-DINA	MIRT
Number of parameters	76	85	99
$-2 \times$ log-likelihood	79,805.49	79,774.96	79,393.43
AIC	79,957.49	79,944.96	79,591.43
BIC	80,415.90	80,457.65	80,188.57

always has a strong penalty for highly parameterized models, G-DINA is regarded as a model that fits the real data better than R-RUM.

Considering the $-2 \times$ log-likelihood, since AIC and BIC are all relative model-data fit statistics, absolute fit indices of MIRT are required to evaluate whether it really fits data well-absolutely: M_2 and the corresponding root mean square error of approximation (RMSEA) are calculated. M_2 is a limited-information GOF statistics, and it has the benefit of being less sensitive to sparseness than the full-information statistics, such as the Pearson's X^2 or the likelihood ratio G^2 (Cai and Hansen, 2012). A statistic unaffected by the sample size and the model complex degree, the RMSEA ranges from 0 to 1. A smaller RMSEA implies a better GOF, and Browne and Cudeck (1992) recommended 0.05 as the threshold for a *close fit*. In this case, the value of M_2 statistics is 2042.59 ($df = 366$, $P < 0.0001$) and M_2 is rejected. The rejection is not surprising considering that M_2 is sensitive to model misspecification. Fortunately, we have the RMSEA index, which is calculated from M_2 and can be used to evaluate the severity of model misspecification (Cai and Hansen, 2012). The value of the RMSEA is 0.04, which suggests that the MIRT model provides a good fit.

In short, based on the log-likelihood and information criteria, the MIRT model is best fitted to the real reading comprehension test data among the three models, followed by G-DINA and then R-RUM; the advantage of MIRT over the other two models is much greater than the improvement from G-DINA to R-RUM. Based on the M_2 and the RMSEA statistics, we conclude that the MIRT model fits the real data well. Therefore, the GOF indices provide strong evidence for the choice of MIRT.

Results With Simulated Data

Comparison of Subskill Scores

Under each research condition, we plotted the estimated subskill scores from the three candidate models against the true subskill scores on each subskill, for all the 75,000 ($3,000 \times 25$) simulated test-takers with every 50th point being plotted. Dozens of scatter plots were produced, because five subskill correlations, three subskills, and three candidate models were used in this study. **Figure 1** presents the scatter plots of the estimated subskill scores against the true subskill scores on only subskill 1, as the same pattern was observed for both subskill 2 and subskill 3 conditions. In addition, we excluded the scatter plots under subskill correlations of 0.3 and 0.7 in **Figure 1** for the sake of simplicity.

The comparison of the estimated subskill scores and the true subskills scores on subskill 1 under subskill correlations of 0.1, 0.5, and 0.9 is displayed in **Figure 1**, with a dual y-axis coordinate system: the lower y-axis represents the estimated subskill scores from MIRT, and the upper y-axis denotes the estimated subskill scores from G-DINA. In the upper part of **Figure 1**, each dark bar (|) represents a pair of scores for one test-taker whose estimated subskill score from G-DINA is the same as that from R-RUM, and each red bar represents that for one test-taker whose estimated subskill score from G-DINA is different to that from R-RUM. Therefore, the relations between the estimated subskill scores from all the three candidate models and their corresponding true values on subskill 1 are shown in **Figure 1**.

Figure 1 displays the relations between the estimated subskill scores from the two kinds of models and the true values. The lower part of **Figure 1** shows that the estimated subskill scores from MIRT tend to increase as the true subskill scores increase, and the correlation between the estimated scores from MIRT and the true scores increases as the correlation between the subskills increases. In addition, the estimated subskill scores from MIRT are approximately centered on their corresponding true subskill scores with estimation bias, and the bias reduces when the subskill correlation is higher. For example, under subskill correlation of 0.1, for all the 305 test-takers with identical true subskill score of 0 (being rounded to two decimal places), their estimated subskill scores from MIRT are centered at 0.08 under normality, with many estimated points being close to the true value.

Regarding the estimated subskill scores from the dichotomous models, R-RUM and G-DINA exhibit similar performances. These two models also have a general tendency that as the true subskill scores increase, the estimated subskill scores tend to rise to score 1 ("mastery"). However, deviation from this tendency clearly exists. The upper part of **Figure 1** illustrates that there is a large overlap in the true subskill scores between the estimated groups of mastery and non-mastery, and test-takers whose true subskill scores are located within this overlapping range may be estimated as mastery or non-mastery of subskill 1, no matter whatever be the scores of their true subskill ability. In addition, the estimated scores from the dichotomous models may be totally different for students who have exactly the same true ability. For example, under subskill correlation of 0.9, for the 307 test-takers with identical true subskill score of 0, 220 of them are scored 1 by G-DINA, and 87 are scored 0 by the same model. In brief, **Figure 1** indicates that with R-RUM and G-DINA, a large number of the true subskill abilities of the test-takers fall into the broadly overlapping ranges in the true subskill score between the groups of mastery and non-mastery estimated by the two models; the large number of test-takers can be subdivided into many subsets with exactly the same true subskill ability, while test-takers in each subset may obtain totally different estimated scores, and entirely different remedial interventions may be provided to them even though actually they should have identical treatment. This situation implies that the estimated subskill scores from the dichotomous models may not be able to represent the true abilities of the test-takers as well as the estimated subskill scores from MIRT.

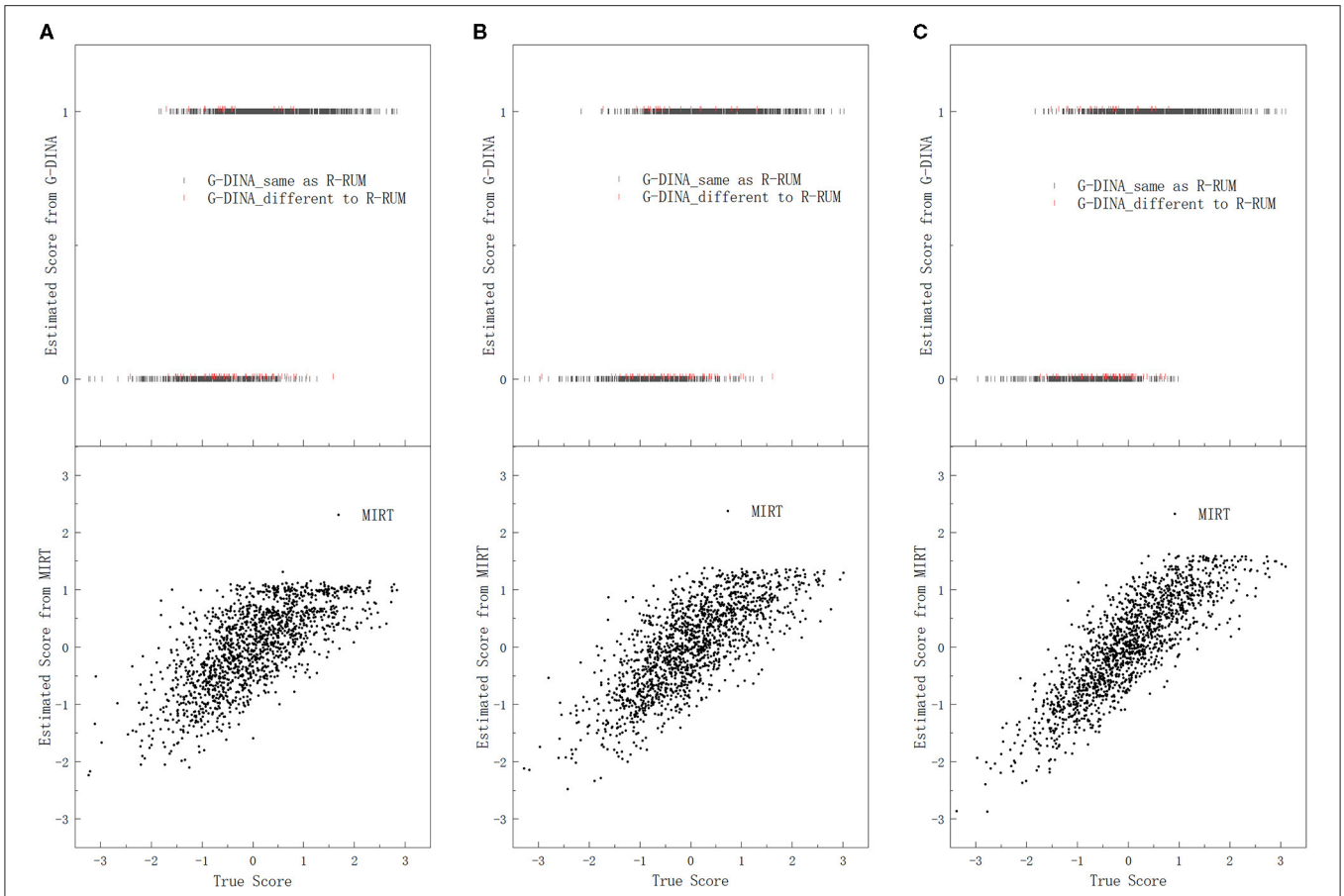


FIGURE 1 | Plots of the estimated subskill scores from the three candidate models against the true subskill scores on subskill 1. **(A)** SC = 0.1, **(B)** SC = 0.5, and **(C)** SC = 0.9.

Correct Classification Rates: SCCR and PCCR

Figure 2 illustrates the SCCRs for R-RUM, G-DINA, and MIRT under all research conditions on only subskill 1, as the same pattern was observed for both subskill 2 and subskill 3 conditions. The specific correct classification rate of subskill 1 ($SCCR_1$) values under different conditions are also given in Table 1A (see **Supplementary Material**).

Differences in the $SCCR_1$s clearly exist among the three models. Across the five simulation conditions, the average $SCCR_1$s includes 0.682–0.699 for R-RUM, 0.688–0.702 for G-DINA, and 0.807–0.865 for MIRT. From **Figure 2**, we can clearly observe that the $SCCR_1$s for MIRT are the highest among the three models under all research conditions and that there are fewer differences across the three models when the subskill correlations are lower. For example, when the subskill correlation is 0.1 and the cut-off point is 0.5, the $SCCR_1$ for MIRT is 0.759, which is higher than the $SCCR_1$ of 0.732 from R-RUM and the $SCCR_1$ of 0.740 from G-DINA by 0.027 and 0.019; when the subskill correlation is 0.5, the $SCCR_1$ for MIRT is also higher than the $SCCR_1$s from R-RUM and G-DINA, and the MIRT model outperforms R-RUM and G-DINA by 0.040 and 0.033, which is more evident than the improvements of 0.027 and 0.019 when the subskill correlation is 0.1. From **Figure 2**, we can also observe that there are fewer differences across the three models when the cut-off points are located near positions where the true mastery proportions from the simulated data are similar to the estimated mastery proportions from R-RUM and G-DINA (the estimated mastery proportions from R-RUM and G-DINA are about 0.66 and 0.63, respectively, which are provided in detail in Table 2A in **Supplementary Material**). For example, when the subskill correlation is 0.3 and the cut-off point is 0.37, the $SCCR_1$ for MIRT is 0.808, which is higher than the value of 0.789 from R-RUM and the value of 0.787 from G-DINA by 0.019 and 0.021; when the cut-off point is 0.57, the $SCCR_1$ for MIRT is also higher than the $SCCR_1$s from R-RUM and G-DINA, and the MIRT model outperforms R-RUM and G-DINA by 0.074 and 0.063, which is more evident than the improvements of 0.019 and 0.021 when the cut-off point is 0.37. In addition, the $SCCR_1$s for R-RUM and G-DINA are close to each, with R-RUM being slightly better when the subskill correlations are lower, G-DINA being slightly better when the subskill correlations are higher, and the average $SCCR_1$s over the five simulation conditions for G-DINA being slightly higher.

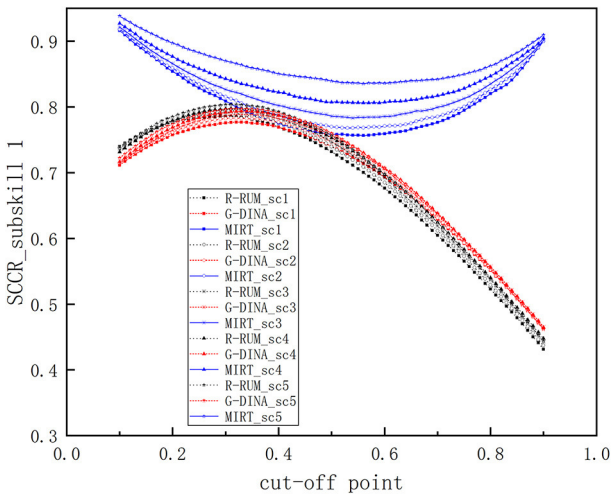


FIGURE 2 | Comparison of $SCCR_1$s among R-RUM, G-DINA, and MIRT.

Regarding the characteristics of the $SCCR_1$s for the models themselves, it is note that the $SCCR_1$s increase as the correlations between the subskills increase; for example, when the cut-off point is 0.6, the $SCCR_1$s for R-RUM increase from 0.676 to 0.698, as the subskill correlations increase from 0.1 to 0.9. The $SCCR_1$s for the MIRT model are the lowest when the cut-off points depart from near the mean and increase as the cut-off point moves farther away; for R-RUM and G-DINA, their $SCCR_1$s are the highest when the cut-off points are located near positions, where the true proportions of masters are similar to the estimated proportions of masters from R-RUM and G-DINA, and their $SCCR_1$s decrease as the cut-off points move farther away. For instance, when the subskill correlation is 0.5, the $SCCR_1$ of 0.783 is obtained when the cut-off point is 0.54, which is the lowest for MIRT, and $SCCR_1$ for MIRT increases as the cut-off point moves from 0.54 to 0.1 or 0.9; under the same simulation condition, the $SCCR_1$ of 0.789 obtained with the cut-off point of 0.34 is the highest for G-DINA and the $SCCR_1$ for G-DINA decreases as the cut-off point moves farther away from this point.

For the three models, the PCCRs exhibit similar performances to the SCCRs: MIRT holds the highest values under all research conditions and G-DINA is slightly better than R-RUM, the differences among them are smaller when the subskill correlations are lower, and in general, the PCCRs for each single model increase as the correlations between the subskills increase, and so on. Because the general change tendencies of the PCCRs and SCCRs are quite similar, we only illustrate the PCCRs in **Figure 1A** (see **Supplementary Material**) without a detailed description.

Frequency Distribution of Subskill Mastery Probability

A comparison between the estimated frequency distributions and the true distribution can show the deviation of the estimated subskill mastery probabilities from the true values in the general value distribution. **Figure 3** shows the frequency counts of the estimated subskill mastery probability vs. the true values, with the subskills mastery probabilities on the x-axis and the frequency counts on the y-axis.

The frequency distribution of the true value is almost a straight line under any simulation condition. The overall frequency distributions of the estimated subskill mastery probability for MIRT are close to their corresponding true values, with bias at the highest and lowest parts, and the bias decreases as the correlation of the subskills increases. The frequency distributions of the estimated subskill mastery probability for R-RUM and G-DINA are similar in shape, which is in line with the U-shaped distribution of Lee and Sawaki (2009b), with overwhelming counts at the highest and lowest ends and very low counts in the wide range of the middle of the x-axis; however, they are somewhat different from the true values. Compared to the true values, R-RUM and G-DINA underestimate a large number of average-scoring and low-scoring test-takers and overestimate many high-scoring test-takers. In addition, the frequency distribution obtained with R-RUM and G-DINA also reminds us that the overall frequency distributions obtained from dichotomous scale models will be U-shaped even if the true values are normally distributed and the corresponding true frequency distribution is a straight line.

DISCUSSION

The MIRT model, which hypothesizes that the latent variables are continuous, has a theoretical advantage in diagnostic analyses of reading comprehension tests. In this research, empirical and simulation studies were conducted to explore whether the model with a theoretical advantage has a practical advantage in reading diagnoses, and the findings demonstrated the practical advantage of MIRT in several ways. First, MIRT and two representatives of the traditionally widely used dichotomous scale models (R-RUM and G-DINA) were compared on model-data fit indices in the empirical analysis, and both the assessments of relative model-data fit (i.e., the $-2 \times$ log-likelihood, AIC, and BIC) and absolute model-data fit (i.e., M_2 and the RMSEA) revealed that MIRT should be the chosen model. Second, a comparison of the relations between the estimated subskill scores from the three models and the true values indicated that the estimated subskill scores from R-RUM and G-DINA could not represent the true subskill abilities of test-takers as well as the estimated scores from MIRT. Third, the correct classification rates for the three models were explored under different research conditions in the simulation section, with the result that MIRT achieved the highest PCCRs and SCCRs under all conditions, and its improvement over the other two models increased as the subskill correlations increased. Finally, the estimated frequency distributions of the subskill mastery probability were compared to the true distribution, and the results revealed that the estimated frequency distributions of the subskill mastery probability from MIRT were more similar to the true values whereas, for R-RUM and G-DINA, the frequency distributions were very different from the corresponding true values. The wide variation between MIRT and the other two models in the frequency distribution of the subskill mastery probability

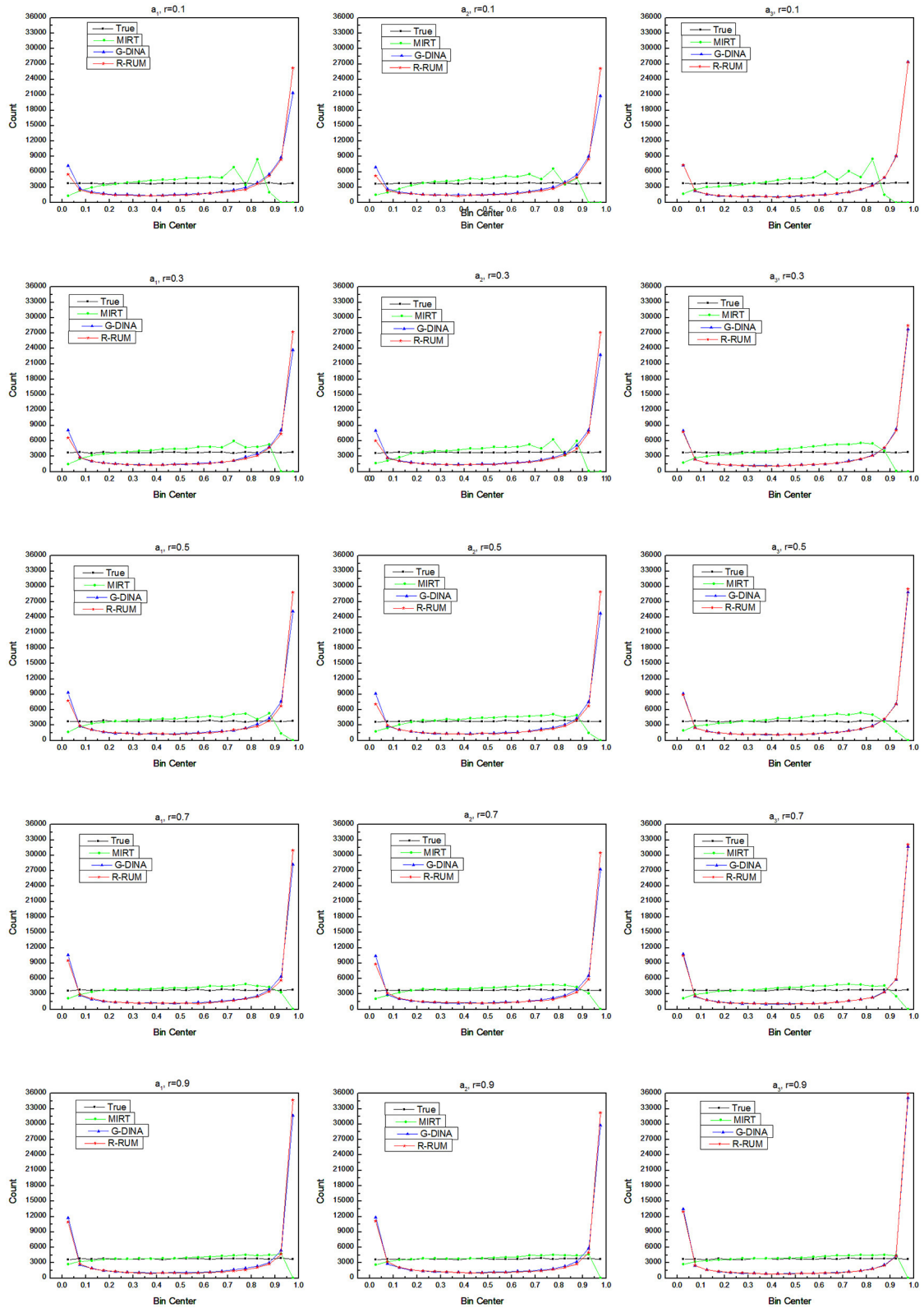


FIGURE 3 | Comparison of frequency distributions of subskill mastery probability.

matched their different hypotheses on latent variables, which also supported the inference that the continuous latent variable hypothesis benefits MIRT in gaining advantages in reading diagnostic analyses. These results were expected because the MIRT model treats the latent variables in question as continuous variables, which is in accordance with the nature of reading subskills; while R-RUM and G-DINA treat the latent variables in question as dichotomous variables, which would lead to information loss correspondingly (Buck and Tatsuoka, 1998; Bonifay and Cai, 2017).

Developing tests with a diagnostic purpose and retrofitting existing tests for diagnostic purposes are the two main methods presently available to obtain diagnostic information (DiBello et al., 2007), and there are pros and cons in both the types of methods. Developing tests with a diagnostic purpose will ultimately provide the user with more fine-grained diagnostic feedback, but the diagnostic reading test is barely in its infancy (Alderson et al., 2015) and few reading tests have been designed to provide diagnostic feedback until recently. Retrofitting existing tests for diagnostic purposes is convenient because the items and response data are already available, but it is certainly not the optimal method to obtain diagnostic information because the cognitive model being obtained in a *post-hoc* analysis depends on the existing test, and a sufficient number of subskills cannot be guaranteed due to the limited number of items in the existing test (Li et al., 2016). In contrast to the latter method, the former develops the diagnostic test with a well-articulated cognitive model from the beginning and can fully reflect the principles that underlie the cognitively diagnostic analysis, so developing tests with a diagnostic purpose is of critical importance to future reading diagnostic study (Jang, 2009). However, for now, retrofitting existing tests for diagnostic uses is currently a feasible approach to address the needs of diagnostic feedback from reading assessments, and almost all the published studies on reading diagnoses used the latter method (Buck et al., 1997; Jang, 2009; Lee and Sawaki, 2009b; Li et al., 2016). The real and simulated data of this study were based on an existing large-scale reading comprehension test, which was not designed to be diagnostic and was developed to measure the reading subskills being generally regarded as continuous latent variables (Griffin and Nix, 1991; Lumley, 1993; Grosjean, 2001; Smith, 2004). In short, similar to previous reading diagnostic studies, the reading diagnostic analysis in this study was conducted by retrofitting an already existing reading test to a diagnostic function. In practice, one of the central functions of reading diagnoses is to offer diagnostic information about the reading abilities of learners that may be used to guide teachers and learners in subsequent remedial teaching and learning. In this study, the empirical and simulation study results suggest that the MIRT model can produce more dependable diagnostic feedback, which will lead to more accurate reports about the weaknesses and strengths of the test-takers. A suitable model is a necessary prerequisite for a reading diagnosis to be useful. If traditional dichotomous scale models had been selected, our inferences about the reading abilities of far more test-takers would have deviated from the truth, and the remedial plans based on such information would be useless or misleading.

We should consider the dependability of feedback before the diagnosis can claim to be successful (Kunnan and Jang, 2009); however, to maintain the practical usefulness of a reading diagnostic analysis, the existence of dependable diagnostic information is not sufficient. The scores should also be understood and used properly; otherwise, even the best test is worthless (Brennan, 2006). In reading diagnoses, the score interpretation will be more reasonable with the MIRT model. With traditional dichotomous scale models, it may be difficult for teachers and students to properly understand the diagnostic results. As noted by Jang (2009), students may require a definition of "master" and should know that "master" cannot be interpreted as "no further action" in reading diagnoses. With dichotomous subskill scales, it may be difficult to explain why the state of "mastery" on a subskill is not equal to "flawless mastery." With continuous subskill scales, however, we tell students that "mastery" means "strong ability," which is related to scores that are higher than the required score but are not equal to "flawless mastery" because "mastery" implies an interval along a continuous scale rather than a point on a dichotomous scale.

In reading instruction, diagnosis can be regarded as the interface between learning and assessment (Alderson, 2005). Without dependable diagnostic information and reasonable score interpretation, however, it will be very difficult for a diagnosis to truly become such an interface (Kunnan and Jang, 2009; Alderson et al., 2015). In this regard, the MIRT model actually matters very much in improving the practical usage of the reading diagnosis.

Moreover, when being used in diagnostic practices, the MIRT model has other strengths, including more informative diagnostic information and being able to extend to model testlets. As a continuous subskill scale diagnostic model, in addition to the traditional diagnostic report of the mastery states of test-takers, MIRT is able to describe their locations on the continuous score scale of each subskill, which can provide the teachers and learners with more detailed diagnostic information. Another important advantage of MIRT is that it is easy to extend to fit data with testlets. Testlets commonly exist in current reading comprehension tests, which are presented as a group of items with the same stimulus developed and administered as a single unit (Wainer and Kiely, 1987). It has been observed that the bifactor model is successful in approximating response data on testlets (McLeod et al., 2001; DeMars, 2006), and the MIRT model is able to combine with the bifactor model to fit multidimensional data with testlets.

Nevertheless, we should know that the mastery states of test-takers from MIRT may sometimes be hard to obtain, because with MIRT, the two-step approach (estimating continuous θ values and then dichotomizing them with cut-off points) should be used to obtain these mastery states, and the establishment of a cut-off point tends to be labor-intensive and time-consuming (Hambleton and Pitoniak, 2006). For cases where the mastery state from MIRT becomes realistically infeasible, the mastery states from traditional diagnostic models in reading diagnoses can be regarded as a viable alternative to provide the diagnostic feedback under certain conditions, because though hypotheses on the latent variable are different, both MIRT and the traditional diagnostic models are actually popular multidimensional models

that can be used for diagnostic purposes. These multidimensional models can help us to provide diagnostic information, while things would be difficult with the unidimensional model. Ma et al. (2020) recently showed that the traditional diagnostic models fitted the data from a cognitive diagnostic assessment better than the unidimensional IRT model, and they could extract useful diagnostic information while the general abilities estimated from the unidimensional model were of little diagnostic utility. Indeed, the diagnostic information previously obtained in reading diagnoses are mostly based on the traditional diagnostic models, and R-RUM and G-DINA are representatives of these models.

Regarding the comparison between R-RUM and G-DINA, G-DINA performed slightly better on the average PCCRs and SCCRs over the five simulation conditions; in addition, when used with empirical reading comprehension test data, G-DINA had slightly lower $-2 \times$ log-likelihood and AIC values and slightly higher BIC values, which coincides with the finding of Li et al. (2016) with the MELAB reading test. It should be noted, however, that the performance of R-RUM was only slightly worse than G-DINA for the most part, and in other circumstances, it had an edge over G-DINA, i.e., differences actually exist between these two models, but there is no huge dissimilarity, and in general, they perform similarly.

Considering that G-DINA is a saturated model, which would obtain a better model-data fit and correct classification rates (de la Torre, 2011), R-RUM actually performed well. Therefore, the R-RUM can be regarded as a good choice when there is a lack of diagnostic models with continuous scale, which also provides some support to previous reading diagnostic research with R-RUM, such as the studies by Jang (2009), Lee and Sawaki (2009b), and Jang et al. (2013).

Finally, the starting point of this study is the contradiction encountered in practice: statistical methods have theoretical assumptions, and typically, they should only be applied to data that meet their assumptions; however, in previous reading diagnostic studies, dichotomous scale models, whose hypothesis on latent variables are not in accordance with the continuous nature of reading subskills, have always been the chosen model. Therefore, the main focus of this study is to compare the performances of MIRT and the traditionally widely used dichotomous scale models when they are used in reading diagnostic analyses. All efforts in this study were oriented toward this practical requirement, including the criteria for selecting the representative models and calibration software programs: representatives of the traditionally widely used models should be widely recognized in previous studies, and the chosen software programs should be well-accredited and conveniently obtainable. For similar reasons, we simulated only datasets consisting of continuous latent variables in this study, with one continuous scale diagnostic model and two dichotomous scale diagnostic models being used to approximate them: using all approximating models as generating models simultaneously seems to be a more popular treatment; however, this study is focused on model selection for reading diagnoses instead of general model comparison, which means continuous latent variables are of particular interest to us, and the discussion about model performance on dichotomous latent variables has little relevance in the aim of this present study: even if R-RUM and G-DINA were used as generating models, the performances of the three models on datasets generated by these two models do nothing to aid in evaluating their abilities in reading diagnoses. In fact, when discussing model selection in reading diagnoses, we know that the theoretical assumption of MIRT on latent variables is in accordance with the continuous nature of reading subskills, and theoretically dichotomous scale models, such as R-RUM and G-DINA are at a disadvantage in the context of reading comprehension test data; so the aim of this present study is to compare the performances of models with the theoretical advantage to those of models traditionally chosen for reading diagnoses.

Similar to the issue that Buck and Tatsuoka (1998) raised more than 20 years ago, there is a longstanding contradiction in reading diagnostic studies: the subskills required by reading comprehension tests are generally regarded as continuous latent variables, whereas dichotomous scale models have been the traditional choice in previous reading diagnostic studies. This study focused on whether the MIRT model, a model with a continuous subskill scale and that can be used for diagnostic purposes, has advantages when used to diagnose reading comprehension tests. Based on comparisons on the model-fit statistics, relations between the estimated subskill scores and the true subskill score, correct classification rates, and frequency distributions of the subskill mastery probability, this study confirmed the MIRT model can produce more dependable diagnostic information, which means more accurate inferences about the weaknesses and strengths of test-takers and greater guidance value for remedial teaching and learning. In addition, with the MIRT model, the reading diagnosis is able to provide more informative diagnostic information and a more reasonable score interpretation and is convenient to extend to fit testlets. These are all important bases for the reading diagnosis to fulfill its practical function and to truly become the interface between learning and assessment. Therefore, the MIRT model warrants broad attention and is recommended for use in future reading diagnoses.

This study was conducted with real and simulated data based on the administration of a single reading comprehension test; therefore, only one set of test items and one Q-matrix was used. More conditions, such as different numbers and grain sizes of subskills (Harding et al., 2015) and different item qualities (being defined by item parameters), should be considered in future studies for a more comprehensive and detailed evaluation. In addition, only the commonly employed correct classification rates based on simulated data were used in the discussion of classification accuracies, while classification accuracies may be empirically examined through approaches used in the validation of diagnostic studies, such as interviews or controlled remediation of students (Tatsuoka, 2009). Future work is required to provide insights about the performances of these empirical approaches as well as the classification accuracies of competing models based on real data. Moreover, no polytomous scale model was chosen as representative of traditionally widely used models to compare with the MIRT model because scarce reading diagnoses with this category of diagnostic model have

been conducted. Although a polytomous scale with more than two ability levels is considered to be a compromise between the continuous scale and the dichotomous scale (Fu, 2005), Haberman et al. (2008) have shown that polytomous scale models are competitive with MIRT when there are four or more ability levels for each subskill; thus, the performance of polytomous scale models in reading diagnoses awaits investigation.

Last but not the least, though subskill scores estimated based on a MIRT model are more accurate and reliable than the estimates based on many other subskill score estimation methods (Fu and Qu, 2018), previous study reveals that if the subskill scores are highly correlated, the subskill score may have no added-value (Sinharay, 2010). Therefore, in practical uses of subskill score estimation methods, we should first determine whether reporting subskill scores is reasonable, especially when the subskill scores are highly correlated.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

HL and YB proposed the original thoughts, designed the study, contributed to the article, and approved the submitted version. HL completed the draft writing, with the support of YB. All authors contributed to the article and approved the submitted version.

FUNDING

This research project is supported by Science Foundation of Beijing Language and Culture University (supported by the Fundamental Research Funds for the Central Universities) (Grant Number: 21YJ140003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.644764/full#supplementary-material

REFERENCES

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment.* New York, NY: Continuum.

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., and Ullakonoja, R. (2015). *The Diagnosis of Reading in a Second or Foreign Language*. New York, NY: Routledge. doi: 10.4324/9780203071830

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

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bernhardt, E. B. (2011). *Understanding Advanced Second-Language Reading*. New York, NY: Routledge. doi: 10.4324/9780203852408

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

Bonifay, W., and Cai, L. (2017). On the complexity of item response theory models. *Multivar. Behav. Res.* 52, 465–484. doi: 10.1080/00273171.2017.1309262

Brennan, R. L. (2006). "Perspectives on the evolution and future of educational measurement," in *Educational Measurement, 4th Edn,* ed R. L. Brennan (Westport, CT: ACE/Praeger Publishers), 1–16.

Browne, M. W., and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociol. Methods Res.* 21, 230–258. doi: 10.1177/0049124192021002005

Buck, G., and Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Langu. Test.* 15, 119–157. doi: 10.1177/026553229801500201

Buck, G., Tatsuoka, K., and Kostin, I. (1997). The subskills of reading: rule-space analysis of a multiple-choice test of second language reading comprehension. *Langu. Learn.* 47, 423–466. doi: 10.1111/0023-8333.00016

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edn.* New York, NY: Springer.

Cai, L. (2015). *flexMIRTTM Version 3.0: A Numerical Engine for Multilevel Item Factor Analysis and Test Scoring [Computer Software]*. Seattle, WA: Vector Psychometric.

Cai, L., and Hansen, M. (2012). Limited-information goodness-of-fit testing of hierarchical item factor models. *Br. J. Math. Statist. Psychol.* 66,245–276. doi: 10.1111/j.2044-8317.2012.02050.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

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

DiBello, L., Roussos, L., and Stout, W. (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. doi: 10.1016/S0169-7161(06)26031-0

DiBello, L., and Stout, W. (2008). *Arpeggio Documentation and Analyst Manual (Ver. 3.1.001) [Computer Software]*. St. Paul: MN: Assessment Systems Corporation.

Droop, M., and Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Read. Res. Q.* 38, 78–103. doi: 10.1598/RRQ.38.1.4

Fu, J. (2005). *A polytomous extension of the fusion model and its Bayesian parameter estimation* (Unpublished doctoral dissertation), University of Wisconsin-Madison, Madison, WI, United States.

Fu, J., and Li, Y. (2007). "Cognitively diagnostic psychometric models: an integrative review," *Paper presented at the Annual Meeting of the National Council on Measurement in Education* (Chicago, IL).

Fu, J., and Qu, Y. (2018). *A Review of Subscore Estimation Methods* (ETS Research Report. No. RR-18-17). Princeton, NJ: ETS. doi: 10.1002/ets2.12203

Grabe, W. (2009). *Reading in a Second Language: Moving from Theory to Practice*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139150484

Griffin, P. E., and Nix, P. (1991). *Educational Assessment and Reporting: A New Approach*. London, Harcourt Brace Jovanovich.

Grosjean, F. (2001). "The bilingual's language modes," in *One Mind, Two Languages: Bilingual Language Processing*, ed J. L. Nicol (Oxford: Basil Blackwell), 1–22.

Guthrie, J. T., and Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *J. Educ. Psychol.* 79, 220–227. doi: 10.1037/0022-0663.79.3.220

Haberman, S. J., and Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika* 75, 209–227. doi: 10.1007/s11336-010-9158-4

Haberman, S. J., von Davier, M., and Lee, Y.-H. (2008). *Comparison of Multidimensional Item Response Models: Multivariate Normal Ability Distributions Versus Multivariate Polytomous Distributions* (ETS Research Rep. No. RR-08-45). Princeton, NJ: ETS. doi: 10.1002/j.2333-8504.2008.tb02131.x

Hambleton, R. K., and Pitoniak, M. J. (2006). "Setting performance standards," in *Educational Measurement, 4th Edn,* ed R. L. Brennan (Westport, CT: ACE/Praeger Publishers), 433–470.

Harding, L., Alderson, J. C., and Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Langu. Test.* 32, 317–336. doi: 10.1177/0265532214564505

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Jang, E. E. (2005). *A validity narrative: effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL* (Unpublished doctoral dissertation), University of Illinois at Urbana-Champaign, Champaign, IL, United States.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for Fusion Model application to *LanguEdge* assessment. *Langu. Test.* 26, 31–73. doi: 10.1177/0265532208097336

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y. H., and Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: roles of length of residence and home language environment. *Langu. Learn.* 63, 400–436. doi: 10.1111/lang.12016

Kunnan, A. J., and Jang, E. E. (2009). "Diagnostic feedback in language assessment," in *The Handbook of Language Teaching,* eds M. Long and C. Doughty (Walden, MA: Wiley-Blackwell), 610–628. doi: 10.1002/9781444315783.ch32

Lee, Y.-W., and Sawaki, Y. (2009a). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Langu. Assess. Q.* 6, 239–263. doi: 10.1080/15434300903079562

Lee, Y.-W., and Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: an overview. *Langu. Assess. Q.* 6, 172–189. doi: 10.1080/15434300902985108

Li, H., Hunter, C. V., and Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Langu. Test.* 33, 391–409. doi: 10.1177/0265532215590848

Lumley, T. (1993). The notion of subskills in reading comprehension tests: an EAP example. *Langu. Test.* 10, 211–234. doi: 10.1177/026553229301000302

Ma, W., Iaconangelo, C., and de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Appl. Psychol. Measur.* 40, 200–217. doi: 10.1177/0146621615621717

Ma, W., Minchen, N., and de la Torre, J. (2020). Choosing between CDM and unidimensional IRT: the proportional reasoning test case. *Measur. Interdisc. Res. Perspect.* 18, 87–96. doi: 10.1080/15366367.2019.1697122

Mackey, A., and Gass, S. M. (2005). *Second Language Research: Methodology and Design*. Mahwah, NJ: Lawrence Erlbaum.

McLeod, L. D., Swygert, K. A., and Thissen, D. (2001). "Factor analysis for items scored in two categories," in *Test Scoring,* eds D. Thissen and H. Wainer (Mahwah, NJ: Lawrence Erlbaum), 189–216.

Mousavi, S. A. (2002). *An Encyclopaedic Dictionary of Language Testing:* Taipei: Tung Hua Book Company.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer. doi: 10.1007/978-0-387-89976-3

Robitzsch, A., Kiefer, T., George, A. C., and Uenlue, A. (2015). *CDM: Cognitive Diagnosis Modeling*. R package version 4.4-1. Available online at: https://cran.r-project.org/packages=CDM (accessed July 24, 2015).

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: The Guilford Press.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *J. Educ. Measur.* 47, 150–174. doi: 10.1111/j.1745-3984.2010.00106.x

Smith, F. (2004). *Understanding Reading: A Psycholinguistic Analysis of Reading and Learning to Read*. Mahwah, NJ: Lawrence Erlbaum.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Read. Res. Q.* 16, 32–71. doi: 10.2307/747348

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *J. Educ. Measur.* 44, 313–324. doi: 10.1111/j.1745-3984.2007.00041.x

Tatsuoka, K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. New York, NY: Routledge. doi: 10.4324/9780203883372

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

Yao, L., and Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Appl. Psychol. Measur.* 31, 83–105. doi: 10.1177/0146621606291559

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 Liu 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(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 August 2021
doi: 10.3389/fpsyg.2021.710497

An Adaptive Design for Item Parameter Online Estimation and Q-Matrix Online Calibration in CD-CAT

Wenyi Wang[1]*, *Yukun Tu*[1], *Lihong Song*[2]*, *Juanjuan Zheng*[1] and *Teng Wang*[1]

[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, [2] School of Education, Jiangxi Normal University, Nanchang, China

The implementation of cognitive diagnostic computerized adaptive testing often depends on a high-quality item bank. How to online estimate the item parameters and calibrate the **Q**-matrix required by items becomes an important problem in the construction of the high-quality item bank for personalized adaptive learning. The related previous research mainly focused on the calibration method with the random design in which the new items were randomly assigned to examinees. Although the way of randomly assigning new items can ensure the randomness of data sampling, some examinees cannot provide enough information about item parameter estimation or Q-matrix calibration for the new items. In order to increase design efficiency, we investigated three adaptive designs under different practical situations: (a) because the non-parametric classification method needs calibrated item attribute vectors, but not item parameters, the first study focused on an optimal design for the calibration of the Q-matrix of the new items based on Shannon entropy; (b) if the Q-matrix of the new items was specified by subject experts, an optimal design was designed for the estimation of item parameters based on Fisher information; and (c) if the Q-matrix and item parameters are unknown for the new items, we developed a hybrid optimal design for simultaneously estimating them. The simulation results showed that, the adaptive designs are better than the random design with a limited number of examinees in terms of the correct recovery rate of attribute vectors and the precision of item parameters.

Keywords: cognitive diagnostic computerized adaptive testing, item bank, item parameter, the Q-matrix, optimal design

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Xiaojian Sun,
Southwest University, China
Shuqun Yang,
Shanghai University of Engineering Sciences, China

***Correspondence:**
Wenyi Wang
wenyiwang@jxnu.edu.cn
Lihong Song
viviansong1981@163.com

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

Received: *16 May 2021*
Accepted: *30 July 2021*
Published: *24 August 2021*

Citation:
Wang W, Tu Y, Song L, Zheng J and Wang T (2021) An Adaptive Design for Item Parameter Online Estimation and Q-Matrix Online Calibration in CD-CAT. Front. Psychol. 12:710497. doi: 10.3389/fpsyg.2021.710497

INTRODUCTION

With the rapid growth of information technology and artificial intelligence in the era of big data, the form of test administration is changing. The paper-and-pencil tests have traditionally been widely used, but they are gradually being replaced by computer-based tests nowadays. The computer adaptive test (CAT) selects test items sequentially based on the examinee's current ability. Compared with CAT based on item response theory (IRT), cognitive diagnostic computerized adaptive test (CD-CAT), based on the cognitive diagnostic model (CDM) combines the dual advantages of computerized adaptive testing and cognitive diagnostic assessment

(Cheng, 2009; Wang and Tu, 2021). With using the idea of CAT and adopting certain item selection strategies, CD-CAT selects the items from the item bank that are most suitable for the examinee's attribute mastery pattern. Thus, CD-CAT not only increases test efficiency, but also can provide the examinee's cognitive strengths and weaknesses (Magis et al., 2017). The diagnosis feedback is useful for personalized learning to customize learning for each examinee's strengths, needs, skills, and interests.

In the past 20 years, many item selection strategies have been developed in CD-CAT, including the Kullback-Leibler (KL) information method and Shannon entropy (SHE) method (Xu et al., 2003), the Posterior-Weighted KL (PWKL) method (Cheng, 2009), the restrictive progressive method (RP; Wang C. et al., 2011), the mutual information (MI) method (Wang, 2013), the modified PWKL (MPWKL) method and the generalized deterministic inputs, noisy "and" gate (G-DINA) model discrimination index (GDI; Kaplan et al., 2015), the Jensen–Shannon divergence (JSD) method (Yigit et al., 2019), and the attribute-balance coverage method (Wang et al., 2020). A modified PWKL method that maximizes item information per time unit was developed by Huang (2019). For classroom assessment with small samples, Chang et al. (2018) proposed non-parametric item selection (NPS) and the weighted non-parametric item selection (WNPS) methods based on the non-parametric classification (NPC) method (Chiu and Douglas, 2013). The advantages of using non-parametric methods are that it only requires the **Q**-matrix to specify the relationship between items and attributes, and does not rely on item parameters which are often required by parametric models. And the study showed that the performance of these two methods is better than the PWKL method when the calibration samples are small. In addition, Yang et al. (2020) proposed three stratified item selection methods based on PWKL, NPS, and WNPS, named S-PWKL, S-NPS, and S-WNPS, respectively. Among them, the S-WNPS and S-NPS methods performed similarly and both of them are better than the S-PWKL method.

All the above methods require an item bank whose **Q**-matrix or item parameters is known to classify the examinees' attribute mastery patterns. Item banks often require new items or raw items to replace retired items. Thus, the specification of the **Q**-matrix or the calibration of item parameters for the new items is very important for the ongoing maintenance of the item bank. If subject experts are invited to specify the item parameters and attribute vectors for the new items, it will not only have big expenses, but also have the subjective components of uncertainty in experts' opinions. In addition, the precision of item parameters and the quality of **Q**-matrix will also directly affect the classification accuracy of the examinees' attribute mastery patterns. For example, Sun et al. (2020) demonstrated the negative effects of the calibration errors during the estimation of item parameters on the measurement accuracy, average test length, and test efficiency for variable-length CD-CAT. If the **Q**-matrix and the item parameters for new items can be automatically calibrated and accurately estimated based on examinees' item response data in the framework of CD-CAT, it can not only reduce labor costs, but also improve the efficiency of expert judgment by providing the results of online calibration (Chen and Xin, 2011).

In recent years, researchers have proposed three types of online estimation methods in IRT-based CAT. The first kind method includes maximum likelihood estimate (MLE), Stocking's Method A and Method B, which are based on the conditional maximum likelihood estimation. Based on the method A and the full-function maximum likelihood estimation (FFMLE) method (Stefanski and Carroll, 1985), Chen (2016) proposed the second method, called the FFMLE-Method A method. The third is the marginal maximum likelihood estimation with one EM cycle (OEM) proposed by Wainer and Mislevy (1990).

There are some online estimation methods in CD-CAT, including the CD-Method A, CD-OEM, CD-MEM method proposed by Chen et al. (2012). The method for online calibration item attribute vectors includes the intersection method proposed by Wang W. Y. et al. (2011). Based on the joint maximum likelihood estimation method, Chen and Xin (2011) proposed the joint estimation algorithm (JEA), which can calibrate item attribute vectors and estimate the item parameters for the new items. Inspired by the JEA algorithm, Chen et al. (2015) proposed the single-item estimation (SIE) method by taking the uncertainty of the attribute mastery pattern estimates into account and the simultaneous item estimation (SimIE) method to calibrate multiple items simultaneously. Results showed that the SIE and SimIE perform better than the JEA method in the calibration of the Q-matrix as well as the estimation of slipping and guessing parameters.

The related previous research in CD-CAT mainly focused on the calibration method with the random design in which the new items were randomly assigned to examinees. Although the way of randomly assigning new items can ensure the randomness of data sampling, some examinees cannot provide enough information about item parameter estimation or Q-matrix calibration for the new items. This becomes extremely important under the situation that the number of examinees is also limited and also would like to optimize the calibration of all new items (Chen et al., 2015). Naturally, we are badly in need of a design problem for how to adaptively assign new items to examinees according to both the current calibration of the new items and the current measurement of the examinees. In order to increase design efficiency, we investigated three adaptive designs under different practical situations: (a) because the non-parametric classification method needs calibrated item attribute vectors, but not item parameters, the first study focused on an optimal design for the calibration of the Q-matrix of the new items based on Shannon entropy; (b) if the Q-matrix of the new items was specified by subject experts, an optimal design was designed for the estimation of item parameters based on Fisher information; and (c) if the Q-matrix and item parameters are unknown for the new items, we developed a hybrid optimal design for simultaneously estimating them.

The rest of this paper is organized as follows: the next section will describe the models and methods in details,

including CDM, attribute mastery pattern estimation method, item parameters estimation method, and three optimal designs for online estimation and online calibration. The third section shows the simulation study about the design for the Q-matrix calibration based on Shannon entropy, the fourth section shows the simulation study about the design for online estimation based on Fisher information, and the fifth section shows the simulation study about the design for online estimation and calibration. The last section presents the summary and discussion, as well as future research directions.

