frontiers | Research Topics

Insights in assessment, testing, and applied measurement 2022

Edited by
Gavin T. L. Brown

Published in
Frontiers in Education



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 978-2-8325-5497-5
DOI 10.3389/978-2-8325-5497-5

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

Insights in assessment, testing, and applied measurement: 2022

Topic editor
Gavin T. L. Brown — The University of Auckland, New Zealand

Citation
Brown, G. T. L., ed. (2024). *Insights in assessment, testing, and applied measurement: 2022*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5497-5

Table of contents

05 Editorial: Insights in assessment, testing, and applied measurement: 2022
Gavin T. L. Brown

08 The cross-cultural suitability analysis of "the Educator Cognitive Sensitivity scale": Empirical exploration from early childhood teachers
Beibei Lu, Dezhi Chen and Congcong Yan

16 Why are certain items more difficult than others in a competency test for music–related argumentation?
Julia Ehninger, Jens Knigge and Christian Rolle

28 Effect of feedback through dynamic assessment on EFL field-dependent and field-independent learners' speaking skill development
Reza Kafipour and Ali Khoshnood

38 Synchronous self-assessment: First experience for higher education instructors
Ana Remesal and Flor G. Estrada

53 Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence
Prosper Kissi, David Baidoo-Anu, Eric Anane and Ruth Keziah Annan-Brew

67 A data-driven multidimensional assessment model for English listening and speaking courses in higher education
Shuwei Xue, Xin Xue, Ye Jun Son, Yaxuan Jiang, Hang Zhou and Shifa Chen

78 Safe and sound: examining the effect of a training targeting psychological safety and trust in peer assessment
Morgane Senden, Dominique De Jaeger and Liesje Coertjens

91 Teacher assessment literacy: a systematic review
Serafina Pastore

116 Challenges and opportunities in score reporting: a panel of personal perspectives
Gavin T. L. Brown, Priya Kannan, Sandip Sinharay, Diego Zapata-Rivera and April L. Zenisky

126 The GRE as a predictor of persistence to a PhD
Brent Bridgeman, Margarita Olivera-Aguilar and Steven Holtzman

135 Factor structure and invariance of the scale to measure teaching performance in the area of social sciences
Patricio Sebastián Henríquez, Juan Carlos Pérez-Morán, Carlos Javier del Cid García and Jesús Enrique Zamora

148 The prevalence of careless response behaviour and its consequences on data quality in self-report questionnaires on student learning
Margot Chauliac, Jonas Willems, David Gijbels and Vincent Donche

161 Predicting formative feedback practices: improving learning and minimising a tendency to ignore feedback
Gavin T. L. Brown, Catarina Andersson, Mikael Winberg and Torulf Palm

171 The dimensionality of the epistemic orientation survey and longitudinal measurement invariance for the short form of EOS (EOS-SF)
Chenchen Ding, Gavin Fulmer, Lesa Hoffman, Brian Hand and Jee K. Suh

181 Unveiling cognitive aspects and accuracy of diagnostic judgments in physical education teachers assessing different tasks in gymnastics
Christian Leukel, Timo Leuders, Flavio Bessi and Katharina Loibl

195 A review of Lawshe's method for calculating content validity in the social sciences
Marcela Romero Jeldres, Elisabet Díaz Costa and Tarik Faouzi Nadim

203 Challenges and opportunities for classroom-based formative assessment and AI: a perspective article
Therese N. Hopfenbeck, Zhonghua Zhang, Sundance Zhihong Sun, Pam Robertson and Joshua A. McGrane

212 The nature of historical knowledge in large-scale assessments – a study of the relationship between item formats and offerings of epistemic cognition in the Swedish national test in history
David Rosenlund

227 Caring assessments: challenges and opportunities
Jesse R. Sparks, Blair Lehman and Diego Zapata-Rivera

235 Note on the radical inflation in the estimates of error variance in measurement models
Jari Metsämuuronen

frontiers | Frontiers in Education

TYPE Editorial
PUBLISHED 18 September 2024
DOI 10.3389/feduc.2024.1488012

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Raona Williams,
Ministry of Education, United Arab Emirates

*CORRESPONDENCE
Gavin T. L. Brown
✉ gt.brown@auckland.ac.nz

RECEIVED 29 August 2024
ACCEPTED 06 September 2024
PUBLISHED 18 September 2024

CITATION
Brown GTL (2024) Editorial: Insights in assessment, testing, and applied measurement: 2022. *Front. Educ.* 9:1488012. doi: 10.3389/feduc.2024.1488012

COPYRIGHT
© 2024 Brown. 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.

Editorial: Insights in assessment, testing, and applied measurement: 2022

Gavin T. L. Brown*

Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand

KEYWORDS

testing, psychometrics, educational assessment, evaluation, methods

Editorial on the Research Topic
Insights in assessment, testing, and applied measurement: 2022

Introduction

This Research Topic focused on new insights, novel developments, current challenges, recent advances, and future perspectives in the field of assessment in education. The goal was to shed light on the progress made in the past decade in the assessment, testing and applied measurement field and on its future challenges to provide a thorough overview of the state of the art of the assessment, testing and applied measurement field. Measurement, assessment, testing, and various classroom or testing protocols matter to quality and justice. These processes are used to inform decision making but as we enter the third decade of the twenty-first century there is increasing complexity in the world in which assessment functions. That means what we already know may not be a good basis for future action, policy, or practice. Consequently, we solicited brief, forward-looking contributions from Assessment, Testing and Applied Measurement editorial board members that either described the state of the art or highlighted changes needed to move the field forward. We expected these authors, based on their contribution to the journal through editing and reviewing manuscripts, let alone their own research agendas, to identify the greatest challenges in their sub-disciplines, and how to address those challenges.

This article Research Topic will inspire, inform, and provide direction and guidance to researchers in the field. From 2022 to 2024, a total of 20 manuscripts were added to the Research Topic. Seven of the papers involved students, nine focused on teachers, and four spoke to concerns of researchers and policymakers. Students and teachers at various levels of the K-12 compulsory school systems were the focus of 12 papers and six papers focused on students at various levels of tertiary or higher education. As an aside, my thanks go out to the many reviewers and editors who helped the authors create good papers. The quality of this RT depended on those folk.

Insights

Two of the papers addressed to researchers were highly technical expositions related to Lawshe's content validity index (Jeldres et al.) and error variance inflation in measurement models (Metsämuuronen), both of which should benefit psychometric researchers needing authoritative sources on those methods.

Similarly, two papers provided review or overview perspectives. Brown, Kannan et al. provide a discursive set of opinions and perspectives about how test developers can better communicate test results to teachers, administrators, and other educational stakeholders. While potentially somewhat repetitive, the voices of five different experts, with varied approaches and contexts, give strong suggestions for future research. Pastore provides a systematic review of the literature on teacher assessment literacy for the most recent 10-year period (2013–2022). Pastore shows that the field has wide variation in how this core classroom teacher competence is defined, understood, and studied; a clear example of the perpetual problem of "jingle-jangle" in educational and psychological research. Nonetheless, Pastore reports that there are foundational components which are contingent upon contextual factors, reinforcing results from previous reviews.