MODELS AND METHODS

The Deterministic Inputs, Noisy, "and" Gate (DINA) Model

The DINA model (Macready and Dayton, 1977; Junker and Sijtsma, 2001) is one of the most commonly used cognitive diagnosis models. The observed item response X_{ij} for examinee i on item j is only right and wrong. If examinee i has mastered attribute k, $\alpha_{ik} = 1$, otherwise $\alpha_{ik} = 0$. The latent response for examinee i on item j is as follows:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}, \tag{1}$$

where K is the number of attributes, and the value of η_{ij} is 0 or 1. $\eta_{ij} = 1$ means that examinee i has mastered all the attributes measured by item j, while $\eta_{ij} = 0$ means that examinee i has not mastered at least one of the attributes of item j. However, it is not certain that if you master all the attributes examined by the item j, you will be able to answer the item correctly. It may be due to the examinees' mistakes that the item will not be answered correctly. Similarly, although they did not master all the attributes of the item, they have the chance to guess the correct answer. Therefore, the combination of slipping and guessing is called noise. In other words, the two item parameters in the DINA model, the slipping parameter s_j and the guessing parameter g_j, represent the probability of noise on item j. They are defined as follows:

$$s_j = P\left(X_{ij} = 0 \,|\eta_{ij} = 1\right), \tag{2}$$

$$g_j = P\left(X_{ij} = 1 \,|\eta_{ij} = 0\right). \tag{3}$$

When the latent response variable η_{ij}, s_j, g_j is known, the item response probability of examinee i on item j under the DINA model can be calculated as follows:

$$P\left(X_{ij} = 1 \,|\boldsymbol{\alpha}_i\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{(1-\eta_{ij})}, \tag{4}$$

where, $P\left(X_{ij} = 1 \,|\boldsymbol{\alpha}_i\right)$ refers to the correct response probability of item j for examinee i whose attribute mastery pattern is $\boldsymbol{\alpha}_i$. A high probability of getting the item right implies that examinees mastered all the required attributes of an item. As long as the examinees have not mastered a certain required attribute of an item, they will answer the item correctly with a low probability.

Attribute Mastery Pattern Estimation

The estimation methods of examinees' attribute mastery pattern mainly include maximum a posteriori (MAP), expected a posteriori (EAP), and maximum likelihood estimation (MLE). This study uses the MLE method. Assuming the length of CD-CAT is fixed at m, under the assumption of local independence, the examinee's conditional likelihood function is:

$$\begin{aligned} L\left(\boldsymbol{X}_i \,|\boldsymbol{\alpha}_c\right) &= \prod_{j=1}^{m} P\left(X_{ij} \big| \boldsymbol{\alpha}_c\right) \\ &= \prod_{j=1}^{m} \left[\left(1 - s_j\right)^{X_{ij}} s_j^{1-X_{ij}}\right]^{\eta_{cj}} \left[g_j^{X_{ij}} \left(1 - g_j\right)^{1-X_{ij}}\right]^{1-\eta_{cj}}. \end{aligned} \tag{5}$$

Then the maximum likelihood estimation of the attribute mastery pattern is the one that maximizes the value of the conditional likelihood function:

$$\hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_c \in Q_S}{\operatorname{argmax}} \left\{L\left(X_i \,|\boldsymbol{\alpha}_c\right)\right\}. \tag{6}$$

Item Parameter Estimation Method

If we know the attribute vectors of the new items, the item parameter estimation method can use the CD-Method A method proposed by Chen et al. (2012). This method is extended from the traditional CAT online calibration method called the Method A to CD-CAT, by using maximum likelihood estimation method to estimate item parameters.

Assuming that n_j independent examinees have answered item j, the logarithmic likelihood function given the observed response x_{ij} on item j is calculated as follow:

$$\ln L_j = \sum_{i=1}^{n} \left(u_{ij} \ln(P_j(\hat{\boldsymbol{\alpha}}_i)) + (1 - u_{ij}) \ln(1 - P_j(\hat{\boldsymbol{\alpha}}_i))\right). \tag{7}$$

We take the partial derivatives of the logarithmic likelihood with respect to g_j and s_j and let them equal to 0

$$\frac{\partial \ln L_j}{\partial g_j} = \sum_{i:\substack{\eta_{ij}=0 \\ u_{ij}=0}} \frac{-g_j}{g_j\left(1 - g_j\right)} + \sum_{i:\substack{\eta_{ij}=0 \\ u_{ij}=1}} \frac{1 - g_j}{g_j\left(1 - g_j\right)} = 0, \tag{8}$$

$$\frac{\partial \ln L_j}{\partial s_j} = \sum_{i:\substack{\eta_{ij}=1 \\ u_{ij}=0}} \frac{1 - s_j}{\left(1 - s_j\right) s_j} + + \sum_{i:\substack{\eta_{ij}=1 \\ u_{ij}=1}} \frac{-s_j}{\left(1 - s_j\right) s_j} = 0. \tag{9}$$

Then the estimated value of the guessing parameter is $\hat{g}_j = n_2/(n_1 + n_2)$, where n_1 represents the number of examinees whose latent response is 0 and the observed response is also 0, and n_2 represents the number of examinees when the latent response is 0 but the observed response is 1. Similarly, the estimated value of the slipping parameter is $\hat{s}_j = n_3/(n_3 + n_4)$, where n_3 represents the number of examinees whose latent response is 1 but the observed response is 0, and n_4 represents the number of examinees when the latent response is 1 and the observed response

is also 1. Therefore, only the value of n_1, n_2, n_3, and n_4 is needed to calculate the estimated values of guessing and slipping parameters.

An Adaptive Design for Q-Matrix Calibration Based on Shannon Entropy

The adaptive design for **Q**-matrix calibration based on Shannon entropy is designed to select the most suitable new item for the examinees to answer, in order to determine the attribute vector of the new item as soon as possible. The steps of the adaptive design algorithm are as follows:

(1) Calculate the posterior probability of the attribute vector of the new item j based on item response data and the examinee's attribute mastery pattern:

$$P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \right.\right) = \frac{P\left(\boldsymbol{q}_r\right) L\left(\boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \left| q_r \right.\right)}{\sum_{\boldsymbol{q}_r \in \mathbf{Q}_r} P\left(\boldsymbol{q}_r\right) L\left(X^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \left| \boldsymbol{q}_r \right.\right)}, \quad (10)$$

where, $\mathbf{Q}_r$ is the set of all possible of attribute vectors, the prior probability $P\left(\boldsymbol{q}_r\right)$ is set to a uniform distribution, $\hat{\boldsymbol{\alpha}}^{(j)} = \left(\hat{\boldsymbol{\alpha}}_1^{(j)}, \hat{\boldsymbol{\alpha}}_2^{(j)}, ..., \hat{\boldsymbol{\alpha}}_{n_j}^{(j)}\right)$ is the attribute mastery pattern matrix estimated by all examinees who has answered item j, $\boldsymbol{X}^{(j)} = \left(X_{1j}, X_{2j}, ..., X_{n_j j}\right)$ is the vector of item responses for all examinees answered item j, n_j is the number of examinees answered item j, and the likelihood function of the attribute vector $\boldsymbol{q}_r$ is:

$$L\left(\boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \left| \boldsymbol{q}_r \right.\right) = \prod_{i=1}^{n_j} P\left(X_{ij} = x \left| \hat{\boldsymbol{\alpha}}_i^{(j)}, \boldsymbol{q}_r \right.\right)^{X_{ij}} \left(1 - P\left(X_{ij} = x \left| \hat{\boldsymbol{\alpha}}_i^{(j)}, \boldsymbol{q}_r \right.\right)\right)^{1 - X_{ij}}. \quad (11)$$

(2) Calculate the Shannon entropy of the current posterior distribution of the attribute vector of the new item j:

$$SHE_j = -\sum_{\boldsymbol{q}_r \in Q_r} P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \right.\right) \log P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \right.\right). \quad (12)$$

Assuming that examinee i, whose attribute mastery pattern is estimated to be $\hat{\boldsymbol{\alpha}}_i$, with item response $X_{ij} = x$ on the candidate new item j, then the posterior distribution of the attribute vector:

$$P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)}, X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \right.\right) = \frac{L\left(X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \left| \boldsymbol{q}_r \right.\right) P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \right.\right)}{\sum_{\boldsymbol{q}_r \in Q_r} L\left(X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \left| \boldsymbol{q}_r \right.\right) P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \right.\right)}, \quad (13)$$

where $L\left(X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \left| \boldsymbol{q}_r \right.\right) = P\left(X_{ij} = x \left| \hat{\boldsymbol{\alpha}}_i, \boldsymbol{q}_r \right.\right)^x \left(1 - P\left(X_{ij} = x \left| \hat{\boldsymbol{\alpha}}_i, \boldsymbol{q}_r \right.\right)\right)^{1-x}$

(3) The Shannon entropy expectation of the item response X_{ij} on the candidate new item j is:

$$SHE_{ij} = -\sum_{x=0}^{1} P\left(X_{ij} = x\right) \left(\sum_{\boldsymbol{q}_r \in \mathbf{Q}_r} P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)}, X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \right.\right) \log P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)}, X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \right.\right)\right), \quad (14)$$

where $P\left(X_{ij} = x\right) = \sum_{\boldsymbol{q}_r \in \mathbf{Q}_r} L\left(X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \left| \boldsymbol{q}_r \right.\right) P\left(\boldsymbol{q}_r \left| \boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)}, X_{ij} = x, \hat{\boldsymbol{\alpha}}_i \right.\right)$

(4) Choose the new item with the smallest difference between SHE_j and SHE_{ij},

$$j = \underset{j \in N^{(i)}}{\operatorname{argmin}} \left(SHE_{ij} - SHE_j\right), \quad (15)$$

where $N^{(i)}$ represents the set of new items that the examinee i has not answered yet. The difference between SHE_j and SHE_{ij} is refered as Mutual information.

(5) Collect the item response X_{ij} on item j and the attribute mastery pattern $\hat{\boldsymbol{\alpha}}_i$ of examinee i, adjoining them to the matrix $\boldsymbol{X}^{(j)}$ and $\hat{\boldsymbol{\alpha}}^{(j)}$, that is $\boldsymbol{X}^{(j)} = (\boldsymbol{X}^{(j)}, X_{ij})$ and $\hat{\boldsymbol{\alpha}}^{(j)} = (\hat{\boldsymbol{\alpha}}^{(j)}, \hat{\boldsymbol{\alpha}}_i)$.

(6) Repeat steps 1 through 5 until the new item meets the required number of examinees.

(7) When the number of examinees of item j meets the specified conditions, the MLE method is used to estimate its **q** vector

$$\hat{\mathbf{q}}_j = \underset{\boldsymbol{q}_r \in \mathbf{Q}_r}{\operatorname{argmax}} \left\{L\left(\boldsymbol{X}^{(j)}, \hat{\boldsymbol{\alpha}}^{(j)} \left| \boldsymbol{q}_r \right.\right)\right\}. \quad (16)$$

An Adaptive Design for Item Parameter Estimation Based on Fisher Information

The most commonly way of measuring the precision of estimated parameters used in IRT is the Fisher information. The more information there is in the sample, the more accurate the estimated parameter is. The calculation formula is as follows:

$$I(\theta) = E\left\{\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right\} = -E\left\{\frac{\partial^2 \ln L(\theta)}{\partial^2 \theta}\right\} = \sum_{i=1}^{n} \frac{\left(P_i'\right)^2}{P_i Q_i}, \quad (17)$$

where $L(\theta)$ is the likelihood function.

According to this theory, we apply it to CD-CAT and propose an adaptive design for estimating item parameters. One cognitive diagnosis model used in this study is the DINA model having slipping and guessing parameters. For the maximum likelihood estimation of the item parameter vector $\hat{\boldsymbol{\gamma}}_j = (\hat{s}_j, \hat{g}_j)^T$, the estimation error of item parameters can usually be described by the item information matrix $I^{(n)}(\hat{\boldsymbol{\gamma}}_j)$. Therefore, we use the form of information matrix to choose the most appropriate item for the examinee to answer, so as to increase the precision of item parameter estimation. That is, the item is selected based on the D-optimal design criteria:

$$j = \underset{j \in N^{(i)}}{\operatorname{argmax}} \frac{\det(I^{(n_j)}(\hat{\boldsymbol{\gamma}}_j) + I(\hat{\boldsymbol{\gamma}}_j; \hat{\boldsymbol{\alpha}}_i)) - \det(I^{(n_j)}(\hat{\boldsymbol{\gamma}}_j))}{\det(I^{(n_j)}(\hat{\boldsymbol{\gamma}}_j)}, \quad (18)$$

where $I^{(n_j)}\left(\hat{\boldsymbol{\gamma}}_j\right)$ represents the current amount of information of item j obtained by the sample size of n_j, $I\left(\hat{\boldsymbol{\gamma}}_j; \hat{\boldsymbol{\alpha}}_i\right)$ represents the amount of information generated by the examinee i after answering item j, det(•) represents the determinant value of the matrix, and $I^{(n_j)}\left(\hat{\boldsymbol{\gamma}}_j\right)$ is calculated as follows:

$$I^{(n_j)}(\hat{\boldsymbol{\gamma}}_j) = \sum_{i=1}^{n_j} I(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\alpha}}_i), \quad (19)$$

$$I(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\alpha}}_i) = \frac{1}{P_{ij}\left(1 - P_{ij}\right)} \begin{bmatrix} \left(\eta_{ij}\left(1 - s_j\right)^{\eta_{ij}-1} g_j^{1-\eta_{ij}}\right)^2 & 0 \\ 0 & \left(\left(1 - \eta_{ij}\right) g_j^{-\eta_{ij}}\left(1 - s_j\right)^{\eta_{ij}}\right)^2 \end{bmatrix}. \quad (20)$$

When $\eta_{ij} = 1$ or $\eta_{ij} = 0$, we have

$$I\left(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\alpha}}_i\right) = \frac{1}{P_{ij}\left(1 - P_{ij}\right)} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \frac{1}{s_j\left(1 - s_j\right)} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad (21)$$

or

$$I\left(\hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\alpha}}_i\right) = \frac{1}{P_{ij}\left(1 - P_{ij}\right)} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{g_j\left(1 - g_j\right)} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \quad (22)$$

Algorithms for Three Adaptive Designs

Item Attribute Vector Online Calibration Algorithm Based on Shannon Entropy

The flow chart of the calibration of item attribute vector based on Shannon entropy is shown in **Figure 1**. The specific steps are as follows:

Step 1. For the examinee *i*, the SHE is used to select the item from the operational item bank, and the item response is collected.
Step 2. The MLE method is used to estimate the attribute mastery pattern of examinee *i*.
Step 3. Repeat steps 1–2 until the examinee *i* has answered 12 operational items, and finally get the estimated attribute mastery pattern.
Step 4. Based on the final estimated attribute mastery pattern, the adaptive design for online calibration is adopted to select 6 new items from the new item bank for examinee *i* and collect the responses on the new items.
Step 5. Update the posterior probability of the **q**-vector of the new item and repeat the previous step.
Step 6. Use the MLE method to calibrate the **q**-vector of each new item until the number of responses to the new item meets the condition.

Item Parameter Online Estimation Algorithm Based on Fisher Information

The flow chart of the estimation of item parameters based on Fisher information is shown in **Figure 2**. The specific steps are as follows:

Step 1. For the examinee *i*, SHE is used to select the item from the operational item bank, and the item response is collected.
Step 2. Using the MLE method to estimate the attribute mastery pattern of the examinee *i*.
Step 3. Repeat steps 1–2 until the examinee *i* has answered 12 basic items, and finally get the estimated attribute mastery pattern.
Step 4. Based on the final estimated attribute mastery pattern, the D-optimal method is adopted to select 6 new items from the new item bank for the examinee *i* and collect item responses on the new items.
Step 5. Estimate the item parameters by the CD-Method A method, update Fisher information for the new items, and repeat the previous steps.
Step 6. Use the CD-Method A method to get the final estimated item parameters until the number of responses to the new item meets the condition.

Adaptive Design Algorithm for Online Estimation and Calibration

The flow of online estimation and calibration of adaptive design was shown in **Figure 3**. The specific algorithm steps are as follows:

Step 1. For the examinee *i*, use the SHE to select items from the operational item bank, and collect the response of the item.
Step 2. Using the MLE method to estimate the attribute mastery pattern of the examinee *i*.
Step 3. Repeat step 1–2 until the examinee *i* has answered 12 operational items, and finally get the estimated attribute mastery pattern of all the examinees.
Step 4. When the examinee answers to the position of the preset new item, judge the value of $\frac{n}{N}$ (n is the number of the current examinees). If it satisfies $n/N < 0.8$, SHE-optimal

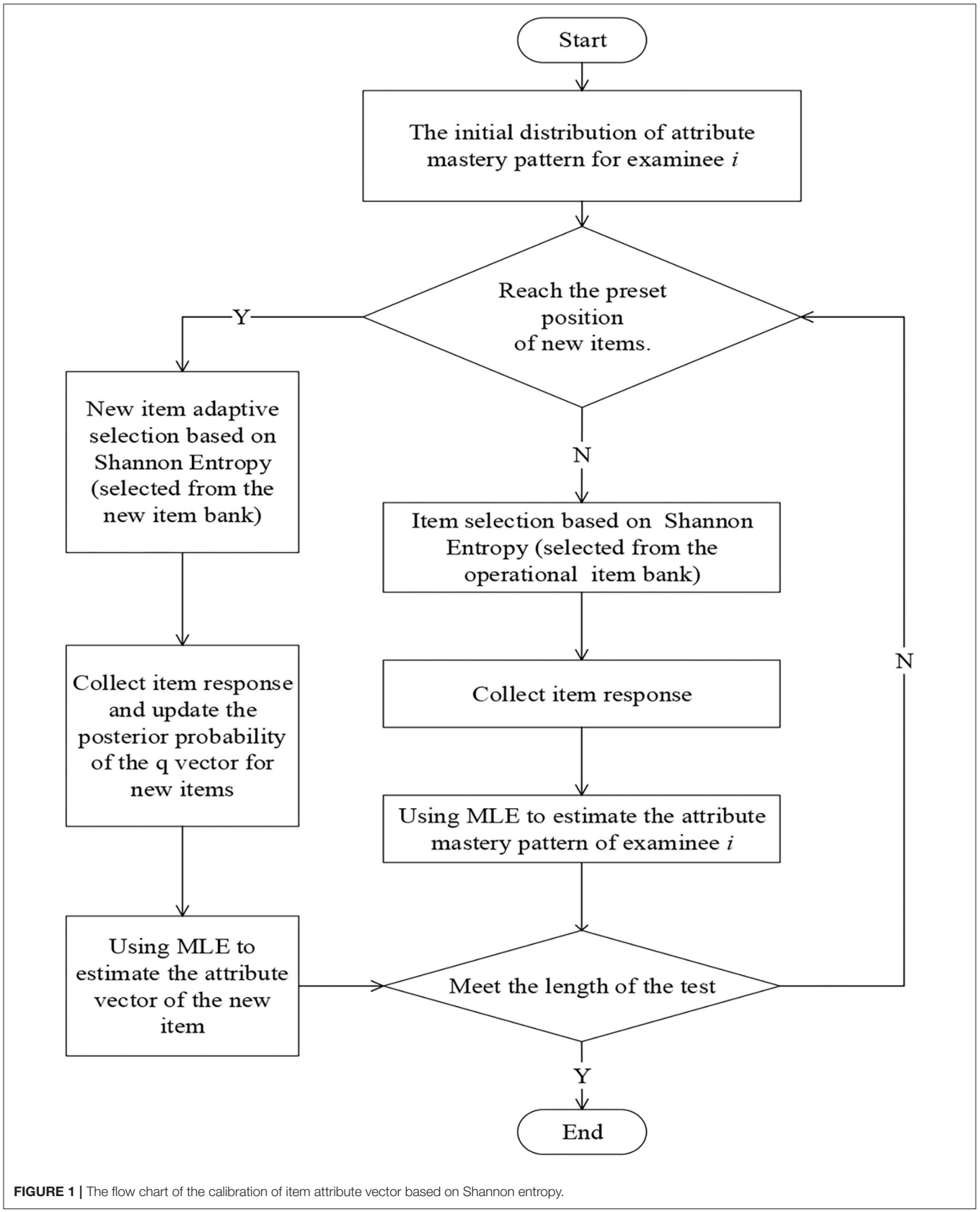


FIGURE 1 | The flow chart of the calibration of item attribute vector based on Shannon entropy.

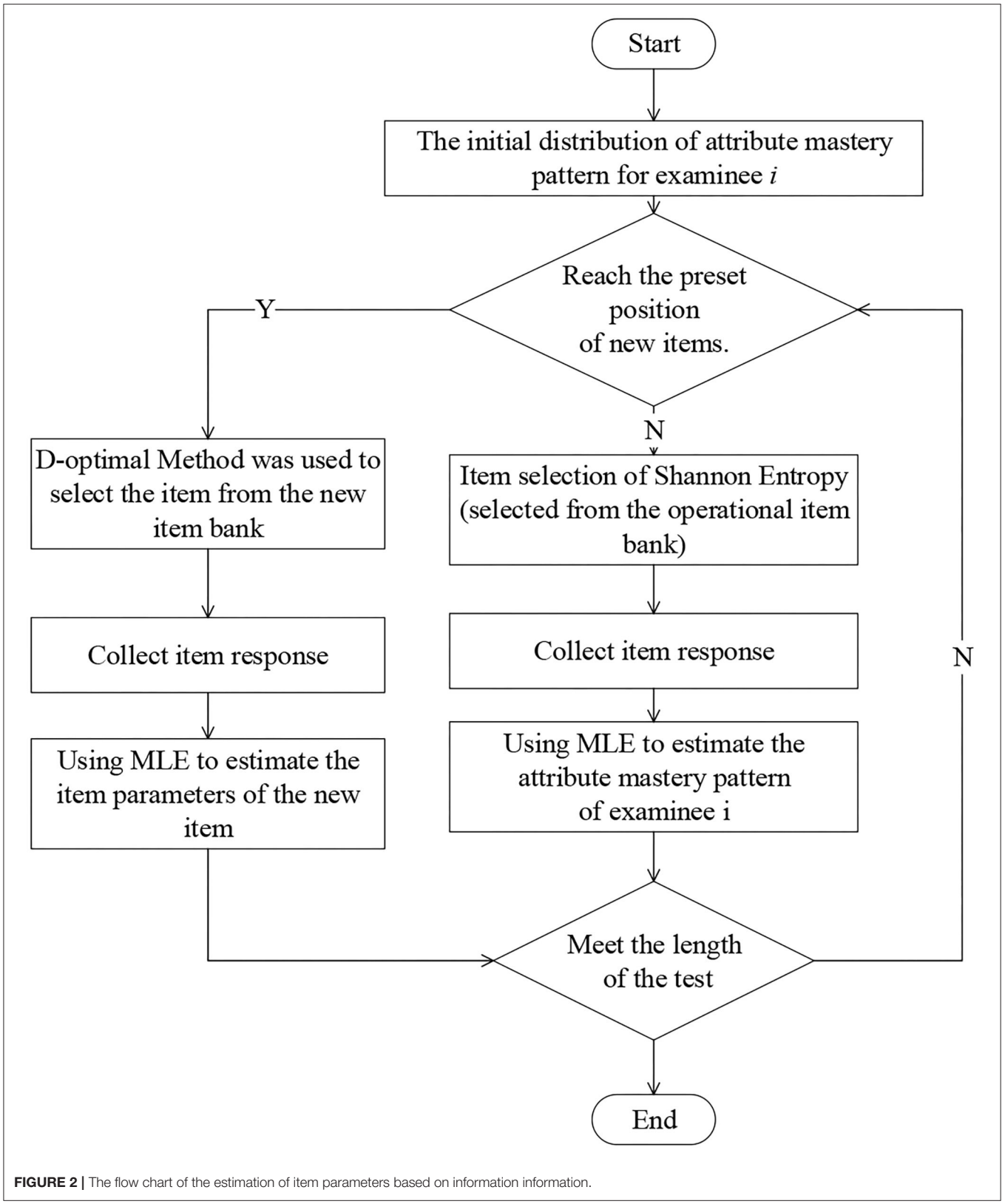


FIGURE 2 | The flow chart of the estimation of item parameters based on information information.

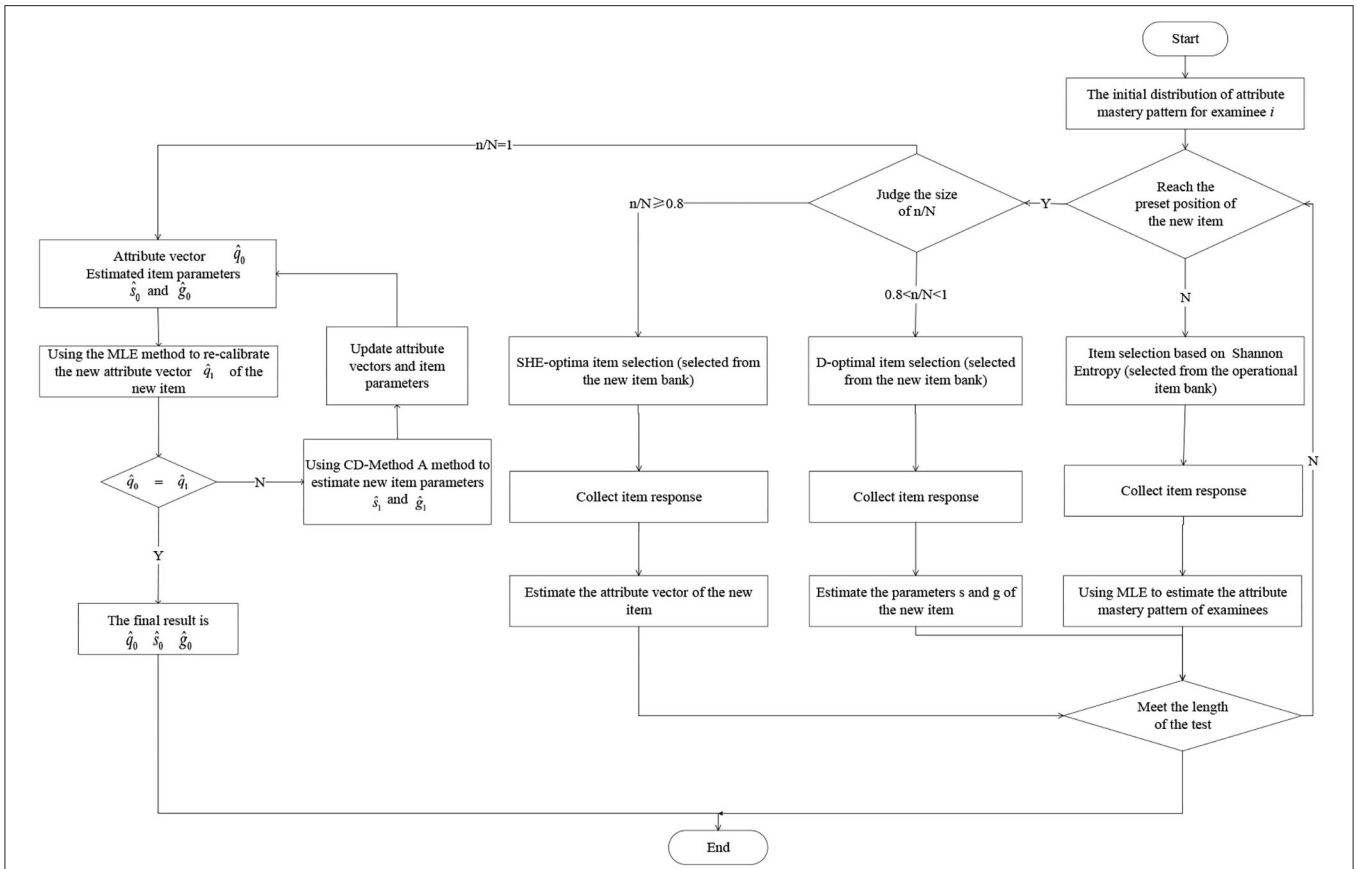


FIGURE 3 | Flow chart of adaptive design for online estimation and calibration.

criterion (or random method) will be used to assign the new item to the examinee, and the response of the examinee to the new item is also collected, and the posterior probability of the attribute vector of the new item is updated. When $n/N = 0.8$, the attribute vector $\hat{q}_0$ is estimated. If $n/N > 0.8$, adopts D-optimal criterion (or random method) to select new items for examinees, and then collects the responses of examinees on the new items, the item parameters of the new item are updated by the CD-Method A method and the attribute vector $\hat{q}_1$ is estimated. When $n/N = 1$, the estimated values of item parameters are $\hat{s}_0$ and $\hat{g}_0$.

Step 5. If $\hat{q}_0 \neq \hat{q}_1$, update the attribute vector $\hat{q}_0$ and then updates the item parameters $\hat{s}_0$ and $\hat{g}_0$.

Step 6. Repeat the above two steps until $\hat{q}_0 = \hat{q}_1$, gets the final attribute vector and item parameter estimation.

TABLE 1 | Item parameters and **Q**-matrix in the new item bank.

New item index	Slipping parameter	Guessing parameter	Q-matrix
1	0.32	0.30	1 0 0 1 0
2	0.18	0.12	0 0 1 0 0
3	0.39	0.32	1 0 1 0 0
4	0.13	0.18	0 1 1 0 0
5	0.38	0.24	0 1 0 0 0
6	0.37	0.20	1 0 1 0 1
7	0.15	0.15	0 0 0 1 0
8	0.16	0.39	0 0 0 1 0
9	0.39	0.13	0 1 0 0 0
10	0.23	0.39	0 1 0 1 1
11	0.30	0.27	1 0 0 0 0
12	0.14	0.18	0 0 0 1 1

SIMULATION STUDY 1

Simulation Design

The purpose of the first simulation study focused on the calibration of Q-matrix to satisfy the requirements of the NPC method. This study mainly discusses the influence of adaptive or random design on the attribute vector calibration. The simulation design in the study of Chen et al. (2015) was used here. Matlab 8.6.0 (R2015b) and R (version 4.1.0) were used in simulation studies. Because the adaptive designs of Fisher information or Shannon entropy were successfully used to sequentially select items based on the status of examinees in CAT (Magis et al., 2017) or CD-CAT (Cheng, 2009). We expect that the

TABLE 2 | The TSR for the two allocation strategies under different initial values of the item parameters.

Sample size	TSR							
	Random allocation				Item allocation based on Shannon entropy			
	0.05	0.15	0.25	0.35	0.05	0.15	0.25	0.35
100	0.61	0.60	0.59	0.61	0.66	0.68	0.64	0.61
200	0.73	0.73	0.71	0.74	0.80	0.78	0.78	0.74
400	0.84	0.83	0.82	0.83	0.86	0.88	0.86	0.83
800	0.91	0.91	0.89	0.91	0.96	0.96	0.91	0.91
1,600	0.97	0.97	0.97	0.95	1.00	0.99	0.99	0.95

proposed adaptive designs can be applied for online calibration of attribute vectors and online estimation of item parameters for new items.

Operational Item Bank

The items in the item bank mainly include **Q**-matrix and item parameters. The number of independent attributes is $K = 5$. Based on the recent research on online calibration, the number of items in the operational item bank is 240, in which composed of 16 $\mathbf{Q}_1$, 8 $\mathbf{Q}_2$ and 8 $\mathbf{Q}_3$. $\mathbf{Q}_1$, $\mathbf{Q}_2$ and $\mathbf{Q}_3$ consist of matrices that examine one, two, and three attributes, respectively, as shown below.

$$\mathbf{Q}_1 = \begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{pmatrix} \mathbf{Q}_2 = \begin{pmatrix} 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 0&1&1&0&0\\ 0&1&0&1&0\\ 0&1&0&0&1\\ 0&0&1&1&0\\ 0&0&1&0&1\\ 0&0&0&1&1 \end{pmatrix} \mathbf{Q}_3 = \begin{pmatrix} 1&1&1&0&0\\ 1&1&0&1&0\\ 1&1&0&0&1\\ 1&0&1&1&0\\ 1&0&1&0&1\\ 1&0&0&1&1\\ 0&1&1&1&0\\ 0&1&1&0&1\\ 0&1&0&1&1\\ 0&0&1&1&1 \end{pmatrix}$$

The slipping and guessing parameters of each item in the operational item bank followed a uniformly distributed $U(0.1, 0.4)$.

New Item Bank

The items in the new item bank also include the **Q**-matrix and the item parameters (slipping and guessing parameters). According to the simulation design in the study of Chen et al. (2015), there are a total of 12 items in the new item bank ($M = 12$). The item parameter distribution of the new items obeys the uniform distribution $U(0.1,\ 0.4)$. The item parameters and the **Q**-matrix are listed in **Table 1**.

Simulation Procedures

In this study, we consider the influence of different sample size on the calibration results. The sample size is set to $N =$ 100, 200, 400, 800, and 1,600. We assume that each examinee mastered each attribute with 50% probability. Each examinee needs to answer 12 operational items and 6 new items ($D =$ 6). The number of examinees who answered each new item is about $C = (N \times D)/12$.

TABLE 3 | The standard deviation of the attribute vector under the two allocation strategies.

Sample size	SD	
	Random allocation	Item allocation based on Shannon entropy
100	0.1280	0.1137
200	0.1039	0.1124
400	0.1030	0.0791
800	0.0782	0.0636
1,600	0.0447	0.0286

For the random design, the balanced incomplete block design (BIBD) is applied to guarantee that each examinee will answer six different new items and the number of examinees to each new item is balanced. For example, we called the function find. BIB (12,100,6) in the R library and *crossdes* to search for balanced incomplete block designs, where the sample size is 100, the total number of new items is 12, and the number of new items answered by each examinee is 6. After item responses are collected, the MLE method is applied to estimate the q-vector for the new item.

For the Shannon entropy allocation strategy, the procedures of the calibration of Q-matrix under are the following: firstly, item parameters and attribute mastery pattern estimates of examinees are required to computer item response functions under each possible attribute vector of the new item; secondly, item response functions and item responses are used to update the posterior distribution of the attribute vector of the new item; thirdly, the mutual information is obtained for adaptively choosing the next new item to the current examinee; finally, the MLE method is applied to estimate the q-vector for the new item.

For the two designs above, item parameters are required for computer item response functions. However, item parameters for the new items are unknown and cannot estimate item response probabilities. Thus, the item parameters of all new items are fixed as the same and four levels are considered to investigate the

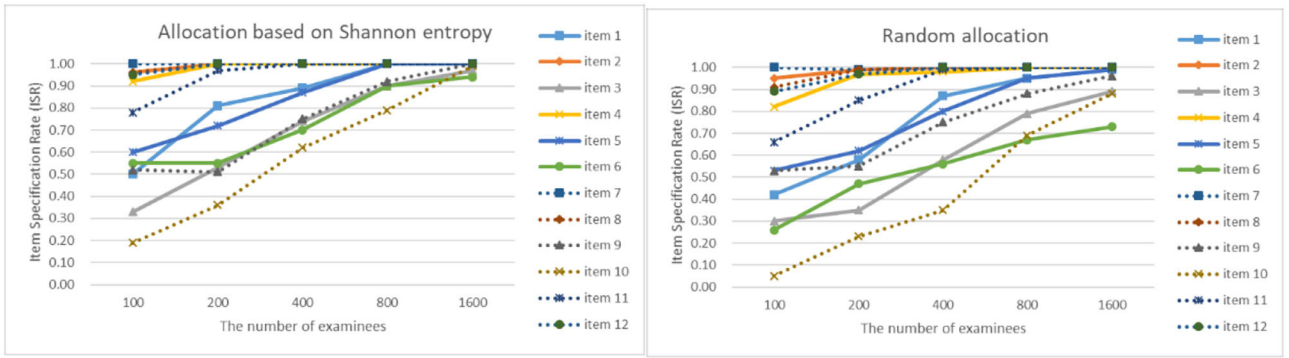


FIGURE 4 | The ISR for two strategies when the initial value of the item parameter is 0.05.

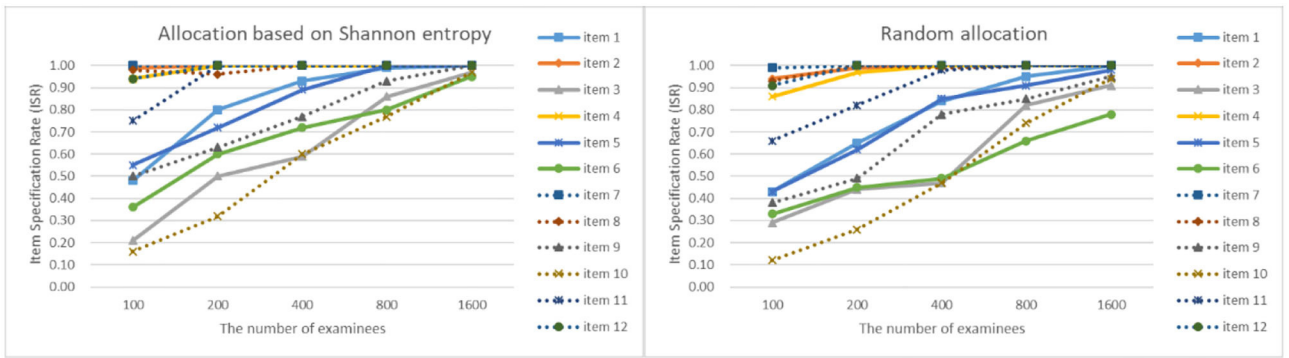


FIGURE 5 | The ISR for two strategies when the initial value of the item parameter is 0.15.

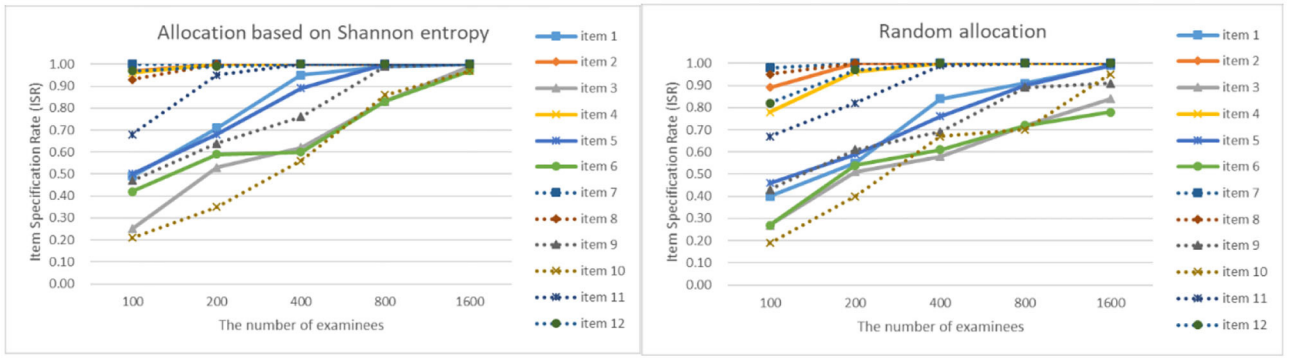


FIGURE 6 | The ISR for two strategies when the initial value of the item parameter is 0.25.

impact of different item parameters on the Q-matrix calibration. The four levels are 0.05, 0.15, 0.25, and 0.35. The reason is that for the NPC method, a frequently used distance measure is the Hamming distance (Chiu and Douglas, 2013), which counts the number of different entries in observed and ideal item response vectors with binary data. For this case, slipping and guessing parameters can be regarded as the same for all items. Thus, item response probabilities are calculated from the DINA model by using the same item parameters in the first design. Repeat $R = 100$ times under each condition.

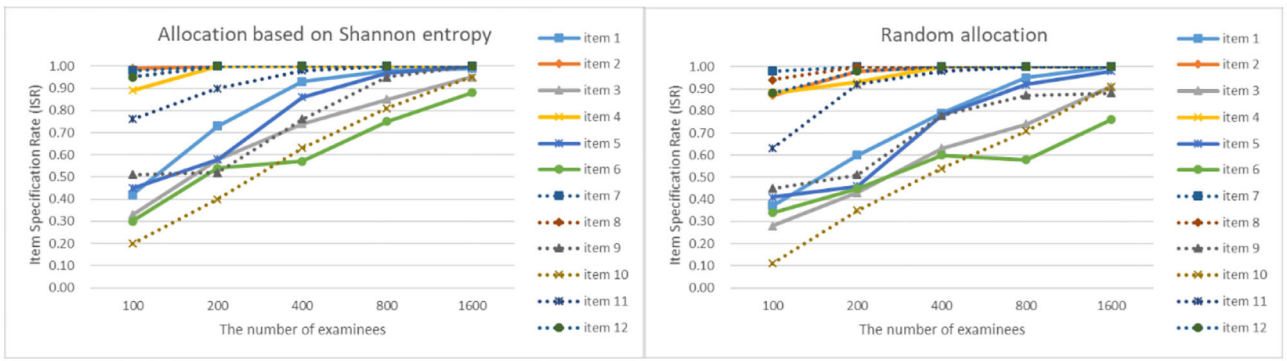


FIGURE 7 | The ISR for two strategies when the initial value of the item parameter is 0.35.

Evaluation Indices

Item specification rate (ISR) refers to the accuracy of estimating **q**-vector for each new item. ISR can be written as

$$ISR_j = \frac{1}{R}\sum_{l=1}^{R} I(\hat{\mathbf{q}}_j^{(l)} = q_j), \tag{23}$$

where R represents the number of repetitions, $\hat{\mathrm{q}}_j^{(l)}$ represents the lth estimation, and the indicator function I (.) takes value 1 when $\hat{\boldsymbol{q}}_j^{(l)} = \boldsymbol{q}_j$ and value 0 when $\hat{\boldsymbol{q}}_j^{(l)} \neq \boldsymbol{q}_j$.

Total specification rate (TSR) refers to the average accuracy of **q**-vectors for all new items. TSR can be written as

$$TSR = \frac{1}{JR}\sum_{j=1}^{J}\sum_{l=1}^{R} I(\hat{\mathbf{q}}_j^{(l)} = q_j). \tag{24}$$

The ISR and TSR are used to compare the performance of the Shannon entropy allocation strategy and random allocation strategy. The higher the ISR and TSR is, the more accurate the calibration of **Q**-matrix is.

Standard deviation (SD) refers to the stability of the method when the estimation accuracy of attribute vectors is similar.

$$SD = \sqrt{\frac{1}{R-1}\sum_{l=1}^{R}(TMR_l - r)^2}. \tag{25}$$

where total misspecification rate (TMR) equals to $1 - TSR$, and r is the average of TMR. The smaller the standard deviation is, the more stable the method is.

Simulation Results

Results about the accuracy of the two allocation designs are shown in **Table 2** for different the initial item parameters used in both item allocation and estimation. From the table, no matter the random or the Shannon entropy allocation strategy, the TSR increases with the increase of the number of examinees. When the sample size reaches a certain value (e.g., 1,600), the TSR is close to 1. The TSR of Shannon entropy allocation strategy is higher than that of the random allocation strategy, especially when the sample size is 100, 200,400, or 800. Although the initial values of item parameters are different, the TSR of the two allocation designs under different item parameters is almost the same. The reason for the small difference results from the different attribute mastery pattern estimates of the examinees under each condition.

Table 3 presents the stability of the two allocation designs under the different initial item parameters. From the standard deviation of the TMR, when the sample size is 400, 800, and 1,600, the SD from the Shannon entropy allocation strategy is much smaller than that of the random allocation strategy. It means that the calibration accuracy of the allocation strategy based on Shannon entropy is more stable than the random allocation strategy.

As can be seen from **Figures 4**–**7**, as the sample size gets larger, the increase of ISR is very obvious. The difference in ISR between the two allocation strategies is relatively small that in TSR. Meanwhile, the ISR of items 1, 3, 5, 6, 9, 10, and 11 changes obviously, and the performance of the two allocation strategies on these items is obviously better than other items. This is due to the fact that the item parameters of these items are larger than that of other items. And when the true item parameters are large, the ISR from the allocation strategy based on Shannon entropy is higher than the random allocation strategy. The larger the item parameters on the new item, the larger the sample size of **Q**-matrix calibration required. But the sample size required by Shannon entropy allocation strategy is less than the random allocation strategy.

Figure 8 shows the distribution of attribute mastery patterns selected by Shannon entropy allocation strategy for each new item for the number of examinees of 1,600 and the initial item parameters of 0.15. Each new item is assigned to ∼800 examinees. From **Table 1**, the true attribute vector is 00100 for item 2. We found that the top five attribute mastery patterns for the item are 01010, 11010, 10110, 10000, and 10001, respectively. Most examinees with the third attribute mastery pattern can answer the item correctly, while examinees with the other attribute mastery patterns cannot answer the item correctly. Intuitively, these attribute mastery patterns can effectively discriminate the true attribute vector with other attribute vectors.

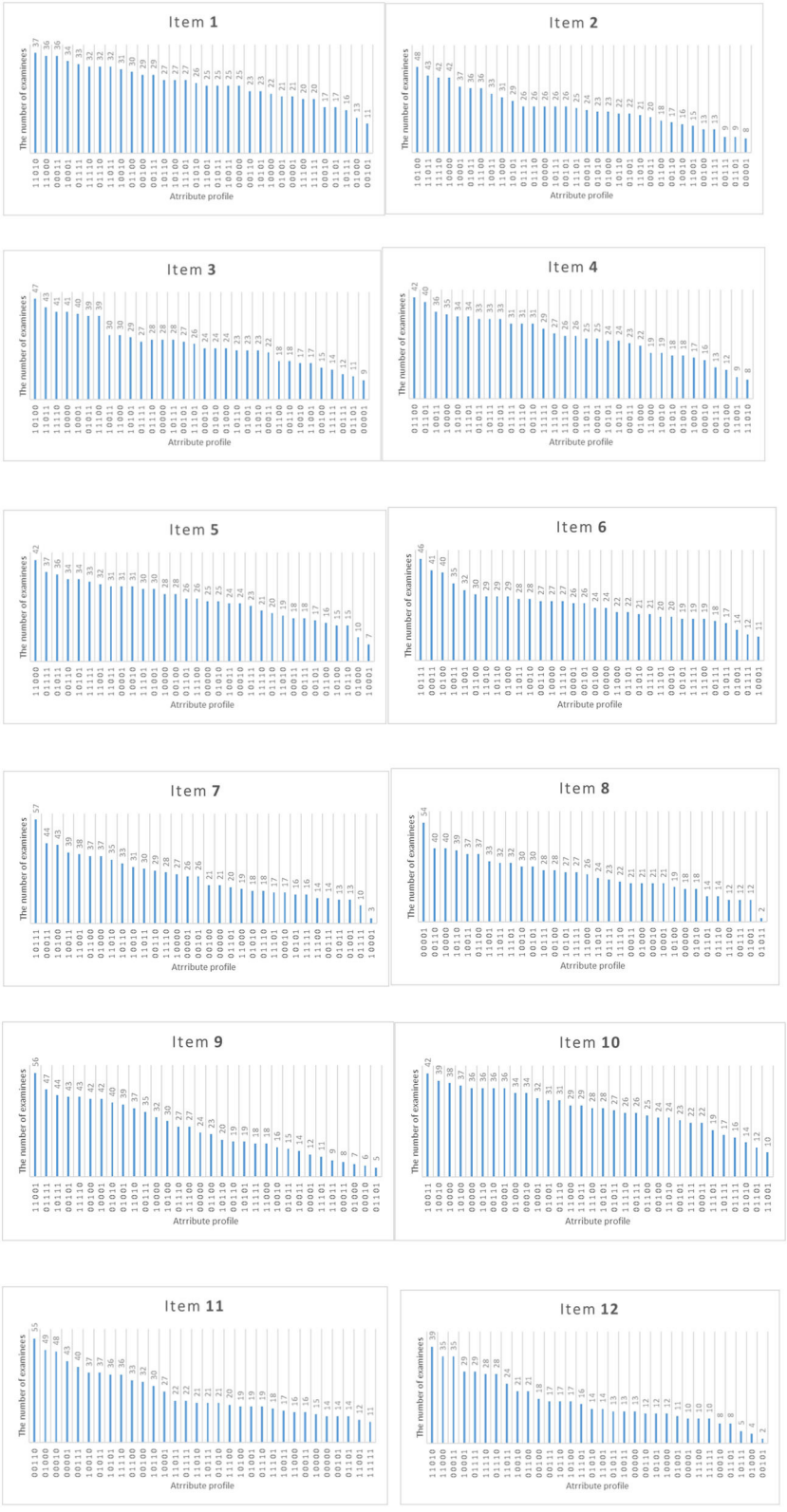


FIGURE 8 | The number of examinees in various attribute mastery pattern assigned to each item under the Shannon entropy allocation strategy.