Two other studies highlighted aspects of teacher assessment competence. Kissi et al. demonstrated the weaknesses Ghanaian teachers had in terms of creating multiple-choice test questions, using a test of item quality and an analysis of actual test forms created by the teachers. Their multimethod study showed, despite well- and long-established guidelines for writing good test questions, teachers could not recognize or create consistently high-quality objectively scored test items. Leukel et al. examined how teachers of gymnastics form quality judgements of student performance compared to more expert trainers. Their study found judgment accuracy, agreement on ratings, and agreement about the temporal structuring of tasks was significantly lower for teachers compared to trainers. Expertise in any domain being evaluated leads to better judgements and grading. Given that teacher assessment literacy is a very broad multifaceted competency, it is highly likely that teacher-made assessments, judgements, or feedback will remain problematic for a long time to come.

Five papers reported scale development or validation studies, drawing on data from Mexico, Sweden, China, USA, and Iran. The studies used complex statistical methods, including multi-group confirmatory factor analysis invariance testing (Henríquez et al.), content analysis of test items (Rosenlund), exploratory factor analysis, hierarchical regression, and multilevel modeling (Lu et al.), WLSMV estimation of longitudinal item factor analysis with invariance testing (Ding et al.), and confirmatory factor analysis and structural equation modeling (Brown, Andersson et al.).

A wide variety of Research Topics have been captured by these studies. Rosenlund examined the epistemic cognition requirements of large-scale tests of history in Sweden, finding an over-emphasis on objective dimensions of historical knowledge, challenging the design of future tests to better evaluate all epistemic requirements of the subject. Henríquez et al. evaluated an inventory with Mexican students for student evaluation of higher education teaching in the social sciences. They reported invariance and good model fit for a three-factor model (i.e., course organization, teaching quality, and evaluation and feedback) of teachers' performance. Lu et al. tested the cross-cultural reliability and validity of a scale concerning educator cognitive sensitivity with a sample of Chinese early childhood educators, concluding that the validation evidence was weak and necessitated further work. Brown, Andersson et al. tested a previously published measure of teacher conceptions of feedback in Sweden and modified it by proposing some of the items constituted a factor of teacher feedback practices. This model had good fit and showed that endorsement of feedback for improvement and that students may ignore feedback both contributed to the feedback practices teachers claimed to make. These studies identify and support further use of measures for practice, research, and possibly even with teacher professional development.

Bridgeman et al.'s analysis of the relationship between test scores for entry into graduate higher education (i.e., Graduate Record Examination, GRE) and doctoral degree completion used multilevel analysis to show that greater persistence was associated with higher verbal and analytical writing scores and inversely with quantitative scores. Despite the odds ratio values being close to 1.00 (grand average = 1.03), the authors recommend keeping GRE scores in the decision matrix, a recommendation that should be taken cautiously.

Innovations

The world of test validity and academic integrity is being threatened by uncontrolled use of AI or LLM technologies, so we need to have insights that might lead to effective use of such technologies in actual teaching and learning practices, let alone its potential to validly create or reliably score high-stakes, large scale assessments. Justice in society is ensured when mean score differences at individual or group levels are supported by well-designed assessments and systems that consider differential opportunity to learn. These are important messages for policymakers and politicians who have responsibility for the design of education systems.

Innovative views of the future of assessment included a paper on how AI can be used in formative assessment (Hopfenbeck et al.) and how AI can be used to design assessments that do not create oppressive outcomes for minority students (Sparks et al.). Hopfenbeck et al. identify ways that AI might be used in classroom contexts (e.g., automated essay scoring, generating feedback, and generating learner profiles and subsequent automated tutoring). However, they point out teachers lack skills to exploit these innovative possibilities and there is considerable work needed to turn the potential of AI into actual formative practices. Sparks et al. have conceived a framework for how AI can be used to develop personalized classroom assessment. Their framework adapts assessment processes to provide "care" for learners before, during, and after testing. The aim is to ensure that contextual information about learners, including their personal characteristics and ways of behaving, are incorporated into the design and administration of assessment. These papers offer visions of how AI can be used to improve the quality of classroom assessments, but this still remains a major challenge for systems.

In terms of new testing or assessment protocols, Kafipour and Khoshnood demonstrated that dynamic assessment in which the instructor assesses student language performance by asking questions and providing hints or prompts had a positive impact on Field Dependent EFL learners in Iran. Field dependence is a cognitive style in which the learner focuses on the overall meaning and the whole field, exhibits more relational behaviors, and needs

more external reinforcements to stay motivated. This alignment makes sense and raises interesting challenges for those of us who rely on traditional assessment processes. Perhaps, AI machines can be programmed to interact with language learners in assessments and reduce workload on teachers?

Remesal and Estrada present a small-scale study of Spanish teacher educators who used an innovative synchronous self-assessment strategy during written exam situations (i.e., during the exam, students select the tasks or questions they will answer and they choose a weighted grading scheme for their successful answers). The four instructors were individually interviewed after the examination and they claimed marking was less tedious because students did not all do the same tasks and that the different weighting choices provided clues to teachers about student competence, potentially informing more effective instruction.

Instead of the traditional focus on determining cognitive difficulty for mainstream subjects, Ehninger et al. focused on predicting the cognitive difficulty of items in a test of music-related argumentation (i.e., MARKO). They found among German high school students that the strongest predictor of harder test questions was "reference to musical attributes," "cross-sentence argumentation," and "dialogical argumentation" features. The study provides validation evidence for the MARKO test and may provide a model for testing of other creative and/or performing arts.

Greater use of peer feedback is advocated, especially in higher education, on the assumption that this helps both feedback recipient and provider. However, lack of psychological safety in the process will lead to faulty communication. Senden et al. created a brief student training program to increase psychological safety and trust in peer feedback, but their experiment with Belgian higher education students in acrobatic sports didactics failed to find any statistically significant effect for safety or grade improvement. Nonetheless, researchers and instructors might want to inspect the treatment design to identify ways in which their own work might improve results.

Xue et al. used artificial neural network (ANN) analysis to predict academic performance in English listening and speaking as a Foreign Language among Chinese university students. Despite a very small sample size ($n = 62$), the data driven ANN found that overall performance seemed to depend on academic performance on the Chinese college entrance examination English test (gao kao), average scores of all peers' assessment covering English abilities, class participation, cooperation and competitiveness, and learning attitude and perseverance, standardized teacher ratings and student self-assessment. Clearly, the results are greatly limited by sample size, but it is encouraging that the ANN system worked with such small numbers, perhaps because there were so many variables per student.

Unsurprisingly, Chauliac et al. studied five different approaches to determining if survey responses are "careless." They reported that notable proportions of Flemish adolescent students (age 15–17) exhibited careless responding when completing self-report surveys (rate ranged from 12 to 31% depending on method). Carelessness clearly mattered to the quality of data. Interestingly, the method of determining carelessness matters, because few participants would be eliminated by two or more methods. Hence, the study provides new options for researchers as to how they might determine whether participants were attentive or not.

Conclusion

I commend this set of papers as a useful contribution to the field. They provide both insights to current methods or findings and innovations concerning the future of assessment, testing, and applied measurement.