TABLE 4 | The MAD and RMSE for two strategies with different initial item parameters.

Sample size	0.05				0.15				0.25				0.35				0.45				Random			
	MAD_s	MAD_g	$RMSE_s$	$RMSE_g$	MAD_s	MAD_g	$RMSE_s$	$RMSE_g$	MAD_s	MAD_g	$RMSE_s$	$RMSE_g$	MAD_s	MAD_g	$RMSE_s$	$RMSE_g$	MAD_s	MAD_g	$RMSE_s$	$RMSE_g$	MAD_s	MAD_g	$RMSE_s$	$RMSE_g$
20	0.1864	0.1440	0.2190	0.1865	0.1602	0.1370	0.1948	0.1804	0.1383	0.1410	0.1754	0.1855	0.1272	0.1478	0.1822	0.1997	0.1456	0.1462	0.1936	0.1958	0.1363	0.1770	0.2263	0.2506
40	0.1183	0.0949	0.1502	0.1193	0.1185	0.0987	0.1500	0.1249	0.1160	0.0979	0.1497	0.1236	0.1175	0.1022	0.1506	0.1293	0.1198	0.0980	0.1525	0.1269	0.1611	0.1020	0.2114	0.1315
80	0.0862	0.0724	0.1103	0.0906	0.0845	0.0760	0.1090	0.0961	0.0841	0.0730	0.1077	0.0922	0.0885	0.0775	0.1159	0.0974	0.0919	0.0795	0.1177	0.0998	0.1131	0.0740	0.1554	0.0940
160	0.0660	0.0553	0.0850	0.0709	0.0668	0.0558	0.0857	0.0701	0.0685	0.0560	0.0880	0.0719	0.0662	0.0520	0.0849	0.0665	0.0696	0.0592	0.0879	0.0739	0.0812	0.0557	0.1086	0.0709
320	0.0508	0.0438	0.0647	0.0551	0.0496	0.0467	0.0631	0.0596	0.0498	0.0395	0.0637	0.0502	0.0538	0.0476	0.0682	0.0603	0.0517	0.0464	0.0654	0.0584	0.0582	0.0422	0.0758	0.0536

SIMULATION STUDY 2

Simulation Design

The second simulation study is using random allocation strategy and Fisher Information allocation strategy to estimate the item parameters of the new items when the item attribute vector is known. The design of this study is similar to the first study, except that there are differences in the number of examinees and the initial setting of item parameters, while other conditions remain unchanged. Five different levels of sample sizes were used to design the number of examinees, which were set to 20, 40, 80, 160, and 320, respectively. The initial values of item parameters are also set at five different levels, which are 0.05, 0.15, 0.25, 0.35, and 0.45, respectively. Repeat $R = 100$ times under each condition.

Evaluation Indices

Mean absolute deviation (MAD) is the average of the absolute value of the difference between the estimated and true value. The average absolute deviation is applied to measure the precision of the online estimation of item parameters. The closer its value is to 0, the better precision is. The formula is as follows:

$$MAD_x = \frac{1}{RJ}\sum_{j-1}^{J}\sum_{l=1}^{R}\left|\hat{x}_j^{(l)} - x_j\right|, \tag{26}$$

where $\hat{x}_j^{(l)}$ and x_t represent the estimated and true values of item parameters (guessing or slipping parameters), respectively.

Item root mean squared error (IRMSE) is used to estimate the precision of a single item parameter. For a single item j, the calculation formula is as follows:

$$IRMSE_x = \sqrt{\frac{1}{R}\sum_{l=1}^{R}\left(\hat{x}_j^{(l)} - x_j\right)^2}, \tag{27}$$

Root mean squared error (RMSE) is used to calculate the estimation precision of all item parameters.

$$RMSE_x = \sqrt{\frac{1}{RJ}\sum_{j=1}^{J}\sum_{l=1}^{R}\left(\hat{x}_j^{(l)} - x_j\right)^2} \tag{28}$$

Simulation Results

The results of online estimation are shown in **Table 4**. It can be analyzed from the table that as the number of examinees increases, the MAD and RMSE of the item parameter estimates are constantly decreasing. When the initial values of the item parameters are different, the MAD and RMSE of the item parameter estimates under the D-optimal criterion are almost the same. It means that the initial values of the item parameters have little influence on the precision of the item parameter estimates. When the number of examinees is 20, 40, 80, and 160, the RMSE of the D-optimal strategy is significantly lower than the random strategy. When the number of examinees is 320, the MAD and RMSE are very similar to the two strategies. For the precision of different item parameters, the error of slipping parameters is

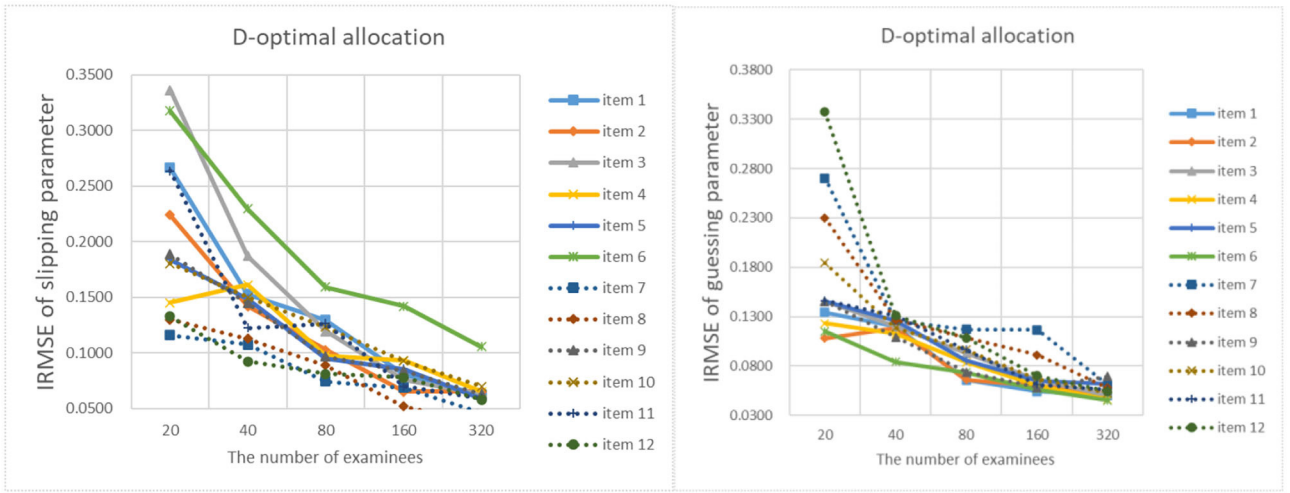


FIGURE 9 | The RMSE of each item parameter of the D-optimal allocation strategy when the initial value of the item parameter is 0.05.

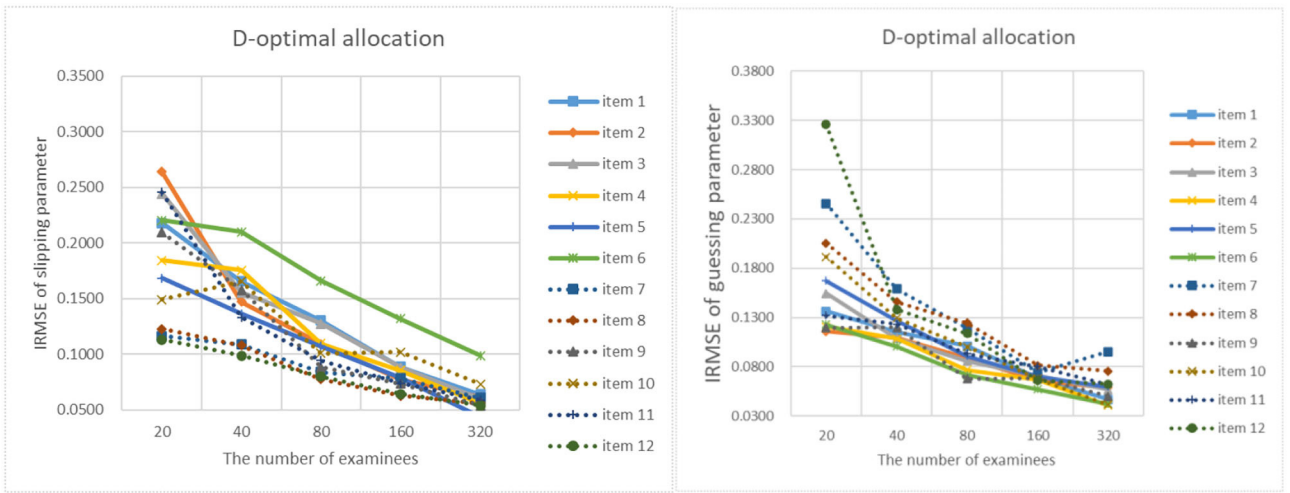


FIGURE 10 | The RMSE of each item parameter of the D-optimal allocation strategy when the initial value of the item parameter is 0.15.

greater than that of guessing parameters. It shows that guessing parameters are easier to estimate than slipping parameters.

Figures 9–13 show the IRMSE for different sample sizes from the D-optimal strategy under different initial values of item parameters. **Figure 14** shows the IRMSE for different sample sizes from the random strategy. It can be seen that the D-optimal strategy is better than the random strategy, especially for the slipping parameters.

Figure 15 shows the distribution of attribute mastery patterns selected by the D-optimal strategy for each new item when the number of examinees is 320. Each new item is assigned to ∼160 examinees. From **Table 1**, the true attribute vector is 00100 for item 2. We found that the top five attribute mastery patterns for the item are 00101, 01111, 11001, 00100, and 01011, respectively. Most examinees with the first, second, and fourth attribute mastery patterns can answer the item correctly, while examinees with the other attribute mastery patterns cannot answer the item correctly. Intuitively, these attribute mastery patterns are very useful for estimating slipping and guessing parameters.

SIMULATION STUDY 3

Adaptive Design of Online Estimation and Online Calibration

When the item parameters and attribute vectors are unknown, the item allocation strategies proposed in the previous two studies are used to estimate the item parameters and calibrate the attribute vector of the new item. From the conclusions of the above two simulation studies, we can see that in order to achieve the same estimation precision, the number of examinees required for item parameter estimation will be less. Thus, the first four fifths of examinees are used to calibrate attribute vectors,

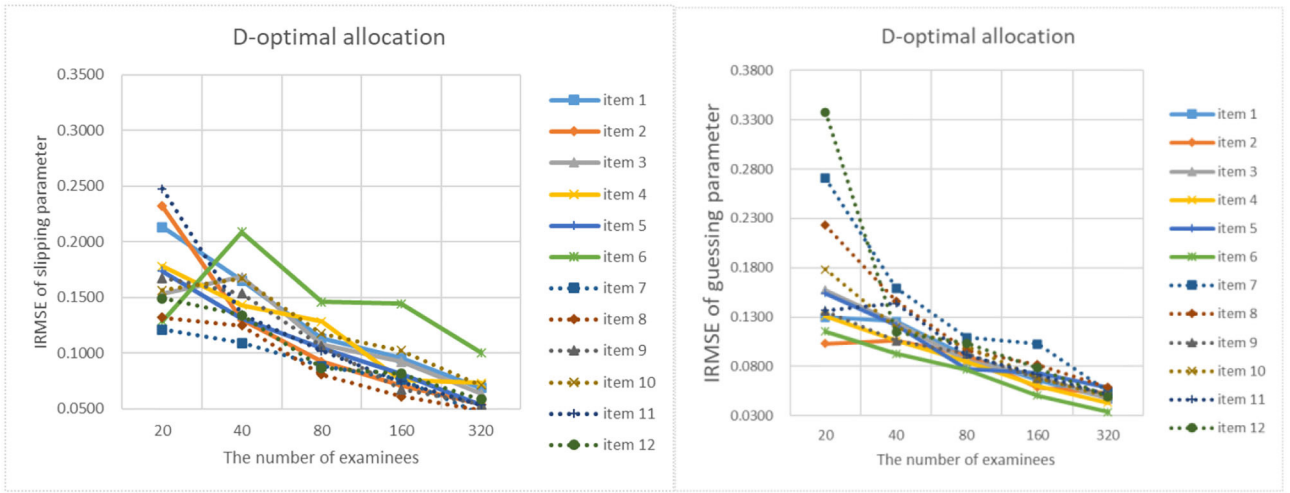


FIGURE 11 | The IRMSE of the D-optimal allocation strategy when the initial value of the item parameter is 0.25.

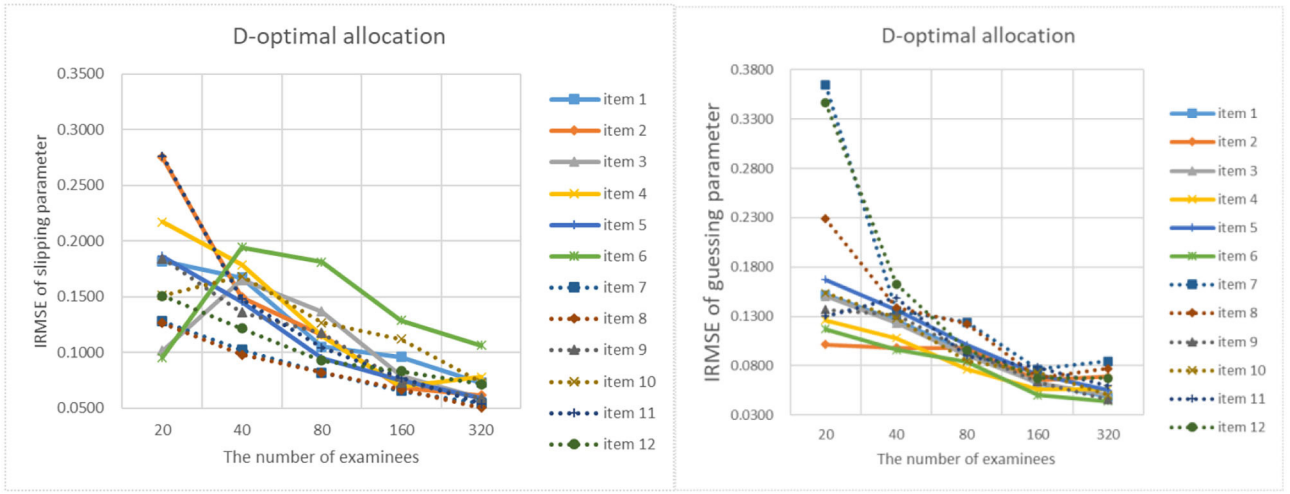


FIGURE 12 | The IRMSE of the D-optimal allocation strategy when the initial value of the item parameter is 0.35.

and the last one fifth are used to estimate item parameters. The initial value of the item parameter for calibrating the item attribute vector or estimating the item parameters is set to 0.15. The five levels of sample sizes as the first study is used here.

The simulation study considers four designs: (a) the Shannon entropy and D-optimal strategy were used for calibrating attribute vectors and estimating item parameters, respectively; (b) the Shannon entropy and random strategy; (b) the random and D-optimal strategy, and two random strategies. After the examinees are all assigned by the D-optimal or random method, the estimation of item parameters and the calibration of attribute vector iterates until the estimated attribute vector unchanged.

Simulation Results

Table 5 show the TSR and the RMSE under different allocation strategies combinations with or without iterations. As the number of examinees increases, the accuracy of attribute vectors and the precision of item parameters are increasing. Whenever the D-optimal or random strategy is used, the Shannon entropy strategy performs better than the random strategy in terms of the TSR under the same sample size. Similarly, we found that the D-optimal strategy can obtain the item parameter estimates more accurately than the random strategy under the same sample size, whenever the Shannon entropy or random strategy is used. The performance of the method with the combination of the Shannon entropy and D-optimal strategies

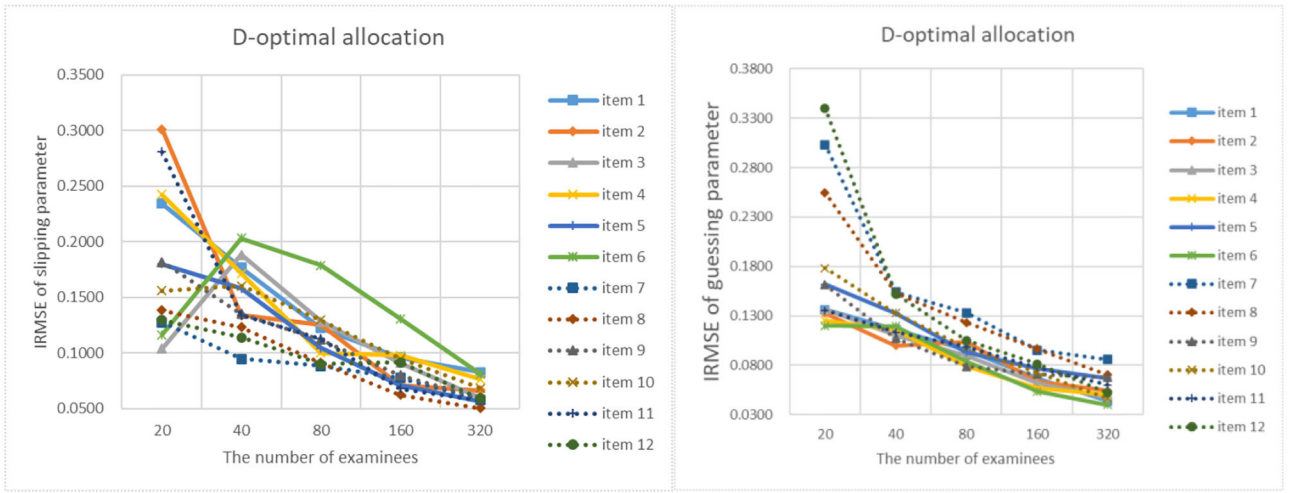


FIGURE 13 | The IRMSE of the D-optimal allocation strategy when the initial value of the item parameter is 0.45.

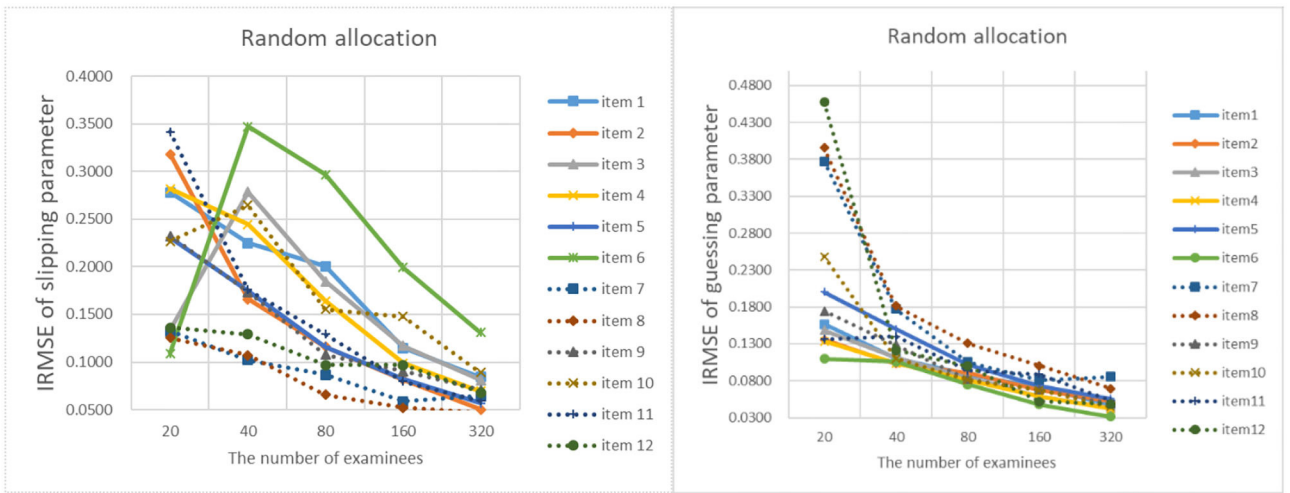


FIGURE 14 | The IRMSE of the random allocation strategy.

is better than the method with only new adaptive strategy or two random strategies.

CONCLUSIONS AND DISCUSSION

The CD-CAT combines the advantages of computerized adaptive testing and cognitive diagnostic assessment. For obtaining higher correct classification rates of attribute mastery patterns, the CD-CAT requires a high-quality calibration item bank. Item replenishing is very important for item bank maintenance in CD-CAT. The related previous research in CD-CAT mainly focused on the calibration method with the random design in which some examinees cannot provide enough information of item parameter estimation or Q-matrix calibration for the new items. In order to implement item replenishing efficiency, we propose the adaptive design for item parameter online estimation and **Q**-matrix online calibration.

We investigated the performance of three adaptive designs under different conditions. The first study showed that the optimal design based on Shannon entropy was better than the random design in the calibration of the Q-matrix of the new items when the number of examinees is $<$1,600. When the number of examinees reaches 1,600, the average estimation accuracy of the two methods is very close. Although the TSR of the two methods are similar, the standard deviation of TMR from Shannon entropy-based allocation method (0.0286) is lower than that of random allocation method (0.0447). It means that the Shannon entropy-based allocation method is more stable than the random allocation method. The second study suggested that given the Q-matrix of the new items, the D-optimal design

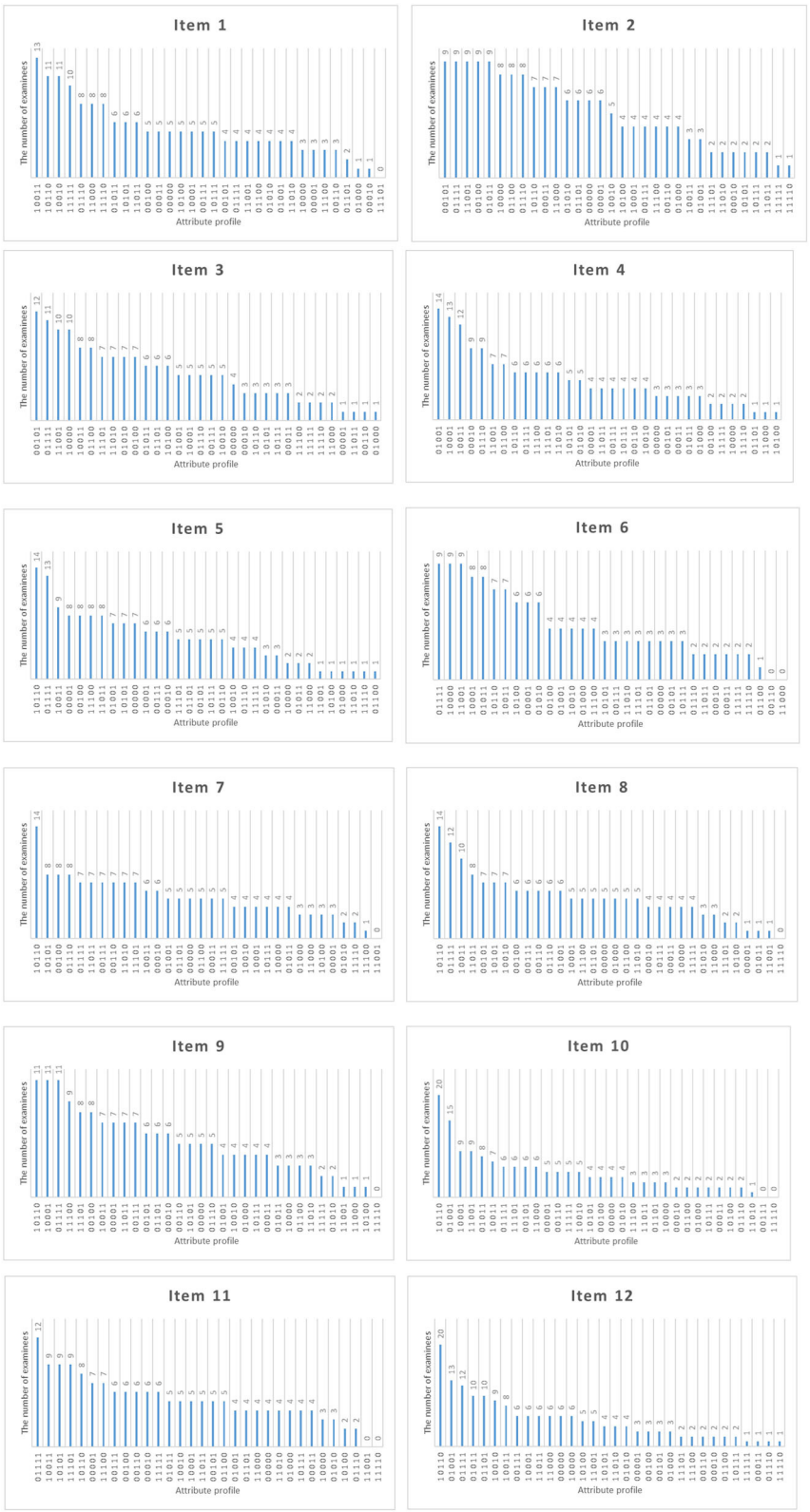


FIGURE 15 | The number of examinees in various attribute mastery pattern assigned to each item under D-optimal strategy.

TABLE 5 | The TSR and the RMSE under different allocation strategies with iterations.

Item allocation method	The number of examinees	D-optimal			Random		
		TSR	$RMSE_s$	$RMSE_g$	TSR	$RMSE_s$	$RMSE_g$
Shannon entropy	80/20/100/100	0.65	0.1266	0.0896	0.61	0.1585	0.1014
	160/40/200/200	0.80	0.0879	0.0699	0.80	0.1018	0.0724
	320/80/400/400	0.90	0.0663	0.0559	0.90	0.0727	0.0548
	640/160/800/800	0.97	0.0505	0.0418	0.97	0.0585	0.0431
	1280/320/1600/1600	0.99	0.0460	0.0440	0.99	0.0455	0.0371
Random	80/20/100/100	0.62	0.1342	0.0856	0.59	0.1537	0.0856
	160/40/200/200	0.75	0.1064	0.0684	0.74	0.1067	0.0676
	320/80/400/400	0.87	0.0731	0.0503	0.86	0.0820	0.0525
	640/160/800/800	0.94	0.0586	0.0428	0.93	0.0649	0.0433
	1280/320/1600/1600	0.99	0.0423	0.0376	0.98	0.0474	0.0301

80/20/100/100 mean that Shannon entropy allocation strategy selected the first 80 examinees for obtaining the first calibration of attribute vector, the last 20 examinees were used for the estimaion of item parameters, and then all examinees' item responses were used to estimated iterately item parameters and attribute vector.

outperforms the random design in terms of the precision of item parameters when the sample size is small. Finally, if the Q-matrix and item parameters are unknown for the new items, the hybrid of two optimal designs could efficiently simultaneously estimate item parameters and calibrate the Q-matrix of the new items. In summary, the adaptive designs for new items is promising in terms of the accuracy of attribute vectors and the precision of item parameters. The conditions for using these designs in CD-CAT are as follows: (a) in the first design, the attribute vector of the new items can be calibrated based on the item responses, so as to meet the needs of the NPC method; (b) the second design requires the determined Q-matrix to estimate the item parameters; (c) the third design is suitable for situations where the Q-matrix and item parameters of the new items are unknown. The new design adaptively assigns new items to examinees according to both the current calibration of the new items and the current status of the examinees. It contributes to the 'adaptive' aspect of CD-CAT. Not only can it accurately estimate the item parameters of the new items and calibrate its Q-matrix when the sample size is small.

There are still some limitations in the study. The independent attribute structure was considered in this study. While Leighton and Hunka (2004) think that the attributes were organized as hierarchical structures, including linear, convergent, divergent, and unstructured. The adaptive designs are worthy of further study under hierarchical structures. At the same time, the study was carried out under the DINA model, a simple non-compensatory cognitive diagnosis model. So far, there are many parametric models, such as the noise input, deterministic "and" gate (NIDA) model (Maris, 1999; Junker and Sijtsma, 2001), the deterministic inputs, noisy, "or" gate (DINO) model (Templin and Henson, 2006), the general diagnostic model (GDM; von Davier, 2005, 2008), the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), the generalized DINA model (G-DINA; de la Torre, 2011), and the CDINA model (Luo et al., 2020). For the non-parametric method, the NPC method has been extended to the general non-parametric classification (GNPC; Chiu et al., 2018). Whether the adaptive designs can be extended to these models is worth studying. Furthermore, the dichotomous item response is only often used in multiple-choice or fill-in-the-blank items. For polytomous scoring items (Gao et al., 2020), the adaptive designs remain to be further studied.

Finally, the termination rule adopted in this study was fixed for balancing the number of examinees assigned to all new items. The variable length rule needs to be studied. For example, when the posterior distribution of the attribute vector of the new item reaches the preset value, the attribute vector will no longer be calibrated. Termination rules for attribute mastery patterns in CD-CAT (Guo and Zheng, 2019) may give you ideas for the variable-length optimal design for online calibration.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

WW and LS designed the study and revised the manuscript. YT, JZ, and TW drafted and revised the manuscript. WW and TW conducted the simulation study. All authors contributed to the article and approved the submitted version.

FUNDING

This research was partially supported by the National Natural Science Foundation of China (Grant No. 62067005), the Social Science Foundation of Jiangxi (Grant No. 17JY10), and the Project of Teaching Reform of Jiangxi Normal University (Grant No. JXSDJG1848). This study received funding from the grant (Grant No. CTI2019B10) of the Chinese Testing International Co., Ltd. These funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

REFERENCES

Chang, Y. P., Chiu, C. Y., and Tsai, R. C. (2018). Nonparametric CAT for CD in educational settings with small samples. *Appl. Psychol. Meas.* 43, 543–561. doi: 10.1177/0146621618813113

Chen, P. (2016). Two new online calibration methods for computerized adaptive testing. *Acta Psychol. Sin.* 48, 1184–1198. doi: 10.3724/SP.J.1041.2016.01184

Chen, P., and Xin, T. (2011). Item replenishing in cognitive diagnostic computerized adaptive testing. *Acta Psychol. Sin.* 43, 836–850. doi: 10.3724/SP.J.1041.2011.00836

Chen, P., Xin, T., Wang, C., and Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika* 77, 201–222. doi: 10.1007/s11336-012-9255-7

Chen, Y., Liu, J., and Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Appl. Psychol. Meas.* 39, 5–15. doi: 10.1177/0146621613513065

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

Chiu, C. Y., and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classif.* 30, 225–250. doi: 10.1007/s00357-013-9132-9

Chiu, C. Y., Sun, Y., and Bian, Y. (2018). Cognitive diagnosis for small educational programs: the general nonparametric classification method. *Psychometrika* 83, 355–375. doi: 10.1007/s11336-017-9595-4

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

Gao, X. L., Wang, D. X., Cai, Y., and Tu, D. B. (2020). Cognitive diagnostic computerized adaptive testing for polytomously scored ite*ms. J. Classif.* 37, 709–729. doi: 10.1007/s00357-019-09357-x

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

Henson, R. A., Templin, J. L., and Willse, J. T. (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

Huang, H. Y. (2019). Utilizing response times in cognitive diagnostic computerized adaptive testing under the higher order deterministic input, noisy 'and' gate model. *Br. J. Math. Stat. Psychol.* 73, 109–141. doi: 10.1111/bmsp.12160

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., de la Torre, J., and Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Appl. Psychol. Meas.* 39, 167–188. doi: 10.1177/0146621614554650

Leighton, J. P., and Hunka, G. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* 41, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Luo, Z. S., Hang, D. D., Qin, C. Y., and Yu, X. F. (2020). The cognitive diagnostic model which can deal with compensation and noncompensation effects: CDINA. *J. Jiangxi Normal Univers.* 44, 441–453. doi: 10.16357/j.cnki.issn1000-5862.2020.05.01

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

Magis, D., Yan, D., and von Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing With R: Using Packages catR and mstR*. Cham: Springer. doi: 10.1007/978-3-319-69218-0

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Stefanski, L. A., and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Stat.* 13, 1335–1351. doi: 10.21236/ADA160277

Sun, X., Liu, Y., Xin, T., and Song, N. (2020). The impact of item calibration error on variable-length cognitive diagnostic computerized adaptive testing. *Front. Psychol.* 11:575141. doi: 10.3389/fpsyg.2020.575141

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

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. (2005). A general diagnostic model applied to language testing data. *ETS Res. Rep. Ser.* 2005, i—35. doi: 10.1002/j.2333-8504.2005.tb01993.x

Wainer, H., and Mislevy, R. J. (1990). "Item response theory, item calibration, and proficiency estimation," in *Computerized Adaptive Testing: A Primer,* ed H. Wainer (Hillsdale, NJ: Lawrence Erlbaum), 65–102.

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 Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *J. Educ. Meas.* 48, 255–273. doi: 10.1111/j.1745-3984.2011.00145.x

Wang, D. X., and Tu, D. B. (2021). The application of cognitive diagnostic computerized adaptive testing on diagnosis and assessment of psychological disorder. *J. Jiangxi Normal Univers.* 45, 111–117. doi: 10.16357/j.cnki.issn1000-5862.2021.02.01

Wang, W. Y., Ding, S. L., and You, X. F. (2011). Attribute calibration of original que stions in computerized adaptive diagnostic test. *Acta Psychol. Sin.* 43, 964–976. doi: 10.3724/SP.J.1041.2011.00964

Wang, Y. T, Sun, X. J, Chong, W. F., and Xin, T. (2020). Attribute discrimination index-based method to balance attribute coverage for short-length cognitive diagnostic computerized adaptive testing. *Front. Psychol.* 11:224. doi: 10.3389/fpsyg.2020.00224

Xu, X. L., Chang, H. H., and Douglas, J. (2003). "A simulation study to compare CAT strategies for cognitive diagnosis," in *Annual Meeting of the American Educational Research Association* (Chicago, IL).

Yang, J., Chang, H.H., Tao, J., and Shi, N. (2020). Stratified item selection methods in cognitive diagnosis computerized adaptive testing. *Appl. Psychol. Meas.* 44, 346–361. doi: 10.1177/0146621619893783

Yigit, H. D., Sorrel, M. A., and de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Appl. Psychol. Meas.* 43, 388–401. doi: 10.1177/0146621618798665

Conflict of Interest: This research was partially supported by the National Natural Science Foundation of China (Grant No. 62067005), the Social Science Foundation of Jiangxi (Grant No. 17JY10), and the Project of Teaching Reform of Jiangxi Normal University (Grant No. JXSDJG1848). This study received funding from the grant (Grant No. CTI2019B10) of the Chinese Testing International Co., Ltd. These funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication. The authors declare no other competing interests.

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 Wang, Tu, Song, Zheng 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: 03 September 2021
doi: 10.3389/fpsyg.2021.704724

A Simplified Method for Predicting Pattern Match Ratio

Xiaojuan Tang[1]*, *Huiqiong Duan*[2], *Shuliang Ding*[3]* and *Mengmeng Mao*[4]

[1] School of Education, Jiangxi Normal University, Nanchang, China, [2] School of Foreign Languages, Nanchang Hangkong University, Nanchang, China, [3] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, [4] School of Public Administration, Nanchang University, Nanchang, China

Cognitive diagnostic test design (CDTD) has a direct impact on the pattern match ratio (PMR) of the classification of examinees. It is more helpful to know the quality of a test during the stage of the test design than after the examination is taken. The theoretical construct validity (TCV) is an index of the test quality that can be calculated without testing, and the relationship between the PMR and the TCV will be revealed. The TCV captures the three aspects of the appeal of the test design as follows: (1) the TCV is a measure of test construct validity, and this index will navigate the processes of item construction and test design toward achieving the goal of measuring the intended objectives, (2) it is the upper bound of the PMR of the knowledge states of examinees, so it can predict the PMR, and (3) it can detect the defects of test design, revise the test in time, improve the efficiency of test design, and save the cost of test design. Furthermore, the TCV is related to the distribution of knowledge states and item categories and has nothing to do with the number of items.

Keywords: cognitive diagnostic test design, pattern match ratio, theoretical construct validity, prediction method, upper bound

OPEN ACCESS

Edited by:
Tao Xin,
Beijing Normal University, China

Reviewed by:
Ren Liu,
University of California, Merced, United States
Chunhua Kang,
Zhejiang Normal University, China

***Correspondence:**
Xiaojuan Tang
psychotang@foxmail.com
Shuliang Ding
ding06026@163.com

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

Received: *03 May 2021*
Accepted: *26 July 2021*
Published: *03 September 2021*

Citation:
Tang X, Duan H, Ding S and Mao M (2021) A Simplified Method for Predicting Pattern Match Ratio. Front. Psychol. 12:704724. doi: 10.3389/fpsyg.2021.704724

INTRODUCTION

Cognitive diagnosis (CD) has received much attention, providing diagnostic information of knowledge or skills (often called "attributes" in the CD literature) to the examinees (de la Torre and Douglas, 2004; de la Torre, 2008; DeCarlo, 2011; Liu et al., 2012; Kang et al., 2017; Huebner et al., 2018). It is critical to ensure that high-quality cognitive diagnostic tests can accurately diagnose the knowledge state (KS, i.e., the latent cognitive states) of examinees. The set of KSs is represented by the Q_S matrix. In fact, cognitive diagnostic test design (CDTD) is the design of a Q matrix, called Q_t, i.e., rows representing attributes and columns representing attribute vectors, namely, items. By anchoring the items with attribute vectors, proposition experts and measurement experts transform items into measurable forms and then diagnose examinees. In a word, the design of the Q_t matrix is the problem of how to match the attribute vectors to achieve a certain predetermined goal.

The CDTDs can be divided into the following aspects based on different dimensions: the dichotomous CDTD (Chiu et al., 2009; Ding et al., 2010) and the polytomous CDTD (Ding et al., 2014a,b,c) according to the scoring methods; Boolean matrix CDTD (Samejima, 1995; Tatsuoka, 1995, 2009; Ding et al., 2011; Cai et al., 2018) and polytomous Q matrix CDTD (Ding et al., 2015; Tu and Cai, 2015) according to the values of elements in the Q_t matrix; model-dependent CDTD (Chiu et al., 2009; Kuo et al., 2016) and model-free CDTD (Shao, 2010) according to whether depending on the cognitive diagnostic models (CDM) or not; cognitive diagnostic computerized adaptive testing (CD-CAT) design (Cheng, 2010; Sun et al., 2019) and cognitive diagnostic testing (CDT) design (Henson and Douglas, 2005; Henson et al., 2008; Ding et al., 2011) according to whether

personalized diagnostic; independent structure CDTD (Cheng, 2009, 2010; Liu et al., 2016) and dependent structure CDTD (Ding et al., 2011; Kuo et al., 2016) according to cognitive structure, and so on. In fact, almost all CDTDs are multidimensional.

Until present, the studies on the CDTD methods are still relatively weak, and they focus on the following two aspects:

(1) CDTD based on the perfect Q matrix

The so-called "perfect Q matrix" refers to the Q_t matrix that makes the ideal response pattern (IRP) and KS correspond one to one. If the Q matrix in tests is a perfect Q matrix, the pattern match ratio (PMR) improves no matter whether the CDTD is either dichotomous or polytomous.

(i) Examples of dichotomous CDTD: For the four attribute hierarchies of Leighton (Leighton et al., 2004), if the Q_t matrix is a Boolean matrix, and there is no compensation between the attributes, then the reachable matrix (or it is equivalent classes) acts as the submatrix of Q_t which can achieve a one-to-one correspondence between the set of IRPs and the set of KSs. The more reachable matrices in the Q_t matrix, the higher the PMR (Ding et al., 2010, 2011). Ding et al. (2010) called such a Q_t matrix a sufficient and necessary matrix, i.e., a perfect Q matrix (Cai et al., 2018). The results are similar to those of Chiu et al. (2009), DeCarlo (2011), and Madison and Bradshaw (2015) on independent structures. With the independent structure and four attributes, Samejima (1995) believed that when the Q_t matrix was the identity matrix (i.e., the identity matrix of independent structure is a reachable matrix), all of the KSs would not be misjudged. Chiu et al. (2009) also found that the Deterministic Input Noisy "AND" Gate (DINA) model and the Deterministic Input Noisy Output "OR" gate (DINO) model could diagnose all potential attribute mastery patterns when the Q_t matrix included the identity matrix. Similar results have been addressed in other studies (DeCarlo, 2011; Madison and Bradshaw, 2015).

(ii) Examples of polytomous CDTD: To achieve the one-to-one correspondence between the set of KSs and the set of IRPs, the rooted tree structure, the independent structure, and the perfect Q matrices of the rhombus structure are introduced under the item score rule that one ideal score is added if mastering one attribute adhering to the item (Ding et al., 2014a). In the initial stage of CD-CAT, each attribute can be diagnosed by using the reachable matrix (Tu et al., 2013). In CD-CAT, the higher the percentage of the examinees is, whose testing items are (or contain) the reachable matrix according to the selection strategy, the higher the PRM is.

(2) CDTD based on the index

The Cognitive Diagnostic Index (CDI) (Henson and Douglas, 2005) and the Attribute-level Discrimination Index (ADI) (Henson et al., 2008) are based on the level of items and attributes for CD. Kuo et al. (2016) indicated that each attribute in the test must be measured at least three times to attain better correct attribute classification, so they proposed modified CDIs and ADIs, namely, MCDI and MADI. The Shannon's entropy (Xu et al., 2003) and posterior-weighted Kullback–Leibler (PWKL) (Cheng, 2009) were introduced in CD-CAT. Cheng (2010) believed that adequate coverage of each attribute could improve the validity of the test scores, and then the attribute-balancing index was proposed. Subsequently, the index was further improved (Yu et al., 2011; Liu et al., 2018; Sun et al., 2019). Adaptive multigroup testing method for cognitive diagnosis (CD-AMGT) (Luo et al., 2018), which selects a group of appropriate items in different diagnosis stages, has the advantages of uniform use of item bank and less time to calculate.

The PMR is the main evaluation index for cognitive diagnostic tests. In CDTD, the pretest evaluation of the PMR is more positive than the posttest evaluation because the designed test can be modified quickly, the designer can make up for possible errors before testing, and material resources and time will be saved. At present, the PMR is the posttest estimation based on the data measured or simulated, so it is impossible to calculate PMR immediately during the design process. Furthermore, it is meaningful to discuss the maximum PMR for the pretest, and the maximum PMR is related to the matching degree between the designed test and the cognitive model, as well as the quality and length of the test.

The rest of the study is organized as follows: First, the TCV used in this study is briefly described. Second, the theoretical proof of the relationships between the TCV and the PMR is introduced in detail. The TCV is then evaluated in a simulation study. The end of the study is the discussion and conclusion.

METHODS

Cognitive Diagnosis

The cognitive model is a prerequisite for CD. It is represented by an attribute hierarchy, which specifies the psychological ordering of the attributes required to solve test items. Attributes are those basic cognitive processes or skills required to solve test items correctly. There are five forms of basic hierarchical structures (Leighton et al., 2004; Cheng, 2010), namely, A, B, C, D and E (**Figure 1**).

Attribute 1 is considered a prerequisite to other attributes, and attribute 5 depends on some attributes in models except the independent model. The adjacency (A), reachability (R), incidence (Q), and reduced incidence (Q_r) matrices are specified by Tatsuoka (1995). The columns of the Q_r matrix indicate that all possible items must be created to reflect the relationships among the attributes in the hierarchy. The possible latent cognitive states (i.e., KS), which is all the columns of the incidence matrix, possess cognitive attributes that are consistent with the hierarchy (when the hierarchy is based on cognitive considerations), and they apply these attributes systematically (when the hierarchy is based on procedural considerations) (Gierl et al., 2007). Let $q_j = \left(q_{j1}, q_{j2}, \cdots, q_{jK}\right)^T \left(j = 1, \cdots, m\right)$ denote the jth dichotomous column vector (i.e., the jth category item) of the Q_r matrix. All KSs are represented by column vectors: $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iK})^T$, where $\alpha_{ik} = 1 (k = 1, \cdots, K)$ indicates that the ith category examinee has mastered attribute k, and $\alpha_{ik} = 0$ otherwise. K is the total number of attributes

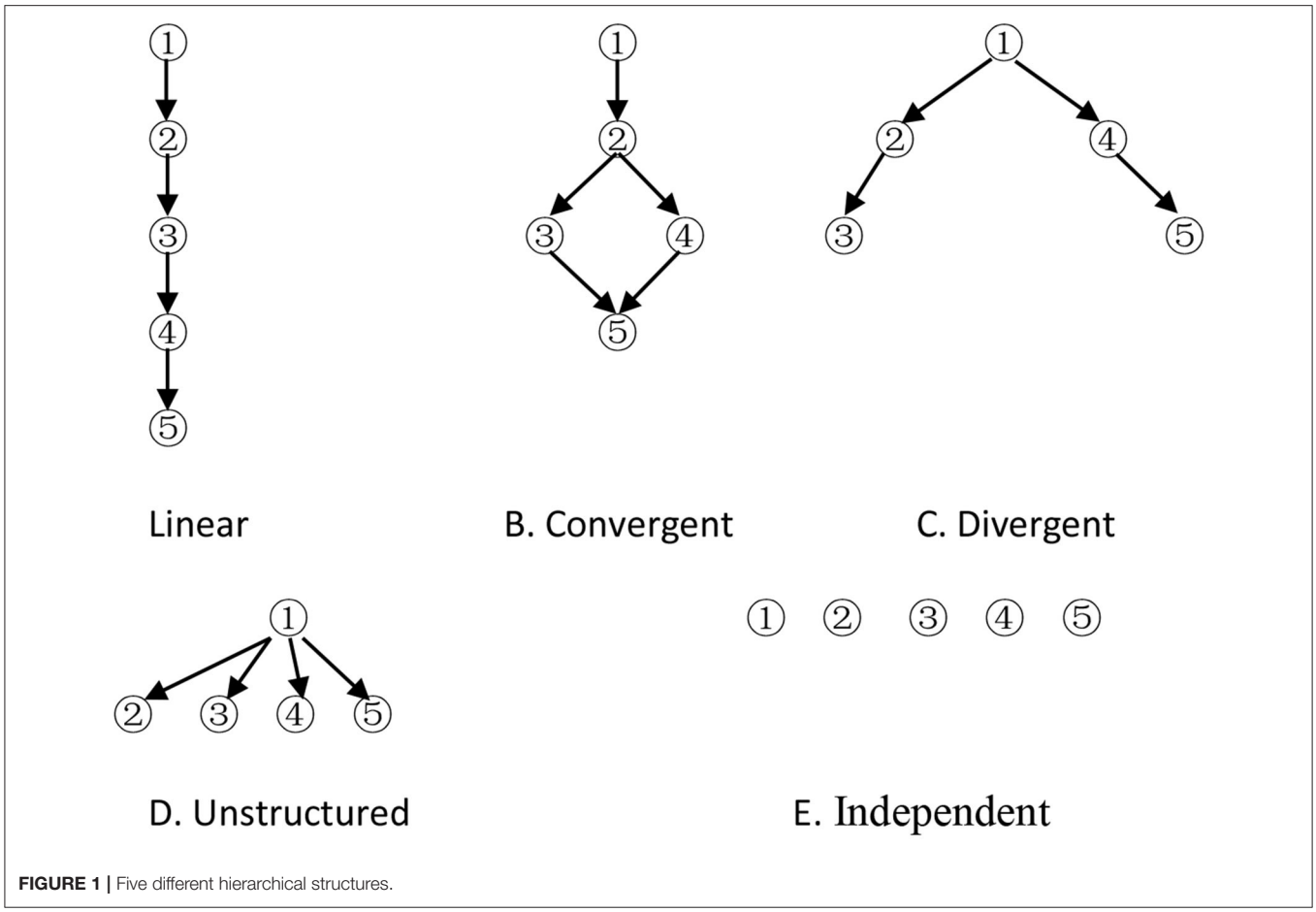


FIGURE 1 | Five different hierarchical structures.

measured by the test. Let the Q_s matrix denote all KSs, in fact, including zero vector (*denoted as* $\bar{0}$, i.e., this kind of examinee does not master any attribute) and the Q_r matrix for cognitive attribute consistency. Thus, α_i and q_j are all K-dimensional vectors. The Q_t matrix consists of some column vectors of the Q_r matrix. Based on the cognitive model (including attributes and hierarchy among them), the Q_r and Q_s matrix can be obtained, that is, all possible items and KSs can be obtained. On the contrary, if the Q_t matrix is known, some KSs can be obtained through the augment algorithm (Ding et al., 2008; Yang et al., 2008), and the cognitive model can be derived by comparing the rows (Tatsuoka, 1995). In general, it is impossible for some items (i.e., the Q_t matrix) to replace all the items (i.e., the Q_r matrix), which express the cognitive structure, so some cognitive structures extracted from the Q_t matrix may be inconsistent with the theoretical one.