Author contributions

GB: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 25 November 2022
DOI 10.3389/feduc.2022.1055847

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Pedro Gil-Madrona,
University of Castilla-La Mancha, Spain
Lihong Song,
Jiangxi Normal University, China

*CORRESPONDENCE
Dezhi Chen
cdezhi@zjnu.cn

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 28 September 2022
ACCEPTED 28 October 2022
PUBLISHED 25 November 2022

CITATION
Lu B, Chen D and Yan C (2022) The cross-cultural suitability analysis of "the Educator Cognitive Sensitivity scale": Empirical exploration from early childhood teachers. *Front. Educ.* 7:1055847. doi: 10.3389/feduc.2022.1055847

COPYRIGHT
© 2022 Lu, Chen and Yan. 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.

The cross-cultural suitability analysis of "the Educator Cognitive Sensitivity scale": Empirical exploration from early childhood teachers

Beibei Lu, Dezhi Chen* and Congcong Yan

Hangzhou Preschool Teachers College, Zhejiang Normal University, Hangzhou, Zhejiang, China

The applicability of *the Educator Cognitive Sensitivity (ECS) scale* is an important prerequisite for promoting teachers'cognitive sensitivity concepts and research. For this reason, we have expanded the study of the measurement properties of the *ECS scale* in the Chinese context of early childhood teachers and promote the development and research of cognitive sensitivity of teachers in the context of Chinese culture. The scale was used to evaluate 100 Early childhood teachers from a province in eastern China. The results showed that the internal consistency of the scale was good. The structural validity analysis results of the scale were similar to the existing research results; taking some items in the curriculum promotion of "*The Path Towards Excellence—Chinese Kindergarten Education Quality Rating Standards*" and children's development scale as the target, the empirical validity analysis results showed that the empirical validity of the scale was not very ideal, which needs further practical exploration in the future.

KEYWORDS

early childhood teachers, cognitive sensitivity, cross-cultural suitability analysis, measurement properties, scale

Introduction

The Educator Cognitive Sensitivity (ECS; Pauker et al., 2018) refers to the ability of educators to create an environment of cognitive stimulation when interacting with a less experienced partner while adapting to the children's inner state, including cognition and emotion. It includes three aspects: back-and-forth interaction, understanding children's thoughts and feelings, and speaking to children using the language they can understand (Landry et al., 1996, 2000; Fox and Hane 2008; Laranjo et al., 2010; Prime et al., 2014, 2016; Pauker et al., 2018). Previous studies had shown that the cognitive sensitivity of kindergarten educators is of great significance to the development of children's cognition, social interaction, sense of achievement and self-efficacy. For example, some studies believed that teachers with high cognitive

sensitivity would provide a more positive atmosphere and high-quality teaching, and children would have more opportunities and support for language and mathematics learning, which was more conducive to their academic growth (Glaser 2000; Hoff 2006; Burchinal et al., 2008). Some studies had shown that the higher the teachers' cognitive sensitivity, the better they were at understanding children's thoughts and feelings, and the more correctly they can assess children's emotional status and gave positive emotional support;this repeated correct emotional measurement can also help children become aware of their own and others' psychological states, and even manage their own behavior, and promote children's social interactions (Dunn 2002; Meins et al., 2002; Carpendale and Lewis 2004; Lundy 2013).Cognitively sensitive teachers could provide timely cognitive and emotional support and assistance to children, and establish positive mutual partnerships with children, promoting children's cognitive and social communication skills while also effectively promoting the development of a sense of achievement and self-efficacy (Bandura 1997; Bernier et al., 2010).

In practice, to explore the cognitive sensitivity of early childhood educators, measurement is an important method and path. At present, the method of observation is mainly adopted in the evaluation of the cognitive sensitivity of educators. The *ECS scale* (Pauker et al., 2018) is one of the most widely used assessment tools. A previous study (Pauker et al., 2018) had been conducted by assessing 350 infant and toddler educators from early childhood care and educational institutions, and conducting reliability and validity analysis based on the assessment results, showing that the scale had good internal consistency (Cronbach's α coefficient value of 0.96) and Inter-rater reliability (0.85); at the same time, Pauker et al., (2018) also used Exploratory Factor Analysis (EFA) to explore the structural validity, and the results of the study concluded that the scale has a single-dimensional structure. In terms of concurrent validity, this scale was related to multiple dimensions of *the Classroom Assessment Scoring System (CLASS,* La Paro et al., 2012) except for negative atmosphere, correlation coefficients ranging from 0.41 to 0.55, and all are significant. And the correlation coefficients with *the Infant/Toddler Environment Rating Scale-Revised (ITERS-R,* Harms et al., 2003*),* were 0.21–0.40 and significantly moderately correlated (Pauker et al., 2018).

The existing studies were all based on the Canadian cultural background of Pauker et al., (2018). In China, awareness and research on teachers' cognitive sensitivities have not yet begun, and the results of related research have hardly been reported. The cross-cultural applicability of the *ECS scale* is an important prerequisite for promoting teachers' cognitive sensitivity concepts and research. Therefore, we expand the study of the measurement properties of the *ECS scale* in the Chinese context of kindergarten educators and promote the development and research of teachers' cognitive sensitivity in the context of Chinese culture.

Materials and methods

Participants

First of all, 100 early childhood teachers from 50 classes were randomly sampled from two cities in an eastern province of China by stratified sampling ($M = 27.5$,SD = 5.2). The 50 classes included 12 pre-school classes (age 3–4), 19 pre-kindergarten (age 4–5) and 19 kindergarten classes (age 5–6) respectively. Secondly, six children (3 males and 3 females) were randomly selected from each sample class to participate in the development evaluation of children's language and mathematics. As a result, a total of 300 young children participated in the child development assessment. After excluding missing and invalid samples, among the 300 young children, the number of valid samples participating in the language assessment was 262 ($M = 4.6$, SD = 0.9), and that in mathematics evaluation was 245 ($M = 4.7$, SD = 0.8).

Measurement instruments

The educator cognitive sensitivity scale

Co-developed by Pauker et al., (2018), the *ECS scale* aimed to evaluate the cognitive sensitivity of educators in early care and educational institutions who were in close contact with children, but also as a tool to promote educators' professional development. The English edition scale and corresponding coding manual were first translated into Chinese by two early-childhood education professionals, and a first Chinese draft was formed after several rounds of discussion, revision and adjustment which were all carried out in accordance with the Chinese grammatical structure, wording and comprehensibility. After unified and rigorous scale evaluation training, 20 graduate students from early-childhood education majors applied the scale and coding manual to observe and evaluate 40 kindergarten educators. After the pilot use of the scale, the interview method was used to collect the suggestions of 20 evaluators on the revision of the scale and coding, and then the Chinese edition of the scale was further revised and refined.

The scale contains a total of 21 items, each of which employs the 5-point Likert scale, where 1 = not at all true,… 5 = very true, which aligns with the original English scale. An example of the item content, scoring method and corresponding operation instructions are shown in Table 1.

The path towards excellence—Chinese kindergarten education quality rating standards

This study used subscale called "Three Field Curriculum Promotion (*TFCP*)" in *the Path Towards Excellence—Chinese Kindergarten Education Quality Rating Standards* (*PTE-CKEQRS*; Chen et al., 2021) with good reliability as the validity criteria to discuss the empirical validity of the *ECS scale*. Former studies have shown that the process quality of the classroom in which children are located, has a significant effect on children's development (Campbell et al., 2001;

TABLE 1 Examples of project content, scoring methods and operation instructions.