The DINA Model

Cognitive diagnostic models have been proposed for many years, including the rule space model (Tatsuoka, 1983), the "Noisy Input Deterministic 'AND' Gate" (NIDA) model (Maris, 1999), the fusion model (Hartz, 2002), the reduced reparameterized unified model (R-RUM; Hartz, 2002), and the DINA model (Haertel, 1989). The DINA model is completely noncompensatory. The DINA model treats slipping and guessing at the item level. Parameter s_j indicates the probability of "slipping," and parameter g_j denotes the probability of "guessing." The item response function, therefore, can be written as follows:

$$P\left(X_{ij} = 1|\alpha_i\right) = \left(1 - s_j\right)^{n_{ij}} g_j^{\,1-n_{ij}} \tag{1}$$

$$n_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{2}$$

When $n_{ij} = 1$, the ith examinee should be able to answer item j correctly, unless he/she "slips." Similarly, when $n_{ij} = 0$, the ith examinee should not be able to answer item j correctly, unless he/she is a lucky guesser (Cheng, 2010).

Theoretical Construct Validity

Theoretical construct validity (TCV) is used to measure the degree of consistency between the theoretical cognitive model and the cognitive model implied in the Q_t matrix (Ding et al., 2012).

Definition 1 Let $\{\alpha_1, \alpha_2, \cdots, \alpha_{N_1}\}$ denote N_1 KS of the theoretical cognitive model given by experts, $\{\beta_1, \beta_2, \cdots, \beta_{N_2}\}$ denote N_2 KS derived from the Q_t matrix, and $\{\gamma_1, \gamma_2, \cdots, \gamma_{N_3}\} = \{\beta_1, \beta_2, \cdots, \beta_{N_2}\} \cap \{\alpha_1, \alpha_2, \cdots, \alpha_{N_1}\}$ denote N_3 KS. when $\gamma_k = \alpha_i$, the TCV for the Q_t matrix can be written as follows:

$$\mathrm{TCV} = \sum_i p_i \tag{3}$$

where p_i represents the probability of the ith category examinees, that is, the ratio of such examinees whose KS is α_i in the total population.

In particular, when all KS ratios in the total population are equal, then

$$\mathrm{TCV} = \begin{cases} \frac{N_3+1}{N_1}; & \bar{0} \notin \{\beta_1, \beta_2, \cdots, \beta_{N_2}\} \\ \frac{N_3}{N_1}; & \text{Otherwise} \end{cases} \tag{4}$$

In fact, the TCV is a measure of the degree to which the Q_t matrix represents the theoretical cognitive model (Ding et al., 2012). The observed response pattern (ORP) and the CDM are necessary for the set of the estimation of KSs of the examinees. The set of IRPs is determined by the set of KSs, the test Q matrix, the element value of the Q_t matrix (the dichotomous or the polytomous), the calculation method of the ideal score, the compensation between attributes, and so on. The ORP is related not only to the above mentioned factors but also to the item quality and random factors. Thus, if there is no random factor, the better the item quality, the closer the ORP is to the IRP. Due to the slipping and the guessing in the answering process of examinees, the PMR of the set of KSs estimated by the ORP is not higher than that estimated by the IRP, that is, $PMR_{ORP} \leq PMR_{IRP}$. *The* PMR_{IRP} acts as the maximum PMR_{ORP}, and the smaller the slipping and the guessing, the more accurate the KSs based on the ORP. How to get the $\mathrm{PMR_{IRP}}$ quickly is an interesting problem.

To clearly solve the interesting problem, a theoretical explanation that makes sense of the complexity is firmly couched within the examples.

Definition 2 Define the relationship between two attribute vectors α_i *and* q_j as $\alpha_i \geq q_j$ if and only if $\alpha_{ik} \geq q_{jk}$, for $k = 1, 2, \ldots, K$. Strict inequality between the attribute vectors is involved (i.e., $\alpha_i > q_j$) if $\alpha_{ik} > q_{jk}$ for at least one k (de la Torre, 2011). $\alpha_i \leq q_j$ and $\alpha_i < q_j$ can be defined similarly as mentioned earlier. If the relationship does not exist, then α_i has nothing to do with q_j. The definition of comparison between column vectors also applies to row vectors.

Examples

The theoretical cognitive model is an independent structure of three attributes, according to the methods suggested by Tatsuoka for calculating the adjacency (A), reachability (R), incidence (Q), and reduced incidence (Q_r) matrices; then, adding zero vector to the Q_r matrix, there are $2^3 = 8$ possible KSs, that is, N_1 is 8. The Q_s matrix is represented by a 3×8 matrix as follows:

$$Q_s = (\alpha_1, \alpha_2, \cdots, \alpha_8) = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \tag{5}$$

where α_i ($i = 1, \cdots, 8$) is the ith category examinees.

Test items, represented by a 3×3 matrix, can be written as follows:

$$Q_t = (q_1, q_2, q_3) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6}$$

where q_j is the jth item when items are not duplicated, otherwise it represents the jth category item.

Calculation of TCV

A new matrix, called the Q_t^+ *matrix*, is made of the Q_t *matrix and the two* new columns. The two new columns based on the augment algorithm (Ding et al., 2008; Yang et al., 2008) are generated from the Q_t *matrix*, $\left(\begin{array}{ccc|c|c} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{array}\right)$ while the non-zero vectors $(0, 1, 1)^T$ and $(0, 0, 1)^T$ in the Q_s matrix cannot be generated as follows:

$$Q_t^+ = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \tag{7}$$

Five KSs in the Q_t^+*matrix* are derived from the Q_t *matrix*, that is, N_2 is 5. There are five same possible latent cognitive states between the theoretical cognitive model and the cognitive model implied in the test design, that is, $\{\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5\} = \{\alpha_2, \alpha_3, \alpha_5, \alpha_6, \alpha_8\}$, N_3 is 5 (N_1, N_2, and N_3 are the same as Definition 1), when adding zero vector ($\bar{0} = (0, 0, 0)^T$).

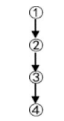
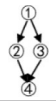
(1) When the probability distribution of the set of KSs in the total population is discrete uniform, then $TCV = (5 + 1)/8 = 3/4$.
(2) Otherwise, suppose the ratios of all α_i are 0.1, 0.1, 0, 1, 0.2, 0.1, 0.2, 0.1, 0.1, respectively, $TCV = 0.1 + 0.1 + 0.1 + 0.2 + 0.1 = 0.6$.

Calculation of $\mathrm{PMR_{IRP}}$

Ideal response (IR) depends on the relationship between α_i and q_j. Let $IR(\alpha_i, q_j) = \alpha_i{}^o q_j = \prod_{k=1}^{K} (\alpha_{ik})^{q_{jk}} = 1$ denote that the ith examinee responses correctly on the jth item, and $IR(\alpha_i, q_j) = 0$ otherwise. Clearly, $IR(\alpha_1, q_1) = IR(\alpha_1, q_2) = IR(\alpha_1, q_3) = 0$ due to $\bar{0} \leq \alpha_1 < q_1 < q_3$ *and* $\bar{0} \leq \alpha_1 < q_2$; $IR(\alpha_2, q_1) = 1, IR(\alpha_2, q_2) = IR(\alpha_2, q_3) = 0$ due to $q_1 \leq \alpha_2 < q_3$; and α_2 *having nothing to do with* q_2. Similarly, the set of IRPs of the Q_s matrix with respect to the Q_t matrix is represented by a 3×8 matrix as follows:

$$IRP = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \tag{8}$$

TABLE 1 | The relationships between the theoretical construct validity (TCV) and the PMR_{IRP} of other structures.

Theoretical cognitive model	Q_s	Q_t	Q_t^+	TCV	IRP	PMR_{IRP}
	$\begin{pmatrix}0&1&1&1&1\\0&0&1&1&1\\0&0&0&1&1\\0&0&0&0&1\end{pmatrix}$	$\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\\0&0&1\end{pmatrix}$	$\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\\0&0&1\end{pmatrix}$	4/5	$\begin{pmatrix}0&1&1&1&1\\0&0&1&1&1\\0&0&0&0&1\end{pmatrix}$	4/5
	$\begin{pmatrix}0&1&1&1&1&1\\0&0&1&0&1&1\\0&0&0&1&1&1\\0&0&0&0&0&1\end{pmatrix}$	$\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\\0&0&1\end{pmatrix}$	$\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\\0&0&1\end{pmatrix}$	2/3	$\begin{pmatrix}0&1&1&1&1&1\\0&0&1&0&1&1\\0&0&0&0&0&1\end{pmatrix}$	2/3
	$\begin{pmatrix}0&1&1&1&1&1&1\\0&0&1&0&1&0&1\\0&0&0&1&1&1&1\\0&0&0&0&0&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1\\1&0&0\\0&1&1\\0&0&1\end{pmatrix}$	$\left(\begin{array}{ccc\|cc}1&1&1&1&1\\1&0&0&1&1\\0&1&1&1&1\\0&0&1&0&1\end{array}\right)$	6/7	$\begin{pmatrix}0&0&1&0&1&0&1\\0&0&0&1&1&1&1\\0&0&0&0&0&1&1\end{pmatrix}$	6/7
	$\begin{pmatrix}0&1&1&1&1&1&1&1&1\\0&0&1&0&0&1&1&0&1\\0&0&0&1&0&1&0&1&1\\0&0&0&0&1&0&1&1&1\end{pmatrix}$	$\begin{pmatrix}1&1&1\\1&0&0\\0&1&1\\0&0&1\end{pmatrix}$	$\left(\begin{array}{ccc\|cc}1&1&1&1&1\\1&0&0&1&1\\0&1&1&1&1\\0&0&1&0&1\end{array}\right)$	2/3	$\begin{pmatrix}0&0&1&0&0&1&1&0&1\\0&0&0&1&0&1&0&1&1\\0&0&0&0&0&0&0&1&1\end{pmatrix}$	2/3

In Equation 8, the row represents the item, and the column represents $\alpha_i^{'}s$ IRP. There are six different IRPs, that is, six KS can be correctly estimated without taking the slipping and the guessing into account. In essence, the estimated five KSs based on five IRPs are the same as vectors in the Q_t^+ *matrix* (five different categories), and adding estimated zero vector (because the IRP is zero vector), there are six categories. $\alpha_4 = (0,0,1)^T$ and $\alpha_7 = (0,1,1)^T$ are the same categories to zero vector ($\alpha_1 = \bar{0}$) and $\alpha_3 = (0,1,0)^T$, respectively; thus, no new categories are generated.

The whole process of dividing the Q_s matrix can be vividly described as follows: the Q_s matrix is similar to a line, and five vectors in the Q_t^+ *matrix* are similar to five dots that classify the line into six categories in which only one KS can be estimated correctly; therefore, $PMR_{IRP} = \frac{6}{8} = \frac{3}{4}$.

From calculations 1 and 2, it can be known that TCV = PMR_{IRP}.

Examples of other structures are shown in **Table 1**.

Although structures are different, convergent, or divergent, the result of the relationship between the TCV and the PMR_{IRP} is the same: the number of vectors of the Q_t^+*matrix* in the convergent structure was 3, and then the Q_s matrix could be classified into four categories; the number of vectors of the Q_t^+ *matrix* in the divergent structure was 5 due to two new columns derived from the Q_t *matrix*, and then the Q_s matrix could be classified into six categories. For the linear structure and the unstructured, the results are similar.

Notably, all items of the Q_t *matrix* are different because the repetition of items does not increase the "coverage" of the cognitive model by the Q_t *matrix*. Repeated items only reduce random errors; thus, in the following discussion, it is not necessary to consider the repeated items in the Q_t *matrix*.

Theoretical Derivation of TCV = PMR_{IRP}

Let R denote reachable matrix, *the* Q_r *matrix* is a set of all possible items that can be written as follows:

$$Q_r = \left\{ q_j^{'} \middle| q_j^{'} = \bigcup_{q \in q_R} q, q_R \subseteq R \right\} \tag{9}$$

In fact, $Q_t \subseteq Q_r$, $Qs = \{\bar{0}, Q_r\}$.

For every $\alpha_i (i = 1, \cdots, n)$ (*except for zero vector*) in the Q_s matrix, there should be a $q_j^{'} (\in Q_r)$ corresponding to it, that is, $\alpha_i = q_j^{'}$.

Based on the augment algorithm, the Q_t^+ *matrix* can be defined as follows:

$$Q_t^+ = \left\{ q_j \middle| q_j = \bigvee_{p \in Q} p, Q \subseteq Q_t, j = 1, \ldots, m \right\} \tag{10}$$

where V represents the Boolean union operation, $p \in Q$ means that p is the item (column) of the Q matrix, and $Q \subseteq Q_t$ indicates that the Q matrix is a subset of the Q_t *matrix* and contains one or more items. New columns of *the* Q_t^+ *matrix* can be obtained by the Boolean union of two or more items in the Q_t *matrix*. There are m columns in *the* Q_t^+*matrix*, adding zero vector, m+1 categories of the KSs are derived from the Q_t *matrix* in total. n is the number of the set of KSs derived from the theoretical

cognitive model, that is, n columns in *the Qs matrix*, so the TCV can be calculated as follows:

$$\mathrm{TCV} = (m+1)/n. \tag{11}$$

The maximum lower bound of α_i can be found in *the* Q_t^+ *matrix* by comparing α_i with q_{j} *in the* Q_t^+ *matrix*, and it can be defined as follows:

$$\begin{aligned} q_{j'} \;=\; & \max\left\{q_j \,|q_j \le \alpha_i, q_j \in Q_t^+, \alpha_i \in Qs,\right. \\ & \left. i = 1,2,\ldots,n; j = 1,2,\ldots,m\right\} \end{aligned} \tag{12}$$

In fact, $j'$is the subscript of the maximum item, that is, $j' = argmax\ \{q_j\}, j' \in \{1,2,\cdots,m\}$.

Let $\{q_{j'}\}$ denote a set of α_i with the same maximum lower bound $q_{j'}$:

$$\{q_{j'}\} = \{\alpha_i | q_j \le \alpha_i, q_{j'} = \max\{q_j\}\} \tag{13}$$

If $q_{j'}$ does not exist, then let $\{\bar{0}\}$ *denote* α_i set as follows:

$$\{\bar{0}\} = \{\alpha_i | q_j > \alpha_i \text{ or } \alpha_i \text{ has nothing to do with } q_j\} \tag{14}$$

All the $\alpha_i's$ with the same IRP will be classified into one category by comparing α_i with all p items in the Q_t *matrix*: based on the definition of $q_{j'}$, if $q_{j'}$ exists, it means that $q_j = \bigvee_{p\in Q} p \le q_{j'} \le \alpha_i$, so the IRs between α_i and p ($p \le \alpha_i$) are 1, that is, $IR(\alpha_i, p) = \alpha_i{}^o p = 1$, the IRs between α_i and the rest of p in the Q_t *matrix* are 0, that is, $IR(\alpha_i, p) = \alpha_i p = 0$. Therefore, all the $\alpha_i's$ in $\{q_{j'}\}$ have the same IR, and these $\alpha_i's$ belong to one category. If $q_{j'}$ does not exist, for all p items in the Q_t *matrix*, $\alpha_i < p$ or α_i *has nothing to do with* p, the IRs between α_i and p is 0, IR is the same with zero vector $(\bar{0})$, and thus, these $\alpha_i's$ are the same category as zero vector.

Proposition 1: All α_is in the Q_s matrix are classified into $\{\mathbf{q}_{j'}\}$ ($\mathbf{q}_{j'} \in \mathbf{Q_t^+}$) or $\{\bar{\mathbf{0}}\}$ ($\mathbf{i = 1,\ldots,n; j' = 1,\ldots,m; m \le n}$).

First, there must be existed a $\alpha_\mathbf{i}$ for every $q_{j'}$ *in the* Q_t^+ *matrix*, so that $\alpha_\mathbf{i}$ *is equal to* $q_{j'}$, so $q_{j'}$ is the maximum lower bound of α_i, α_i is an element of a *set* $\{q_{j'}\}$. m $\alpha_i's$ are divided into m sets $\{q_{j'}\}$.

Second, for the remaining n-m $\alpha_i's$,

(1) For every p in the Q_t *matrix*, if $\alpha_i < p$ or α_i *has nothing to do with* p, then $q_{j'}$ does not exist, so α_i belongs to set $\{\bar{0}\}$;

(2) If $p \le \alpha_i$, there must be existed $q_{j'}$ acted as the maximum lower bound of α_i, so α_i belongs to set $\{q_{j'}\}$.

Combining (1) and (2), Proposition 1 is proved.

Proposition 2: If the number of $\mathbf{q_j}$ in the $\mathbf{Q_t^+}$ matrix is m, all $\alpha_\mathbf{i}'$s in the $\mathbf{Q_s}$ matrix are classified into m+1 categories.

From Proposition 1, the conclusion is clearly true, that is, m+1 categories of the set of KSs can be estimated correctly. Thus, $PMR_{IRP} = \frac{m+1}{n}$. The result of TCV = PMR_{IRP} shows that the TCV is equal to the PMR estimated by the set of IRPs. For $PMR_{ORP} \le PMR_{IRP} = TCV$, the TCV is the upper bound of the PMR estimated by the ORP. When k is smaller, such as $k \le 5$, the TCV can be calculated by pen, otherwise, it is easily derived by using a computer.

SIMULATION STUDY

A simulation study was carried out to evaluate the relationships between the TCV and the PMR.

Five attribute hierarchical structures were studied, namely, independent, linear, convergent, divergent, and unstructured. The number of attributes was set at 4, that is, $K = 4$. The study needed to consider the influence of the distribution of examinees, item attribute vector, and their proportions on the TCV. Two kinds of distribution of the KSs of examinees were discussed as follows: the average distribution (30 persons for every KS) and the normal distribution. In particular, the standard multivariate normal distributions in the independent structure were investigated. The total number of examinees was the same. In contrast, there were six Q_t matrices for each structure, items would be selected from the Q_r matrix, and its proportions were different. The test length was 20. The descriptive statistics of the examinees and the Q_t matrices are reported in **Table 2**.

To compare the effects of different slips on the TCV and the PMR, the slips were 0.15 and 0.02, respectively. The set of IRPs was obtained by the items of the Q_t matrix and the set of KSs of the Q_s matrix. Let x denoted the IR score of an examinee on an item, r randomly generated from Uniform (0, 1), if $r > 1 - s$, x (x was dichotomous) would be changed to 1–x, and x otherwise.

The DINA model and the maximum-likelihood estimation method were used to estimate the KS. Considering the differences in the distribution of examinees, the Q_t matrix, and the slips, there were 116 levels in total, and each level was tested 30 times. The final PRM was an average of 30 PMRs.

The PMR index can be defined as follows:

$$PMR = \frac{\sum_{i=1}^{N} \alpha_{i-correct}}{N} \tag{15}$$

where N is the number of examinees. $\alpha_{i-correct} = 1$ *represents* that the ith examinee is estimated correctly.

RESULTS

Table 3 compares the TCV and the PMR obtained from the linear structure. The first column shows the different distribution of examinees, and the other columns show the results of the different Q_t matrices.

Clearly, the TCV was superior: the TCV was uniformly higher than the PMR regardless of the distribution of examinees and the Q_t matrices. Although the repetition of items in the Q_t matrices, the TCV was not changed when the distribution of examinees and the category of items in the Q_t matrices remain unchanged. Therefore, this helped in explaining why repeated items were not necessary to count. As is known to all, the smaller the slip is, the higher the PMR is. But the TCV had nothing to do with the slip, so the smaller the slip, the smaller the gap between the TCV and the PMR. For all the attribute structures, when the

TABLE 2 | The distributions of examinees and the proportions of items for five different hierarchical structures.

Attribute structure	Distribution of examinees	The ratio of examinees	The Q_t matrix	The ratio of items
Linear	The average distribution	Each KS has the same ratio	An item category: (1100)	
			Two item categories: (1000) (1110)	①1:1 ②1:3
	The normal distribution	(0000):(1111):(1000):(1110): (1100) = 1:1:2:2:4	Four item categories: (1000, 1100, 1110, 1111)	①1:1:1:1②2:8:8:2
Convergent	The average distribution	Each KS has the same ratio	Two item category: (1100) (1010)	①1:1 ②1:3
	The normal distribution	(0000):(1111): (1000):(1110): (1100):(1010) = 1:1:2:2:7:7	Four item categories: (1000) (1100) (1010) (1110)	①1:1:1:1 ②2:8:8:2
			Five item categories: (1000) (1100) (1010) (1110) (1111)	①1:1:1:1:1②2:4:8:4:2
Divergent	The average distribution	Each KS has the same ratio	Two item category: (1100) (1011)	①1:1 ②1:3
	The normal distribution	(0000):(1111): (1000):(1110): (1100):(1011): (1010) = 10:10:21:21:42:42:64	Four item categories: (1000) (1100) (1010) (1111)	①1:1:1:1 ②2:8:8:2
			Six item categories: (1000) (1100) (1010) (1011) (1110) (1111)	①2:4:4:4:4:2②2:3:5:5:3:2
Unstructured	The average distribution	Each KS has the same ratio	Four item categories: (1000) (1010) (1101) (1011)	
	The normal distribution	(0000):(1111): (1000):(1011): (1100):(1101): (1010):(1110): (1001) = 16:16:22:22:27:27:43:43:54	Six item categories: (1000) (1100) (1001) (1101) (1011) (1111)	①1:1:1:1 ②2:8:8:2 ①2:4:4:4:4:2②2:3:5:5:3:2
			Eight item categories: (1000) (1100) (1010) (1001) (1110) (1101) (1011) (1111)	①2:2:2:4:4:2:2:2②1::1:3:5:5:3:1:1
Independent	The average distribution	Each KS has the same ratio	Eight item categories: (1000) (0010) (1100) (1001) (1110) (1101) (1011) (1111)	①2:2:2:4:4:2:2:2 ②1::1:3:5:5:3:1:1
	The normal distribution	(0000):(1111):(1000):(0111): (0100):(1011):(0010):(1101): (0001):(1110):(1100):(0011): (1010):(0101): (1001):(0110) = 5:5:14:14:24:24:34:34:34 :38:38:43:43:48:48	Twelve item categories: (1000) (0100) (0010) (0001) (1100) (1010) (1001) (0101) (0110) (0011) (1110) (1111)	①1:1:2:2:2:2:2:2:2:2:1:1 ②1:1:1:1:3:3:3:3:1:1:1:1
	The standard multivariate normal distribution	(0000):(1111):(1000):(0111): (0100):(1011):(0010):(1101): (0001):(1110):(1100):(0011): (1010):(0101): (1001):(0110) = 64:50:77:3:19:3:8:17:0:103: 90:0:28:3:13:2	Fifteen item categories: (1000) (0100) (0010) (0001) (1100) (1010) (1001) (0110) (0101) (0011) (1110) (1101) (1011) (0111) (1111)	①1:1:1:1:1:1:2:4:2:1:1:1:1:1:1 ②1:1:1:2:2:2:2:2:2:2:2:1:1:1

The third column is the ratio of the knowledge state (KS); ① and ② are two different ratios of items in the Q_t matrix; thus, there should be two different Q_t matrices.

TCV was low, the PMR was also low and vice versa. Notably, the more the item categories were, the larger the TCV would be. In particular, if the Q_t matrix contained the reachable matrix that could augment all possible item categories, then TCV = 1, regardless of the distribution of examinees. In other words, when the reachable matrix was a submatrix of the Q_t matrix, the PMR would be higher than that of the Q_t matrix that did not include the reachable matrix if the other conditions were the same.

From **Tables 4**–**7**, the data of other structures show the same results as linear. In addition, the lesser the structure, the greater the difference between the TCV and the PMR.

DISCUSSION AND CONCLUSION

Guided by a cognitive model, the CD can detect how well the examinees have mastered certain knowledge or skills. All CDTDs aim at diagnosing examinees as much as possible, and the main evaluation index is the PMR. The higher the accuracy rate of the KSs, the higher the test construct validity. It is more meaningful to be able to calculate the PMR during CDTD. Tatsuoka (2009, p. 78–79) believed that the sufficient Q matrix can improve the test construct validity. However, how to measure the construct validity? Inspired by the

TABLE 3 | The comparison between the TCV and the PMR_{ORP} of the linear structure.

Distribution of examinees	1 category	2 categories		4 categories	
	①	①	②	①	②
The average distribution	TCV = 0.4 PMR = 0.4 (PMR = 0.4)	TCV = 0.6 PMR = 0.5956 (PMR = 0.6)	TCV = 0.6 PMR = 0.5780 (PMR = 0.6)	TCV = 1 PMR = 0.9413 (PMR = 1)	TCV = 1 PMR = 0.7867 (PMR = 0.9827)
The normal distribution	TCV = 0.5 PMR = 0.5 (PMR = 0.5)	TCV = 0.5 PMR = 0.4971 (PMR = 0.5)	TCV = 0.5 PMR = 0.4851 (PMR = 0.5)	TCV = 1 PMR = 0.9164 (PMR = 0.9998)	TCV = 1 PMR = 0.8353 (PMR = 0.9849)

Values in brackets were the average of PMR when the slip was 0.02, and the following tables were similar.

TABLE 4 | The comparison between the TCV and the PMR_{ORP} of the convergent structure.

Distribution of examinees	2 categories		4 categories		5 categories	
	①	②	①	②	①	②
The average distribution	TCV = 0.6667 PMR = 0.6583 (PMR = 0.6667)	TCV = 0.6667 PMR = 0.6398 (PMR = 0.6667)	TCV = 0.8333 PMR = 0.7726 (PMR = 0.8332)	TCV = 0.8333 PMR = 0.7370 (PMR = 0.8226)	TCV = 1 PMR = 0.8056 (PMR = 0.9935)	TCV = 1 PMR = 0.7600 (PMR = 0.9765)
The normal distribution	TCV = 0.85 PMR = 0.8374 (PMR = 0.85)	TCV = 0.85 PMR = 0.8135 (PMR = 0.8498)	TCV = 0.95 PMR = 0.8748 (PMR = 0.95)	TCV = 0.95 PMR = 0.8954 (PMR = 0.9444)	TCV = 1 PMR = 0.8057 (PMR = 0.9987)	TCV = 1 PMR = 0.8167 (PMR = 0.9898)

TABLE 5 | The comparison between the TCV and the PMR_{ORP} of the divergent structure.

Distribution of examinees	2 categories		4 categories		6 categories	
	①	②	①	②	①	②
The average distribution	TCV = 0.5714 PMR = 0.5610 (PMR = 0.5714)	TCV = 0.5714 PMR = 0.5316 (PMR = 0.5714)	TCV = 0.8571 PMR = 0.7773 (PMR = 0.8570)	TCV = 0.8571 PMR = 0.7095 (PMR = 0.8329)	TCV = 1 PMR = 0.8083 (PMR = 0.9740)	TCV = 1 PMR = 0.8390 (PMR = 0.9848)
The normal distribution	TCV = 0.4952 PMR = 0.4883 (PMR = 0.4952)	TCV = 0.4952 PMR = 0.4835 (PMR = 0.4952)	TCV = 0.8 PMR = 0.6737 (PMR = 0.7997)	TCV = 0.8 PMR = 0.6694 (PMR = 0.7930)	TCV = 1 PMR = 0.7681 (PMR = 0.9937)	TCV = 1 PMR = 0.8300 (PMR = 0.9962)

TABLE 6 | The comparison between the TCV and the PMR_{ORP} of the unstructured structure.

Distribution of examinees	4 categories		6 categories		8 categories	
	①	②	①	②	①	②
The average distribution	TCV = 0.6667 PMR = 0.6042 (PMR = 0.6665)	TCV = 0.6667 PMR = 0.5463 (PMR = 0.6548)	TCV = 0.7778 PMR = 0.6446 (PMR = 0.7727)	TCV = 0.7778 PMR = 0.6384 (PMR = 0.7683)	TCV = 1 PMR = 0.7078 (PMR = 0.9774)	TCV = 1 PMR = 0.7363 (PMR = 0.9641)
The normal distribution	TCV = 0.5407 PMR = 0.4846 (PMR = 0.5407)	TCV = 0.5407 PMR = 0.4491 (PMR = 0.5362)	TCV = 0.6815 PMR = 0.5945 (PMR = 0.6780)	TCV = 0.6815 PMR = 0.5483 (PMR = 0.6785)	TCV = 1 PMR = 0.5662 (PMR = 0.9848)	TCV = 1 PMR = 0.7418 (PMR = 0.9815)

evaluation of the sufficient Q matrix by Tatsuoka (1995, 2009), an evaluation index for cognitive diagnostic test (design) was developed, i.e., TCV, which made up for the defects of Tatsuoka's idea (Tatsuoka, 1995, 2009).

This study proposes a simplified method for predicting the PMR, namely, the TCV method for CD. The TCV intuitive meaning is as follows: the set of KSs is derived from the Q_t matrix through the augment algorithm (i.e., this design can inspire some latent cognitive states), and if the probability distribution of the examinees in the population is known, then $TCV = \sum_j p_j$. In particular, when the probability distribution of the set of KSs in the total population is discrete uniform, the TCV is equal to the sum, which is the number of categories of the set of KSs derived from the Q_t matrix plus 1, divided by the number of categories of the set of KSs in the population. In general, the TCV measures the degree of consistency between the cognitive

TABLE 7 | The comparison between the TCV and the PMR_{ORP} of the independent structure.

Distribution of examinees	8 categories		12 categories		15 categories	
	①	②	①	②	①	②
The average distribution	TCV = 0.625 PMR = 0.4242 (PMR = 0.6094)	TCV = 0.625 PMR = 0.4515 (PMR = 0.6093)	TCV = 1 PMR = 0.6592 (PMR = 0.9597)	TCV = 1 PMR = 0.6417 (PMR = 0.9737)	TCV = 1 PMR = 0.6110 (PMR = 0.9433)	TCV = 1 PMR = 0.6285 (PMR = 0.9489)
The normal distribution	TCV = 0.5813 PMR = 0.3987 (PMR = 0.5631)	TCV = 0.5813 PMR = 0.4223 (PMR = 0.5628)	TCV = 1 PMR = 0.6953 (PMR = 0.9520)	TCV = 1 PMR = 0.6390 (PMR = 0.9837)	TCV = 1 PMR = 0.6035 (PMR = 0.9472)	TCV = 1 PMR = 0.6438 (PMR = 0.9583)
The standard Multivariate normal distribution	TCV = 0.9438 PMR = 0.5458 (PMR = 0.9014)	TCV = 0.9438 PMR = 0.6421 (PMR = 0.9254)	TCV = 1 PMR = 0.7148 (PMR = 0.9699)	TCV = 1 PMR = 0.7448 (PMR = 0.9792)	TCV = 1 PMR = 0.5848 (PMR = 0.9687)	TCV = 1 PMR = 0.6508 (PMR = 0.9490)

model derived from matrix Q_t and the theoretical cognitive model (Ding et al., 2012).

As the proof and the simulation showed, $PMR_{ORP} \leq PMR_{IRP} = TCV$. Therefore, the TCV can be used to predict the PMR. Notably, the TCV is related to the distribution of examinees and item category, not related to the proportion of items. In other words, when calculating the TCV, repeated items should be treated as one item.

The TCV is numerically equal to the PMR based on the set of IRPs, and the factors that affect the set of IRPs are as follows: the cognitive model (e.g., the number of attributes, attribute hierarchy, and compensation between attributes), the composition of the test matrix (e.g., Boolean matrix and multivalued Q matrix), the item score (e.g., 0–1 score or multilevel score). Whatever has an effect on the set of IRPs influences the TCV. When the test Q matrix (Q_t) is a Boolean matrix, the score is 0 or 1, and the IR is 1 if and only if $\alpha_i \geq q_j$, the TCV is the upper bound of the PMR. The TCV has nothing to do with the CDM (i.e., classification method); therefore, the TCV is calculated by CDM-free. Thus, the conclusion is the same for the DINA model, the AHM (Attribute Hierarchy Method, Gierl et al., 2007) model, the RSM (Rule Space Method, Tatsuoka, 2009) model, and the GDD (Generalized Distance Discrimination, Sun et al., 2011) model.

The number of attributes has an effect on the TCV. For example, independent structure, if the probability distribution of the set of KSs in the total population is equal, different items containing only two attributes are selected, then when the number of attributes K is 3, and the TCV is 5/8; when K is 4, the TCV is 3/4; and when K is 5, the TCV is 27/32. However, under the same conditions, the number of attributes does not affect the conclusion that the TCV is the upper bound of the PMR at all (as shown by the proof). Furthermore, the lower the number of attributes, the higher the PMR. Therefore, the simulation study selected fewer attributes ($K = 4$). Similarly, the smaller the random in the ORP is, the higher the PMR is. To prove that the TCV is the upper bound of the PMR, in the simulation study, the random is relatively small ($s = 0.02$). According to the abovementioned logic, the result that TCV is the upper bound of the PMR is also true when the random is larger.

An interesting question arises as follows: the TCV is not equal to the PMR, why the TCV is useful for predicting the PMR? There are three reasons: First, the most important reason is that the TCV can be obtained during CDTD, which is instructive to adjust selected items at any time and to timely judge the test quality. Second, the TCV is the upper bound of the PMR, the smaller the slip, the smaller the gap between the TCV and the PMR. The TCV does not change with the slip. If the TCV is high, the PMR is also higher; therefore, it is feasible to use the TCV as an index of the PMR to predict the test quality. Third, the TCV is easy to calculate according to the formula.

The TCV can be used not only to predict the PMR but also, more importantly, to detect the defects of CDTD. By using the augment algorithm, the set of KSs can be derived from the Q_t matrix, and then, the TCV can be calculated. Under the same conditions, if the TCV value is lower, it means that there are fewer kinds of attribute vectors (i.e., items) of the reachable matrix in the Q_t matrix, and thus, the more KSs cannot be accurately estimated. At this time, test designers can modify the test Q matrix (i.e., the Q_t matrix) before testing (not posttest evaluation), that is, modify the test (such as filling the columns of the reachable matrix or filling the columns expanded by the reachable matrix through the augment algorithm). Adjusting the selected items according to the TCV value at any time is not only beneficial to evaluate the test quality in time in CDTD but also can save cost and improve efficiency, which has the effect of two times the result with half the effort. This method undoubtedly has great advantages in CDTD.

If the test contains the reachable matrix, the cognitive model derived from the test is consistent with the theoretical cognitive model, and the TCV is 1. At this time, as long as the item quality is good (i.e., the slip is low) and attributes are measured a certain number of times, then the PMR is relatively high. In most cases, however, the PMR is not equal to 1 because the test is short, the quality of the items is poor, or the examinees do not answer carefully. At this time, although the result is rough when the TCV is used to predict the PMR, even so, under the same cases, the test, which contained the reachable matrix (in this case, the Q_t matrix is complete Q matrix, Cai et al., 2018), has the higher PMR.

Although this study shows that the TCV method works successfully with CD, it has limitations in several aspects: (1) Since the TCV is determined by the Q_t matrix, the Q_t matrix must be complete and reliable, which is the premise of using the TCV. In some cases, this condition may be quite harsh. But the RUM model allows the Q_t matrix to be incomplete, and the conclusion of this study cannot be applied. Furthermore, the complete and accurate calibration of the Q_t matrix is still a very difficult problem. (2) If the score is 0 or 1 and IR is 1 if $\alpha_i \geq q_j$, other IR rules are not applicable in this case. Nor does it apply if there is compensation between attributes. (3) Only the dichotomous and non-compensable attributes are considered, a natural question that arises is how to get the TCV when the scoring is polytomous and attributes are compensable. These will be the interesting topics for future studies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

XT designed the study, conducted the simulation study, and wrote the manuscript. SD, HD, and MM revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was partially supported by the National Natural Science Foundation of China (Grant No. 31360237, 31500909, 61967009, 62067005), the National Social Science Foundation of China (Grant No. 13BYY087), the Project of Teaching Reform of Jiangxi (Grant No. JXJG-19-2-16), the Project of Education Sciences Planning of Jiangxi (Grant No. 20YB028), the Project of Humanity and Social Science Youth Foundation of the Ministry of Education (Grant No. 16yjc190016), and the Project of Graduate Innovation Special Foundation of Jiangxi (Grant No. YC2019033).

REFERENCES

Cai, Y., Tu, D., and Ding, S. (2018). Theorems and methods of a complete Q matrix with attribute hierarchies under restricted q-matrix design. *Front. Psychol.* 9:1413. doi: 10.3389/fpsyg.2018.01413

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method. *Educ. Psychol. Measur.* 70, 902–913. doi: 10.1177/0013164410366693

Chiu, C. Y., Douglas, J. A., and Li, X. D. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika* 74, 633–665. doi: 10.1007/s11336-009-9125-0

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: development and applications. *J. Educ. Measure.* 45, 343–362. doi: 10.1111/j.1745-3984.2008.00069.x

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. A. (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. Measure.* 35, 8–26. doi: 10.1177/0146621610377081

Ding, S. L., Luo, F., Cai, Y., Lin, J., and Wang, X. B. (2008). "Complement to Tatsuoka's Q matrix theory," in *New Trends in Psychometrics*, eds K. shigemasu, A. Okada, T. Imaizumi, and T. Hoshino (Tokyo, Japan: Universal Academy Press, Inc.), 417–423.

Ding, S. L., Mao, M. M., Wang, W. Y., Luo, F., and Cui, Y. (2012). Evaluating the consistency of test items relative to the cognitive model for educational cognitive diagnosis. *Acta Psychol. Sinica* 44, 1535–1546. doi: 10.3724/SP.J.1041.2012.01535

Ding, S. L., Luo, F., and Wang, W. Y. (2014a). Design of polytomous cognitively diagnostic test blueprint-for the independent and the rhombus attribute hierarchies. *J. Jiangxi Normal Univ.* 38, 265–268. doi: 10.16357/j.cnki.issn1000-5862.2014.03.012

Ding, S. L., Luo, F., Wang, W. Y., and Xiong, J. H. (2014c). The designing cognitive diagnostic test with dichotomous scoring. *J. Jiangxi Normal Univ.* 43, 441–447. doi: 10.16357/j.cnki.issn1000-5862.2019.05.01

Ding, S. L., Wang, W. Y., and andYang, S. Q. (2011). The design of cognitive diagnostic test blueprints. *J. Psychol. Sci.* 34, 258–265.

Ding, S. L., Wang, W. Y., and Luo, F. (2014b). The word frequency effect of fovea and its effect on the preview effect of parafovea in Tibetan reading. *J. Jiangxi Normal Univ.* 38, 111–118.

Ding, S. L., Wang, W. Y., Luo, F., and Xiong, J. H. (2015). The polytomous Q- matrix theory. *J. Jiangxi Normal Univ.* 39, 365–370. doi: 10.16357/j.cnki.issn1000-5862.2015.04.07

Ding, S. L., Yang, S. Q., and Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *J. Jiangxi Normal Univ.* 34, 490–494. doi: 10.16357/j.cnki.issn1000-5862.2010.05.023

Gierl, M. J., Leighton, J. P., and Hunka, S. M. (2007). "Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills," in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. P. Leighton and M. J. Gierl (New York, NY: Cambridge University Press), 242–274.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Measure.* 26, 333–352. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hartz, S. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*, Unpublished doctoral dissertation University of Illinois at Urbana-Champaign: Champaign, IL.

Henson, R. A., and Douglas, J. (2005). Test construction for cognitive diagnostics. *Appl. Psychol. Measure.* 29, 262–277. doi: 10.1177/0146621604272623

Henson, R. A., Roussos, L., Douglas, J., and He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Appl. Psychol. Measure.* 32, 275–288. doi: 10.1177/0146621607302478

Huebner, A., Finkelman, M. D., and Weissman, A. (2018). Factors affecting the classification accuracy and average length of a variable-length cognitive diagnostic computerized test. *J. Comput. Adapt. Test.* 6, 1–14. doi: 10.7333/1802-060101

Kang, H. A., Zhang, S. S., and Chang, H. H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *J. Educ. Measure.* 54, 165–183. doi: 10.1111/jedm.12139

Kuo, B. C., Pai, H. S., and de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Appl. Psychol. Measure.* 40, 315–330. doi: 10.1177/0146621616638643

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *J. Educ. Measure.* 41, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Liu, J. C., Xu, G. J., and Ying, Z. L. (2012). Data-driven learning of Q-matrix. *Appl. Psychol. Measure.* 36, 548–564. doi: 10.1177/0146621612456591

Liu, R., Huggins-Manley, A. C., and Bradshaw, L. (2016). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educ. Psychol. Measure.* 77, 220–240. doi: 10.1177/0013164416645636

Liu, S. C., Tu, D. B., Cai, Y., and Zhao, Y. (2018). Four new item selection strategies based on attribute balancing in CD-CAT. *Psychol. Sci.* 41, 976–981. doi: 10.16719/j.cnki.1671-6981.20180432

Luo, F., Wang, X. Q., Ding, S. L., and Xiong, J. H. (2018). The design and selection strategies of adaptive multi-group testing for cognitive diagnosis. *Psychol. Sci.* 41,720–726. doi: 10.16719/j.cnki.1671-6981.20180332

Madison, M. J., and Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educ. Psychol. Measure.* 75, 491–511. doi: 10.1177/0013164414539162

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Samejima, F. (1995). "A cognitive diagnosis method using latent trait models: competency space approach and it s relationship with DiBelloand Stout' s Unified Cognitive-Psychometric Diagnosis Model," in *Cognitively Diagnostic Assessment*, eds P. D. Nichols, S. F. Chipman, and R. L. Brennan (Hillsdale, NJ: Erlbaum), 391–410.

Shao, H. (2010). *Cognitive Diagnosis and its Application for the Surface Similarity Test of Children's Analogical Reasoning* (Unpublished master's thesis). Jiang xi Normal University.

Sun, J. N., Zhang, S. M., Xin, T., and Bao, Y. (2011). A cognitive diagnosis method based on Q-Matrix and generalized distance. *Acta Psychol. Sin.* 43, 1095–1102. doi: 10.3724/SP.J.1041.2011.01095

Sun, X. J., Wang, Y. T., Zhang, S. Y., and Xin, T. (2019). New methods to balance attribute coverage for cognitive diagnostic computerized adaptive testing. *Psychol. Sci.* 42, 1236–1244. doi: 10.16719/j.cnki.1671-6981.20190531

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Measure.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K. K. (1995). "Architecture of knowledge structures and cognitive diagnosis: a statistical pattern classification approach," in *Cognitively Diagnostic Assessments*, eds P. D. Nichols, S. F. Chipman, R. L. Brennan (Hillsdale, NJ: Erlbaum), 327–359.

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. New York, NY: Taylor and Francis Group.

Tu, D. B., and Cai, Y. (2015). The development of CD-CAT with polytomous attributes. *Acta Psychol. Sinica* 47, 1405–1414. doi: 10.3724/SP.J.1041.2015.01405

Tu, D. B., Cai, Y., and Dai, H. Q. (2013). Item selection strategies and initial items selection methods of CD_CAT. *J. Psychol. Sci.* 36, 469–474. doi: 10.16719/j.cnki.1671-6981.2013.02.040

Xu, X., Chang, H., and Douglas, J. (2003). "Computerized adaptive testing strategies for cognitive diagnosis," in *Paper Presented at the Annual Meeting of National Council on Measurement in Education, Montreal, Canada*.

Yang, S. Q., Cai, S. Z., Ding, S. L., Lin, H. J., and Ding, Q. L. (2008). Augment algorithm for reduced Q-matrix. *J. Lanzhou Univ.* 44, 87–91, 96. doi: 10.13885/j.issn.0455-2059.2008.03.027

Yu, D., Pan, Y. R., Ding, S. L., and Yang, Q. H. (2011). A new items election strategy of computerized adaptive testing with cognitive diagnosis. *J. Jiangxi Normal Univ.* 35, 548–550. doi: 10.16357/j.cnki.issn1000-5862.2011.05.015

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 Tang, Duan, Ding and Mao. 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: 16 December 2021
doi: 10.3389/fpsyg.2021.786612

Validating a Reading Assessment Within the Cognitive Diagnostic Assessment Framework: Q-Matrix Construction and Model Comparisons for Different Primary Grades

*Yan Li[1], Miaomiao Zhen[1] and Jia Liu[2]**

[1]Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China, [2]Tsinghua Laboratory of Brain and Intelligence, Department of Psychology, Tsinghua University, Beijing, China

OPEN ACCESS

Edited by:
Liu Yanlou,
Qufu Normal University, China

Reviewed by:
Huilin Chen,
Shanghai International Studies University, China
Sebastian Weirich,
Institute for Educational Quality Improvement (IQB), Germany

****Correspondence:***
Jia Liu
liujiathu@tsinghua.edu.cn

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

Received: *30 September 2021*
Accepted: *15 November 2021*
Published: *16 December 2021*

Citation:
Li Y, Zhen M and Liu J (2021) Validating a Reading Assessment Within the Cognitive Diagnostic Assessment Framework: Q-Matrix Construction and Model Comparisons for Different Primary Grades. Front. Psychol. 12:786612. doi: 10.3389/fpsyg.2021.786612

Cognitive diagnostic assessment (CDA) has been developed rapidly to provide fine-grained diagnostic feedback on students' subskills and to provide insights on remedial instructions in specific domains. To date, most cognitive diagnostic studies on reading tests have focused on retrofitting a single booklet from a large-scale assessment (e.g., PISA and PIRLS). Critical issues in CDA involve the scarcity of research to develop diagnostic tests and the lack of reliability and validity evidence. This study explored the development and validation of the Diagnostic Chinese Reading Comprehension Assessment (DCRCA) for primary students under the CDA framework. Reading attributes were synthesized based on a literature review, the national curriculum criteria, the results of expert panel judgments, and student think-aloud protocols. Then, the tentative attributes were used to construct three booklets of reading comprehension items for 2–6 graders at three key stages. The assessment was administered to a large population of students ($N = 21{,}466$) in grades 2–6 from 20 schools in a district of Changchun City, China. Q-matrices were compared and refined using the model-data fit and an empirical validation procedure, and five representative cognitive diagnostic models (CDMs) were compared for optimal performance. The fit indices suggested that a six-attribute structure and the G-DINA model were best fitted for the reading comprehension assessment. In addition, diagnostic reliability, construct, internal and external validity results were provided, supporting CDM classifications as reliable, accurate, and useful. Such diagnostic information could be utilized by students, teachers, and administrators of reading programs and instructions.

Keywords: cognitive diagnostic assessment, cognitive diagnostic models, G-DINA, Q-matrix validation, reading comprehension assessment, primary students

INTRODUCTION

Many existing learning and assessment systems generate scores, levels, and ranks to evaluate students' learning outcomes. This single outcome evaluation form has caused many problems, such as hurting students' self-esteem, heightening excessive competition, and increasing the learning burden, which are not conducive to the overall development of students (Lei, 2020). Therefore, new approaches are needed to improve outcome evaluation in the stage of basic education by keeping the evaluation content consistent with the curriculum criteria, providing diagnostic information on students' strengths and weaknesses in learning, and offering evidence for schools to implement intervention measures.

Cognitive diagnostic models (CDMs) are confirmatory latent class models that combine cognitive theory and psychometric models to reveal the innate structure of a given ability by estimating an individual's knowledge and skill mastery state (Leighton and Gierl, 2007). CDMs can group examinees into similar latent classes and thus can compensate for the deficiency of single outcome results generated *via* classical test theory and traditional item response theory (Ravand and Robitzsch, 2018). Due to the need for formative evaluation and instructions, CDMs have become popular in educational settings. However, Ravand and Baghaei (2020) noted that over 95% of CDM studies are methodological or simulation-oriented, approximately 4% are retrofitting, and less than 1% focus on real diagnostic test development in recent decades. Therefore, real CDM application studies have rarely found their ways into educational systems, probably because of the lack of reliability and validity evidence and thus limited confidence in the information provided by CDMs (Sessoms and Henson, 2018). There is still a wide gap between CDMs and educational practices, and true CDM studies to develop diagnostic tests from scratch are urgently needed (Alderson, 2010; Sessoms and Henson, 2018; Ravand and Baghaei, 2020).