Item 6	(Not at all true) (very true)	Coding manual
This educator is responsive to children's request for help, even those that are subtle and/ or nonverbal.	1 2 3 4 5	When asked verbally or non-verbally by a child for assistance, this educator responds immediately and appropriately in any form. (Examples: reassures child verbally or non-verbally, approaches child, guides child in task.)

Peisner-Feinberg et al., 2001; Nores et al., 2005; Burchinal et al., 2008, 2015; Burchinal et al., 2011; Li et al., 2016). The process quality mainly includes educator-child interaction and curriculum (activities), etc., and emphasizes the factors related to educators in activities (Mashburn et al., 2008; Vandell et al., 2010; Li and Hu 2012). The *TFCP* subscale in the *PTE-CKEQRS* was used to measure the process quality of the class, based on the five fields of child's health, language, society, science and the arts. In this study, it was used as a benchmark to measure the effectiveness of the *ECS scale*. After calculation, the internal consistency coefficient of the five fields was 0.748.

The child development assessment tool

Child development assessment was mainly based on two aspects: language, mathematics. Children's language adopts *the Peabody Picture Vocabulary Test (Revised edition—A type, PPVT;* Dunn & Dunn and Dunn 1981) with widely used and good measurement properties. The item uses 0–1 scoring method, and the correct answer was marked "1" and the wrong answer was "0." The internal consistency coefficient of *PPVT* in this paper was 0.980. Children's mathematical developments were mainly measured based on *the Research-based Early Maths Assessment-Short Form* (*REMA-SF;* Weiland et al., 2012), some studies had confirmed that the *REMA-SF* has good reliability and validity (Clements et al., 2008; Weiland et al., 2012; Sarama and Clements 2017). A score of "1" for correct answer and "0" for error or no answer in this scale, the internal consistency coefficient of this scale was 0.872.

Procedure

All raters were uniformly and rigorously trained in assessment, each sample class is assigned five raters, of which two were responsible for the *ECS scale* assessment, the other two were responsible for the evaluation of the *TFCP* subscale of *PTE-CKEQRS*, and the last one was responsible for the child development assessment.

For the evaluation of the *ECS scale* and the *TFCP* subscale of *PTE-CKEQRS*, non-participatory observation under natural scenarios was used, and the time was generally from 8 a.m. to 4 p.m. The two raters score individually and then discuss together, with the discussion score as the final result, and the scoring consistency was calculated to be 0.765 and 0.822, respectively.

For the child development assessment, each child was tested with *PPVT* and *REMA-SF*. The assessment guidelines and procedures were strictly in accordance with the instructions of each scale. All measurement was accomplished in approximately 20 to 35 min for one child. The child had a break when he or she was tired and inattentive during the assessment, and then the testing was continued under the child's consent. In addition, the written consent of the children's parents or guardians was obtained before the child development assessment, and the children's research ethics were strictly observed. Participants had the option to quit the testing at any moment.

Data analysis

Data analysis was processed and analyzed by SPSS 26.0 and HLM 8.0 software.

Results

Descriptive statistics and internal consistency

The results of descriptive statistics and internal consistency were shown in Table 2. The minimum of the item score was 1, the maximum was 5 and the mean of the item score was between 3.290 and 4.200. The value of SD and SE were all below 1. The item-total correlation was all significant ($p < 0.05$) and the minimum item-total correlation was item 11 ($r = 0.209$) The internal consistency used the Cronbach's coefficient, the results of the current study showed that the internal consistency of the scale was 0.884.

Structure validity

First, the results of the applicable condition analysis of EFA showed that it was feasible to adopt this method (KMO = 0.769, Barlett's = 765.351, df = 210, $p < 0.01$). Secondly, combined with the parallel analysis method (see Figure 1), the result was extracted from a common factor, and its variance contribution rate was 28.101%. The results of the optimal oblique rotation method show that all but the 6th and 11th items had the factor loadings coefficient of 0.4 or above (see Table 3). This result further verified the one-factor structure of *ECS scale*, but the variance contribution and the factor loading coefficient of individual items were both lower than the Pauker et al.'s (2018) research results.

Empirical validity

Take the evaluation results of TFCP as the effectiveness criterion

The basic situation of the assessment results of the five items in the *TFCP* section of *PTE-CKEQRS* and the correlation analysis with the cognitive sensitivity of early childhood teachers are shown in Table 4. At the same time, the sum of these five items was used as an effectiveness criterion to further explore the empirical validity of the *ECS scale*. The results showed that there was no significant correlation between the evaluation results of teachers' cognitive sensitivity and the five items of class process quality, or with the sum of these five items. Among them, there was no significant correlation with the two items of society and science.

TABLE 2 Basic assessment conditions, internal consistency and Item-total correlation of the *ECS scale*.

	Min.	Max.	Mean	SD	SE	Item-total correlation
Iteam1	2	5	4.200	0.586	0.059	0.608**
Item2	2	5	3.290	0.868	0.087	0.556**
Item3	2	5	3.740	0.613	0.061	0.552**
Item4	2	5	3.450	0.687	0.069	0.481**
Item5	2	5	3.940	0.633	0.063	0.588**
Item6	2	5	3.760	0.588	0.059	0.443**
Item7	2	5	3.730	0.750	0.075	0.697**
Item8	1	5	3.490	0.870	0.087	0.541**
Item9	2	5	3.510	0.628	0.063	0.500**
Item10	2	5	3.650	0.869	0.087	0.652**
Item11	3	5	4.100	0.628	0.063	0.209*
Item12	2	5	4.030	0.643	0.064	0.602**
Item13	2	5	4.080	0.825	0.082	0.688**
Item14	2	5	3.940	0.776	0.078	0.497**
Item15	2	5	3.950	0.730	0.073	0.691**
Item16	2	5	3.460	0.673	0.067	0.487**
Item17	2	5	3.350	0.744	0.074	0.484**
Item18	2	5	4.000	0.682	0.068	0.530**
Item19	1	5	3.670	0.637	0.064	0.594**
Item20	2	5	3.380	0.801	0.080	0.556**
Item21	2	5	3.790	0.782	0.078	0.567**
Total scale	55	97	78.510	8.307	0.831	0.884

* represents significant at the level of 0.05; ** represents significant at the level of 0.01.

Take the results of the child development as the effectiveness criterion

Analytical methods

A basic descriptive statistical overview of child development measurement was first provided in the current study. Secondly, Multilevel Linear Model (MLM) was used to analyze and verified whether the cognitive sensitivity of teachers promoted children's development. Specifically, a two-level linear model was constructed: level 1 was the child individual, and the predictor variables were the gender and age; level 2 was the class level, and the predictor variables were mainly the cognitive sensitivity of the teachers. Outcome variable was the result of child development (*PPVT* and *REMA-SF*, respectively). The specific modeling process was as follows.

The Null Model: Acted as a baseline model and did not contain any predictors. Based on the Intra-class Correlation Coefficient (ICC), i.e., the ratio of the difference between the groups to the total variation, it was judged whether it could be analyzed by MLM.