CDA Framework

One of the ultimate purposes of CDMs is to make inferences about what attributes an examinee has mastered using a diagnostic assessment. That is, CDA offers valuable information on the diagnostic quality of test items as well as the skill mastery patterns of test-takers, classifying those who have not mastered the item's required skills, named non-masters, as distinct from those who have, named masters. The CDA frameworks have been proposed and optimized since Rupp and Templin (2008) published the first didactic introduction (de la Torre and Chiu, 2016; Ravand and Baghaei, 2020). In general, the construction of CDA depends on two major elements: the implicit theory section and the CDM section.

The first step in CDA is to specify the implicit attributes that a test-taker must possess to solve an item. The generic term "attribute" is defined as posited knowledge and thinking skill (de la Torre and Douglas, 2004) or a description of the processes, subskills, and strategies that are vital for the successful execution of a particular test (Leighton et al., 2004). Once the target attributes are defined *via* domain experts or think-aloud protocols, individual test items can be coded at the point of item development as a Q-matrix, an incidence matrix that transforms cognitive attributes into observable item response patterns (Tatsuoka, 1990; Li, 2011). It is essential to point out that diagnostic feedback is valid only when the attribute specification is complete, the items effectively measure the targeted attributes, and the attributes are correctly specified in the Q-matrix (Ravand and Baghaei, 2020). The quality of inferences about students is unlikely to be ensured in retrofitting studies, as they commonly include items that fail to adequately tap specific cognitive characteristics (Gierl and Cui, 2008; Chen and de la Torre, 2014).

Then, CDMs are utilized to group examinees with similar skill mastery profiles, to evaluate the diagnostic capacity of items and tests and thus to reveal the degree to which they can measure the postulated attributes (Ravand and Robitzsch, 2018). CDMs make various assumptions to reveal the innate structure of a given ability by estimating the interactions among attributes (Leighton and Gierl, 2007). That is, representative CDMs can mainly be classified into three types: compensatory, non-compensatory, and general models. In compensatory CDMs, mastering one or more targeted attributes can compensate for other attributes that are not mastered. The deterministic input noisy-or-gate model (DINO; Templin and Henson, 2006) and the additive CDM (A-CDM; de la Torre, 2011) are the most representative compensatory CDMs. In contrast, if an attribute has not been mastered, the probability of a correct response in the non-compensatory CDM would be low, as other mastered attributes cannot fully compensate for it. Representative non-compensatory CDMs include the deterministic input noisy-and-gate model (DINA; Haertel, 1989) and the reduced reparameterized unified model (R-RUM; Hartz, 2002). General CDMs allow the estimation of both compensatory and non-compensatory interactions among attributes within the same test, which has influentially led to the unification of various CDMs. The most famous general model is the general DINA model (G-DINA; de la Torre, 2011), which can be transformed into the abovementioned CDMs simply by setting specific constraints to zero or changing link functions.

Like other statistical models, a CDM has no value if it fits the data poorly (de La Torre and Lee, 2010). Specifically, the fitness of CDMs can be ascertained in two ways. Relative fit indices evaluate whether the fit of one model differs significantly from that of another, and the model with smaller relative fit values is judged to better fit the data (Lei and Li, 2016). According to previous research, three well-known relative fit indices are also applicable to CDM studies, including −2 log-likelihood (−2LL), Akaike's information criterion (AIC), and Bayesian information criterion (BIC; Lei and Li, 2016). In addition, absolute fit indices examine the adequacy of a single model (Liu et al., 2017). For instance, a model can be considered a good fit only if the value of the standardized root mean square residual (SRMSR) is less than 0.05 (Maydeu-Olivares, 2013; George and Robitzsch, 2015). In addition, the max χ^2, which is the mean of the χ^2 test statistics of independence for all item pairs, was found to be sensitive in specifying model misfit (Chen and Thissen, 1997; Lei and Li, 2016).

A significant value of p of max χ^2 suggests that the model fits poorly (George and Robitzsch, 2015).

CDM Applications in Reading Tests

As one of the most frequently assessed skills, reading is considered a prerequisite for success in school and life (Kim and Wagner, 2015). As complex and multiple-task abilities, the innate characteristics of reading comprehension have been widely discussed (Barnes, 2015). For example, the construction-integration model regards reading as a meaning-construction process that involves interaction between both reader and text and is influenced strongly by background knowledge (Kintsch, 1991; Snow, 2002). This model characterized reading as an iterative and context-dependent process by which readers integrate information from a text (Compton and Pearson, 2016). In contrast, theorists of component models have pointed out that some important language knowledge, cognitive processes, and reading strategies make relatively independent contributions to reading comprehension (Cain et al., 2004; Cain, 2009). These models indicate that subcomponents of reading, including but not limited to vocabulary, syntax, morphology, semantics, inference, reasoning, discourse comprehension, working memory, and comprehension monitoring, are strong and persistent predictors for readers from children to adults (Aaron et al., 2008; Kim, 2017). Although many studies found that Chinese reading and English reading shared significantly in common (Mo, 1992; Chen et al., 1993), a consensus has not been reached on the number of subcomponents involved at different developmental stages. For example, Mo (1992) proposed that the structure of Chinese language reading displayed a "replacing developmental pattern." Factor analysis results of a reading test battery suggested that 75% of the variance in grade-6 students' reading ability was explained by six factors, including word decoding, integration and coherence, inference, memory and storage, fast reading, and transfer ability. As grades increased to the secondary and high school levels, the influences of the abovementioned factors remained important but were partly replaced by newly emerged, higher-level factors such as generalization ability, evaluation ability, and semantic inference ability.

Early research on reading cognitive diagnosis tried to explore the separability of reading ability and identify whether there are relatively independent cognitive components, processes, or skills in reading ability. For example, Jang (2009) found that evidence in Markov chain Monte Carlo aggregation supported the separability of reading into 9 attributes, and most LanguEdge test items have good diagnostic and discrimination power to measure the attributes well. Then, CDMs have been applied to retrofit the data of large-scale reading assessments such as the Progress in International Reading Literacy Study (PIRLS), the Programme for International Student Assessment (PISA), the Test of English as a Foreign Language (TOEFL), the Michigan English Language Assessment Battery (MELAB), and the Iranian National University Entrance Examination (e.g., Jang, 2005; Sawaki et al., 2009; Li, 2011; Chen and de la Torre, 2014; Chen and Chen, 2016; Ravand, 2016; Ravand and Robitzsch, 2018; Javidanmehr and Sarab, 2019; George and Robitzsch, 2021; Toprak-Yildiz, 2021). Many studies have used one preset CDM for reading tests, including DINA (George and Robitzsch, 2021), Fusion (Jang, 2009; Li, 2011), LCDM (Toprak and Cakir, 2021), or G-DINA (Ravand, 2016) models. Only a few compared multiple CDMs and found that general models, such as G-DINA or LCDM, had better fits for reading assessment data (Chen and Chen, 2016; Li et al., 2016; Ravand and Robitzsch, 2018; Javidanmehr and Sarab, 2019). In some cases, compensatory models such as A-CDM or LLM have shown a relatively close fit to those of general models (Li et al., 2016; Chen and de la Torre, 2014). Therefore, researchers called for further comparison of general and reduced CDMs for optimal performance and for an understanding of the interaction mechanism among reading attributes.

In the context of real CDA applications in reading assessment, research is relatively scarce. One notable effort was conducted by Xie (2014), in which a reading comprehension assessment of modern Chinese prose for junior high school students was developed and validated. Fusion model results revealed an unstructured attribute hierarchy of Chinese reading, which was composed of word decoding, formal schema, information extraction, information deduced, content analysis, content generalization, and text evaluation. In addition, Toprak and Cakir (2021) examined the second language reading comprehension ability of Turkish adults with a cognitive diagnostic reading test using the CDA framework.

We collected a total of 15 relevant empirical reading studies in diverse age groups with various language backgrounds and summarized a list of candidate attributes (see **Supplementary Table 1** for details) and CDMs for the next phases of test development and analysis. This detailed review yielded 6 commonly specified cognitive attributes, including vocabulary, syntax, retrieving information, making inferences, integration, and evaluation. Text-related attributes, such as narrative text, expository text, and discontinuous text, were also specified in studies of PIRLS and PISA. However, the abovementioned large-scale reading assessments were generally designed and developed under a unidimensional item response theory approach. CDM implementations to extract diagnostic feedback may raise severe issues with model fit, item characteristics, and diagnostic inferences for retrofitting data (Rupp and Templin, 2008; Gierl et al., 2010; Sessoms and Henson, 2018).

Primary students are in the key stages of reading development, during which they need to transition from "learning to read" to "reading to learn," and begin to encounter difficulties in new comprehension requirements (Carlson et al., 2014). The need for suitable instructions and reading materials as scaffolding is felt mostly at the primary level; therefore, assessing the extent to which the reading ability and subskills of students grow is valuable during their primary school years. However, students' reading ability grows so much over the course of their schooling that a single-booklet testing design for all grades is beset with problems (Brennan, 2006). Multilevel booklet designs are typically adopted, of which the contents and difficulty can be purposefully differed to balance test precision and efficiency. However, to the best of our knowledge, all CDM implementations were conducted on a single reading booklet for second language learners or grade 4 students and above. Several authors (e.g., Ravand, 2016; Sessoms and Henson, 2018) have briefly noted

that CDM applications might be specific to different characteristics of items or students. The construct equivalence of reading attributes and the generalizability of CDMs to other key developmental stages of reading remain unproven.

To address these issues, this study had three goals: (a) to illustrate how the cognitive diagnostic assessment (CDA) framework can be applied to develop the Diagnostic Chinese Reading Comprehension Assessment (DCRCA) for primary students at various key stages, (b) to evaluate the attribute equivalence and model fit adequacy of the CDMs for different developmental stages, and (c) to validate the diagnostic inferences of the DCRCA about primary students' reading subskills. To answer these questions, the study was mostly concerned with the construction of cognitive models of Chinese reading, the model-data fit evaluation of CDMs for three reading booklets, the validation of diagnostic psychometric properties, and the skill mastery profiles of primary students. This process can shed light on the limited CDA applications in reading test development and provide new methodologies for exploring reading skill structure. To the best of our knowledge, this is the first reading assessment whose CDM model fitness, diagnostic reliability and validity were examined at various developmental stages.

MATERIALS AND METHODS

The development and validation of the reading assessment followed the guidelines of the CDA framework (Ravand and Baghaei, 2020). The research processes are outlined in **Figure 1**.

Attributes Specification

Reading attributes were specified through multiple steps, involving domain experts and test-takers who participated in the determination of the core reading features for further curricular use.

a. **Literature review**: Candidate attributes were summarized by reviewing 15 empirically validated studies (see **Supplementary Table 1**), particularly based on those of Chinese reading and native language reading of primary students (Xie, 2014; Yun, 2017; George and Robitzsch, 2021; Toprak-Yildiz, 2021). This detailed review yielded 6 commonly specified cognitive attributes, including retrieving information, making inferences, integration and summation, evaluation, vocabulary, and syntax, as well as three text-related attributes, including narrative text, expository text and discontinuous text.
b. **Expert panel's judgments**: As reading attributes are highly dependent on the characteristics of Chinese reading and the framework of reading education, researchers invited five experts in reading assessment or education to obtain their judgments of large-scale reading assessments and the Chinese Language Curriculum Criterion for Compulsory Education (abbreviated as the curriculum criterion). The "syntax" attribute was first excluded because the curriculum criterion does not advocate any grammar teaching or evaluation at the primary school level but emphasizes helping students comprehend naturally occurring materials in a real language environment (Ministry of Education, 2011). Vocabulary is considered as important as reading comprehension at the primary level, and therefore, this skill was excluded and evaluated by the Chinese Character Recognition Assessment in the test battery. Infrequent attributes were also discussed case by case. For example, formal schema (Xie, 2014) was excluded because it might blend text evaluation with text-type attributes. The importance of literary text (i.e., narrative text and poetry) at the primary level has been emphasized by the curriculum criterion as well as large-scale assessments, including the PIRLS and PISA. However, inconsistencies in other text types have been observed. The curriculum criterion merges expository text (extracted from PIRLS) and

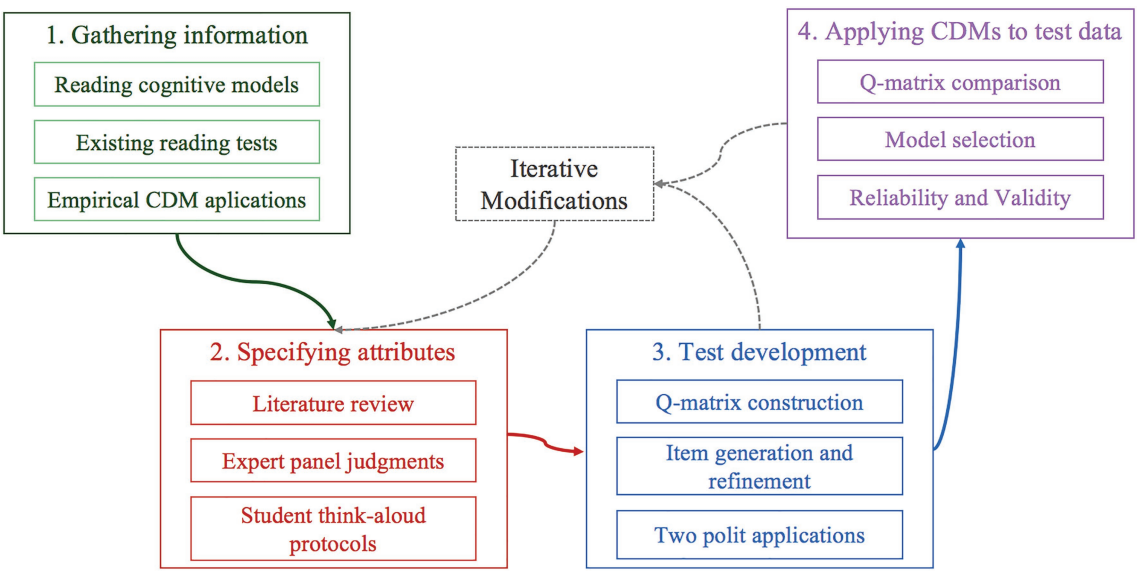


FIGURE 1 | An overview of the research processes.

discontinuous text (extracted from PISA) into practical text, as they have similarities in their reading objectives and strategies (Compulsory Education Curriculum and Textbook Committee of the Ministry of Education, 2012). After discussion, all experts agreed that this inconsistency was worth further evaluation *via* empirical results.

c. **Student think-aloud protocols**: To clarify the cognitive procedures that test-takers went through, 15 students from grades 2 to 6 were selected for think-aloud protocols. These students verbalized their thoughts when solving sample items. According to their answers and oral explanations, researchers identified clues to cognitive processes with an eye on the attributes inferred from the previous procedures. Overall, researchers specified and defined an initial set of eight attributes that might be crucial for primary school students (**Table 1**).

Test Development

According to the curriculum criterion, reading education can be divided into three key stages at the primary level. Key stage one is for grades 1 to 2, key stage two is for grades 3 to 4, and key stage three is for grades 5 to 6. Therefore, three booklets of reading diagnosis items were compiled for students at each key stage. An initial common Q-matrix for the three booklets was intentionally designed, as each item reflects one of the four cognitive processes of reading comprehension (α_1–α_4) and one text-related attribute (α_5, α_{6a}, and α_{6b}). The genre and complexity of texts were controlled, as they were important factors in assessing reading comprehension (Collins et al., 2020). Fragments of literary texts (including fairy tales, stories, fables, narratives, novels, and children's poems) and practical texts (including explanatory texts, simple argumentative articles, and discontinuous texts) were carefully selected and modified as item stems. A Chinese readability formula (Liu et al., 2021) was adopted to calculate the length, token types, lexical difficulty, function word ratio, and overall difficulty of each text. The average text length of the three booklets ranges from 150.60 to 278.57 characters, and the average text difficulty levels for the three booklets are 3.38, 3.69, and 4.40 (for details, please see **Supplementary Table 2**). Therefore, the three booklets are composed of conceptually appropriate short texts with increased complexity.

The item generation procedures were as follows: mapping cognitive and text-type attributes to compile 73 draft multiple-choice items, an expert review to cross-validate the Q-matrix, and item refinement following the expert review. Then, after the first pilot using two booklets for grade 1–2 and 3–6 students ($n = 378$), 17 problematic items were removed according to the item discrimination index (item-total correlation <0.19), and several items were modified. Grade 1 students were excluded from further study because they could not adapt to the computer assessment procedures. The second pilot included 56 items in three booklets, and each booklet consisted of 18–20 items. Pilot data were obtained from 5,949 grade 2–6 students. Both classical test theory and a 2PL item response model analysis were conducted. Five items with unsatisfactory discrimination (item-total correlation <0.30 or IRT discrimination <0.50) and three items with moderate to large differential item functioning issues on gender (effect size >0.88) were removed. A total of 48 items were retained, and four items were modified or rearranged for facility (passing rates by grade < 0.20 or > 0.90). The four cognitive attributes were intentionally balanced in testing frequency (4 to 5 times each attribute), and the proportion of literacy and practical texts were similar in the three booklets. Therefore, as shown in the last line of **Table 2**, the total testing frequencies of the attributes were similar in the three final booklets, with slight differences in item order and proportions of text type.

Measures

The Diagnostic Chinese Reading Comprehension Assessment (DCRCA)

DCRCA was developed as a multiple-choice, computer-based, online reading comprehension assessment to identify cognitive processes used during understanding literacy or practical short passages. The final DCRCA for grades 2 to 6 comprises 3 booklets, and each booklet contains 16 items. These items required students to answer multiple-choice questions on their

TABLE 1 | Definitions of the initial reading attributes.

No.	Attribute	Definition
α_1	Retrieving information	Retrieving information requires the abilities to understand a text literally and match the micro/macrolevel propositions to relevant parts of the text (Kintsch, 1991; O'Reilly and Sheehan, 2009; Xie, 2014).
α_2	Making inferences	Making inferences require combining reader background knowledge with contextual clues to determine implicit meaning and form a beyond surface-level understanding of the text (van Dijk and Kintsch, 1983; Toprak and Cakir, 2021).
α_3	Integration and summation	Integration and summation require an understanding of relationships across sentences and paragraphs as well as an understanding of the comparative importance of information (main and supporting; Grabe, 2009; O'Reilly and Sheehan, 2009).
α_4	Reflective evaluation	Reflective evaluation requires an understanding of the author's purpose, mood, tone, and stance toward the subject as well as evaluating the quality or appropriateness of a text (Chen and de la Torre, 2014; Xie, 2014; Toprak and Cakir, 2021).
α_5	Literary text	Literary text includes stories, folktales, legends, fables, simple fiction, nursery rhymes, narrative poem, limerick, and shallow ancient poetry (Common Core State Standards Initiative, 2010; Ministry of Education, 2011).
α_6	Practical text	Practical text contains shallow expository text and discontinuous text at the primary school level (Common Core State Standards Initiative, 2010; Ministry of Education, 2011).
α_{6a}	Expository text	Expository text includes illustrative text and simple argumentative text (Yun, 2017).
α_{6b}	Discontinuous text	Discontinuous text displays digital sources on a range of topics and information in charts, graphs, or maps (Chen and Chen, 2016).

TABLE 2 | Initial Q-Matrices.

Item	Booklet KS1								Booklet KS2								Booklet KS3							
	α_1	α_2	α_3	α_4	α_5	α_6	α_{6a}	α_{6b}	α_1	α_2	α_3	α_4	α_5	α_6	α_{6a}	α_{6b}	α_1	α_2	α_3	α_4	α_5	α_6	α_{6a}	α_{6b}
1	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	1	0
2	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0
3	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0
4	0	1	0	0	0	1	1	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
5	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0
6	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0
7	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	1	1	0	0	0	1	0	1
8	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	1	0	1
9	1	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0
10	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0
11	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
12	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0
13	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	1	1	0	0	0
14	0	1	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0
15	0	0	0	1	0	1	0	1	0	0	1	0	0	1	1	0	0	0	0	1	1	0	0	0
16	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1	0	0	0
Total	4	4	4	4	13	3	1	2	4	4	4	4	13	3	2	1	4	5	4	5	13	3	1	2

α_1 = *Retrieving information;* α_2 = *Making inferences;* α_3 = *Integration and summation;* α_4 = *Reflective evaluation;* α_5 = *Literary text;* α_6 = *Practical text;* α_{6a} = *Expository text;* α_{6b} = *Discontinuous text.*

comprehension of short passages. Students' responses were scored dichotomously (0 = incorrect, 1 = correct) for each item. As already described, each item was intentionally constructed by experts to align with precisely one of the four processes of reading comprehension (α_1–α_4) and one text-related attribute (α_5–α_6). The total testing frequencies of the attributes were similar in the three final booklets, while the short passages in the three booklets were compiled with increased complexity. Cronbach's α values for the assessment of the three booklets were 0.82, 0.71, and 0.64.

The Chinese Word Recognition Assessment

The Chinese word recognition assessment was adopted for validation purposes, and it was adapted from the Chinese character recognition task (Li et al., 2012) to measure students' word recognition skills. Students listened to the sound of a word composed of a given Chinese character and then chose the correct character from three distracting character options. A total of 150 character items were collected based on Chinese language textbooks (Shu et al., 2003). The maximum score of this assessment was 150. The internal reliability of the assessment was 0.91.

Sample

The study was conducted for a regional reading education project in Changchun City, China. The project aims to investigate the development of primary students' reading ability, recommend books suitable for reading, and provide them with corresponding reading courses. A total of 21,466 grade 2 to grade 6 students from 20 primary schools completed the assessments in November 2020, accounting for 94.1% of the total sample. Students were aged from 7.3 to 13.2 years, and the proportion of male students was 52.4% in total.

Procedure

Considering the large number of students participating in the DCRCA, the organization and implementation were completed by Chinese teachers and computer teachers of each class. Researchers trained all teachers and provided them with standardized assessment manuals. The assessments were administered collectively *via* an online web page, which presented one item at a time to students. The web page set all items as compulsory, so there was no missing value in the formal test as long as the student submitted successfully. Considering primary students' computer proficiency, students only needed to click medium-size options with mice to answer all questions. Students took approximately 20 min to successively complete the test battery, including the Chinese Word Recognition Assessment and the DCRCA. All students received an assessment analysis report with a recommended reading list and learning suggestions 1 month after the testing.

Analysis

Data were analyzed using R studio (R Core Team, 2021). As a correctly specified Q-matrix is considered a prerequisite of model-data fitness and low bias in diagnostic classifications (Rupp and Templin, 2008; Kunina-Habenicht et al., 2012), both theoretical and empirical procedures (de la Torre and Chiu, 2016) were applied iteratively to obtain the best attribute numbers and the best item-attribute relationships using the "GDINA" package, version 2.8.0 (Ma and de la Torre, 2020). The "CDM" package, version 7.5–15, was used for fitting CDMs (e.g., DINA, DINO, R-RUM, A-CDM, and G-DINA) based on the MMLE/EM algorithm (George et al., 2016; Robitzsch and George, 2019). The CDM package allows the estimation of rich sets of models, fit indices, and diagnostic validity with various emphases, which can help researchers find the most appropriate model. Two-parameter logistic item response theory (2PL-IRT) statistics were calculated using the ltm package (Rizopoulos, 2006).

RESULTS

Q-Matrix Validation

Three types of Q-matrices were created for each booklet to evaluate the applicability of attributes. Q1 contained only the four commonly agreed-upon cognitive attributes (α_1–α_4), Q2 added two text-type attributes (α_5 and α_6) to Q1 with reference to the curriculum criterion, and Q3 added three text-type attributes (α_5, α_{6a}, and α_{6b}) to Q1 with reference to PISA and PIRLS. These Q-matrices were compared based on the model-data fit of the G-DINA model and likelihood ratio test (see **Table 3**).

The SRMSR values of all Q-matrices were acceptable (below the 0.05 rule of thumb suggested by Maydeu-Olivares, 2013), while none of Q1 could be accepted based on the max χ^2. The -2LL and AIC values suggested a direction of improvement from Q1 to Q2, while the fit values of Q2 and Q3 were close in all booklets. Likelihood ratio tests were adopted between the adjacent Q-matrices within each booklet. We found that (1) all Q2 and Q3 values were significantly better than Q1 values ($p < 0.001$); (2) the -2LL and AIC differences between Q2 and Q3 were small and unstable, as p values fluctuated around significance boundaries for booklets KS1 to KS3 ($p \approx 0.006$, 1.00 and 0.049 respectively); and (3) the BIC consistently favored Q2 over Q3, as it was more compact and efficient. In summary, the fit indices showed similarities across booklets, suggesting that the attribute structure was the same across key stages. Based on the above results, we chose Q2 as a basis to finalize the item-attribute relationship.

An empirical Q-matrix validation procedure was conducted on all Q2s to compare the proportion of variance accounted for (PVAF) by plausible q-vectors for a given item (de la Torre and Chiu, 2016). A given q-vector was deemed correct if it was the simplest vector with a PVAF above 0.95. The validation results suggested no modification for booklet KS2 or KS3 and generated suggested Q-vectors for items 6 and 15 in booklet KS1. This indicated a relatively high attribute-wise agreement between the provisional and data-driven Q-matrices across all booklets. After expert revisions and iterative modeling, researchers concluded that the suggested changes in the Q-matrix were consistent with what the item truly assessed. The likelihood ratio test suggested that the fit of finalized Q2 was significantly better than that of the initial Q2 and was slightly better than that of Q3 for booklet KS1. The final Q-matrices are given in **Table 4**.

Model Comparison

To select the optimal CDM for the whole assessment and to reveal the relationships among reading attributes, we compared five representative CDMs including DINA, DINO, R-RUM, A-CDM, and G-DINA models, for each booklet using the final Q-matrices. As **Table 5** shows, the five CDMs performed stably across booklets. The AIC and -2LL values for the G-DINA models were the lowest in the three booklets, followed by the A-CDM and the R-RUM models, while the values of the more parsimonious DINO and DINA models were observably worse. The BIC favored A-CDM, G-DINA, and A-CDM in booklets KS1 to KS3. Likelihood ratio tests suggested that none of the other CDMs fit as good as the G-DINA model. For the absolute fit values, the SRMSR values of all CDMs were below 0.05. However, only the G-DINA had insignificant max χ^2 values in all cases, indicating a good fit to the data, while the DINO and DINA models were stably rejected by the significance of max χ^2 in all cases. It is evident that the G-DINA model fits the entire assessment data reasonably better than the more parsimonious reduced models.

Reliabilities and Validity

Pattern accuracy (Pa) and pattern consistency (Pc) indices show the degree to which the examinees were accurately and consistently classified as masters and non-masters (Cui et al., 2012). Therefore, they were adopted as indicators of reliability in **Table 6**. The Pa values for each separate attribute were between 0.68 and 0.95, and the Pc values were between 0.63 and 0.92. Despite a lack of consensus on general guidelines for what constitutes a high or acceptable reliability (Templin and Bradshaw, 2013), these results indicated an above acceptable capacity of measuring students' reading attributes.

Evidence of internal validity was provided using item mastery plots to quantify the various discriminatory and

TABLE 3 | Model-data fitting results for Q-matrix validation.

Booklet	Q-matrix	Npars	Relative fit			Absolute fit		
			-2LL	AIC	BIC	SRMSR	max χ^2	p (max χ^2)
KS1	Q1[a]	45	−38109.4	76308.7	76594.7	0.041	219.23	<0.001
	Q2[b]	90	−37547.7	75275.5	**75847.4**	0.016	11.30	0.09
	Q3[c]	97	−37540.7	75275.4	75891.8	0.015	9.32	0.27
	Sug Q2[c]	94	**−37537.9**	**75263.8**	75861.2	0.015	7.45	0.76
KS2	Q1[a]	43	−86989.8	174065.6	174370.0	0.021	26.65	<0.001
	Q2[b]	86	**−86569.9**	**173311.7**	**173920.5**	0.012	8.92	0.339
	Q3[b]	93	−86570.9	173327.8	173986.1	0.012	8.16	0.514
KS3	Q1[a]	47	−85552.4	171198.8	**171529.2**	0.015	44.05	<0.001
	Q2[c]	94	**−85370.6**	**170929.2**	171590.0	0.011	5.84	1
	Q3[b]	101	−85371.6	170945.2	171655.3	0.010	6.00	1

Different letter superscripts in column 2 indicate significant model fit improvement between adjacent Q-matrices by likelihood ratio test within each booklet. The best relative fit results within each booklet are shown in bold.

TABLE 4 | Final Q-Matrices.

Item	Booklet KS1						Booklet KS2						Booklet KS3					
	α_1	α_2	α_3	α_4	α_5	α_6	α_1	α_2	α_3	α_4	α_5	α_6	α_1	α_2	α_3	α_4	α_5	α_6
1	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1
2	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
3	1	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	1	0
4	0	1	0	0	0	1	1	0	0	0	1	0	1	0	0	0	1	0
5	0	0	1	0	1	0	1	0	0	0	0	1	0	0	1	0	1	0
6	0	1	0	1*	1	0	0	0	1	0	1	0	0	0	1	0	1	0
7	0	0	1	0	1	0	0	0	1	0	1	0	1	1	0	0	0	1
8	0	0	1	0	1	0	0	0	0	1	1	0	0	1	0	0	0	1
9	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	1	0
10	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	1	1	0
11	0	0	1	0	1	0	1	0	0	0	1	0	0	0	0	1	1	0
12	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0
13	0	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	1	0
14	0	1	0	0	0	1	0	0	0	1	1	0	0	1	0	0	1	0
15	1*	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	1	0
16	0	1	0	0	1	0	0	0	0	1	1	0	1	0	0	0	1	0
Total	5	4	4	5	13	3	4	4	4	4	13	3	4	5	4	5	13	3

*α_1 = Retrieving information; α_2 = Making inferences; α_3 = Integration and summation; α_4 = Reflective evaluation; α_5 = Literary text; α_6 = Practical text. The Q-matrix modifications were denoted by *.*

TABLE 5 | Model fit comparison of CDMs using the final Q-matrices.

Booklet	CDM	Npars	Relative fit			Absolute fit		
			-2LL	AIC	BIC	SRMSR	max χ^2	p (max χ^2)
	DINA[a]	54	−38147.9	76403.8	76747.0	0.042	269.83	<0.001
	DINO[b]	54	−38101.0	76310.0	76653.2	0.040	225.18	<0.001
	RRUM[c]	72	−37600.7	75345.3	75802.9	0.019	11.56	0.08
	A-CDM[c]	72	−37592.3	75328.6	**75786.1**	0.017	10.51	0.14
KS1	G-DINA[d]	94	**−37537.9**	**75263.8**	75861.2	0.015	7.45	0.76
	DINA[a]	54	−86967.4	174042.8	174425.0	0.020	24.75	<0.001
	DINO[a]	54	−86992.3	174092.6	174474.8	0.021	24.27	<0.001
	RRUM[b]	70	−86653.9	173447.8	173943.3	0.013	13.97	0.02
	A-CDM[b]	70	−86675.2	173490.5	173986.0	0.018	39.49	<0.001
KS2	G-DINA[c]	86	**−86569.9**	**173311.7**	**173920.5**	0.012	8.92	0.339
	DINA[a]	54	−85647.9	171403.7	171783.4	0.017	21.80	<0.001
	DINO[b]	54	−85641.5	171391.0	171770.7	0.017	26.65	<0.001
	RRUM[c]	72	−85520.2	171184.3	171690.5	0.014	10.51	0.14
	A-CDM[d]	72	−85484.2	171112.3	**171618.5**	0.012	9.38	0.26
KS3	G-DINA[e]	94	**−85424.1**	**171036.2**	171697.0	0.011	8.59	0.41

Different letter superscripts in column 2 indicate significant model-fit improvement among CDMs by likelihood ratio test within each booklet. The best relative fit results within each booklet are shown in bold.

diagnostic capacities of test items (Roussos et al., 2007; von Davier and Lee, 2019). **Figure 2** shows the item correct proportions for the masters versus the non-masters. The average item proportion correct difference was 0.53, and the differences in 41 out of the 48 items were greater than 0.40. This high value indicates a good fit between models and data, suggesting a strong diagnostic power of items and the DCRCA. In addition, this provided a valuable tool for finding poor items. For example, the differences of items 5 and 9 in booklet KS2 were smaller than 0.30. An in-depth examination suggested that these items were difficult; therefore, the item proportion correct for masters tended to be close to that for non-masters.

To further verify the external validity, the correlations between the scores on the DCRCA and the Chinese word recognition test were calculated. Word recognition scores were positively correlated with reading scores [KS1, r (4251) = 0.69, $p < 0.001$, KS2, r (8863) = 0.65, $p < 0.001$, KS3, r (8352) = 0.57, $p < 0.001$]. To summarize, the results suggested that the reliability and validity of the DCRCA were satisfactory.

Skill Profiles

CDMs classify test-takers into latent classes, which represent skill mastery/non-mastery profiles for attributes specified in the Q-matrix. With the six-attribute Q-matrix structure, 64 theoretically existing latent classes (2^k) were identified. For

TABLE 6 | Mastery classification reliability.

Attributes	Booklet KS1		Booklet KS2		Booklet KS3	
	Pa	Pc	Pa	Pc	Pa	Pc
α_1	0.86	0.81	0.93	0.90	0.71	0.63
α_2	0.90	0.86	0.89	0.88	0.80	0.78
α_3	0.90	0.86	0.93	0.90	0.70	0.63
α_4	0.91	0.88	0.92	0.87	0.68	0.63
α_5	0.95	0.92	0.83	0.74	0.86	0.82
α_6	0.86	0.85	0.87	0.88	0.77	0.85

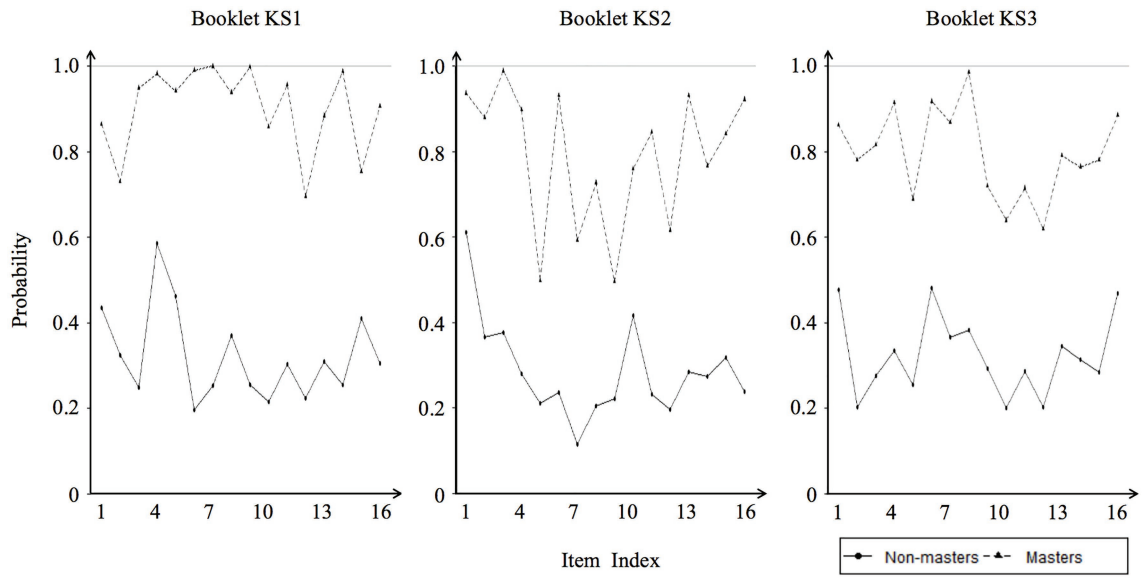


FIGURE 2 | Item mastery plots.

space considerations, only 15 skill profiles of the grade 2 students are presented in **Table 7**, as 49 classes showed lower posterior probabilities than 0.1%, suggesting that these skill classes may not be relevant to the data. Among the remaining 15 classes, the latent class [111111], mastery of all the subskills, had the highest posterior probability, followed by [000000], mastery of none of the subskills. CDM revealed that other dominant latent classes were [000011] and [111100], to which 27.15% of the test-takers belong. The profile [000011] might reflect children's knowledge and experiences in reading specific text genres in the given items, while the profile [111100] might reflect children's skills and experiences in answering specific reading tasks. This result supported the RAND report (RAND Reading Study Group, 2002) that mastery of the first four cognitive attributes and the last two text attributes may be relatively independent sources of variance in different reading comprehension scores.

TABLE 7 | Latent classes and posterior probabilities.

#	Latent class	Posterior probability (%)	#	Latent class	Posterior probability (%)
1	111111	32.84	9	111000	0.56
2	000000	26.14	10	011111	0.52
3	000011	19.85	11	101100	0.45
4	111100	7.29	12	111010	0.40
5	100000	4.47	13	011011	0.26
6	010011	2.68	14	110000	0.21
7	110011	2.21	15	101000	0.16
8	111011	1.48		Total	99.53

DISCUSSION AND CONCLUSION

This study developed and validated an instrument for diagnosing the strengths and weaknesses of Chinese reading comprehension ability at the primary level. Due to the criticism about a lack of true CDA research for educational purposes, the DCRCA was designed to meet the requirements of the Chinese curriculum criterion under the CDA framework proposed by Ravand and Baghaei (2020). Multiple steps were applied to maximize the diagnostic capacity and effectiveness of the DCRCA, including (1) gathering information about previous reading models and assessments; (2) specifying attribute lists based on the literature, student think-aloud protocols and expert review; (3) standardized test development and pilots; (4) empirical comparisons and refinements of Q-matrices and CDMs; and (5) reliability and

validity analyses using the formal test data. The results indicate that the overall quality of the DCRCA is satisfactory and that the diagnostic classifications are reliable, accurate, and valid.

Following multiple procedures of attribute specification, model-data fit comparison, and empirical validation, the Q-matrix construction results yielded six final reading attributes, including four cognitive attributes that are consistent with cognitive processing and previous empirical studies of reading and two text-related attributes that were synthesized from large-scale assessment frameworks and the Chinese curricular criterion. Adding text-related attributes significantly improved the model-data fits of Q-matrices, implying that pragmatic or background knowledge of different text types might be vital in successful reading. The literacy text attribute is consistent with previous research, while the practical text attribute is a newly extracted attribute in CDM studies on reading. Our attempts to combine expository text with discontinuous text attributes may reveal their similarity in reading strategies and worth further investigation. The validation of text-related attributes also improved the application value and scope of the DCRCA because these attributes come from the experiences of educators and thus might be easier to recognize and train (Perfetti et al., 2005). Besides, the six-attribute structure has been scrutinized as a theoretical framework of reading comprehension for students at different developmental stages. This result provides evidence regarding the construct of primary-level Chinese reading and the DCRCA from theoretical and empirical perspectives.

The selection of the CDMs is critical in all CDA studies, as the optimal model not only caters to the diagnostic demands of the assessment but also reveals the interrelationships of attributes in the given domain. Five representative CDMs were compared, and the superiority of the G-DINA model was supported by all booklets and model-data fit. Therefore, it is safe to analyze the DCRCA with the saturated G-DINA model, which appeared to be flexible in accommodating various relationships among reading skills (Chen and Chen, 2016; Li et al., 2016; Ravand, 2016). The A-CDM model performed the closest level of fit indices to the G-DINA model. From a theoretical perspective, the A-CDM model could be a special case of the G-DINA model by only estimating the main effects of attributes, as the difference between the two models is that G-DINA allows additional estimation of interactions among latent skills (de la Torre, 2011). Therefore, given that the majority of the DCRCA items were designed to map one of the cognitive processes and one text type of reading, our findings support Stanovich's (1980) interactive view of reading that holds both cognitive processes and text-related attributes to be crucial and interactive in successful execution of reading comprehension.

In addition, our results showed that the absolute fit indices preferred neither compensatory (A-CDM and DINO) nor non-compensatory (R-RUM and DINA) types of CDM, and max χ^2 rejected all the reduced models in booklet KS2. Consequently, current results are not enough to assert that the relationship of reading attributes is either compensatory or non-compensatory. This is consistent with the findings of Jang (2009), Li et al. (2016), and Javidanmehr and Sarab (2019), who also voted for the co-existence of compensatory and non-compensatory relationships among the latent reading subcomponents.

The present study examined the diagnostic reliability and validity of the DCRCA. Reliability evidence is generally considered essential support for interpreting test results. The pattern accuracy and consistency index (Cui et al., 2012) suggested that the DCRCA reliably measures multiple reading attributes. Validity analyses are rarely conducted, with less than 22% of studies providing such information according to the literature review (Sessoms and Henson, 2018). Therefore, construct, internal, and external validities are provided for the Q-matrix and the DCRCA. The Q-matrix validation results suggest that the provisional Q-matrices have an approximately 95% attribute-wise agreement rate across booklets, which provides strong evidence for the construct validity of Q-matrix constructions (Deonovic et al., 2019). The internal validity evidence showed that the average proportion correct differences for each item were sufficiently large for most of the test items, indicating that these items have satisfying diagnostic capacity to differentiate masters from non-masters of reading. The mean score differences of only 4% of the items were less than 0.3, much lower than the proportion of 23% in retrofitted studies (Jang, 2009). This might be because retrofitting studies had to include many items that were weakly associated with targeted attributes. The possible presence of nondiagnostic items could lead to critical issues in the validity of measures of skill competencies, and thus, the test inferences might be limited.

The present study contributes to instructional practices at the elementary school level, as the assessment can provide reliable, valid and useful diagnostic information. This is the first empirical study that attempts to provide evidence in construct invariance of diagnosing Chinese reading attributes at different primary grades. As reading assessment can function as formative assessment, such diagnostic feedback could be further utilized by teachers and educators for monitoring learning progressions, providing remedial instructions for reading courses and programs. However, some limitations are also worth enumerating. First, the present research did not examine how diagnostic feedback is perceived and utilized by students and teachers in a classroom setting. More studies are needed to reveal the influences of CDA applications. Second, as the DCRCA was not equated vertically, the attribute mastery states can be compared only within each key stage. Future studies are needed to apply appropriate longitudinal CDMs (Zhan et al., 2019) or vertical equating methods (von Davier et al., 2008) for CDA to investigate the developmental course of students' reading attributes. Third, the present study did not include a sufficient number of items to assess attributes a_{6a} and a_{6b}. The Q-matrices may not be exhaustive to capture all reading comprehension and likely lead to limitations of the present study. Therefore, caution should be taken in interpreting our final results, and explorations of a more balanced Q-matrix construction are needed in the future. Last, although the results related to model fit and item parameters were fairly acceptable, future research should seek to improve the psychometric properties to make the report inferences more reliable. Therefore, the study was only a start. A deeper understanding of CDM application may be deduced by interpreting the dominant skill classes as learning states and the combination of skill classes as

learning paths and learning progressions (Wu et al., 2020). Future studies are needed to help instructors design suitable learning plans with fine-grained diagnostic reports of students. In addition, more well-designed items can be generated and scaled as formative and summative assessments to satisfy expectations from the curriculum criterion. With the help of the DCRCA, teachers could design their own classroom reading materials and assessments as learning objectives that they wish students to attain.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics and Human Safety Committee, Faculty of Psychology, Beijing Normal University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

YL and JL conceived the study. YL and MZ organized the pilots and analyzed the original data. MZ developed the test items and conducted the think-aloud protocols. YL collected the formal test, analyzed the data, and wrote the manuscript. JL provided technical advices. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (31861143039) and the National Key R&D Program of China (2019YFA0709503).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.786612/full#supplementary-material

REFERENCES

Aaron, P. G., Joshi, R. M., Gooden, R., and Bentum, K. E. (2008). Diagnosis and treatment of Reading disabilities based on the component model of reading: an alternative to the discrepancy model of LD. *J. Learn. Disabil.* 41, 67–84. doi: 10.1177/0022219407310838

Alderson, J. C. (2010). "Cognitive diagnosis and Q-matrices in language assessment": a commentary. *Lang. Assess. Q.* 7, 96–103. doi: 10.1080/15434300903426748

Barnes, M. A. (2015). "What do models of Reading comprehension and its development have to contribute to a science of comprehension instruction and assessment for adolescents?" in *Improving Reading Comprehension of Middle and High School Students*. eds. K. L. Santi and D. K. Reed (Cham: Springer International Publishing), 1–18.

Brennan, R. L. (2006). *Educational Measurement. 4th Edn.* Rowman & Littlefield Publishers.

Cain, K. (2009). *Making Sense of Text: Skills that Support Text Comprehension and Its Development, Perspectives on Language and Literacy*. Springer.

Cain, K., Oakhill, J., and Bryant, P. (2004). Children's Reading comprehension ability: concurrent prediction by working memory, verbal ability, and component skills. *J. Educ. Psychol.* 96, 31–42. doi: 10.1037/0022-0663.96.1.31

Carlson, S. E., Seipel, B., and McMaster, K. (2014). Development of a new reading comprehension assessment: identifying comprehension differences among readers. *Learn. Individ. Differ.* 32, 40–53. doi: 10.1016/j.lindif.2014.03.003

Chen, H., and Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment Under the generalized DINA model framework. *Lang. Assess. Q.* 13, 218–230. doi: 10.1080/15434303.2016.1210610

Chen, J., and de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: the case of the programme for international student assessment in Reading. *Psychology* 05:1967. doi: 10.4236/psych.2014.518200

Chen, M. J., Lau, L. L., and Yung, Y. F. (1993). Development of component skills in reading Chinese. *Int. J. Psychol.* 28, 481–507. doi: 10.1080/00207599308246933

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

Collins, A. A., Compton, D. L., Lindström, E. R., and Gilbert, J. K. (2020). Performance variations across reading comprehension assessments: examining the unique contributions of text, activity, and reader. *Read. Writ.* 33, 605–634. doi: 10.1007/s11145-019-09972-5

Common Core State Standards Initiative (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. In Common Core State Standards Initiative.

Compulsory Education Curriculum and Textbook Committee of the Ministry of Education (2012). *Interpretation of Chinese language curriculum standard for compulsory education (2011 Edition)* (in Chinese). Beijing: Higher Education Press.

Compton, D. L., and Pearson, P. D. (2016). Identifying robust variations associated with reading comprehension skill: the search for pressure points. *J. Res. Educ. Effect* 9, 223–231. doi: 10.1080/19345747.2016.1149007

Cui, Y., Gierl, M. J., and Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Educ. Meas.* 49, 19–38. doi: 10.1111/j.1745-3984.2011.00158.x

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 Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika* 81, 253–273. doi: 10.1007/s11336-015-9467-8

de la Torre, J., and Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

de La Torre, J., and Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *J. Educ. Meas.* 47, 115–127. doi: 10.1111/j.1745-3984.2009.00102.x

Deonovic, B., Chopade, P., Yudelson, M., de la Torre, J., and von Davier, A. A. (2019). "Application of cognitive diagnostic models to learning and assessment systems," in *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. eds. M. von Davier and Y.-S. Lee. (Springer, Cham: Springer International Publishing).

George, A. C., and Robitzsch, A. (2015). Cognitive diagnosis models in R: a didactic. *Quant. Meth. Psych.* 11, 189–205. doi: 10.20982/tqmp.11.3.p189

George, A. C., and Robitzsch, A. (2021). Validating theoretical assumptions about reading with cognitive diagnosis models. *Int. J. Test.* 21, 105–129. doi: 10.1080/15305058.2021.1931238

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., and Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *J. Stat. Softw.* 74, 1–24. doi: 10.18637/jss.v074.i02

Gierl, M. J., Alves, C., and Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: an operational implementation of cognitive diagnostic assessment. *Int. J. Test.* 10, 318–341. doi: 10.1080/15305058.2010.509554

Gierl, M. J., and Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement* 6, 263–268. doi: 10.1080/15366360802497762

Grabe, W. (2009). *Reading in a Second Language: Moving From Theory to Practice*. Cambridge University Press; Cambridge Core.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26, 301–321. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hartz, S. M. (2002). A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality. [doctoral dissertation]. University of Illinois at Urbana-Champaign.

Jang, E. E. (2005). A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL. [doctoral dissertation]. University of Illinois at Urbana-Champaign.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: validity arguments for fusion model application to LanguEdge assessment. *Lang. Test.* 26, 031–073. doi: 10.1177/0265532208097336

Javidanmehr, Z., and Sarab, M. R. A. (2019). Retrofitting non-diagnostic Reading comprehension assessment: application of the G-DINA model to a high stakes reading comprehension test. *Lang. Assess. Q.* 16, 294–311. doi: 10.1080/15434303.2019.1654479

Kim, Y.-S. G. (2017). Why the simple view of Reading is not simplistic: unpacking component skills of reading using a direct and indirect effect model of Reading (DIER). *Sci. Stud. Read.* 21, 310–333. doi: 10.1080/10888438.2017.1291643

Kim, Y.-S. G., and Wagner, R. K. (2015). Text (Oral) Reading fluency as a construct in Reading development: an investigation of its mediating role for children from grades 1 to 4. *Sci. Stud. Read.* 19, 224–242. doi: 10.1080/10888438.2015.1007375

Kintsch, W. (1991). "The role of knowledge in discourse comprehension: a construction-integration model," in *Advances in Psychology*. *Vol. 79*. eds. G. E. Stelmach and P. A. Vroon (North-Holland), 107–153.

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *J. Educ. Meas.* 49, 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Lei, X. (2020). On the Issues of Educational Evaluation Reform in China. China Examinations, No. 341(09), 13–17.

Lei, P.-W., and Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Appl. Psychol. Meas.* 40, 405–417. doi: 10.1177/0146621616647954

Leighton, J. P., and Gierl, M. J. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications (1st Version)*. Cambridge University Press.

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* 41, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Li, H. (2011). Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach. [doctoral dissertation]. The Pennsylvania State University. Available at: http://www.pqdtcn.com/thesisDetails/99641D9398A8DB6ACEAE84A5915F5CBF (Accessed September 20, 2021).

Li, H., Hunter, C. V., and Lei, P. -W. (2016). he selection of cognitive diagnostic models for a reading comprehension test. *Language Testing.* 33, 391–409. doi: 10.1z177/0265532215590848

Li, H., Shu, H., McBride-Chang, C., Liu, H., and Peng, H. (2012). Chinese children's character recognition: visuo-orthographic, phonological processing and morphological skills. *J. Res. Read.* 35, 287–307. doi: 10.1111/j.1467-9817.2010.01460.x

Liu, R., Huggins-Manley, A. C., and Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educ. Psychol. Meas.* 78, 357–383. doi: 10.1177/0013164416685599

Liu, M., Li, Y., Wang, X., Gan, L., and Li, H. (2021). Leveled Reading for primary students: construction and evaluation of Chinese readability formulas based on textbooks. *Appl. Linguis.* 2, 116–126. doi: 10.16499/j.cnki.1003-5397.2021.02.010

Ma, W., and de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis Modeling. *J. Stat. Softw.* 93, 1–26. doi: 10.18637/jss.v093.i14

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement* 11, 71–101. doi: 10.1080/15366367.2013.831680

Ministry of Education (2011). *The Chinese Language Curriculum Criterion for Compulsory Education*. *2011th Edn*. Beijing Normal University Press.

Mo, L. (1992). Study on the characteristics of the development Of Chinese reading ability structure of middle and primary school students. *Acta Psychol. Sin.* 24, 12–20.

O'Reilly, T., and Sheehan, K. M. (2009). Cognitively based assessment of, for, and as learning: a framework for assessing reading competency. *ETS Res. Rep. Ser.* 2009, i–43. doi: 10.1002/j.2333-8504.2009.tb02183.x

Perfetti, C. A., Landi, N., and Oakhill, J. (2005). "The Acquisition of Reading Comprehension Skill," in *The Science of Reading: A Handbook*. eds. M. J. Snowling and C. Hulme (Blackwell Publishing)

R Core Team (2021). R: The R Project for Statistical Computing [Computer software]. Available at: https://www.r-project.org/ (Accessed August 31, 2021).

RAND Reading Study Group (2002). *Reading for Understanding: Toward an R&D Program in Reading Comprehension*. Santa Monica, CA: RAND.

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

Ravand, H., and Baghaei, P. (2020). Diagnostic classification models: recent developments, practical issues, and prospects. *Int. J. Test.* 20, 24–56. doi: 10.1080/15305058.2019.1588278

Ravand, H., and Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educ. Psychol.* 38, 1255–1277. doi: 10.1080/01443410.2018.1489524

Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *J. Stat. Softw.* 17, 1–25. doi: 10.18637/jss.v017.i05

Robitzsch, A., and George, A. C. (2019). "The R package CDM for diagnostic Modeling," in *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages,* 549–572. Cham: Springer International Publishing.

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 Assessment for Education: Theory and Applications*. eds. J. Leighton and M. Gierl (Cambridge Core: Cambridge University Press), 275–318.

Rupp, A. A., and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Measurement* 6, 219–262. doi: 10.1080/15366360802490866

Sawaki, Y., Kim, H.-J., and Gentile, C. (2009). Q-matrix construction: defining the link Between constructs and test items in large-scale reading and listening comprehension assessments. *Lang. Assess. Q.* 6, 190–209. doi: 10.1080/15434300902801917

Sessoms, J., and Henson, R. A. (2018). Applications of diagnostic classification models: a literature review and critical commentary. *Measurement* 16, 1–17. doi: 10.1080/15366367.2018.1435104

Shu, H., Chen, X., Anderson, R. C., Wu, N., and Xuan, Y. (2003). Properties of school Chinese: implications for learning to read. *Child Dev.* 74, 27–47. doi: 10.1111/1467-8624.00519

Snow, C. (2002). Reading for Understanding: Toward an R&D Program in Reading Comprehension. RAND Corporation. Available at: https://www.rand.org/pubs/monograph_reports/MR1465.html (Accessed September 20, 2021).

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Read. Res. Q.* 16, 32–71. doi: 10.2307/747348

Tatsuoka, K. K. (1990). "Toward an integration of item-response theory and cognitive error diagnosis," in *Diagnostic Monitoring of Skill and Knowledge Acquisition*. eds. N. Frederiksen, R. Glaser, A. Lesgold and M. G. Shafto (Hillsdale, NJ: Erlbaum), 453–488.

Templin, J., and Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *J. Classif.* 30, 251–275. doi: 10.1007/s00357-013-9129-4

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

Toprak, T. E., and Cakir, A. (2021). Examining the L2 reading comprehension ability of adult ELLs: developing a diagnostic test within the cognitive diagnostic assessment framework. *Lang. Test.* 38, 106–131. doi: 10.1177/0265532220941470

Toprak-Yildiz, T. E. (2021). An international comparison using cognitive diagnostic assessment: fourth graders' diagnostic profile of reading skills on PIRLS 2016. *Stud. Educ. Eval.* 70:101057. doi: 10.1016/j.stueduc.2021.101057

van Dijk, T. A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

von Davier, A., Carstensen, C. H., and von Davier, M. (2008). "Linking competencies in horizontal, vertical, and longitudinal settings and measuring growth," in *Assessment of Competencies in Educational Contexts*. eds. J. Hartig, E. Klieme and D. Leutner (New York: Hogrefe & Huber), 121–149.

von Davier, M., and Lee, Y.-S. (eds.) (2019). *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. Springer International Publishing. Cham.

Wu, X., Wu, R., Chang, H.-H., Kong, Q., and Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Front. Psychol.* 11:2230. doi: 10.3389/fpsyg.2020.02230

Xie, M. (2014). The cognitive diagnostic assessment of junior high school students' reading comprehension ability of modern Chinese Prose. [doctoral dissertation]. Jiangxi Normal University. Available at: http://cdmd.cnki.com.cn/Article/CDMD-10414-1015402211.htm (Accessed September 20, 2021).

Yun, J. (2017). Investigating Structures of Reading Comprehension Attributes at Different Proficiency Levels: Applying Cognitive Diagnosis Models and Factor Analyses. [doctoral dissertation]. The Florida State University. Available at: https://www.proquest.com/docview/2017341448/EFEAC191350D4116PQ/14?accountid=8554 (Accessed September 20, 2021).

Zhan, P., Jiao, H., Liao, D., and Li, F. (2019). A longitudinal higher-order diagnostic classification model. *J. Educ. Behav. Stat.* 44, 251–281. doi: 10.3102/1076998619827593

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 Li, Zhen 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

TECHNOLOGY AND CODE
published: 11 February 2022
doi: 10.3389/fpsyg.2022.786772

Two Test Assembly Methods With Two Statistical Targets

*Zheng Huijing, Li Junjie, Zeng Pingfei and Kang Chunhua**

Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China

In educational measurement, exploring the method of generating multiple high-quality parallel tests has become a research hotspot. One purpose of this research is to construct parallel forms item by item according to a seed test, using two proposed item selection heuristic methods [minimum parameters–information–distance method (MPID) and minimum information–parameters–distance method (MIPD)]. Moreover, previous research addressing test assembly issues has been limited mainly to situations in which the information curve of the item pool or seed test has a normal or skewed distribution. However, in practice, the distributions of information curves for tests are diverse. These include multimodal distributions, the most common type of which is the bimodal distribution. Therefore, another main aim of this article is to extend the information curves of unimodal distributions to bimodal distributions. Thus, this study adopts simulation research to compare the results of two item, response, theory (IRT)-based item matching methods (MPID and MIPD) using different information curve distributions for item pools or seed tests. The results show that the MPID and MIPD methods yield rather good performance in terms of both two statistical targets when the information curve has a unimodal distribution, and two new methods yield better performance than two existing methods in terms of test information functions target when the information curve has a bimodal distribution.

Keywords: bimodal distribution, item matching test assembly methods, item response theory, information curve, parallel forms of tests

OPEN ACCESS

Edited by:
Tao Xin,
Beijing Normal University, China

Reviewed by:
Tour Liu,
Tianjin Normal University, China
Yu Bao,
James Madison University, United States

***Correspondence:**
Kang Chunhua
akang@zjnu.cn

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

Received: *30 September 2021*
Accepted: *10 January 2022*
Published: *11 February 2022*

Citation:
Huijing Z, Junjie L, Pingfei Z and Chunhua K (2022) Two Test Assembly Methods With Two Statistical Targets. Front. Psychol. 13:786772. doi: 10.3389/fpsyg.2022.786772

INTRODUCTION

Constructing multiple equivalent forms with higher quality to be administered at different timepoints and locations has always posed a challenge for developers of educational assessments and licensure tests. The application of automated test assembly (ATA) procedures benefits test developers in that it dramatically reduces their workload and ensures the quality of parallel test forms. Over the past two decades, researchers have successfully implemented optimization-based automated test assembly techniques such as mixed integer programming (MIP; Cor et al., 2009; Finkelman et al., 2010) and enumerative heuristics (Armstrong et al., 1992; Finkelman et al., 2009; Brusco et al., 2013).

The MIP methods convert test specifications (the test blueprint) into mathematical expressions from which a globally optimal solution can usually be derived using available software packages (Chen, 2014). Heuristics methods following stepwise procedures are of great influence even though they often yield a locally optimal solution at each step, not a globally optimal one (Chen, 2014).

Because of the nondeterministic polynomial (NP)-hard nature of MIP problems, heuristic methods can improve both the performance of MIP solvers and the quality of solutions (Chen, 2015).

Mixed integer programming approaches look for the optimal solution, so time is longer. Besides, many solvers are commercially available and costly. For users with a weak mathematical background, MIP approaches are not easily accessible (Chen, 2014). Heuristic methods avoid the above shortcomings. Although heuristic methods find the suboptimal solution, the suboptimal solution is acceptable for test assembly, so this article focuses on heuristic methods. There are many heuristics (Armstrong et al., 1992; Finkelman et al., 2009; Brusco et al., 2013), but most of them like greedy algorithms, random and sampling algorithms are relatively old algorithms, which are difficult to meet today's demand for test papers with diverse constraints. With the development of test theory, the trend of test assembly is to assemble high-quality test papers that meet the constraints under the test theory framework based on seed test. Minimum information distance method (MID) and minimum parameters distance method (MPD) are two classical test assembly methods based on seed test under item response theory.

When the seed test is available, one of the targets of test assembly is to make test information curve of generated tests similar to test information curve of the seed test, because an important indicator for testing whether two tests are parallel tests, is the similarity of test information curves of the two tests (Ali and Van Rijn, 2016). The more similar they are, the more they can be regarded as parallel tests. The core idea of the MID method is to match item information curve item by item, so that the test information curve of the seed test and generated tests will be identical (Armstrong et al., 1992). Another indicator is the test characteristic curve of two tests (Ali and Van Rijn, 2016). Similar test information does not necessarily guarantee that the test characteristic curve is the same. So the advantage of MID is that the generated tests are similar to the seed test in terms of test information curve, but the disadvantage is that the test characteristic curve is not necessarily similar.

In order to meet both the two indicators, Armstrong et al. (1992) have attempted to use the MPD method for directly matching the item's parameters, because test information function and test characteristic function are both functions composed of some parameters, which will inevitably be decided by parameters.

In general, MID only focuses on test information curve, while MPD has a wide range of influence. It can be inferred that MID is better than MPD on test information curve matching target, while MPD is better than MID on test characteristic curve matching target (Armstrong et al., 1992). On the basis of MID and MID, can new test assembly methods be produced to make both test information curve and test characteristic curve matching targets achieve more satisfactorily?

Moreover, the majority of previous research addressing test assembly problems has focused on the condition when the information curve of the item pool or the reference test has a unimodal distribution by default (Chen et al., 2012; Chen, 2014, 2015; Ali and Van Rijn, 2016; Shao et al., 2019). However, information curves vary greatly in practice, and they include both unimodal and multimodal distributions. The bimodal distribution is a simple and typical representative of the multimodal distribution. Accordingly, the present study explores both unimodal and bimodal distributions. In sum, this study's goal is to develop two novel item, response, theory (IRT)-based item matching test assembly methods based on the two previously mentioned methods and then compare the four, using different information curve distributions for the item pools and seed tests.

The article is organized as follows. First, we briefly review two extant item matching test assembly methods (the minimum parameters–distance method and the minimum information–distance method), explaining their limitations and proposing two new methods. Subsequently, we introduce information curves for unimodal and bimodal distributions. Finally, we compare the proposed methods with the two traditional item matching methods, using different information curve distributions for the item pools and seed tests based on several criteria.

TWO TRADITIONAL HEURISTIC METHODS

Minimum Information Distance Method

The idea of the MID method is to find one item in the item pool that is most similar with the item in the seed test in terms of item information curve. The figure given below (**Figure 1**) is the information curve of all the items in the item pool (gray curve) and an item in the seed test (red curve).

According to the image, it is hard to judge which item in the item pool has the closest information curve to the item in the seed test, so it is necessary to calculate the information curve distance (ID) between each item in the item pool and the item in the seed test to find the minimum ID (MID), and this is in line with the original intention of the MID method to assemble tests.

$$ID_{ij}^{2} = \sum_{m=1}^{M} \omega_m \left(f_i\left(\theta_m\right) - f_j\left(\theta_m\right)\right)^2$$

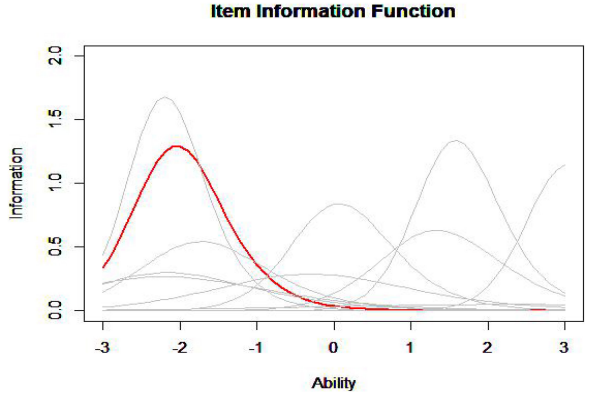


FIGURE 1 | Item information function.

The information distance is estimated by the following equation: where ID_{ij} is the information distance between item i (in the seed test) and item j (in the item pool). $f_i(\theta)$ is the information for item i at the ability level θ and $f_j(\theta)$ is the information for item j at the ability level θ. M is the number of the ability levels of interest, ω_m is weight coefficient and $\sum_{m=1}^{M} \omega_m = 1$, $\omega_m > 0$, ω_m is selected by the practitioner (Armstrong et al., 1992). The ability level θ selected in this study are −2, −1, 0, 1, 2, and the weight of each ability point is 0.2, the same for both.

The IRT model used in this study is a three-parameter logistic model, so the calculation formula for item information curve $f(\theta)$ is as follows:

$$f(\theta) = \frac{1.7^2 * a^2 * (1-c)}{\left(c + e^{1.7a(\theta-b)}\right)\left(1 + e^{-1.7a(\theta-b)}\right)^2}$$

where a, b and c are discrimination, difficulty, and guessing parameters of an item, respectively.

The core objective of the MID method is to minimize the differences in information statistics at crucial ability points between the assembled test forms and the seed test, which directly meets the criterion of test information matching.

Minimum Parameters Distance Method

In addition to matching the test information curve (TIC) of the seed test, matching the test characteristic curve (TCC) of the seed test is another important target of test assembly (Ali and Van Rijn, 2016). Constraining the test information curve to be equal does not necessarily guarantee similarity of the test characteristic curve (Ali and Van Rijn, 2016). It can be concluded that the MID method can only meet the matching requirements of TIC but cannot meet the matching requirements of TCC.

The IRT model used in this study is a three-parameter logistic model, so the calculation formula for item characteristic function is as follows:

$$P(\theta) = c + \frac{1-c}{1 + e^{[-1.7a(\theta-b)]}}$$

where a, b, and c are discrimination, difficulty, and guessing parameters of an item, respectively.

It can be concluded from the calculation formulas of item characteristic curve (ICC) and item information curve (IIC) that they are both functions of three parameters. The idea of the MPD method is to find one item in the item pool that is most similar with the item in the seed test in terms of item's parameters. Tests that match based on collective indices such as test, information, function (TIF) may not be presumed to exhibit stable, similar properties any more than can those based on item matching. Tests built by matching item parameters (MIP) directly capture the main properties of the items in the seed test, thereby ensuring the satisfaction of all cumulative indices, including TIFs and TCCs (Chen, 2015).

The figure below (**Figure 2**) is the item's parameters of all the items in the item pool (gray dot) and one item in the seed test (red dot).

According to the image, it is difficult to judge which item in the item pool has the closest item's parameters to the item in the seed test, so it is necessary to calculate the item's parameters distance (PD) between each item in the item pool and the item in the seed test to find the minimum PD (MPD), and this is in line with the original intention of the MPD method to assemble tests (Wang et al., 2016). The IRT model in this study is a commonly used three-parameters logistic model. The PD is estimated by the following equation:

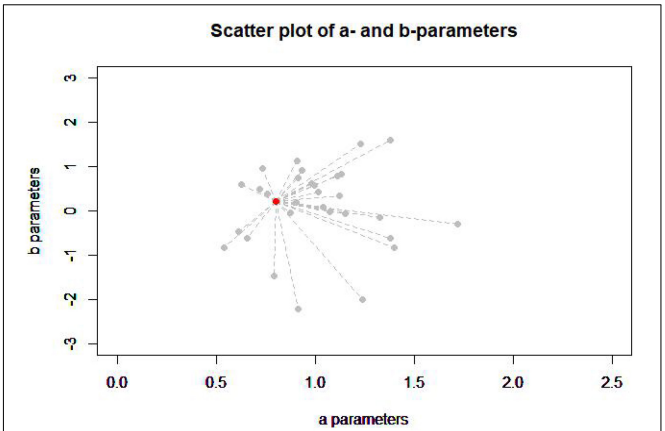


FIGURE 2 | Scatter plot of *a*- and *b*-parameters.

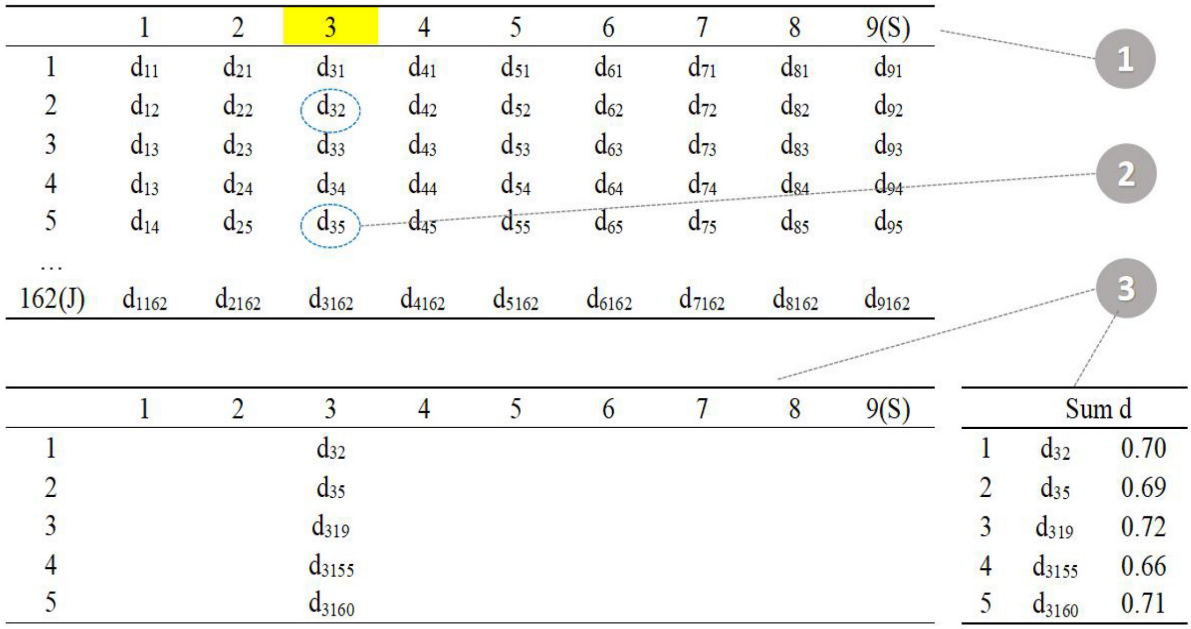
$$PD_{ij}^2 = \varphi_1 \left(a_i - a_j\right)^2 + \varphi_2 \left(b_i - b_j\right)^2 + \varphi_3 \left(c_i - c_j\right)^2$$

where PD_{ij} is the parameter's distance between item i and item j; a_i, b_i and c_i are the discrimination, difficulty and guessing parameters, respectively, of item i in the seed test; and a_j, b_j, and c_j are the discrimination, difficulty, and guessing parameters, respectively, of item j in the item pool.

φ_1, φ_2, and φ_3 are weight coefficient. $\varphi_1 \geq 0$, $\varphi_2 \geq 0$, $\varphi_3 \geq 0$ and $\varphi_1 + \varphi_2 + \varphi_3 = 1$. They are selected by the practitioner (Armstrong et al., 1992; Chen, 2015; Wang et al., 2021). Different parameters have different effects on the test information function and test characteristic function. Taking the three-parameters logistic model as an example, for test information function, the degree of discrimination and guessing parameters have a greater impact on it, while for test characteristic function, the degree of discrimination has the greatest influence, followed by the difficulty and guessing parameters. Therefore, when calculating the parameter distance, different weights are generally given to the parameters. Chen (2017) found that these weights ($\varphi_1 = 0.5$, $\varphi_2 = 0.25$, and $\varphi_3 = 0.25$) were used to represent the relative importance of a parameter to the information function after examination of the TIC and TCC resulting from the unweighted and weighted versions. The weights used in this study are the same.

Test Assembly Procedure

1. Randomly select an item in the seed test.
2. Choose five items (the number of items is determined by the number of parallel tests) from the item pool according to MID or MPD.



	1	2	3	4	5	6	7	8	9(S)
1	d_{11}	d_{21}	d_{31}	d_{41}	d_{51}	d_{61}	d_{71}	d_{81}	d_{91}
2	d_{12}	d_{22}	d_{32}	d_{42}	d_{52}	d_{62}	d_{72}	d_{82}	d_{92}
3	d_{13}	d_{23}	d_{33}	d_{43}	d_{53}	d_{63}	d_{73}	d_{83}	d_{93}
4	d_{13}	d_{24}	d_{34}	d_{44}	d_{54}	d_{64}	d_{74}	d_{84}	d_{94}
5	d_{14}	d_{25}	d_{35}	d_{45}	d_{55}	d_{65}	d_{75}	d_{85}	d_{95}
…									
162(J)	d_{1162}	d_{2162}	d_{3162}	d_{4162}	d_{5162}	d_{6162}	d_{7162}	d_{8162}	d_{9162}

	1	2	3	4	5	6	7	8	9(S)
1			d_{32}						
2			d_{35}						
3			d_{319}						
4			d_{3155}						
5			d_{3160}						

Sum d		
1	d_{32}	0.70
2	d_{35}	0.69
3	d_{319}	0.72
4	d_{3155}	0.66
5	d_{3160}	0.71

FIGURE 3 | Procedure 1–3.

3. Five items are randomly assigned to five parallel tests and calculate the sum of the distances between the selected items of the five parallel tests and the seed test.
4. Delete the selected item from the item pool to prevent repeated selection.
5. Randomly select another item in the seed test again and choose five items from the new item pool according to MID or MPD.
6. The five items are allocated to five parallel tests based on the sum of distances (procedure three). The principle is that the greater the sum of the distances of parallel paper, the more priority items with a smaller MID or MPD are to be assigned to it, so as to reduce the difference between parallel tests.
7. Repeat 4–6 until all the items in the seed test have been selected.

As shown in the figure above (**Figure 3**), the upper left corner is the distance matrix between the item in the item pool (row) and the item in the seed test (column); the lower left corner is the distance matrix of five parallel tests; the lower right corner calculates the sum of the current distances of each parallel test.

The first step is to randomly select one item in the seed test (item 3 in the seed test), the second step is to find the five items with the smallest *d* value in the item pool (item 2, item 5, item 19, item 155, and item 160 in the item pool), and the third step is randomly assigned to five parallel tests, and the total distance is calculated.

The fourth step is to randomly select one item in the seed test (item 6). The fifth step is to find the five items with the smallest value of d in the item pool (item 1, item 66, item 68, item 142, and item 149 in the item pool). The sixth step is to assign five items. The total distance calculated in the third step is allocated to the five parallel tests in reverse order (the smaller distance item is assigned to the test with larger total distance) to reduce the difference between parallel tests (**Figure 4**).

Repeat steps 3–6 until all items in the seed test have been selected (**Figure 5**).

Two New Heuristic Methods

The MID method aims to make generated tests similar to the seed test in terms of the test information curve, but fails to take test characteristic curve into account. MPD method of directly matching parameters expects to achieve two targets, but the result of test information curve is inferior to method MID. The two methods have their own strengths, so why not combine the two distances together to construct a new distance index to assemble tests?

Minimum Parameters–Information–Distance Method

In order to achieve the best result of TIF and TCC target matching, the two methods are combined when constructing the distance moment. The parameters–information–distance (PID) is estimated by the following equation:

$$PID_{ij}^{2} = (1 - \lambda) * (PD_{ij}) + \lambda * (ID_{ij})$$

$$\lambda = \frac{s - 1}{test_length}$$

where PID_{ij} is the combined distance between item i (in the seed test) and item j (in the item pool); PD_{ij} is the parameter's distance

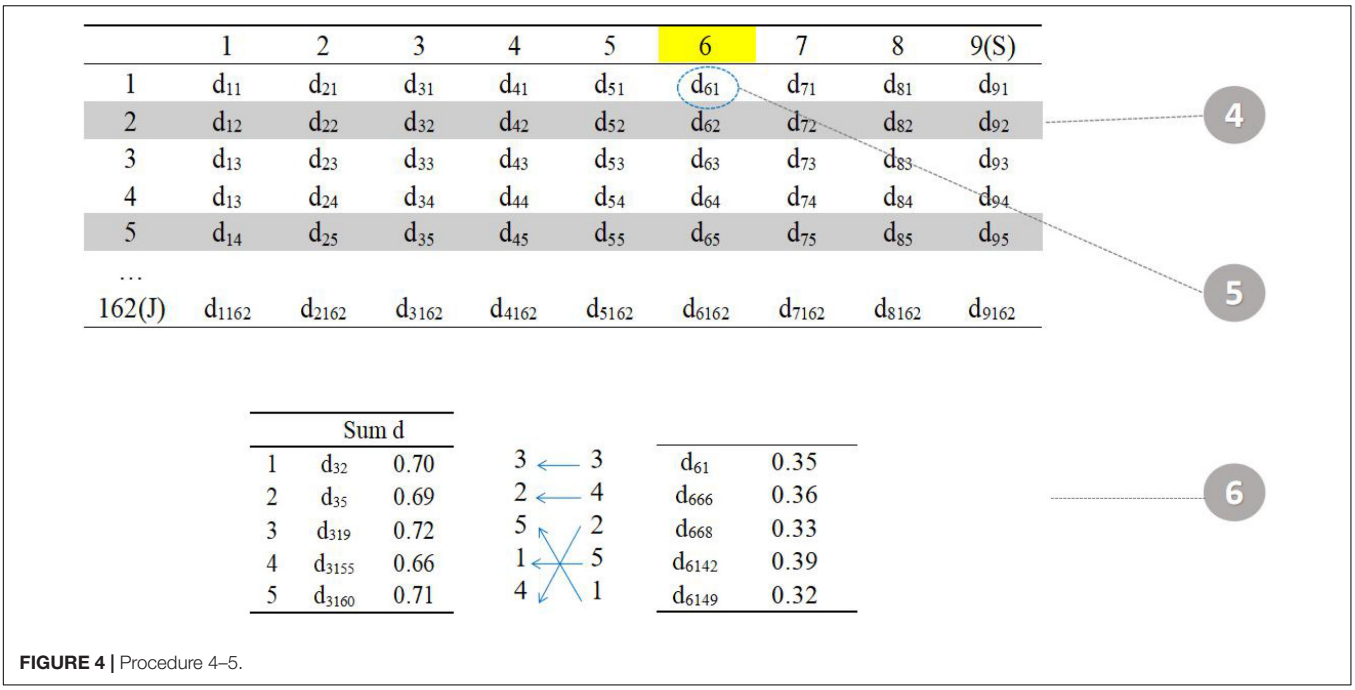


FIGURE 4 | Procedure 4–5.

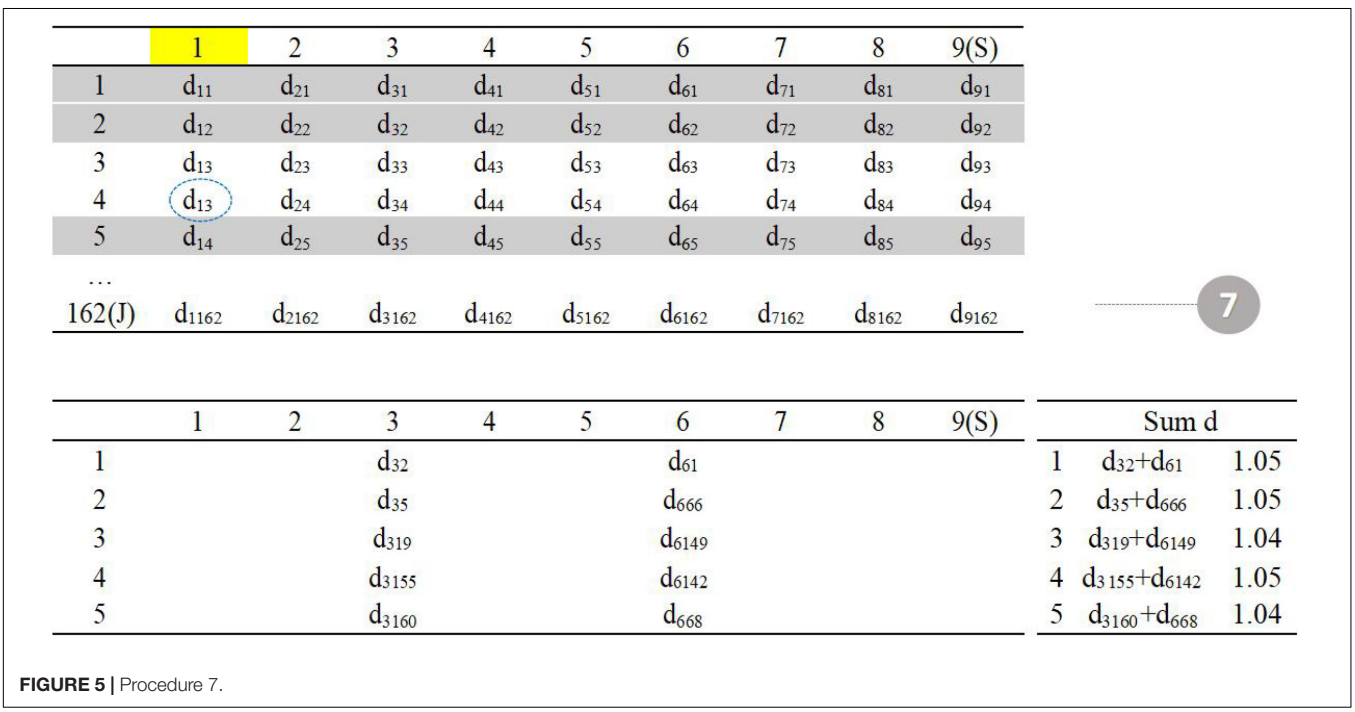


FIGURE 5 | Procedure 7.

between item i and item j; ID_{ij} is the information distance between item i and item j; λ is the adjustment factor, and s is the number of items that have been selected so far.

The progressive method of Revuelta and Ponsoda (1998) is used as a template for our holistic item selection index. The role of λ is to select and generate papers in the previous stage in order to highlight the advantages of the MPD method and match the TCC. As s increases, it becomes larger and 1-s decreases, highlighting the advantages of the MID method and matching the TIC at the later stage.

Minimum Information–Parameters–Distance Method

It remains unknown whether the two methods' sequence affects test assembly results. It is feasible to reverse the order, producing

a minimum information–parameters–distance (MIPD) method for meeting a variety of practical demands. At the first stage, the purpose of selecting items is to obtain smaller TIC differences, and during the next stage, the aim is to minimize parameters' distances. The procedure is exactly the opposite of the MPID method. The information, parameters, and distance (IPD) is estimated by the following equation:

$$IPD_{ij}^{2} = (1 - \lambda) * (ID_{ij}) + \lambda * (PD_{ij})$$

$$\lambda = \frac{s - 1}{test_length}$$

The meaning of the letters in the formula is the same as above.

Bimodally Distributed Test Information Curves

Bimodal distributions often appear in the fields of biology, life sciences, geology, and so on. For example, in a clinical context, the highest incidence of fibrolamellar hepatocellular carcinoma (FLC) occurs between ages 15 and 19 and between ages 70 and 74; that is, the curve representing the age of onset is bimodal (Ramai et al., 2021). Of course, bimodal distributions are not uncommon in the fields of psychology and pedagogy. Bimodal distributions appear in many psychological tests (Steinley and McDonald, 2007). This often occurs with education examinations. Tang (2018) has found that students' English subject test scores in each semester exhibit abnormal bimodal distributions (based on the Academic Quality Monitoring and Evaluation Department).

Different disciplines have different definitions of bimodal distributions. In this article, we are referring to a distribution showing two obvious peaks—that is, a mixed distribution composed of two unimodal distributions—where the two peaks need not be equal.

As we can see from the formula of the item information function, it is not surprising that the item information curve has one peak, such as that shown in **Figure 6**. An item measures the ability with greatest precision at the ability level corresponding to the item's difficulty parameter. The amount of item information decreases as the ability level departs from the item difficulty and approaches zero at the extremes of the ability scale.

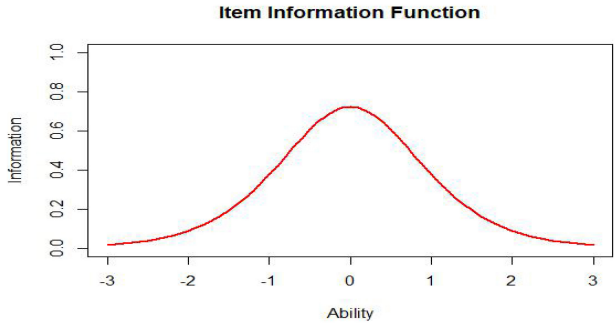


FIGURE 6 | Item information function (a = 1, b = 0).

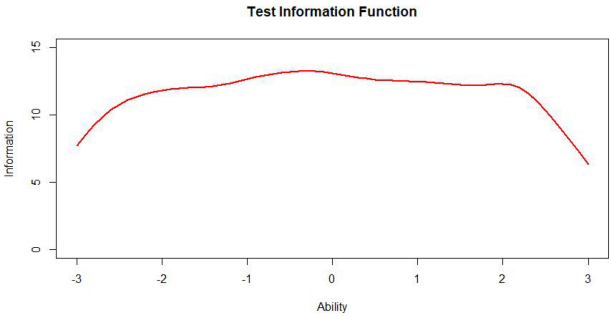


FIGURE 7 | Unimodally distributed test, information, function (TIF).

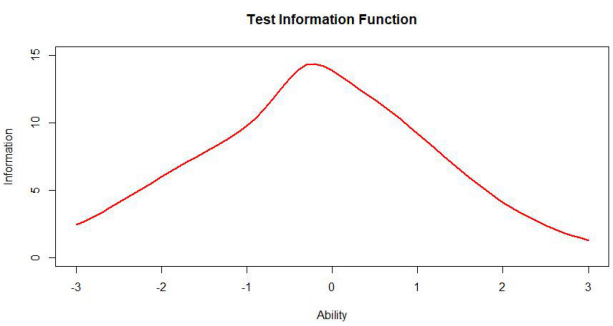


FIGURE 8 | Uniformly distributed test, information, function (TIF).

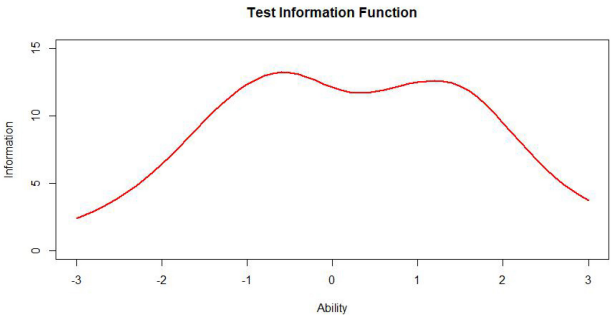


FIGURE 9 | Bimodally distributed test, information, function (TIF).

Because a test is used to estimate an examinee's ability, we can also obtain the amount of information yielded by the test at any ability level. A test is a set of items; therefore, the test information at a given ability level is simply the sum of the item information values at that level. Consequently, the test information function is defined as

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta)$$

where $I(\theta)$ is the amount of test information at ability level θ, $I_j(\theta)$ is the amount of information for item j at ability level θ, and J is the number of items in the test.

TABLE 1 | MSD_{TIC}.

	Unimodal (D = 0)	Unimodal (D = 1)	Bimodal (D = 2)	Bimodal (D = 3)
MPD	0.401	0.408	0.426	0.438
MID	0.302	0.331	0.363	0.406
MPID	0.296	0.315	0.339	0.348
MIPD	0.286	0.309	0.335	0.350

The test information function is an extremely useful feature of item response theory. It provides a metric of how well the test is doing in estimating ability over the range of ability scores (Xiong et al., 2002). While the ideal test information function often may be a horizontal line (**Figure 7**, n represents test length), it may not be optimal for meeting specific demands. For example, if one aims to construct a test to award scholarships, this ideal function may not be appropriate. In this situation, one aims to measure ability with considerable precision at ability levels near that used to separate those who will receive the scholarship from those who will not. The best test information function in this case would have a peak at the cutoff score (**Figure 8**; Baker and Kim, 2017). Other specialized uses of tests could require different test information functions. For example, for a test provided to award scholarships at several levels, the satisfactory test information function would have multiple peaks at the cutoff scores (a multimodal distribution). The bimodal distribution is one of the simplest types (**Figure 9**).

Nevertheless, to our knowledge, there is little research specifically on the information curves of bimodal distributions in the context of automated test assembly. In some educational measurement, not only is it required to have a demarcation score with small error and strong discrimination at the boundary between qualified and unqualified, but also hope to have another demarcation score with small error and strong discrimination at the boundary between excellent and non-excellent. This requires that the target state of the test information function be designed as a bimodal curve (Chen and Wang, 2010). It is undoubtedly worthwhile to investigate the performance of test assembly methods based on the item pool information curves of bimodal distributions.

METHOD

The goal of the simulation study was to investigate the performance of four item selection methods under various conditions:

Pool size, test length, and number of forms: The size of the item pool was 540, the test length was 30, and the number of parallel tests was 5.

Item parameters: Each item was subject to 3PLM, the discrimination parameter had a normal distribution, with a mean value of 1 and a standard deviation of 0.3; the difficulty parameter had a bimodal distribution, and the guessing parameter had a [0, 0.3] uniform distribution.

Non-statistical constraints. The items in the item pool covered three content areas A, B, and C, whose proportions of the total content were 40, 30, and 30%, respectively. The seed test consisted of 30 items (content proportions identical to this in item pool).

Variables for bimodal curve of test information. The most common bimodal distribution is a combination of two normal distributions. Xu et al. (2013) first used Excel to randomly generate two normally distributed datasets with a seed size of 1,000 and then extracted $n \times 1{,}000$ random datapoints from the first normal distribution and $1{,}000 - n \times 1{,}000$ random datapoints from the second normal distribution (n is a ratio ranging from 0 to 1). They extracted random datapoints from the seed and then created a scatterplot and a histogram based on the extracted data to obtain the bimodal distribution's shape. The procedure for producing bimodally distributed TIC is similar to the preceding process, except it is a combination of two unimodally distributed TICs. Different values of a will generate bimodal distributions of different shapes. In Kim and Lee's (2020) study, the ratios of extraction from the normal component $\theta \sim N(-1.8, 0.8)$ and the normal distribution $\theta \sim N(0.8, 0.8)$ were 3:7, 5:5, and 7:3, composing three different bimodal distributions. These three ratios can effectively represent the different forms of the bimodal distribution. Therefore, in this study, we set the mixing ratios of the bimodally distributed TICs to 3:7, 5:5, and 7:3. In addition to the mixing ratio, the horizontal spacing between the double peaks will also affect the shape of the bimodal curve. For this reason, we included the bimodal horizontal distances of the TICs in our estimates, which we set to 0, 1, 2, and 3. Simply put, the main variables of bimodal TICs observed in this study were the bimodal mixing ratio and bimodal horizontal spacing.

We repeated the test for R (1,000) times, each time randomly generating the item pool parameters and seed tests that met the preceding requirements, using the four item selection methods to generate 5 parallel test papers. To accommodate the content constraints of the test, we only directly determined the most matching items from each content sub-item pool and did not use weighting factors.

Evaluation Criteria

1. Mean square deviation indicator of test information function (MSD_{TIC})
   We used this indicator to evaluate the difference between the assembled test and the seed test in terms of their TICs. We calculated it using the following formula:

$$MSD_{TIC} = \left(\sum_{n=1}^{N} \left(I_{(\theta_n)} - I_{s(\theta_n)}\right)^2\right)/N$$

   where $I_{(\theta_n)}$ and $I_{s(\theta_n)}$ represent test information of the assembled test and the seed test at ability point θ_n (n = 1, 2,..., N), respectively. The number of capability nodes N was set to 61, the capability range was −3 to 3, and the step size was 0.1.
2. Mean square deviation indicator of test characteristic curve (MSD_{TCC})

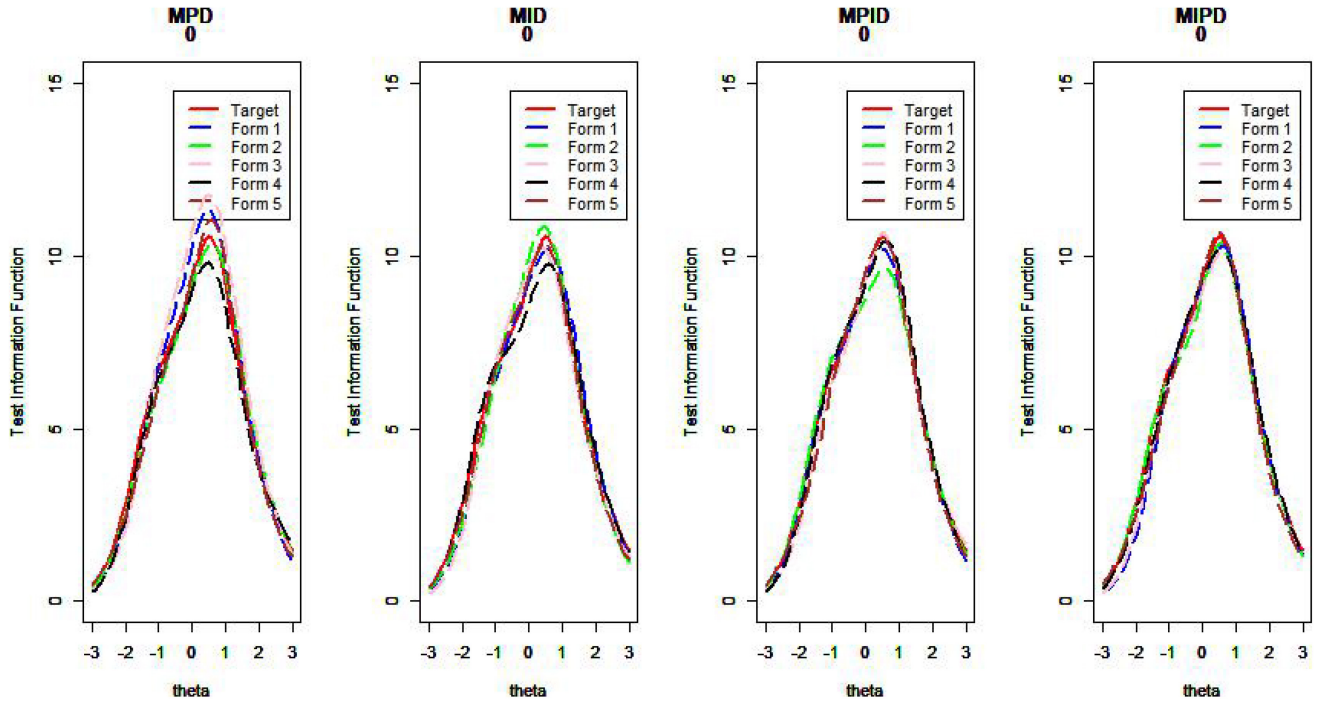


FIGURE 10 | Test information curve ($D = 0$).