Model I: On the basis of the null model, individual predictors, e.g., demographic variables such as gender and age, were included to discuss the impact of individual-level variables on child development. The expression was as the following E (2):

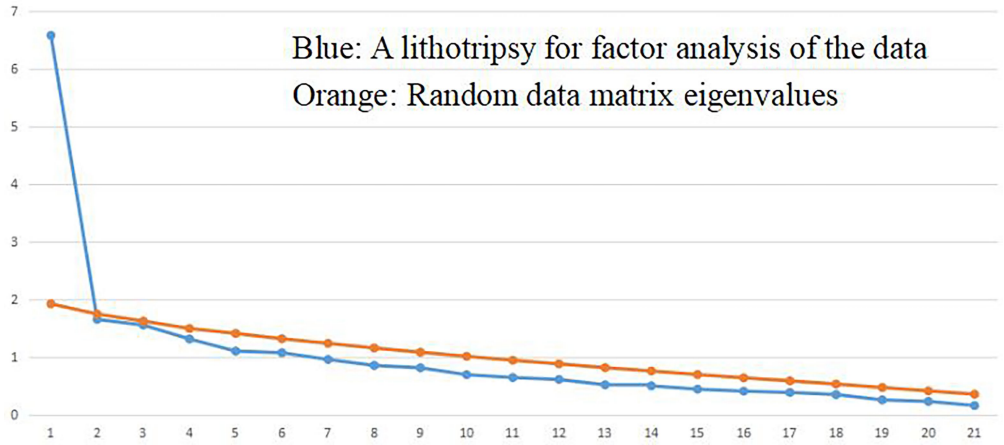


FIGURE 1
Exploratory factor analysis versus parallel analysis results.

TABLE 3 The common factor loading coefficient.

Item	Coefficient	Item	Coefficient	Item	Coefficient
Item1	0.613	Item8	0.486	Item15	0.690
Item2	0.529	Item9	0.488	Item16	0.432
Item3	0.516	Item10	0.651	Item17	0.458
Item4	0.426	Item11	0.127	Item18	0.496
Item5	0.555	Item12	0.589	Item19	0.562
Item6	0.386	Item13	0.694	Item20	0.497
Item7	0.683	Item14	0.432	Item21	0.512

TABLE 4 The basic situation of the assessment of *TFCP* and the correlation with *ECS*.

	Min	Max	M	SD	Correlation
Health	2.50	5.50	3.725	0.726	0.046
Language	1.50	4.50	3.265	0.785	0.041
Society	2.20	5.80	3.788	0.796	−0.059
Science	1.40	4.20	2.645	0.748	−0.122
Arts	1.00	5.00	2.296	0.770	0.199
Sum of the five projects	9.90	20.50	15.718	2.699	0.030

$$Level1: y_{ij} = \beta_{0j} + \beta_{1j}(gender) + \beta_{2j}(age) + e_{ij}.$$

$$Level2: \beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20} \quad \text{E(2)}$$

Model II: Based on Model I, class-level predictor, teachers' cognitive sensitivity, was incorporated to discuss the impact of class-level predictor on child development. The expression was as the following E (3):

$$Level1: y_{ij} = \beta_{0j} + \beta_{1j}(gender) + \beta_{2j}(age) + e_{ij}$$

$$Level2: \beta_{0j} = \gamma_{00} + \gamma_{01}(teachers'\ cognitive\ sensitivity) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20} \quad \text{E(3)}$$

Model III: Based on Model II, the cross-layer interaction of two level predictors was discussed. The analysis results of Model I (see Table 5) showed that gender did not play a significant predictive role in child development, so Model III only discussed the cross-layer interaction between cognitive sensitivity of teachers and age. The expression was as the following E (4):

$$Level1: y_{ij} = \beta_{0j} + \beta_{1j} * gender + \beta_{2j} * age + e_{ij}$$

$$Level2: \beta_{0j} = \gamma_{00} + \gamma_{01} * (teachers'\ cognitive\ sensitivity) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} * (teachers'\ cognitive\ sensitivity) \quad \text{E(4)}$$

Analytical results

(1) Results of the Null Model

According to the null model results (see Table 5), the ICC values of *PPVT* and *REMA-SF* were obtained as 0.604 and 0.545, respectively. Depending on the value of ICC was greater than 0.138, it was necessary to use MLM to analysis the data (Cohen et al., 1990).

(2) Results of Model I

The analytical results of Model I were shown in Table 5 (Model I column). It can be seen from the table that gender had a significant predictive effect on children's *REMA-SF*, but not on *PPVT*; Age had a significant predictive effect on *PPVT* and *REMA-SF* development in children. According to the S&B (Snnijders & Bosker) method, the proportion of variance explained by individual-level predictors was calculated (Snijders and Bosker 1994, 2011). As shown in Table 5, individual-level predictors (gender and age) explained 47.9% of the variance variation of *PPVT* in children, 40.5% of the variance variation in children's *REMA-SF*. The model fit indices showed a decrease in deviance.

(3) Results of Model II

The analytical results of Model II were shown in Table 5 (Model II column). It can be seen from the table that teachers' cognitive sensitivity did not have a significant predictive effect on *PPVT* and *REMA-SF* development in children. The S&B method was also used to calculate the proportion of variance explained by adding class-level predictive variables (see Table 5). At this time, the variance variation explained by the cognitive sensitivity of teachers at the class level for children's *PPVT* was 73.8%; the variance variation interpreted for children's *REMA-SF* was 63.1%. Compared with the variation explained by Model I, there was an increase. The results of model fit indices appeared that the value of deviance barely changed from Model I to Model II.

(4) Results of Model III

The analysis results of model 3 were shown in Table 5 (Mode III column). It can be seen from the table that there was no significant cross-layer interaction between teachers' cognitive sensitivity and children's age. In addition, after the addition of interaction terms, the values of τ_{00} and σ_2 in Model III were increased compared with those in Model II, but the increase was

TABLE 5 Analysis results of two-level linear model.

	Level 1 dependent variable($M + SD$)	Null Model	Model I	Model II	Model III
Fixed effect					
Intercept (γ_{00})	PPVT(56.78 + 24.04)	57.029**	−18.090**	−38.298*	−10.503
	REMA-SF(15.98 + 8.76)	15.521**	−10.499**	−12.863	−15.863
Gender (γ_{10})	PPVT	/	1.058	1.095	1.131
	REMA-SF	/	2.201**	2.202**	2.198**
Age (γ_{20})	PPVT	/	16.315**	16.545**	10.563
	REMA-SF	/	5.438**	5.419**	6.059
ECS (γ_{01})	PPVT	/	/	0.240	−0.101
	REMA-SF	/	/	0.031	0.068
Interaction effect					
ECS*Age(γ_{21})	PPVT	/	/	/	0.073
	REMA-SF	/	/	/	−0.008
Random effect					
Intercept (τ_{00})	PPVT	347.741	62.021	58.691	59.605
	REMA-SF	41.906	10.975	11.336	11.483
Residual error (σ_2)	PPVT	227.894	237.825	238.524	239.018
	REMA-SF	35.185	34.910	34.900	35.005
Model fit					
Deviance	PPVT	2,264.044	2,205.614	2,207.278	2,206.642
	REMA-SF	1,656.444	1,604.673	1,606.010	1,610.871
Number of estimated parameters		2	2	2	2

* represents significant at the level of 0.05; ** represents significant at the level of 0.01.

not very large. The Deviance value of the model fit index also had little change between Model II and Model III.

Discussion

Reliability discussion

The method of observation and evaluation is mainly adopted in the cognitive sensitivity measurement of early childhood teachers. Generally, two observers enter the one-day activity site of classes, and make dynamic on-site observation and corresponding evaluation of class teachers according to the scale. The reliability of the *ECS scale* for teachers mainly includes the raters' and the scale's. In this study, the consistency between the two raters was used for the reliability of the raters, and the results showed that the mean consistency between the raters was 0.765. In order to avoid the difference of evaluation results caused by the consistency between raters, although the discussion results of the two raters were adopted as the final score of *ECS* in the study. In the evaluation process, the two raters observed teachers according to the scale and gave subjective assessment, the rater's error was an important factor that could not be ignored and affected the measurement results. Therefore, further exploration would be needed to verify the raters' reliability of using observation method to evaluate *ECS*. In addition, the study used the method of internal consistency coefficient to measure the reliability of the scale. The results showed that the internal consistency coefficient of the scale was, α = 0.884 which meant that the internal consistency of the scale was good. Usually, the closer the internal consistency coefficient is to 1, the more desirable it will be (Shrout and Fleiss 1979; McGraw and Wong 1996; Weir 2005). The study of Pauker et al. (2018) showed that the internal consistency of the scale was, α = 0.96 and in comparison, the internal consistency of this scale in the context of Chinese culture needs to be improved. However, the value of α was mainly related to the length of the scale and the quality of the items (Cronbach 1951). In other words, the more items contained in the scale and the higher the quality of the items, the higher can be improved. Therefore, the reliability of cross-cultural adaptability of the scale needs to be further improved, and consideration could be given to increasing the number of items or revising the content of the items.

Validity discussion

The validity analysis of cross-cultural adaptability of the *ECS scale* was mainly carried out from the aspects of structural validity and empirical validity. The results of structural validity verified the research of Pauker et al. (2018), that was the scale contained a single common factor structure. But different from Pauker et al.'s (2018) results, the variance contribution rate of single factor explanation was not high (28.101%); Secondly, the factor loadings coefficient of

some items (such as the 11th item) was not high. The main reasons for the differences in these research results may be that the existing samples were mainly from one province and two cities in East China, and the sample size was not large enough. On the other hand, these differences may be caused by the fact that the interpretation and evaluation of the scale items differ in different cultural contexts. In addition, EFA was employed to explore the structure of ECS in order to compare with the existing results. Due to the sample size limitations, Confirmatory Factor Analysis (CFA) was not used to confirm the structure. More participants will be sampled to validate the sturcture of the result in the future research.

Previous studies had used *ITERS-R* (Harms et al., 2003) and *CLASS* (La Paro et al., 2012) as the validity criteria to discuss the concurrent validity of the *ECS scale*, and the results showed that the scale had good concurrent validity. In the present study, The *TFCP* subscale of the *PTE-CKEQRS* with Chinese cultural background and the results of child development were used as the validity criteria to verify the empirical validity of the *ECS scale*. In conclusion, the results of the analysis showed that there was no significant correlation between the scale and the *TFCP*; the direct and interactive effects of teachers' cognitive sensitivity on children's development were not significant. On the one hand, these different results may be due to the different research tools and methods used, such as different criteria and measurement tools. Secondly, the cultural background and research object may be another reason. These results indicated that the appropriateness and promotion of the *ECS scale* in Chinese cultural background need to be further discussed, especially the measurement content, process and method of each item need to be continuously revised and practiced in the context of Chinese culture.

Conclusion

In summary, through the actual assessment and analysis of the cognitive sensitivity of 100 early childhood teachers from 50 classes in a certain province in eastern China, the results showed that: the internal consistency of the *ECS scale* was good; the results of the structural validity analysis of the scale were similar to those of existing studies. Based on the empirical validity analysis results where the ECS score had no associations with child development scales in the curriculum promotion field of *PTE-CKEQRS*, it was concluded that the empirical validity of the scale was not ideal, which needed to be further explored in practice for the future.

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 the Ethics Committee of the Institute of Psychology, Zhejiang Normal University (ID.ZSRT2022055). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

BL: first draft, data analysis, revised manuscript. DC: research design, data analysis, first draft, revised manuscript, translation. CY: translation. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Zhejiang Provincial Foundation of Philosophy and Social Sciences (No: 20NDJC07Z).

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.

References

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York: Freeman.

Bernier, A., Carlson, S. M., and Whipple, N. (2010). From external regulation to self-regulation: Early parenting precursors of young Children's executive functioning. *Child Dev.* 81, 326–339. doi: 10.1111/j.1467-8624.2009.01397.x

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Appl. Dev. Sci.* 12, 140–153. doi: 10.1080/10888690802199418

Burchinal, M., Kainz, K., and Cai, Y. (2011). "How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings" in *Quality Measurement in Early Childhood Settings*. eds. M. Zaslow, I. Martinez-Beck, K. Tout and T. Halle (Baltimore, MD: Paul H. Brookes Publishing Co), 11–31.

Burchinal, M., Magnuson, K., Powell, D., and Hong, S. S. (2015). "Early childcare and education," in *Handbook of Child Psychology and Developmental Science: Ecological Settings and Processes*. eds. M. H. Bornstein, T. Leventhal and R. M. Lerner, vol. *Vol. 4. 7th Edn.* (Hoboken, NJ: John Wiley & Sons, Inc), 223–267.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., and Ramey, C. T. (2001). The development of cognitive and academic abilities: growth curves from an early childhood educational experiment. *Dev. Psychol.* 37, 231–242. doi: 10.1037/0012-1649.37.2.231

Carpendale, J. I. M., and Lewis, C. (2004). Constructing an understanding of mind: the development of children's social understanding within social interaction. *Behav. Brain Sci.* 27, 79–151. doi: 10.1017/S0140525X04000032

Chen, D. Z., Li, K. J., and Zhou, J. (2021). Psychometric property analysis on "the path towards excellence Chinese kindergarten education quality rating standards (PTE-CKEQRS)" based on the data about 100 kindergartens and 1670 preschoolers in China. *Stud. Early Childhood Educ.* 01, 3–16. doi: 10.13861/j.cnki.sece.2021.01.001

Clements, D. H., Sarama, J. H., and Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: the research-based Early Maths assessment. *Educ. Psychol.* 28, 457–482. doi: 10.1080/01443410701777272

Cohen, P., Cohen, J., Teresi, J., Marchi, M. L., and Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Appl. Psychol. Meas.* 14, 183–196. doi: 10.1177/014662169001400207

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Dunn, J. (2002). "Sibling relationships" in *Blackwell Handbook of Childhood social Development*. eds. P. K. Smith and C. H. Hart (New York, NY: Wiley), 223–237.

Dunn, M., and Dunn, L.M. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.