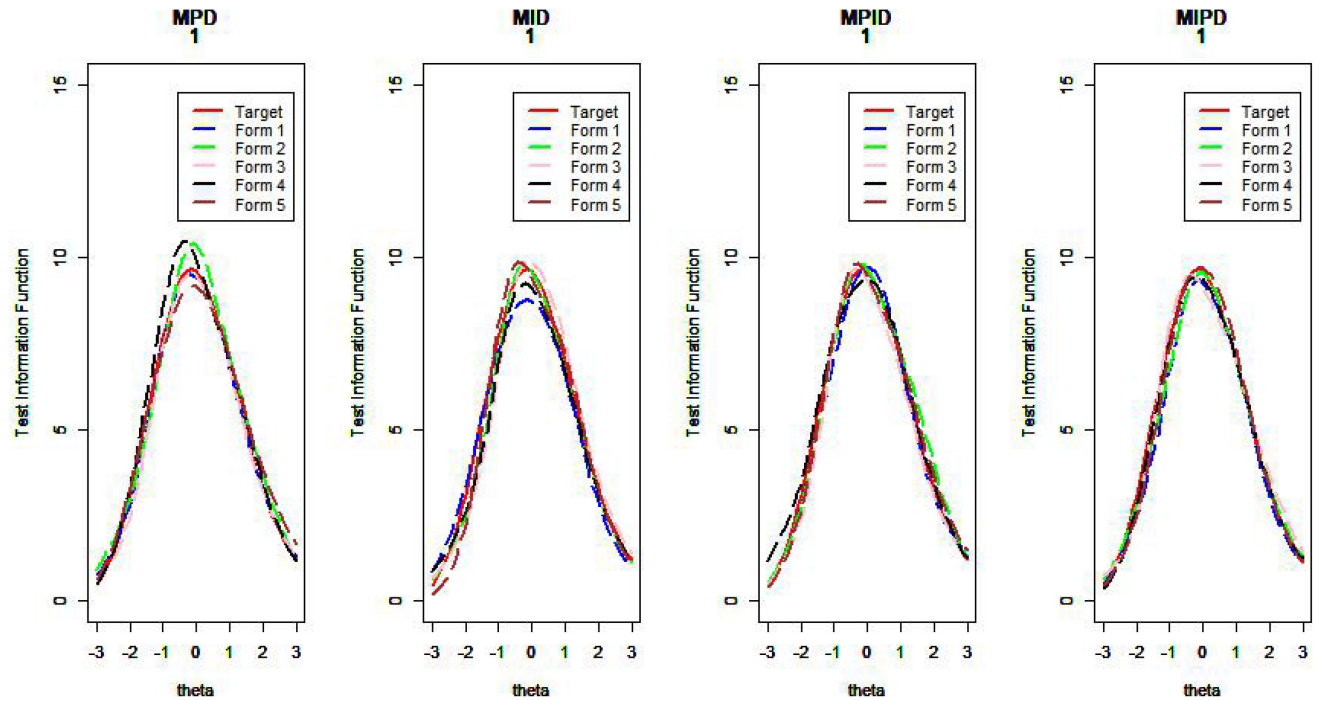


FIGURE 11 | Test information curve ($D = 1$).

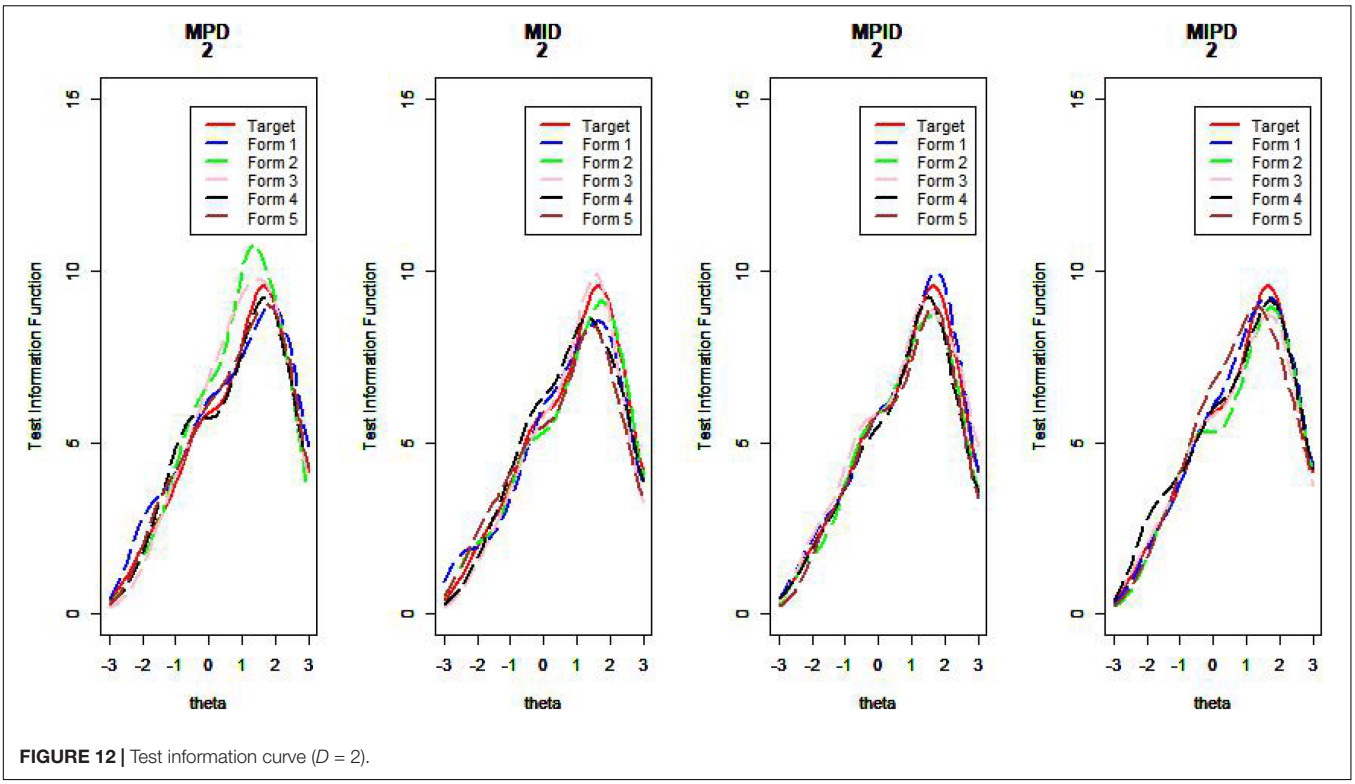


FIGURE 12 | Test information curve ($D = 2$).

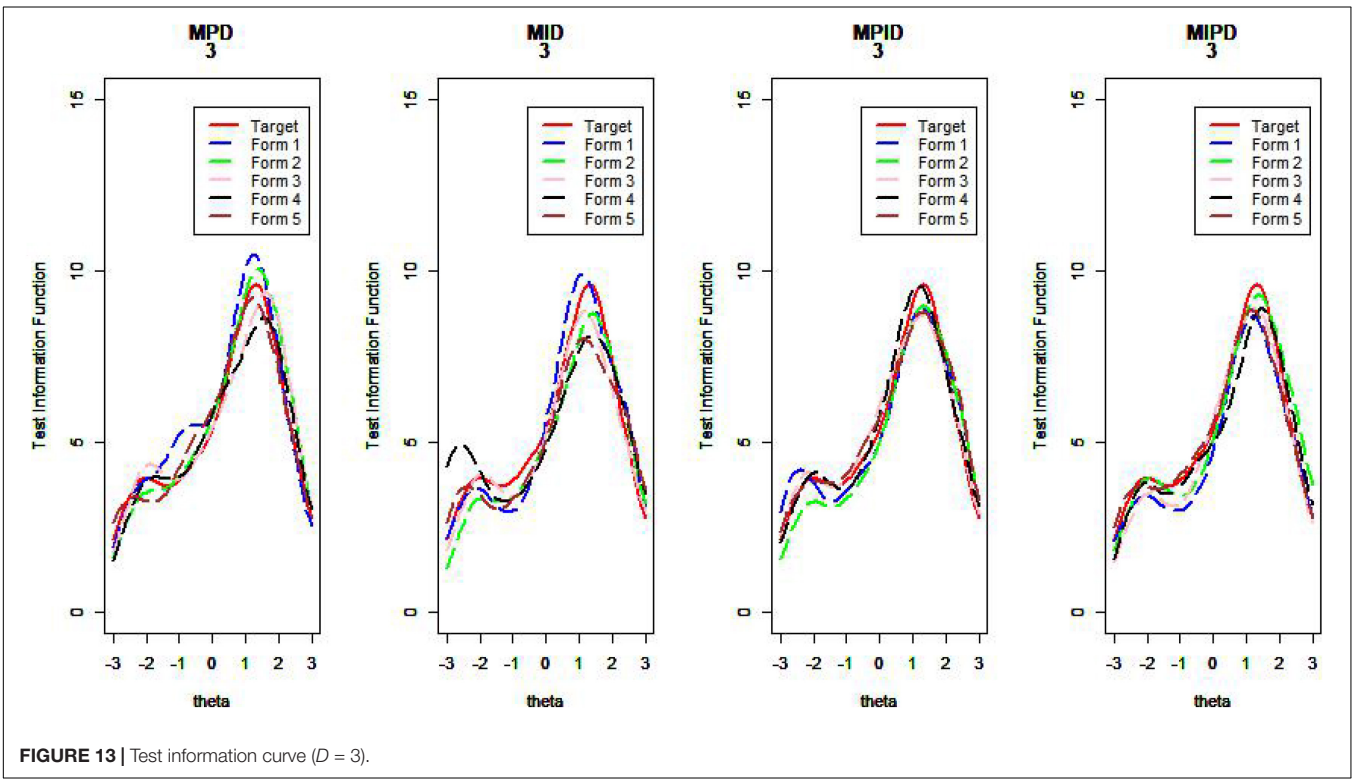


FIGURE 13 | Test information curve ($D = 3$).

TABLE 2 | MSD_{TCC}.

	Unimodal (*D* = 0)	Unimodal (*D* = 1)	Bimodal (*D* = 2)	Bimodal (*D* = 3)
MPD	0.227	0.232	0.249	0.268
MID	0.255	0.288	0.368	0.485
MPID	0.193	0.208	0.233	0.247
MIPD	0.197	0.215	0.233	0.263

This indicator was used to evaluate the difference between the TCCs of the assembled test and the seed test. We calculated it using the following formula:

$$MSD_{TCC} = (\sum_{n=1}^{N} (C_{(\theta_n)} - C_{s(\theta_n)})^2)/N$$

where $C_{(\theta_n)}$and $C_{s(\theta_n)}$represent TCCs of the assembled test and the seed test at ability point θ_n (n = 1, 2,..., N), respectively. The number of capability nodes N was set to 61, the capability range was −3 to 3, and the step size was 0.1.

RESULTS

The mixing ratio of the bimodal distribution has little effect on the results, so to avoid cluttering the presentations, the following only shows the results with a bimodal mixing ratio of 3:7. The **Supplementary Appendix** presents the rest of the results for interested readers. *D* represents the two peaks' horizontal spacing.

Test Information Curve

Table 1 lists the mean values for the mean square deviation of the five forms from the target test information at 61 ability points for all test assembly methods.

As **Table 1** illustrates, when the TIC has a unimodal distribution, the MID method performs better than the MPD methods and the MPID and MIPD methods achieve the same *MSD* as the MID method. Furthermore, the MPD method rivals the MID methods gradually and the MPID and MIPD method perform best when the TIC has a bimodal distribution.

In sum, the MPID and MIPD method—regardless of the bimodal horizontal distance—perform the best among all the four methods. The performance of the MID method when the TIC has two peaks is not as good as when the TIC has a single peak, indicating that the MPID and MIPD method (especially the former) are much more suited for use with bimodally distributed TIC than is the MID method. The MPD method has no advantage in TIC.

van der Linden (2005) argues that if the information function curves of the two tests are very similar—that is, when the difference in the amount of information between the assembled test and the seed test at different abilities is small—then the two tests can be considered statistically equivalent. The proximity of the TCCs can also be used as an evaluation criterion for the quality of the assembled test. Plotting the test information function and test characteristic curve for the assembled test and the seed test, one can intuitively judge the pros and cons of the item selection methods (Wang et al., 2021). Due to limited space, we show only some of the results here.

Figures 10–13 show the test information curves resulting from the four methods.

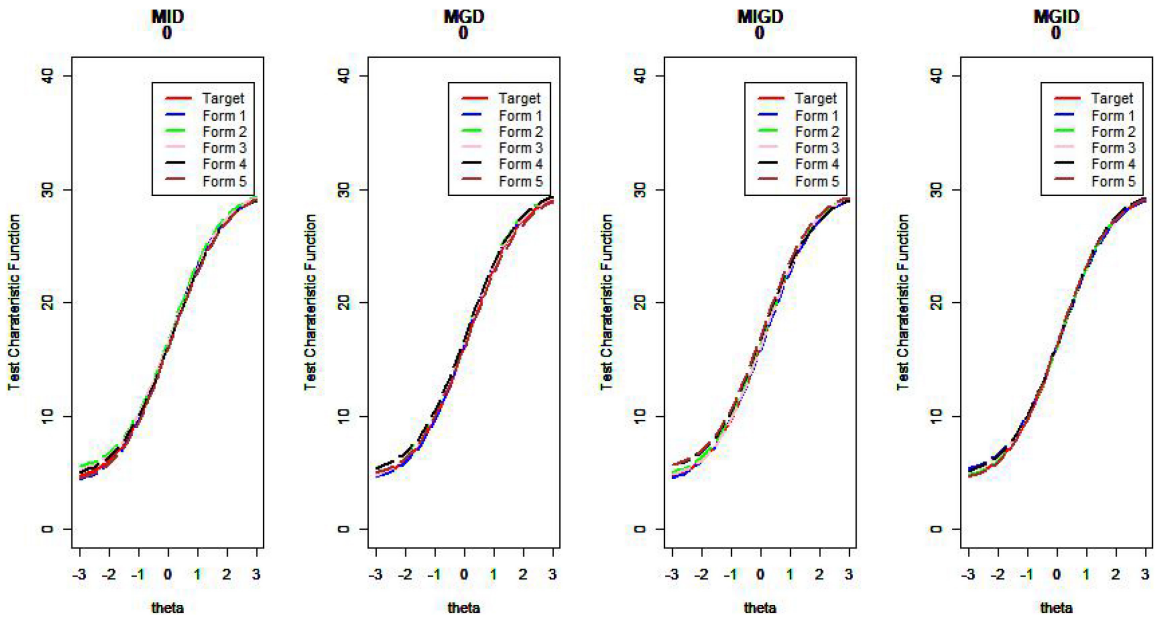


FIGURE 14 | Test characteristic curve (*D* = 0).

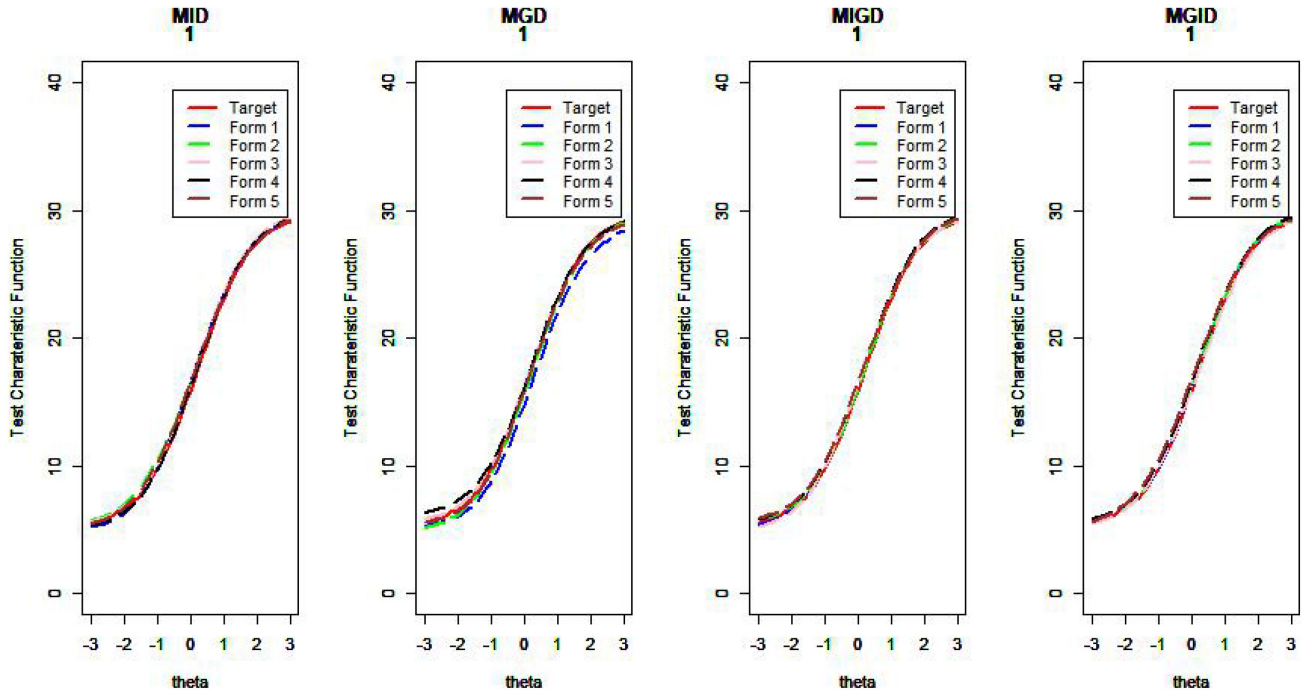


FIGURE 15 | Test characteristic curve ($D = 1$).

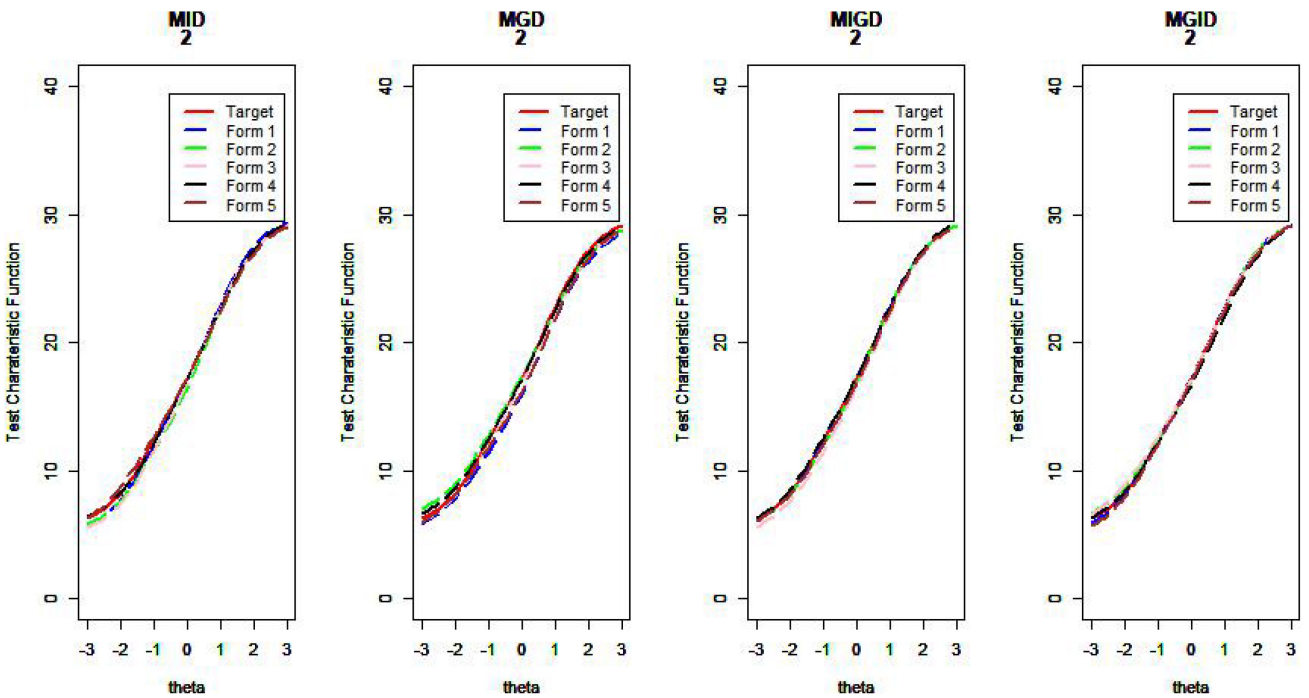


FIGURE 16 | Test characteristic curve ($D = 2$).

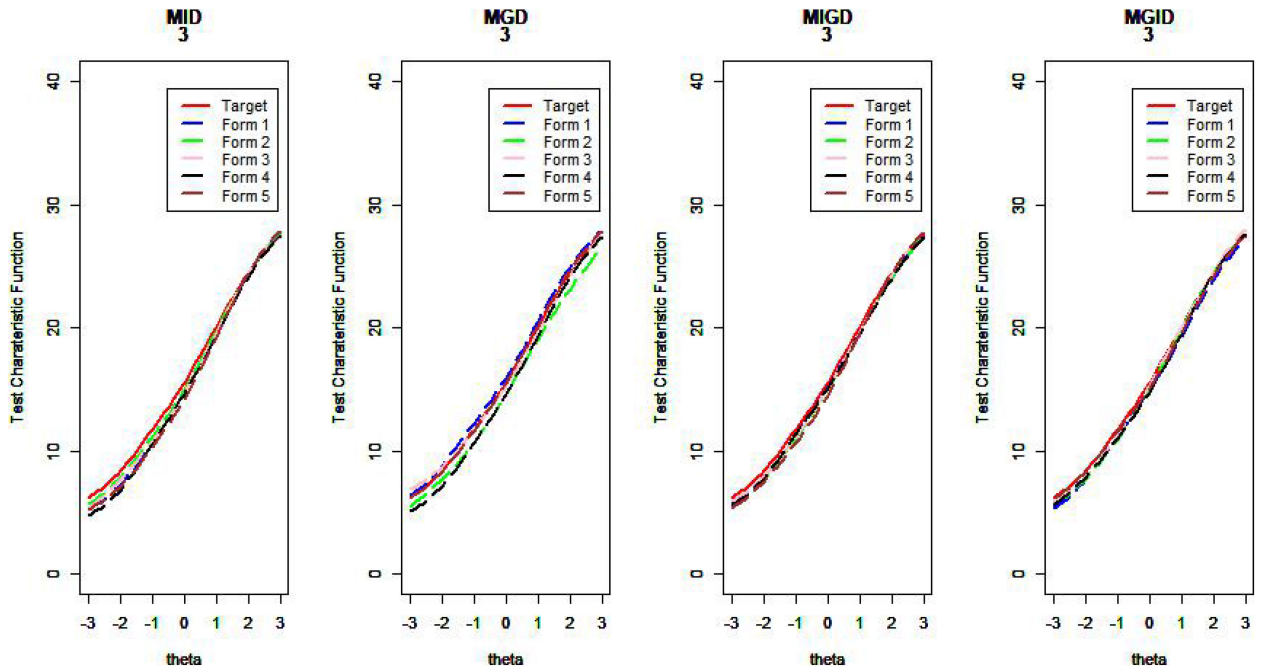


FIGURE 17 | Test characteristic curve (D = 3).

Test Characteristic Curve

Table 2 lists the mean values for the mean square deviation of the five forms from the target test characteristic curve at 61 ability points for all test assembly methods.

Regarding the MSD_{TCC}, the MPD method shows its strength of lowering the disparity between target test and assembly tests, resulting in smaller MSD_{TCC} and outperforming the other methods. The MPID and MIPD methods' performance is close to that of MPD. Obviously, the MID method has no advantage in TCC.

Figures 14–17 show the test characteristic curves resulting from the four methods.

DISCUSSION AND CONCLUSION

As far as the two existing methods are concerned, MPD has advantages in matching TCCs, while MID is superior in matching TICs. Two new methods combining the two methods (the MPID and MIPD method) can not only better match TCCs but also TICs. Although the new method only combines the original method with progressive coefficients, we contend that this research may help test agencies needing to generate multiple test forms for the sake of maintaining test security when administering multiple tests at various locations and times.

There are several reasons for our argument. First, it is undeniable that the two indicators are very important, and they have distinct meanings. The similar TCC results hold for forms that are similar in difficulty; test forms with the same TIF are similar in terms of precision (Ali and Van Rijn, 2016). Hence, we expect that all two indicators will be satisfactory (Chen, 2014). Then, the advantages of the new method are especially reflected in the bimodal distribution conditions. The MID method is susceptible to distribution. Under the bimodal condition, the TCCs matching effect of the two new methods is significantly better than the existing method. Additionally, new methods separating item selection phase into several stages and applying various methods in each stage offer a simple perspective on how to integrate diverse methods' merits. Finally, this study's consideration of the different distributions of test information closely matches the reality of test data.

This study has several limitations. The MPID and MIPD methods presented here are simplified versions, and further modification would make them more practical. Other important issues must be addressed in future research, including the setting of λ parameter, to take full advantage of each approach (Liu and Chang, 2018). In addition, it is common knowledge that the ability point specifications can influence the MID method's results. The issue of whether ability points (–2, –1, 0, 1, 2) suitable for information of unimodal distributions are as appropriate for bimodal distributions deserves additional attention (Chen, 2015). Finally, the item, response, theory (IRT)-based ATA methods proposed in this study focus on information curves of bimodal distributions. Whether the results can be extended to test designs with information curves of other multimodal distributions needs further investigation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZH prepared the first draft. LJ, ZP, and KC provided insightful comments that critically improved the quality of the manuscript. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.786772/full#supplementary-material

REFERENCES

Ali, U. S., and Van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Appl. Psychol. Meas.* 40, 163–179. doi: 10.1177/0146621615613308

Armstrong, R., Jones, D., and Wu, I. L. (1992). An automated test development of parallel tests from a seed test. *Psychometrika* 57, 271–288. doi: 10.1007/BF02294509

Baker, F. B., and Kim, S. H. (2017). *The Basics of Item Response Theory Using R*. New York, NY: Springer International Publishing. doi: 10.1007/978-3-319-54205-8

Brusco, M. J., Koehn, H. F., and Steinley, D. (2013). Exact and approximate methods for a one-dimensional minimax bin-packing problem. *Ann. Oper. Res.* 206: 611–626. doi: 10.1007/s10479-012-1175-5

Shao, C., Liu, S., Yang, H. W., and Tsai, T.-H. (2019). Automated test assembly using SAS operations research software in a medical licensing examination. *Appl. Psychol. Meas.* 44, 234–248.

Chen, P. H. (2014). A sampling and classification item selection approach with content balancing. *Behav. Res. Methods* 47, 98–106.

Chen, P. H. (2015). Three-element item selection procedures for multiple forms assembly:an item machine approach. *Appl. Psychol. Meas.* 40, 114–127. doi: 10.1177/0146621615605307

Chen, P. H. (2017). Should we stop developing heuristics and only rely on mixed integer programming solvers in automated test assembly? A rejoinder to van der Linden and Li (2016). *Appl. Psychol. Meas.* 41, 227–240. doi: 10.1177/0146621617695523

Chen, P. H., Chang, H. H., and Wu, H. (2012). Item selection for the development of parallel forms from an IRT-Based seed test using a sampling and classification approach. *Educ. Psychol. Meas.* 72, 933–953. doi: 10.1177/0013164412443688

Chen, M., and Wang, C. R. (2010). Algorithm design based on IRT-guided double boundary test. *J. Minzu Univ. China Nat. Sci. Ed.* 19, 53–55. doi: 10.3969/j.issn.1005-8036.2010.03.010

Cor, K., Alves, C., and Gierl, M. (2009). Three applications of automated test assembly within a user-friendly modeling environment. *Pract. Assess. Res. Eval.* 14:23.

Finkelman, M., Kim, W., and Roussos, L. A. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *J. Educ. Meas.* 46, 273–292. doi: 10.1111/j.1745-3984.2009.00081.x

Finkelman, M., Kim, W., Roussos, L., and Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Appl. Psychol. Meas.* 34, 310–326. doi: 10.1177/0146621609344846

Kim, S. Y., and Lee, W. (2020). Classification consistency and accuracy with atypical score distributions. *J. Educ. Meas.* 57, 286–310. doi: 10.1111/jedm.12250

Liu, C. J., and Chang, H. H. (2018). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Edu. Psychol. Meas.* 79, 335–357. doi: 10.1177/0013164418790634

Ramai, D., Ofosu, A., Lai, J. K., Gao, Z. H., and Adler, D. G. (2021). Fibrolamellar hepatocellular carcinoma: a population-based observational study. *Dig. Dis. Sci.* 66, 308–314.

Revuelta, J., and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *J. Educ. Meas.* 35, 311–327.

Steinley, D., and McDonald, R. P. (2007). Examining factor score distributions to determine the nature of latent spaces. *Multivar. Behav. Res.* 42, 133–156.

Tang, H. (2018). Research on the causes of the bimodal distribution of students' academic achievement based on QMAS. *Comput. Fan* 109:232.

van der Linden, W. (2005). *Linear Models for Optimal Test Design*. New York, N Y: Springer, 408. doi: 10.1198/jasa.2006.s148

Wang, S., Yi, Z., Zheng, C., Su, Y. H., and Li, P. (2016). An automated test assembly design for a large-scale chinese proficiency test. *Appl. Psychol. Meas.* 40, 233–237. doi: 10.1177/0146621616628503

Wang, W. Y., Xiong, J., Song, L. H., Zheng, J. J., and Hu, H. Y. (2021). MPI method of double-question matching and its application in parallel test paper generation. *J. Jiangxi Norm. Univ. Nat. Sci. Ed.* 45, 118–125. doi: 10.16357/j.cnki.issn1000-5862.2021.02.02

Xiong, J. H., Ding, S. L., Qi, S. Q., and Dai, H. Q. (2002). Use test information to analyze test paper quality. *J. Jiangxi Norm. Univ.* 26, 225–228.

Xu, X., Guo, M. Z., and Shi, F. L. (2013). Simulation of bimodal data distribution. *J. Yunan Norm. Univ.* 33, 46–51.

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 © 2022 Huijing, Junjie, Pingfei and Chunhua. 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 March 2022
doi: 10.3389/fpsyg.2022.851378

A Multi-level Remedial Teaching Design Based on Cognitive Diagnostic Assessment: Taking the Electromagnetic Induction as an Example

*Rui Huang[1,2], Zengze Liu[2], Defu Zi[3], Qinmei Huang[3] and Sudong Pan[2]**

[1]School of Physics and Electronic Science, East China Normal University, Shanghai, China, [2]College of Teacher Education, Faculty of Education, East China Normal University, Shanghai, China, [3]Dali No. 1 High School, Dali, China

Multi-level teaching has been proven to be more effective than a one-size-fits-all learning approach. This study aimed to develop and implement a multi-level remedial teaching scheme in various high school classes containing students of a wide range of learning levels and to determine its effect of their learning. The deterministic inputs noisy and gate model of cognitive diagnosis theory was used to classify students at multiple levels according to their knowledge and desired learning outcomes. A total of 680 senior high school students from central provinces in China participated in the initial cognitive diagnostic test, and 1,615 high school sophomores from seven high schools in China participated in a formal cognitive diagnosis test. Thirty-six high school students from Southwestern China participated in the think-aloud protocols, and 258 seniors from three high schools in southwest China participated in the remedial teaching experiment. Through an analysis of students' think-aloud protocols, cognitive errors of students at all levels were determined, and multi-level remedial teaching programs were designed to address these common cognitive errors. The remedial teaching programs were then implemented in three schools and compared with a control group. The results indicated that the students in the experimental group showed a more significant improvement. In this study, the steps of designing multi-level remedial teaching include assessment, classification, and preparing a teaching scheme, which are feasible and can have remarkable teaching effects. This process can be used for reference by teachers of various subjects.

Keywords: multi-level teaching, remedial teaching, electromagnetic induction, DINA model, cognitive diagnostic assessment

OPEN ACCESS

Edited by:
Tao Xin,
Beijing Normal University, China

Reviewed by:
Ling Ling Wang,
Shenyang Normal University, China
Xiaojian Sun,
Southwest University, China

****Correspondence:***
Sudong Pan
sdpan@phy.ecnu.edu.cn

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

Received: *09 January 2022*
Accepted: *04 February 2022*
Published: *23 March 2022*

Citation:
Huang R, Liu Z, Zi D, Huang Q and Pan S (2022) A Multi-level Remedial Teaching Design Based on Cognitive Diagnostic Assessment: Taking the Electromagnetic Induction as an Example. Front. Psychol. 13:851378. doi: 10.3389/fpsyg.2022.851378

INTRODUCTION

Psychometry-based cognitive diagnostic assessment (CDA) is a method that can be used by frontline teachers to determine students' learning outcomes and classify them based on their diagnostic results for more personalized remedial teaching. At the core of the new generation of test theory are cognitive diagnosis models (CDMs)—models which are able to measure

more detailed information about participants' knowledge, skills, and strategies. Currently, research on cognitive diagnosis has received worldwide attention from researchers, teachers, and evaluators alike, with a number of studies having implemented individualized remedial teaching for students based on diagnostic classifications.

One study by Wang et al. (2021) applied the CDM to evaluate and classify urban and rural middle school students based on their mastery of a topic on linear equations. Remedial teaching based on cognitive results was implemented for the test group, while the traditional "answer-explanation" style of remedial teaching was implemented for the control group. Results showed that remedial teaching based on cognitive diagnosis results significantly improved the learning effect. Similarly, Ren et al. (2021) used the CDM to learn about the poorly mastered attributes of "data distribution characteristics" in teaching math to middle schoolers and found that verified cognitive diagnoses can be used for targeted interventions to improve students' abilities more effectively. Consistent with this, Fan et al. (2021) proposed an integrative framework of diagnosis which connects CDA to feedback and remediation, and they empirically demonstrated the application of the framework in an English as a Foreign Language (EFL) context.

These studies suggest that CDA can effectively diagnose students' learning outcomes and be used to conduct personalized remedial teaching. However, the basis of remedial teaching in these studies was only whether or not attributes were mastered, but the level of mastery of attributes was not graded. With this being said, in order to be truly personalized, the basis for classification in remedial teaching should not only be dependent on whether the attribute is mastered, but also on the mastery level of knowledge, skills, and cognitive processes. Hence, in this study, multi-level remedial teaching refers to the remedial teaching for students based on their respective levels of knowledge or skills mastery for a specific topic after they have learned it in class.

Multi-level teaching and learning are widely regarded as an important way of improving teaching efficiency. In recent years, many studies have designed and developed methods to stratify students' learning outcomes, with many of these methods in previous literature being based on computer algorithms (He et al., 2016; Zhang et al., 2016). For instance, Wu (2019) used CDA to classify fourth grade students based on their learning scores in order to provide personalized online remedial guidance, and results showed that the online personalized tutor program was superior to the traditional tutorial program. You et al. (2019) applied CDA's deterministic inputs noisy and gate (DINA) model to develop a cognitive diagnostic system for Chinese learning, which was applied in all subjects of an experimental high school to provide personalized learning feedback for students and teachers. The study found that through this the efficiency and self-efficacy of students improved. Additionally, Shute et al. (2008) developed a multivariate probit model for CDA and applied it to the data from the Adaptive Content with Evidence-Based Diagnosis (ACED) evaluation study to verify the validity of the new model. These aforementioned studies developed adaptive learning systems based on cognitive CDA. However, to be applied in schools, this requires the purchase of hardware and software which may be difficult to use for some frontline teachers.

Computer-based adaptive learning is a process dependent on the use of a technological device, whereas teacher-student interaction is much more common in Chinese high schools. Therefore, our study provides frontline teachers with examples of multi-level teaching designs based on CDA without the need for large hardware or paid software services and integrates experiments and discussions into multi-level remedial teaching that human-computer interactions cannot provide.

Most of the abovementioned CDA application cases in educational practice are conducted in subjects such as mathematics (Groß et al., 2016) and second foreign languages (Liu et al., 2013), but its application in the sciences have not yet been thoroughly studied. Zhan et al. (2019b) developed a new CDM which realized the assessment of scientific literacy and filled the gap in the application of CDA in scientific disciplines. Practical data from an eighth-grade physical circuit topic were used to examine the performance of the newly developed model (Zhan et al., 2019a). However, there is currently almost no empirical remedial teaching research on applying CDMs to the physics curriculum in high school education.

The present study aimed to design and implement a micro multi-level remedial teaching plan based on CDA that is easy to use in the classroom-setting and which may be applied to high school physics remedial teaching plans. Electromagnetic induction is a very important chapter in high school physics. It is inherently logical and difficult to learn. This study takes the topic of electromagnetic induction as an example. This study intended to solve the following problems:

1. How can multi-level classifying be performed for students learning a certain topic?
2. How can a multi-level remedial teaching plan be developed according to the created classifications?
3. Does the implementation of multi-level remedial teaching effectively improve students' learning results?

MATERIALS AND METHODS

In this study, CDM was used to assess and classify students' learning of electromagnetic induction topics in high school. To design multi-level remedial teaching, think-aloud protocols of students at different levels were collected to analyze their thinking and determine common cognitive errors of students at each level. Multi-level remedial teaching plans were then designed. After the multi-level remedial teaching experiment, the students completed a post-test, which was used to test the intervention effect. The research process is shown in **Figure 1**.

Assessment and Classification

The CDM was used to assess students' learning on electromagnetic induction. Based on the results of the assessment, the students

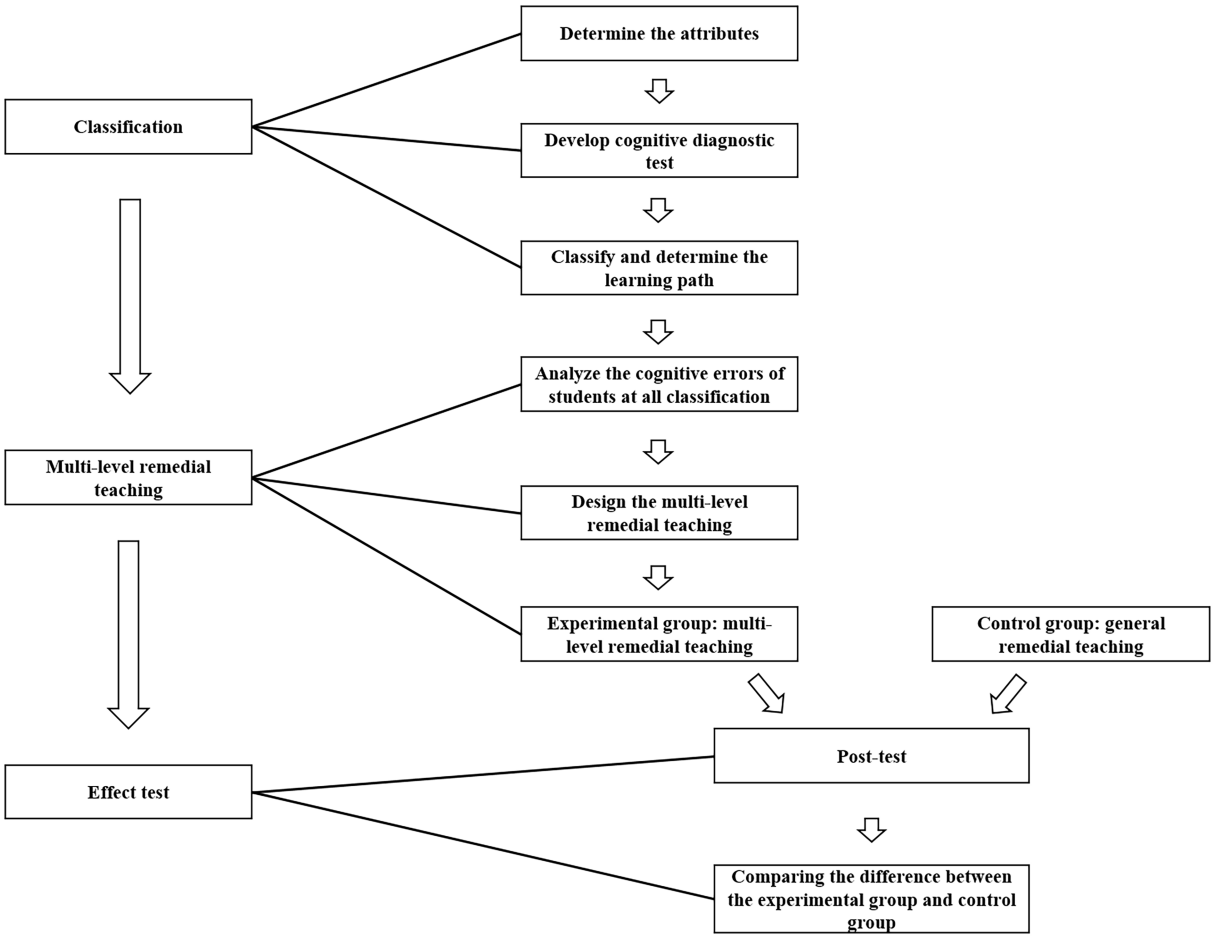


FIGURE 1 | The multi-level remedial teaching research process.

were classified according to their level of learning of concepts and rules.

Multi-level Attributes

For the hierarchical diagnostic classification of learning result, attributes will be identified by two dimensions: knowledge content and learning level.

In this study, to determine the attributes, five physics teachers with over 10 years of teaching experience in high school physics were invited to analyze the content and the learning result levels determined based on the questions in the question bank. The bank contained a total of 53 multiple-choice items on electromagnetic induction from five recent versions of the Chinese Higher Education Entrance Examination (2014–2018) and academic level examination review papers. According to the teachers, the question bank contents related to the topic of electromagnetic induction were the following: (1) electromagnetic induction phenomenon, (2) conditions to generate induced current, (3) Lenz's law, (4) the right-hand rule, and (5) Faraday's law. These concepts and laws comprise the first dimension of attributes (i.e., knowledge content).

Based on Benjamin Bloom's research, Anderson et al. (2001) divided learning results in the cognitive domain into several progressive levels based on explicit behaviors that correspond to the degree to which students understand the subject. The learning results in the cognitive domain were divided into six levels: knowledge, comprehension, application, analysis, synthesis, and evaluation. In line with this, Lee et al. (2011) divided the cognitive dimensions measured in the math items of Trends in Mathematics and Science Study (TIMSS) test into three levels: knowing, applying, and reasoning. In the present study, the five experts determined that the relevant questions could be divided into four levels according to the learning results of the examination: (1) knowledge, (2) understanding, (3) application, and (4) integrated application.

Knowledge

If students were able to correctly answer questions that examine concepts, content of laws, and corresponding physical phenomena, their learning results were defined as "knowledge." For example, the following question examines the attribute "electromagnetic induction phenomenon: knowledge."

Which of the following phenomena is electromagnetic induction?
A. A current is subjected to a force in a magnetic field.
B. There is a magnetic field around the current.

C. A soft iron bar can be magnetized by a magnetic field.

D. The changing magnetic field causes an electric current to be generated in a closed conductor.

Understanding

If students were able to correctly answer questions that use laws for calculation and reasoning, their learning results were defined as "understanding." For example, the following question examines the attribute "conditions to generate induced current: understanding."

Which situation can induce current?

A. The conductor moves in a cutting magnetic field line.

B. A part of the closed circuit moves parallel to the magnetic field.

C. A part of the closed circuit cuts the magnetic field lines in a magnetic field.

D. None of this is true.

Application

If students were able to correctly answer questions regarding two-related physical processes and two-related physical objects which do not involve knowledge beyond electromagnetism, their learning results were defined as "application." For example, the following question examines the attribute "Lenz's law: application."

*As shown in the **Figure 2**, two coils are wound around an iron core. One coil is connected to the switch and the power supply, and the other coil is connected in a loop with a straight wire placed horizontally in a north–south direction. A magnetic needle is suspended directly over the straight wire and stands still when the switch is off. Which choice is true?*

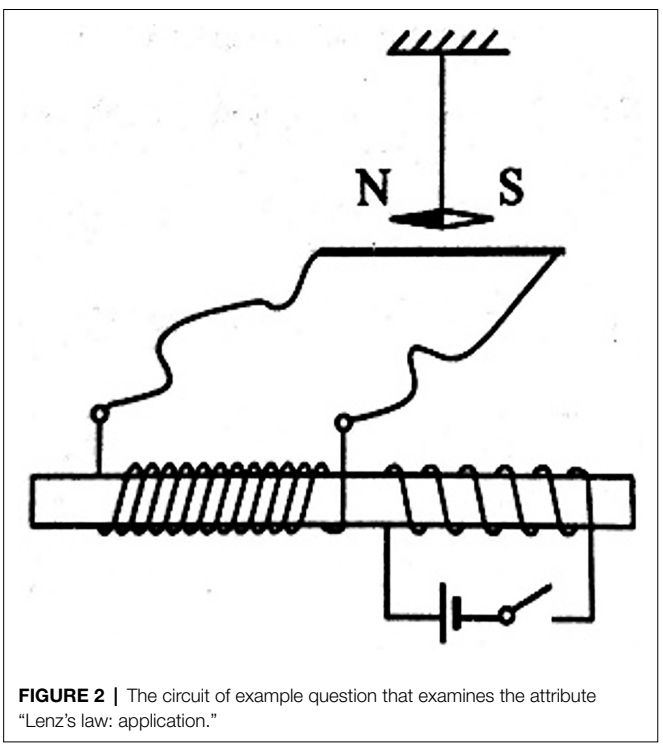


FIGURE 2 | The circuit of example question that examines the attribute "Lenz's law: application."

A. At the instant after the switch is on, the N pole of the needle points inward to the paper.

B. After the switch is on and held for a period of time, the N pole of the magnetic needle points inward to the paper.

C. After the switch is on and held for a period of time, the N pole of the magnetic needle points out of the paper.

D. When the switch is closed for a period of time and then opened, the N pole of the magnetic needle points outward to the paper.

Integrated Application

If students were able to correctly answer questions regarding more than two-related physical processes and more than two-related physical objects which involve knowledge beyond electromagnetism, their learning results were defined as "integrated application." For example, the following question examines the attribute "Faraday's law: integrated application."

*As shown in the **Figure 3**, the smooth parallel metal with no resistance and spacing of L is horizontally placed in the uniform magnetic field with a magnetic induction intensity of B and direction of vertical downward, and the left end of the rail is connected with a resistance R. The metal bar MN with mass m and resistance R is placed on the rail and moves from rest under the action of the horizontal external force F perpendicular to the metal bar. The relationship between the F and the speed v of the metal bar is $F = F_0 + kv$ (F_0 and k are constant). The metal bar and the rail are always vertical and in good contact. The induced current in the metal bar is i, the ampere force is F_A, the voltage at both ends of the resistance R is U_R, and the power of the induced current is P. Which graph might correctly represent the trend of physical quantities over time?* (The options are shown in **Figure 3**).

On average, the consistency among the five experts for the knowledge content and learning result of each question was 83.0%. Based on the cognitive attributes marked by the experts, if three or more experts agreed on the level of learning results examined by the item, they were included as cognitive attributes. Ultimately, 12 attributes were determined, which are shown in **Table 1**.

Concepts or rules that were not within the examination scope of college entrance exams and academic achievement tests (i.e., higher levels) were considered beyond the scope of this study and were thereby not included.

Preparing the Test

A Q-matrix which conforms to the goal of cognitive diagnostic was developed prior to compiling the cognitive diagnostic test. Based on the principle that each attribute must be tested at least thrice, the Q-matrix was developed for two sets of parallel tests (**Supplementary Table S1**). Using the cognitive attributes and question bank items, two sets of 36 items which follow the measurement patterns were selected from the question bank for Form A (pre-test) and Form B (post-test): Electromagnetic Induction Cognitive Diagnostic Test (EICD). In the EICD test, the items that examined the attributes of "knowledge" were fill-in-the-blank questions; the other items were multiple-choice questions.

To test the EICD's quality, 680 senior high school students were recruited by convenience sampling to participate in the

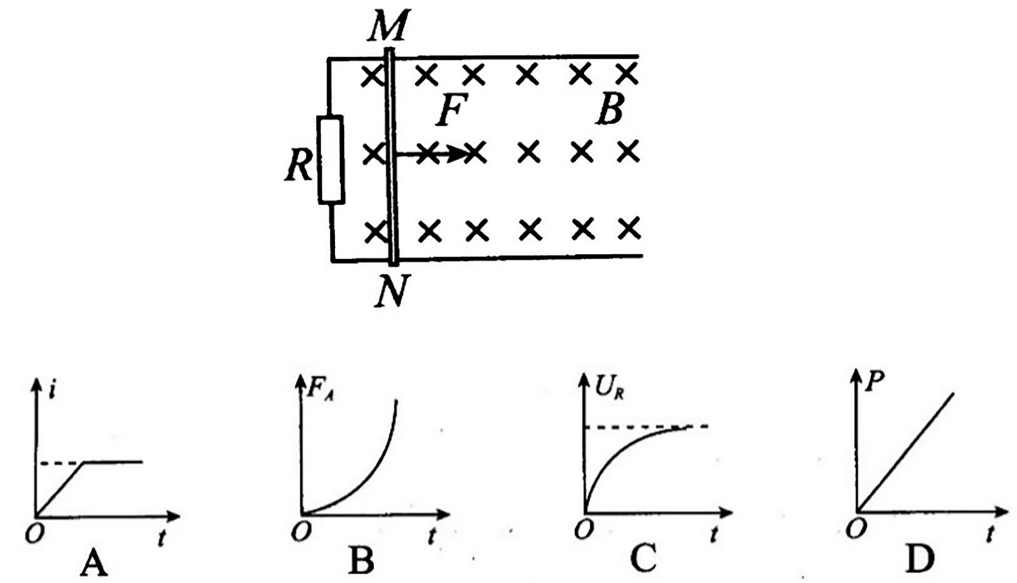


FIGURE 3 | The circuit and options of example question that examines the attribute "Faraday's law: integrated application."