Fox, N. A., and Hane, A. A. (2008). "Studying the biology of human attachment" in *Handbook of Attachment: Theory, Research, and Clinical Applications*. eds. J. Cassidy and P. R. Shaver. *2nd Edn.* (New York, NY: The Guilford Press), 217–240.

Glaser, D. (2000). Child abuse and neglect and the brain—a review. *J. Child Psychol. Psychiatry* 41, 97–116. doi: 10.1111/1469-7610.00551

Harms, T., Cryer, D., and Clifford, R. M. (2003). *Infant/Toddler Environment Rating Scale Revised Edition (ITERS-R)*. New York, NY: Teachers College Press.

Hoff, E. (2006). How social contexts support and shape language development. *Dev. Rev.* 26, 55–88. doi: 10.1016/j.dr.2005.11.002

La Paro, K. M., Hamre, B. K., and Pianta, R. C. (2012). *Classroom Assessment Scoring System (CLASS) Manual, Toddler*. Baltimore, MD: Brookes.

Landry, S. H., Garner, P. W., Swank, P. R., and Baldwin, C. D. (1996). Effects of maternal scaffolding during joint toy play with preterm and full-term infants. *Merrill-Palmer Q.* 42, 177–199.

Landry, S. H., Smith, K. E., Landry, S. H., Smith, K. E., Swank, P. R., and Miller-Loncar, C. L. (2000). Early maternal and child influences on children's later independent cognitive and social functioning. *Child Dev.* 71, 358–375. doi: 10.1111/1467-8624.00150

Laranjo, J., Bernier, A., Meins, E., and Carlson, S. M. (2010). Early manifestations of children's theory of mind: the roles of maternal mind-mindedness and infant security of attachment. *Infancy* 15, 300–323. doi: 10.1111/j.1532-7078.2009.00014.x

Li, K., and Hu, B. (2012). On quality assessment of Early childhood programs in the international context and reconstructing the related conceptions in China. *Int. Comp. Educ.* 34, 15–20. doi: 10.20013/j.cnki.ICE.2012.07.04

Li, K., Pan, Y., Hu, B., Burchinal, M., De Marco, A., Fan, X., et al. (2016). Early childhood education quality and child outcomes in China: evidence from Zhejiang Province. *Early Child. Res. Q.* 36, 427–438. doi: 10.1016/j.ecresq.2016.01.009

Lundy, B. L. (2013). Paternal and maternal mind-mindedness and preschoolers' theory of mind: the mediating role of interactional attunement. *Soc. Dev.* 22, 58–74. doi: 10.1111/sode.12009

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., et al. (2008). Measures of classroom quality in prekindergarten and Children's development of academic, language, and social skills. *Child Dev.* 79, 732–749. doi: 10.1111/j.1467-8624.2008.01154.x

McGraw, K. O., and Wong, S. P. (1996). "Forming inferences about some intraclass correlations coefficients": correction. *Psychol. Methods* 1:390. doi: 10.1037/1082-989X.1.4.390

Meins, E., Fernyhough, C., Wainwright, R., Gupta, M. D., Fradley, E., and Tuckey, M. (2002). Maternal mind-mindedness and attachment security as predictors of theory of mind understanding. *Child Dev.* 73, 1715–1726. doi: 10.1111/1467-8624.00501

Nores, M., Belfield, C. R., Barnett, W. S., and Schweinhart, L. (2005). Updating the economic impacts of the high/scope Perry preschool program. *Educ. Eval. Policy Anal.* 27, 245–261. doi: 10.3102/01623737027003245

Pauker, S., Perlman, M., Prime, H., and Jenkins, J. (2018). Caregiver cognitive sensitivity: measure development and validation in Early childhood education and care (ECEC) settings. *Early Child. Res. Q.* 45, 45–57. doi: 10.1016/j.ecresq.2018.05.001

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Dev.* 72, 1534–1553. doi: 10.1111/1467-8624.00364

Prime, H., Perlman, M., Tackett, J., and Jenkins, J. (2014). Cognitive sensitivity in sibling interactions: development of the construct and comparison of two coding methodologies. *Early Educ. Dev.* 25, 240–258. doi: 10.1080/10409289.2013.821313

Prime, H., Plamondon, A., Pauker, S., Perlman, M., and Jenkins, J. M. (2016). Sibling cognitive sensitivity as a moderator of the relationship between sibship size and children's theory of mind: a longitudinal analysis. *Cogn. Dev.* 39, 93–102. doi: 10.1016/j.cogdev.2016.03.005

Sarama, J., and Clements, D. H. (2017). "Interventions in early mathematics: avoiding pollution and dilution," in *The Development of Early Childhood Mathematics Education*. eds. J. Sarama, D. H. Clements, C. Germeroth and C. Day-Hess, *Vol. 53,* (Cambridge, MA: Elsevier Academic Press), 95–126. doi: 10.1016/bs.acdb.2017.03.003

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86, 420–428. doi: 10.1037/0033-2909.86.2.420

Snijders, T. A. B., and Bosker, R. J. (1994). Modeled variance in two-level models. *Sociol. Methods Res.* 22, 342–363. doi: 10.1177/0049124194022003004

Snijders, T.A., and Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: SAGE Publications Ltd, Limited.

Vandell, D. L., Belsky, J., Burchinal, M., Steinberg, L., and Vandergrift, N. (2010). Do effects of early child care extend to age 15 years? Results from the NICHD study of early child care and youth development. *Child Dev.* 81, 737–756. doi: 10.1111/j.1467-8624.2010.01431.x

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., and Yoshikawa, H. (2012). Early mathematics assessment: validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educ. Psychol.* 32, 311–333. doi: 10.1080/01443410.2011.654190

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J. Strength Cond. Res.* 19, 231–240. doi: 10.1519/15184.1

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 28 November 2022
DOI 10.3389/feduc.2022.1013841

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Quan Zhang,
Jiaxing University,
China
Anna Wolf,
University of Hamburg,
Germany

*CORRESPONDENCE
Julia Ehninger
info@juliaehninger.de
Jens Knigge
jens.knigge@nord.no

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 07 August 2022
ACCEPTED 28 October 2022
PUBLISHED 28 November 2022

CITATION
Ehninger J, Knigge J and Rolle C (2022) Why are certain items more difficult than others in a competency test for music–related argumentation? *Front. Educ.* 7:1013841. doi: 10.3389/feduc.2022.1013841

COPYRIGHT
© 2022 Ehninger, Knigge and Rolle. 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.

Why are certain items more difficult than others in a competency test for music–related argumentation?

Julia Ehninger[1*†], Jens Knigge[2*†] and Christian Rolle[1]

[1]Institute for Music Education, University of Cologne, Cologne, Germany, [2]Department for Arts and Culture, Nord University, Levanger, Norway

This paper examines why certain items in a competency test for music-related argumentation are more difficult than others. Based on previous studies on school-related achievement tests, the authors assume that differences in item difficulty are related to different item characteristics or combinations of characteristics. In this study, the item characteristics of a test for music-related argumentation were first identified and coded. Three domains were identified as contributing to item difficulty: cognitive requirements, knowledge, and formal item features. Second, multiple linear regression analyses were conducted with the item characteristics as predictors of item difficulty, which had been estimated in a prior study. A comparison of three regression models confirmed that the model holding four predictors of the domain "cognitive requirements" best fit the study data $R_{adj}^2 = 0.71$. The strongest predictor in the final model was "reference to musical attributes" $(\beta = 0.46, \beta = 0.51)$, followed by "cross-sentence argumentation" $(\beta = 0.37)$ and "dialogical argumentation" $(\beta = 0.20)$. These results indicate that the difficulty of an item increased most when participants had to refer to musical attributes to solve the task. The items that required the participants to provide cross-sentence or dialogical argumentation were more challenging as well. The findings regarding the relations between item characteristics and item difficulty contribute to a better understanding of music-related argumentative competence, with important implications for the music classroom.