TABLE 1 | Multi-level attributes.

Attributes	Detailed description
EIP	Electromagnetic induction phenomenon: Knowledge
CIC1	Conditions to generate induced current: Knowledge
CIC2	Conditions to generate induced current: Understanding
LL1	Lenz's law: Knowledge
LL2	Lenz's law: Understanding
LL3	Lenz's law: Application
RHR1	Right-hand rule: Knowledge
RHR2	Right-hand rule: Understanding
FL1	Faraday's law: Knowledge
FL2	Faraday's law: Understanding
FL3	Faraday's law: Application
FL4	Faraday's law: Integrated application

initial test: 442 answered Form A and 238 answered Form B. Before the test, the students and their guardians provided written informed consent for the study. The students were told that participation was voluntary and that they were allowed to withdraw at any time. In addition, participants were also told that the tests must be accomplished independently and that their results will not be part of their physics class evaluation.

The overall quality of test and items was judged according to Classical Measurement Theory (CTT). The reliability of the EICD test (initial) was measured by the CTT-based Cronbach's α coefficient. The CTT score is calculated as 2 points for correct answers at the "knowledge" level items, 3 points to the "understanding" level items, 4 points to the "application" level items, and 5 points for the "integrated application" level items. The α coefficients of EICD test (initial) Forms A and B were 0.7634 and 0.7364, respectively. Except for Item 3 and Item 11, the difficulty coefficient of the items was basically consistent with the learning level dimension of the attributes examined by the item. The difficulty coefficients ranged from [0.17, 0.97]. Item 3 and Item 11 were both "knowledge" level items, but the difficulty coefficients indicate that they were too difficult. Except for Item 11, the discrimination of items is between [0.21, 1.0]. Items with a difficulty coefficient above 0.4 accounted for more than half of the total items. Item 3 and Item 11 were replaced by items of the same level of assessment in the item bank. So far, after the initial test, the quality of the EICD test had been optimized, and the EICD test (formal) had been developed, with 36 items for each of the Form A and Form B, which are used for the formal test.

Formal Test

Because CDA requires a sufficient number of participants, approximately 1,000 responses were needed prior to form the response matrix with the test of the teaching experiment. In this section, the method for diagnostic classification will be discussed.

Participants

Using stratified sampling, 1,615 senior high school students from seven high schools in Eastern, Northwest, Southwest, and Central China participated in the formal cognitive diagnostic test, which was conducted in June and July 2019. In total, 861 participants effectively completed Form A, while 849 participants effectively completed Form B. Of these, 95 participants completed both Form A and B. Participants were given 1 h to complete the test. Written informed consent was provided by the participants and their guardians prior to the study.

Data Analysis

The quality of the EICD test(formal) was measured by CTT score. To examine whether the EICD test (formal) can reflect the real learning situation of students, 267 participants that completed the same Academic Level Test (2019 High School Academic Level Test in Yunnan Province) were selected. The grades of the Academic Level Test were A, B, and C from high to low. Correlations between Academic Level Test grades and EICD average scores were compared.

In order to explore whether Form A and B meet the requirements of parallel papers, the mean (*M*), standard deviation (SD), item difficulty (*P*), and item discrimination (*D*) of students' scores in Form A and B were compared.

Model Selection

Selecting an appropriate CDM was a necessary condition for obtaining reasonable diagnostic feedback. In this study, because the attributes correspond to specific mastery levels, students were considered to have mastered an attribute (attaining a commanded mastery level) if the items were answered correctly. The hierarchical structure between attributes was not considered. Therefore, the DINA model and the GDM model, which were both non-compensatory and also do not consider hierarchical structure, were selected as alternative models in this study. The relative fitting parameters showed that the theoretical relative fitting degree of GDM model is the better fit (AIC = 32673.6, BIC= 33392.1), than DINA model (AIC= 39938.5, BIC= 59765.5). During the formal test, when assessed with the DINA and GDM models, it was found that the same participant can have different patterns of attribute mastery. When participants with different diagnostic results were interviewed about the answered items and when the attribute lists were compared, the diagnostic accuracy rate of the DINA model appeared to be higher than that of the GDM model (**Supplementary Table S2**). Cai et al. (2013) used the method of data simulation to compare the diagnostic accuracy of the five commonly used models for mastering patterns in different situations. The results showed that under any knowledge state distribution, with a large sample size of about 1,000, and in the case of a large number of cognitive attributes, the accuracy of the DINA model diagnosis is relatively good. Hence, the present study adopted the DINA model for CDA. The DINA model was first proposed by Macready and Mitchell (1977) and was subsequently improved by Haertel (1989). Currently, it has now become a comparative, basic, and commonly used model in research (Junker and Klaas, 2001; Templin and Henson, 2006). The answer matrix was analyzed on the flexCDMs.[1]

The cognitive diagnosis reliability of the 12 attributes was measured by the consistency index of the attribute test–retest reliability. Assuming that the students' mastery probability of attributes remains unchanged, the two-by-two contingency tables on the correlation of the mastery of attributes in the test and retest may be obtained and used as the index of attribute reliability in the cognitive diagnostic test (Templin and Laine, 2013).

Diagnostic Classification Method

The DINA model gives feedback on the mastery pattern of each student for the attributes. The mastery pattern of a student is a vector composed of 12 elements that equal 0 (attribute not mastered) or 1 (mastered attribute). In this study, the vector of mastery pattern was divided into five small mastery pattern vectors according to the five knowledge contents: (EIP), (CIC1 and CIC2), (LL1, LL2, and LL3), (RHR1 and RHR2), (FL1, FL2, FL3, and FL4). (See the notes to **Table 1** for the full meaning of the acronyms). The five small mastery patterns were classified, as shown in **Figure 4**. The purpose of the diagnostic classification during the pre-test (Form A) was to identify the students' level of learning regarding the five knowledge contents of electromagnetic induction. Different mastery patterns of the attributes imply different learning levels. Then, according to the results of diagnosis and classification, multi-level remedial teaching can be carried out.

The purpose of post-test diagnostic classification was to test whether the student has achieved the target mastery pattern. **Figure 4** shows the method of diagnostic classification and

[1]http://www.psychometrics-studio.cn/

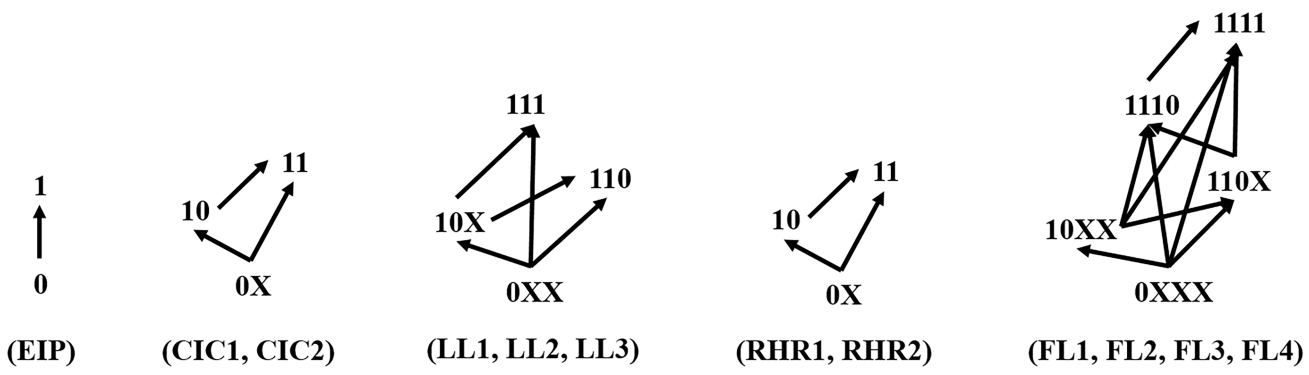


FIGURE 4 | Classification schemes and learning paths based on attribute mastery pattern. The 12 attributes are divided into five groups according to the knowledge content. Students are categorized by their mastery pattern for each attribute group. 1 represents mastered the attribute, 0 represents not mastered the attribute, and X represents 1 or 0. The arrow represents the learning path after diagnosis and classification, the tail of the arrow represents the attribute mastery pattern of the pre-test, and the tip refers to the target attribute mastery pattern after the layered remedial teaching.

learning paths for multi-level remedial teaching, with X representing either 0 or 1. Each student was classified five times according to the attribute mastery pattern of the five knowledge contents. In a knowledge content mastery pattern, students are classified according to the lowest level attributes that they have not mastered, and students with the same minimum level attributes are classified into one category. The end of the arrow is the diagnostic classification result of the attribute mastery pattern in the pre-test, and the tip of the arrow represents the target attribute mastery pattern in the post-test.

Multi-level Remedial Teaching Design

Before designing a multi-level remedial teaching plan, it is necessary to understand the cognitive errors of students at each mastery pattern classification. According to the diagnosis classification method, each student in the experimental group was classified into five groups in five knowledge contents. As shown in **Figure 2**, there were 17 groups in total. If a student was classified in groups without a 0 element, they were no longer included for remedial teaching. There was a total of 12 mastery pattern classifications in which corresponding cognitive errors had to be determined. In order to obtain the cognitive errors of students under each classification, three students were selected from each group for a total of 36 students. The 36 students were then asked to restate their thoughts when they completed either form (A or B). The researchers recorded and analyzed the cognitive errors, and then summarized the common characteristics of the cognitive errors made at each classification. Afterward, remedial teaching schemes were then designed with the aim of addressing these cognitive errors.

Think-Aloud Protocols

Before the think-aloud protocols, the 36 students were informed by their physics teacher that this was a teaching research in which students would be asked to repeat their thoughts during the test, that participation was voluntary, and that they could quit at any time during the process. The students were then asked to go into a classroom one-by-one and to repeat their thoughts aloud regarding Form A or Form B. The three researchers then noted the participants' cognitive errors. Finally, the three researchers discussed the common cognitive errors made by the students, as shown in **Table 2**.

Multi-level Remedial Teaching Scheme

As **Table 2** shows, students whose attribute mastery pattern was (EIP) = 0, (CIC1, CIC2) = (0X), (LL1, LL2, LL3) = (0XX), (RHR1, RHR2) = (0X), (FL1, FL2, FL3, FL4) = (0XXX) made errors in identifying related phenomena and concepts, repeating laws, and comprehending the essence of laws. Therefore, the remedial teaching scheme for these attribute mastery patterns focused on experiment and discussion. For students, whose attribute mastery pattern was (CIC1, CIC2) = (10), (LL1, LL2, LL3) = (10X), (RHR1, RHR2) = (10), (FL1, FL2, FL3, FL4) = (10XX), their cognitive errors involved the misuse of laws and not understanding the essence of the law. Therefore, the remedial teaching scheme for these attribute mastery patterns focused on experiment, speculation, and discussion. Lastly, for students whose attribute mastery pattern was (LL1, LL2, LL3) = (110), (FL1, FL2, FL3, FL4) = (110X) or (111X), their cognitive error involved not finding the relationship between multiple physical processes and quantities in the physical situation, or lacking a deep knowledge beyond the topic of electromagnetic induction. Therefore, the remedial teaching scheme for these levels of classification included guiding students in discussing the aforementioned topics together. Due to manuscript limitations, only three different levels of remedial teaching schemes are listed below. If readers are interested, the authors may be contacted for all of the multi-level remedial teaching schemes.

(a) EIP = 0 level remedial teaching scheme

Based on the results of the think-aloud protocols, multi-level remedial teaching should not only make students aware of electromagnetic induction, but it should also allow them to correctly distinguish electromagnetic induction from other electromagnetic phenomena. Based on this, the following remedial teaching plan was designed as:

[Teaching method] Experiment and discussion.

[Teaching aim] To be able to distinguish three types of electromagnetic interactions.

[Experimental material] A core, two solenoids, a magnetic needle, a DC power supply, a switch, a slide rheostat, wires, a sensitive galvanometer, and a bar magnet.

[Experimental circuit i is shown in the **Figure 5**]

[Discussion topic i] Corresponding to cognitive error 1 in **Table 2**: *Is there a power supply in the circuit? How does the pointer of the sensitive galvanometer deflect at the moment when the bar magnet enters and leaves the solenoid? Why is the pointer of the sensitive galvanometer deflected? What phenomenon does this process belong to?*

[Experimental circuit ii is shown in the **Figure 6**]

[Discussion topic ii] Corresponding to cognitive error 1 in **Table 2**: *Observe the change of the magnetic needle as the switch closes and opens. What kind of electromagnetic interaction is involved in each link of the experiment?*

[Discussion topic iii] Corresponding to cognitive error 2 in **Table 2**: *Discuss the difference and connection between the generator and motor.*

(b) (LL1, LL2) = (10) level remedial teaching scheme

According to the results of the think-aloud protocols, the following remedial teaching plans were designed to address the cognitive errors of students in order to reach the attribute mastery pattern of (LL1, LL2) = (11):

[Teaching method] Experiment, speculation, and discussion.

[Teaching aim] To use Lenz's law to determine the direction of the induced current in a simple situation.

[Experimental material] A bar magnet and a Lenz's law demonstrator.

[Experimental circuit i is shown in the **Figure 7**]

TABLE 2 | The cognitive errors of each mastery pattern classification.

Mastery pattern	No.	Cognitive errors
EIP = 0	1	The interaction of electricity and magnetism was incorrectly considered as electromagnetic induction
	2	Confusion over the physical principles on which generators and motors are based
(CIC1, CIC2) = (0X)	3	Misunderstanding that magnetic flux changes if the conductor cuts the magnetic field lines
(CIC1, CIC2) = (10)	4	Misunderstanding that an induced current can be generated if the conductor cuts the magnetic field lines
	5	Misunderstanding that the unclosed coil has no magnetic flux, and the magnetic flux changes when it is closed again
	6	Did not understand the magnetic field distribution around the bar magnet
(LL1, LL2, LL3) = (0XX)	7	Did not understand the concept of flux change
	8	Could not describe Lenz's law
(LL1, LL2, LL3) = (10X)	9	Could not translate a change in the physical situation into a change in the magnetic flux, thereby not knowing how to use Lenz's law
	10	Lenz's law was not understood in terms of energy, and the transformation of functional relations in electromagnetic induction was not understood.
	11	In the process of using the formula, the formula was mistaken for Lenz's law
(LL1, LL2, LL3) = (110)	12	In physical situations, the hindrances of Lenz's law were not used to determine the direction of the induced current
(RHR1, RHR2) = (0X)	13	Students could not distinguish between the right-hand rule, left-hand rule, and ampere rule
	14	Misunderstanding that the direction of the conductor cutting magnetic field lines is the direction of the conductor force
	15	It was incorrectly believed that the direction of the induced electromotive force is opposite to that of the induced current
(RHR1, RHR2) = (10)	16	Incorrectly used the left-hand rule to solve the electromagnetic induction problem
	17	Did not understand the direction of the current or magnetic field lines in the diagram
(FL1, FL2, FL3, FL4) = (0XXX)	18	Did not know the basic concept of Faraday's law
	19	Could write Faraday's law formula but could not explain Faraday's law
	20	Inability to distinguish between the rate of change and quantity of change
(FL1, FL2, FL3, FL4) = (10XX)	21	The left-hand rule was used
	22	Did not understand the circuit knowledge. Misjudged the direction of current and potential inside the source
	23	Did not know that the conductor cutting the magnetic field lines is equivalent to the power supply
(FL1, FL2, FL3, FL4) = (110X)	24	Did not know that the conductor cutting the magnetic field lines is equivalent to the power supply
	25	Did not know the terminal voltage
	26	Did not know that the uniform increase of B and $\delta B/\delta t$ have the same meaning, and mistook the former for the magnetic flux uniform increase
(FL1, FL2, FL3, FL4) = (1110)	27	Could not find the functional relationship between each physical quantity and t
	28	No in-depth understanding of the concept of acceleration
	29	Did not consider that the width of the field is larger than the edge of the wire
	30	Calculation error

EIP, Electromagnetic induction phenomenon: knowledge; CIC1, Conditions to generate induced current: knowledge; CIC2, Conditions to generate induced current: understanding; LL1, Lenz's Law: knowledge; LL2, Lenz's Law: understanding; LL3, Lenz's Law: application; RHR1, Right-hand rule: knowledge; RHR2, Right-hand rule: understanding; FL1, Faraday's law: knowledge; FL2, Faraday's law: understanding; FL3, Faraday's law: application; and FL4, Faraday's law: integrated application.

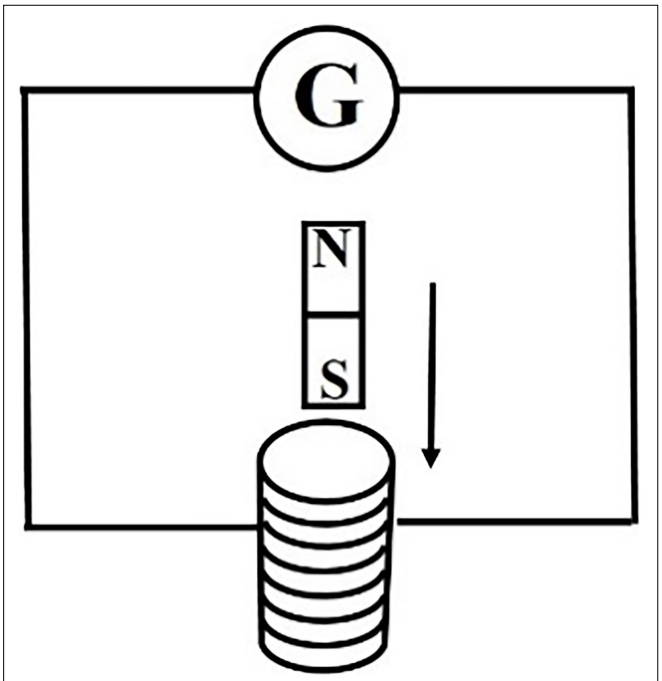


FIGURE 5 | Experimental circuit i of EIP = 0 level remedial teaching scheme.

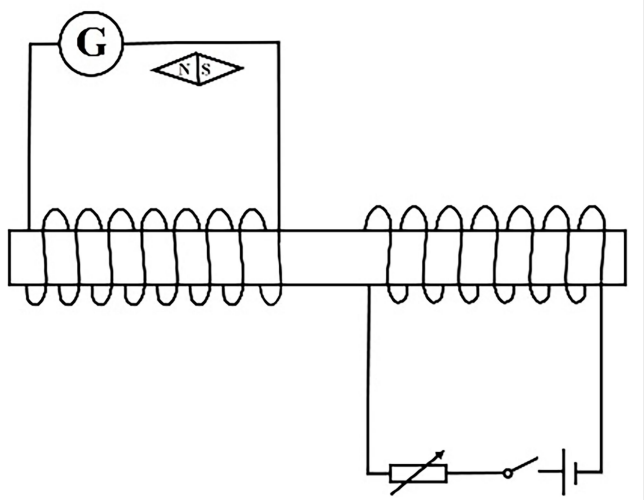


FIGURE 6 | Experimental circuit ii of EIP = 0 level remedial teaching scheme.

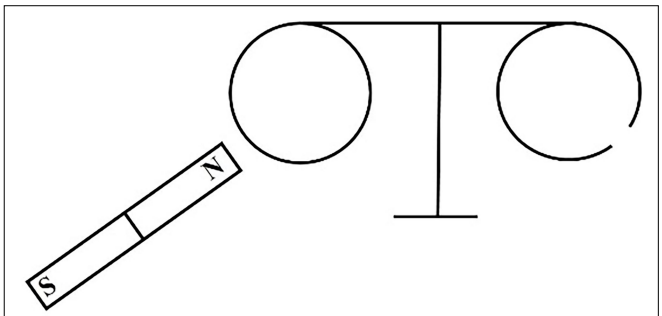


FIGURE 7 | Experimental circuit i of (LL1, LL2) = (10) level remedial teaching scheme.

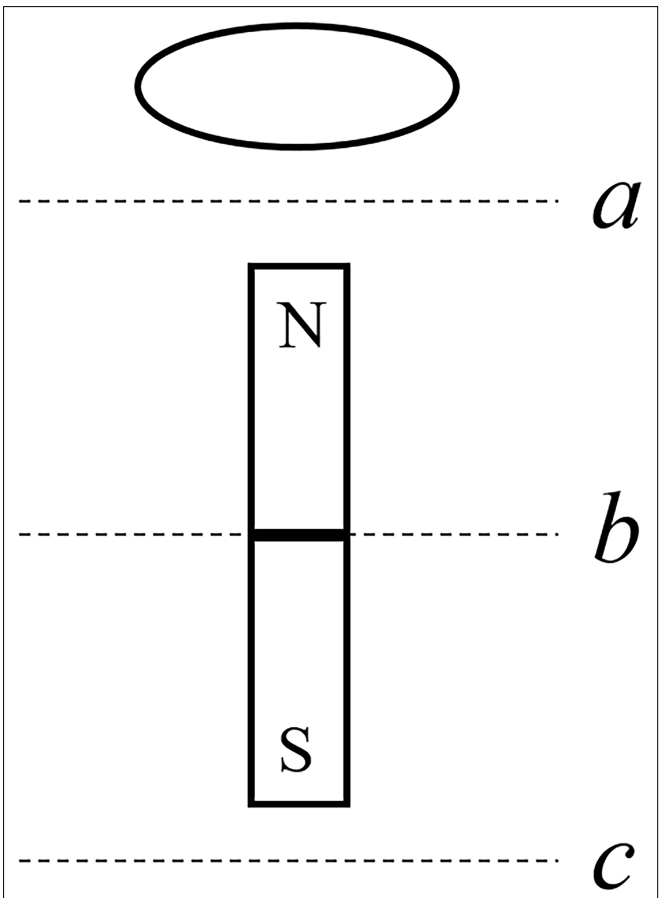


FIGURE 8 | The physical situation circuit of (LL1, LL2) = (10) level remedial teaching scheme.

[Discussion topic i] *Enter and leave the two aluminum rings in a bar magnet; which will generate the electromagnetic induction and why?*

[Teaching aim i] Corresponding to cognitive error 9 in **Table 2**: To enable students to judge the change of magnetic flux in a closed loop in a physical situation, so as to judge the direction of the induced current.

[Speculation ii : The physical situation is shown in **Figure 8**]

[Discussion topic ii] *Imagine the physical process of a speculative drawing shown in **Figure 8**. Hold the bar magnet and let the coil move. The induced current directions of the coil at positions a, b, and c are discussed. Is the force exerted on the coil at a and c by the bar magnet dynamic or resistance? Is the work done by the force of gravity on the coil fully converted to kinetic energy as if it were in free fall?*

[Teaching aim ii] Corresponding to cognitive error 10 in **Table 2**: To clarify the relationship between work and energy in electromagnetic induction.

[Discussion topic iii] *What is the fundamental reason for the use of the formula of "increase, reverse, decrease, same" and "come, go, refuse, and stay"?*

[Teaching aim iii] Corresponding to cognitive error 11 in **Table 2**: To develop an in-depth understanding of Lenz's law behind the formula.

(c) (LL1, LL2, LL3) = (110), (FL1, FL2, FL3, FL4) = (110X) or (1110) levels remedial teaching scheme

The results of the think-aloud protocols showed that students usually fail to reach the "application" and "integrated application" level when they lack further physics knowledge and are unable to correctly relate two or more physical processes and laws. For these students, the remedial teaching method involved the teaching assistant leading the students in a discussion of the corresponding problems in classifications so that the students may better understand the connection between the physical processes from the analysis of the physical processes and the applied physical laws. With this, long-term remedial teaching should strengthen the application of additional physical knowledge. To improve the learning of the students, the remedial teaching plan was designed as follows:

[Teaching method] Discussion.

[Teaching material] Questions at the corresponding level in Form A.

[Discussion topic] Each student explains the reasons for their choice through a speech, other students give their opinions, and one student summarizes the results of the discussion.

[Teaching assistant (TA) guidance] The TA guides the students in analyzing the physical process involved in the questions and the physical laws corresponding to the physical process. Then, they find a physical quantity that connects multiple physical processes.

Implementing Multi-level Remedial Teaching and Post-test

Participants

Among the seven schools that participated in the formal test, the authors selected three schools of different levels from Ministry of Education evaluations in the same city to conduct the experiment. School 1 (S1) was a private high school belonging to the lowest level in the local area; school 2 (S2) was a general public high school belonging to the middle level in the local area; and school 3 (S3) was a Level 1 high school, which is considered a high-level high school in the local area. In each school, two parallel classes were selected as the experimental group and control group. Experimental groups S1, S2, and S3 had 30, 52, and 48 participants, respectively, while the matching control groups had 29, 50, and 49 participants, respectively.

Multi-level Remedial Teaching Process

The students of the experimental groups and control groups were given 1 h to complete Form A. For the response matrix, the results of the experimental and control groups were then combined with those of the students who participated in the formal test of Form A. The DINA model was then used for cognitive diagnosis, and the experimental group was classified according to the method in stated in the "Diagnostic classification method" section. In the experimental group, researchers and teaching assistants completed multi-level remedial teaching according to the classification results, and each attribute mastery pattern classification group spent approximately 30 min learning. Each student in the experimental group could be classified into multiple groups and could receive remedial teaching for up to 2.5 h. In the control group, in order to minimize the influence of the number of students on the teaching effect, the students in the control group were randomly divided into a group of approximately 10 students and received approximately 2.5 h of instruction on the correct answers and solutions to the Form A questions. **Table 3** shows the process of the multi-level remedial teaching experiment.

Post-test

After remedial teaching was completed, students in both control and experimental groups spent 1 h simultaneously completing Form B. The response results were then combined with those of the students who participated in the formal test of Form B into the response matrix.

Data Analysis

The DINA model was used for cognitive diagnosis in post-test. We focused on whether students of each pre-test-based classification achieved the target attribute mastery pattern in the post-test, following the path in **Figure 4**. To test the effect of multi-level remedial teaching, the researchers defined a target attribute mastery pattern achievement rate index to determine whether post-test targets were reached. This rate was calculated as follows:

$$\eta_i = \frac{n_i}{N_i}, \tag{1}$$

where η_i is the achievement rate of the No. i attributes mastery pattern classification; N_i is the number of students in the No. i attributes mastery pattern classification in the pre-test; and n_i is the number of students who had achieved the target attribute mastery pattern in the post-test and were in the No. i attributes mastery pattern classification in pre-test.

To compare the differences in remedial teaching among the three schools, ANOVA was performed with the method of remedial teaching as the independent variable, while the students' target attribute mastery pattern achievement rate was considered as the dependent variable. Microsoft Excel 2016 was used to perform a single factor ANOVA.

RESULTS

Assessment and Classification

Quality of the EICD Test

Figure 9 shows the relationship between the EICD test CTT scores of the participants and the grades of the Academic Level

TABLE 3 | The process of the multi-level remedial teaching experiment.

Day-period	Experimental group				Control group		
Day1	Pre-test (Form A)						
Day2	There is no task for students, while researchers complete the classification						
	Classification group/remedial teaching	S1(N)	S2(N)	S3(N)	Remedial teaching method	S1(N)	S2(N)	S3(N)
Day3-1	EIP = 0	22	38	45	Explain the correct answers in Form A	8	13	13
Day3-2	(CIC1, CIC2) = (0X)	21	23	28				
Day3-3	(CIC1, CIC2) = (10)	3	29	22				
Day4-1	(LL1, LL2, LL3) = (0XX)	19	29	28	Explain the correct answers in Form A	7	13	12
Day4-2	(LL1, LL2, LL3) = (10X)	5	45	48				
Day4-3	(LL1, LL2, LL3) = (110)	6	8	7				
Day5-1	(RHR1, RHR2) = (0X)	17	11	3	Explain the correct answers in Form A	7	12	12
Day5-2	(RHR1, RHR2) = (10)	10	22	25				
Day5-3	(FL1, FL2, FL3, FL4) = (0XXX)	22	17	20				
Day6-1	(FL1, FL2, FL3, FL4) = (10XX)	5	26	24	Explain the correct answers in Form A	7	12	12
Day6-2	(FL1, FL2, FL3, FL4) = (110X)	2	18	6				
Day6-3	(FL1, FL2, FL3, FL4) = (1110)	0	7	20				
Day7	Post-test (Form B)							

Test. The grades of the Academic Level Test are A, B, and C from high to low, which are marked on the abscissa of **Figure 9**. The ordinate is the average score in the EICD test(formal) of the students who obtained the A, B, or C grade. The average score of the EICD (formal) test is basically linear and positively related to the Academic Level Test grades, indicating that the EICD test(formal) can reflect the real learning situation of students.

In order to explore whether Form A and B meet the requirements of parallel papers, **Table 4** presents the mean (*M*), standard deviation (SD), item difficulty (*P*), and item discrimination (*D*) of students' scores in Form A and B. It was found that the descriptive statistical indicators of the two Forms were basically the same. In addition, the one-way ANOVA with the test paper as the independent variable and the score as the dependent variable showed ($p = 0.286$) that the difference of the test paper had no significant effect on the total score. Finally, because the two Forms measured the same knowledge contents and learning level, containing the same number of questions, question types, and the same test Q-matrix, the Form A and Form B in this study can be considered to meet the requirements of parallel tests.

The cognitive diagnostic reliability of the 12 cognitive attributes of Forms A and B is shown in **Table 5**. The cognitive diagnostic reliability of all cognitive attributes is above 0.9, which indicates that the reliability of cognitive attributes was sufficient and that the test truly reflects the learning results.

Classification Results

The students participating in the formal test were classified based on the classification method in the **Figure 4**. The classification results are shown in **Table 6**.

As shown in **Table 6**, if the attribute mastery pattern is all 1 (i.e., 1111), it means that the student reached the highest learning level and does not need to participate in remedial teaching. The proportion of the other 12 classifications was 0.0702–0.3526, indicating that the classification scheme was able to basically stratify students according to the learning level of each knowledge content. However, the distinction between FL3 and FL4 was not obvious. In the attribute mastery pattern related to Faraday's law (FL1, FL2, FL3, and FL4), the proportion of 1111 was 0.0702, while the proportion of 1111 was 0.3263. This indicates that if students are able to achieve the "Faraday's Law: Application" level, most of them would also be able to reach the level of "Faraday's Law: Integrated Application." In solving Faraday's law, most students who were able to solve two physical processes would also be able to solve more than two physical processes and correlate other knowledge problems.

Effect of Multi-level Remedial Teaching

After receiving multi-level remedial teaching (experimental group) and the explanation of the answers to the test questions (control group), both groups completed Form B as the post-test.

Figures 10A–D show the target attribute mastery pattern achievement rate of 12 attribute mastery pattern classifications in the experimental group and control group in S1, S2, and S3, and the average of the three schools, respectively. **Figure 10D** shows that both multi-level remedial teaching and the traditional method of explaining answers are able to improve the learning level of students in 12 classifications, but multi-level remedial teaching was found to be more effective and more efficient. The target achievement rates of the experimental group and the control group of the three schools also had their own characteristics. It should be mentioned that we focused on the classification where the experimental group had a greater than 20% target achievement rate compared to the control group. The effect of multi-level remedial teaching on the level of "knowledge" in school S1 was more obvious, but in school S2 and school S3, the difference between the experimental group and the control group of "electromagnetic induction phenomenon: knowledge" and "conditions for generating induced current: knowledge" was not as obvious. The experimental group in the (LL1, LL2, LL3) = (10X) and (110), (RHR1, RHR2) = (0X) and (10), (FL1, FL2, FL3, FL4) = (0XXX), (110X) and (1110) categories of the school S2 had a more obvious effect than the control group. In addition, the experimental group of (RHR1, RHR2) = (0X) and (10), (FL1, FL2, FL3, FL4) = (1110) of school S3 had more obvious effects of remedial teaching. Therefore, these results may be related to the varied levels of the schools.

Since the three schools were of different levels and had large differences between them, the ANOVA of multi-level remedial teaching was not analyzed (**Figure 10**).

To compare the differences in remedial teaching among the three schools, ANOVA was performed with the method of remedial teaching as the independent variable. The students' target attribute mastery pattern achievement rate was considered as the dependent variable. The results are shown in **Table 7**.

For the seniors at schools S1 ($F = 4.77 > F\ crit = 4.41$, $p = 0.0424 < 0.05$) and S3 ($F = 7.80 > F\ crit = 4.301$, $p = 0.011 < 0.05$), **Table 7** reports that different remedial teaching method led to significant differences in the target attribute mastery pattern achievement rate. However, for school S2, the effect of remedial teaching method was not significant ($F = 1.77 < F\ crit = 4.30$, $p = 0.197 > 0.05$), indicating that there was no significant difference in the target level achievement rate among participants with different remedial methods.

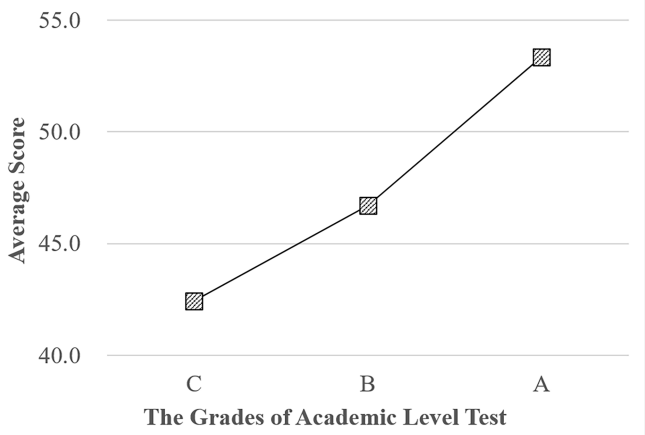


FIGURE 9 | The relationship between the EICD test CTT scores and the grades of the Academic Level Test. The ordinate is the average score in the EICD test(formal) of the students who obtained the A, B, or C grade.

TABLE 4 | The statistical description results of Form A and B.

	M + SD	*P*	*D*
Form A	59.3 + 18.3	0.58	0.48
Form B	61.3 + 21.9	0.60	0.48

M, Mean; SD, Standard deviation; P, difficulty; and D, discrimination.

DISCUSSION AND CONCLUSION

In order to study the effect of multi-level remedial teaching for students, we used the CDA method to develop the EICD test and to carry out the multi-level classification for students on "electromagnetic induction"—a topic that is a part of the usual high school physics curriculum. The EICD test reliability was high, and the consistency of Forms A and B was good; thus, these could be effectively used for the multi-level classification of the students. As for the cognitive errors analyzed from the think-aloud protocols, remedial teaching schemes for the different classified levels were then designed. The results of the experiment showed that the multi-level remedial teaching design used in this study was more effective for the improvement of students' learning than the traditional way. In a region, the effect of CDA-based multi-level remedial teaching in promoting learning was found to be more significant in high-level and low-level schools, but not in secondary schools. Although this study only used the multi-level remedial teaching design method for one topic in high school physics, the developed multi-level remedial teaching method of assessment, classification, and remedial teaching design can theoretically be used for any high school physics topic. The method designed in this study does not rely on purchasing hardware and software and costs an acceptable amount. For teachers, all progress requires a significant amount of time.

TABLE 5 | The reliability of cognitive attributes.

	EIP	CIC1	CIC2	FL1	FL2	FL3	RHR1	RH2	FL1	FL2	FL3	FL4	Mean
Form A	0.9766	0.9847	1	0.9797	1	0.9829	1	0.9908	0.9972	0.9961	1	1	0.9923
Form B	0.9714	0.9649	0.9311	0.9369	1	0.9482	0.9734	0.9624	0.9896	0.9772	0.9812	0.9851	0.9684

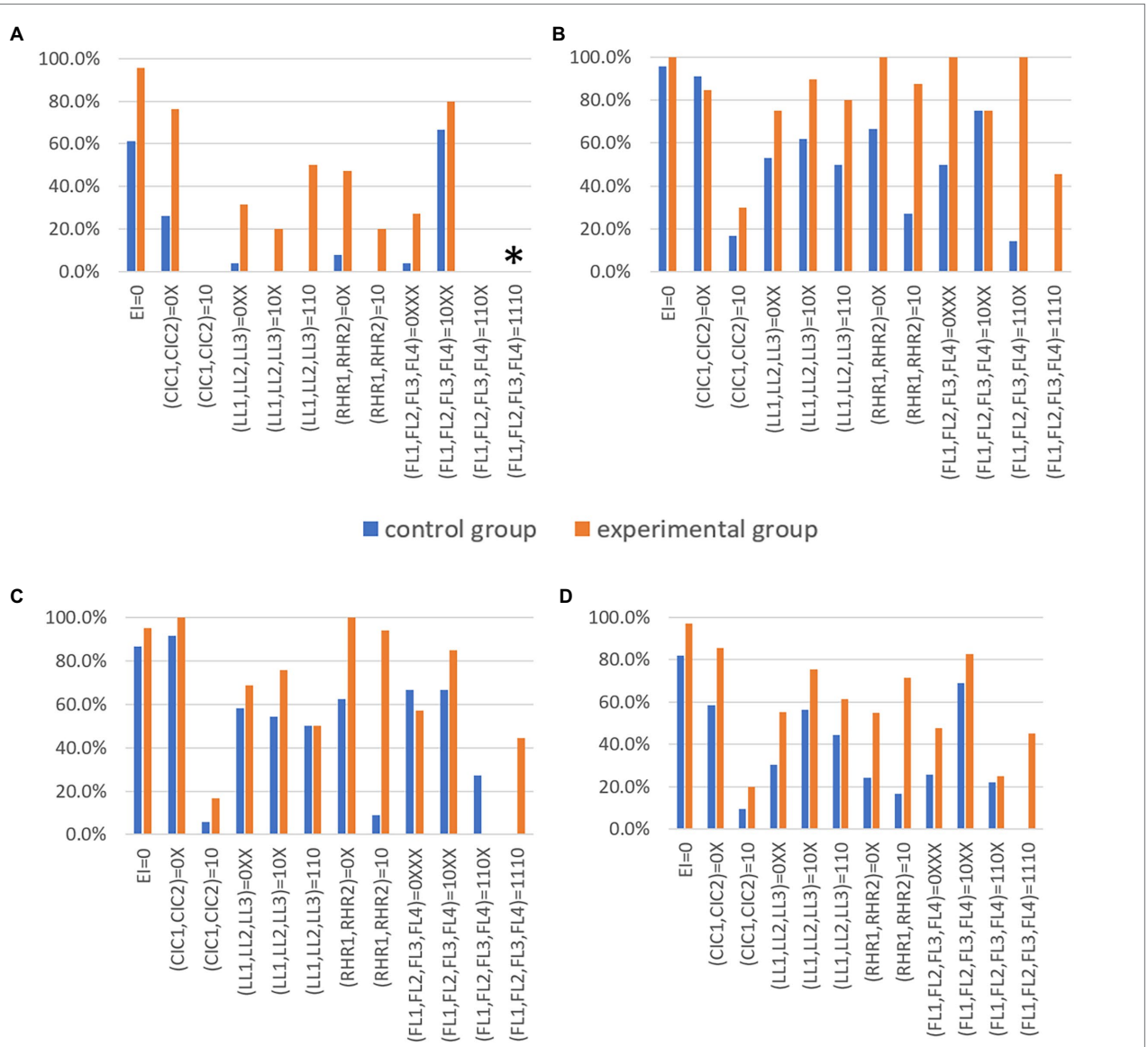


FIGURE 10 | Comparison of the achievement rate of 12 target attribute mastery patterns between the control and experimental groups in schools S1, S2, and S3. **(A)** is a bar graph comparing the achievement rate of 12 target levels between the control and experimental groups of participants at School S1. The achievement rate of (FL1,FL2,FL3,FL4) = (1110) was 0 because no student had mastered the attributive FL3, so no students should regard (FL1,FL2,FL3,FL4) = (1110) as the target attribute mastery. **(B,C)** show seniors at Schools S2 and S3. **(D)** Shows a bar comparison of the average achievement rate of 12 target levels between the control and experimental groups in Schools S1, S2, and S3.

However, it enables students to study more efficiently, which benefits teaching efficiency. To further promote teaching efficiency, the completed teaching scheme can be reused or revised for multiple classes.

Limitations

With the continuous development of research on CDMs, more refined and improved performance models are being developed by researchers, such as a series of longitudinal diagnostic classification models developed by Zhan et al. wherein attribute-level growth may be quantified in a more refined manner (Zhan et al., 2019a, Zhan, 2021; Zhan and He, 2021). A mixed model method can also be considered so that different items in a test are calculated using different models. This is one of the limitations of this study; thus, a model with better fit should be tried in future studies. Second, in order to analyze the cognitive errors of students learning about electromagnetic induction, a limited amount of think-aloud material was collected given the limited time. Hence, it is highly likely that not all cognitive errors were covered. Therefore, teachers on the frontline of teaching may be able to accumulate more comprehensive and accurate cognitive errors during the teaching process, which can then be used to design other multi-level remedial teaching

TABLE 6 | Classification results.

Attributes group	Attribute mastery pattern	Proportion
EI	0	0.3082
	1	0.6918
CIC1, CIC2	0X	0.2509
	10	0.3526
	11	0.3965
LL1, LL2, LL3	0XX	0.2772
	10X	0.2374
	110	0.1275
	111	0.3579
RHR, RHR2	0X	0.2076
	10	0.1906
	11	0.6018
FL1, FL2, FL3, FL4	0XXX	0.2509
	10XX	0.1380
	110X	0.2146
	1111	0.0702
	1111	0.3263

Table 6 shows the classification result of the formal cognitive diagnostic tests. The first column is the attributes group. The vectors in the second column represent the attribute mastery patterns at the same group of knowledge contents. X means 0 or 1. The data in the third column are the proportion of the classified population. EIP, Electromagnetic induction phenomenon: knowledge; CIC1, Conditions to generate induced current: knowledge; CIC2, Conditions to generate induced current: understanding; LL1, Lenz's Law: knowledge; LL2, Lenz's Law: understanding; LL3, Lenz's Law: application; RHR1, Right-hand rule: knowledge; RHR2, Right-hand rule: understanding; FL1, Faraday's law: knowledge; FL2, Faraday's law: understanding; FL3, Faraday's law: application; and FL4, Faraday's law: integrated application.

TABLE 7 | ANOVA results of S1, S2, and S3.

Source of difference	School	SS	df	MS	*F*	value of *p*	F crit
Remedial measures	S1	0.388	1	0.388	4.77	0.0424*	4.41
	S2	0.181	1	0.181	1.77	0.197	4.30
	S3	0.557	1	0.557	7.80	0.011*	4.301

*SS, Sum of squares, df, degree of freedom, MS, Mean square, F, Test statistic, and F crit, Critical value of the test. *p < 0.05.*

plans. Third, the classification of students in this study was limited to knowledge content and learning level, so future research should focus on improving the cognitive structure and developing potential traits as cognitive attributes.

Although the scheme of multi-level remedial teaching in this study revolved around the topic of electromagnetic induction in high school physics, the method of developing the design scheme may be applied to other topics and other subjects as well. In terms of teaching, the program can be constantly revised and improved to further enhance the teaching effect.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplemental Material,** further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Committee on Human Research Protection, East China Normal University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

RH and SP conceived the study. RH developed the cognitive model and the test items. DZ, QH, and RH conducted the think-aloud protocols. DZ, QH, and RH conducted the teaching experiment and collected the test results. RH analyzed the data and wrote the manuscript. ZL and RH revised the manuscript. SP provided technical advice. All authors contributed to the article and approved of the submitted version.

FUNDING

This work was supported by the National Education Science Planning Key Projects of the Ministry of Education under Grant DHA170352.

ACKNOWLEDGMENTS

We would like to thank Dongbo Tu's team for designing the cognitive diagnosis data analysis platform (http://www.psychometrics-studio.cn), which helped us complete the data processing. We would also like to thank Editage for English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.851378/full#supplementary-material

REFERENCES

Anderson, L. W., and David, R. K. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Boston, MA: Allyn & Bacon (Pearson Education Group).

Cai, Y., Tu, D., and Ding, S. (2013). A simulation study to compare five cognitive diagnostic Models. *Acta Psychol. Sin.* 45, 1295–1304. doi: 10.3724/SP.J.1041.2013.01295

Fan, T., Song, J., and Guan, Z. (2021). Integrating diagnostic assessment into curriculum: a theoretical framework and teaching practices. *Lang. Test. Asia* 11, 1–23. doi: 10.1186/s40468-020-00117-y

Groß, J., Robitzsch, A., and George, A. C. (2016). Cognitive diagnosis models for baseline testing of educational standards in math. *J. Appl. Stat.* 43, 229–243. doi: 10.1080/02664763.2014.1000841

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26, 301–321. doi: 10.1111/j.1745-3984.1989.tb00336.x

He, H., Zhang, H., and Luo, H. (2016). Design of multi-level teaching system based on association rule mining. *Int. J. Emerg. Technol. Learn.* 11, 43–48. doi: 10.3991/ijet.v11i11.6252

Junker, B. W., and Klaas, S. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Lee, Y., Park, Y. S., and Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. National Sample Using the TIMSS 2007. *Int. J. Test.* 11, 144–177. doi: 10.1080/15305058.2010.534571

Liu, H., You, X., Wang, W., Ding, S., and Chang, H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an english achievement test in China. *Classif.* 30, 152–172. doi: 10.1007/ s00357-013-9128-5

Macready, G. B., and Mitchell, D. C. (1977). The use of probabilistic models in the assessment of mastery. *J. Educ. Stat.* 2, 99–120. doi: 10.2307/1164802

Ren, H., Xu, N., Lin, Y., Zhang, S., and Yang, T. (2021). Remedial teaching and learning from a cognitive diagnostic model perspective: taking the data distribution characteristics as an example. *Front. Psychol.* 12:628607. doi: 10.3389/fpsyg.2021.628607

Shute, V. J., Hansen, E. G., and Almond, R. G. (2008). You can't fatten a hog by weighing it-or can you? Evaluating an assessment for learning system called ACED. *Int. J. Artif. Intell. Educ.* 18, 289–316.

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989X.11.3.287

Templin, J., and Laine, B. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *J. Classif.* 30, 251–275. doi: 10.1007/s00357-013-9129-4

Wang, L., Tang, F., and Zhan, P. (2021). Effect analysis of individualized remedial teaching based on cognitive diagnostic assessment: taking "linear equation with one unknown" as an example. *J. Psychol. Sci.* 2020, 1490–1497. doi: 10.16719/j.cnki.1671-6981.20200630

Wu, H. (2019). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. *Educ. Psychol.* 39, 1218–1232. doi: 10.1080/01443410.2018.1494819

You, X., Li, M., and Liu, H. (2019). The feedback of the Chinese learning diagnosis system for personalized learning in classrooms. *Front. Psychol.* 10:1751. doi: 10.3389/fpsyg.2019.01751

Zhan, P. (2021). Refined learning tracking with a longitudinal probabilistic diagnostic model. *Educ. Meas. Issues Pract.* 40, 44–58. doi: 10.1111/emip.12397

Zhan, P., and He, K. (2021). A longitudinal diagnostic model with hierarchical learning trajectories. *Educ. Meas. Issues Pract.* 40, 18–30. doi: 10.1111/emip.12422

Zhan, P., Jiao, H., Liao, D., and Li, F. (2019a). A longitudinal higher-order diagnostic classification model. *J. Educ. Behav. Stat.* 44, 251–281. doi: 10.3102/1076998619827593

Zhan, P., Yu, Z., Li, F., and Wang, L. (2019b). Using a multi-order cognitive diagnosis model to assess scientific literacy. *Acta Psychol. Sin.* 51, 734–746. doi: 10.3724/SP.J.1041.2019.00734

Zhang, Y., Lin, J., and Zhang, H. (2016). A hierarchical teaching mode of college computer basic application course based on K-means and improved PSO algorithm. *Int. J. Emerg. Technol. Learn.* 11, 53–58. doi: 10.3991/ijet.v11i10.5909

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 © 2022 Huang, Liu, Zi, Huang and Pan. 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.

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