KEYWORDS

assessment and education, competency testing, music education, music–related argumentation, item characteristics, empirical research

Introduction

Argumentation is an essential part of our everyday lives. We are familiar with it from discussions at work and debates in court, and it is an integral part of the democratic process. Argumentation also plays a major role—whether intended or not—in the classroom. Students require argumentative competence in order to engage in classroom activities such as group discussions. Furthermore, discourse practices such as arguing and explaining

contribute to the acquisition and negotiation of knowledge (Kuhn, 2005; Morek and Heller, 2012; Morek et al., 2017). Therefore, it is not surprising that argumentative skills are considered a key competency for students' overall educational success (Quasthoff et al., 2020b). Although it has become evident in recent years that language is constitutive of learning in all school subjects (Lazarou et al., 2016; Rapanta, 2018; Quasthoff et al., 2020a), there has been little theoretical and empirical research on the role of language competence, such as argumentative competence, in music as a school subject (Bossen, 2017).

Language is an important medium of communication, including in the music classroom. When rehearsing, musicians often verbally negotiate how music should sound, whether it is a band working on a song or members of a string quartet who must agree on the interpretation of the musical piece they are rehearsing. In music lessons, verbal engagement with music plays a major role and music-related argumentative competence is an integral part of German school curricula (Kultusministerkonferenz, 2005). In a prior study, we empirically modeled music-related argumentative competence and developed the MARKO test, a competency test for *music-related argumentative competence* (Musikbezogene ARgumentationsKOmpetenz; German for music-related argumentative competence).

In this paper, we explore the question of why certain items of the competency test were more difficult than others. Prior research on competency tests has shown that certain item characteristics can increase an item's difficulty (e.g., Knigge, 2010). For example, items containing a great deal of text can be more challenging for students to solve (e.g., Prenzel et al., 2002). The complexity of a musical piece can also contribute to the difficulty of an item (e.g., Knigge, 2010, pp. 228–231). Therefore, we present in-depth analyses on the item characteristics of the items of the MARKO test for music-related argumentative competence. We analyze which item characteristics contribute to item difficulty and closely examine the requisite competencies for solving the items in the competency test.

Theoretical background

Music–related argumentative competence

Music-related argumentative competence can be defined as the "context-specific cognitive disposition that is acquired and needed to justify and defend esthetic judgments about music in a comprehensive, plausible, and differentiated way" (Ehninger et al., 2021, pp. 2–3). The competence to reflect on and justify judgments about music is relevant in school curricula in many countries (e.g., Germany: Kultusministerkonferenz, 2005; Norway: Utdanningsdirektoratet, 2020); however, until now, there has been little research on the requirements for engaging in music-related argumentation.

Rolle (2013) proposed a theoretical competency model on music-related argumentation and distinguished between several competency levels. This model assumes that it is easier to refer to subjective impressions of music and personal taste than the cultural and social context of music or esthetic conventions. People on higher competency levels are better able to reflect on their own judgment about music and can integrate criticism and other people's opinions into their reasoning (see also Knörzer et al., 2016). Based on Rolle's theoretical assumptions, the MARKO competency test for music-related argumentation was developed and validated.

Competency test for music-related argumentation (MARKO)

The MARKO test is in German and includes 25 open-ended items distributed online. It was designed for ninth to twelfth grade high school students as well as university students. During the test, the participants worked individually on computers and used headphones to listen to music (and sometimes watch videos) of various musical genres. In the test, they were asked to justify their esthetic judgment in a written answer. For example, they were asked why they thought a musical piece created a certain atmosphere or were prompted to comment on a discussion below a YouTube video or a concert review in a newspaper.

The validation of the test as well as a competency model resulting from data collected from 440 participants were presented in a prior study (Ehninger et al., 2021; Ehninger, 2022). Two sample items of the test are presented below to provide an insight into the test. Figure 1 shows the sample item "Star Wars," which was developed to assess how the participants referred to the atmosphere and musical attributes of a musical piece. In this item, the participants were asked whether a musical piece illustrated the atmosphere in outer space. To solve the item, the participants produced texts that were rated in accordance with a coding scheme (Table 1).

While some items in the test were aimed at assessing how the participants referred to musical attributes or their subjective impressions of the music, others were designed to measure the dialogical dimension of argumentation. In the item "Eurovision Song Contest," the participants were asked to comment on a discussion on YouTube about the winner of the *Eurovision Song Contest* (Figure 2). The participants' answers were also coded with a coding scheme (Table 2).

The two sample items differed in various respects. While the "Star Wars" item (Figure 1) contained little text, the participants had to read a great deal of text before solving the "Eurovision Song Contest" item (Figure 2). Furthermore, the cognitive requirements for solving the items seemed to differ. For the "Eurovision Song Contest" item, the participants had to consider the social and cultural contexts, such as feminism and social justice. In comparison to the "Eurovision Song Contest" item, a much more differentiated reference to musical attributes was required in the

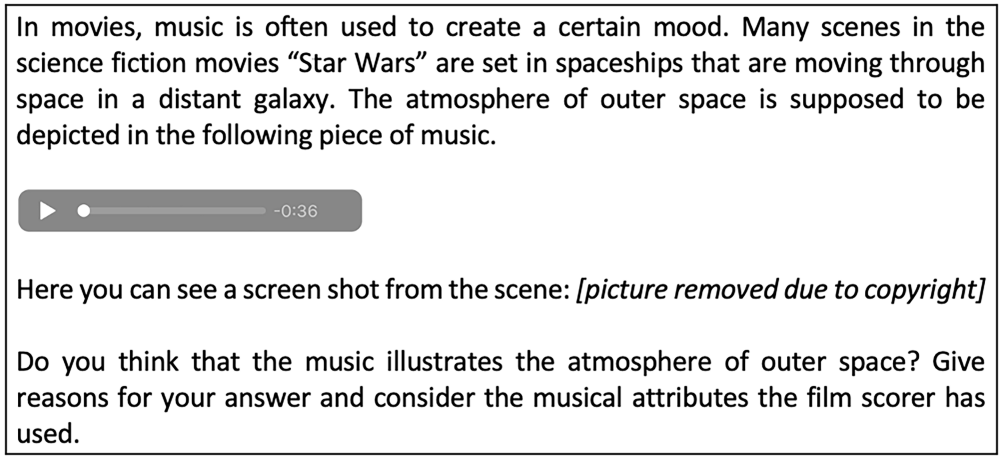


FIGURE 1
Test item "Star Wars" (English translation). Note: The participants listened to an excerpt from the film score (Arrival at Naboo, Episode I). A screenshot from the scene was shown in the item but had to be omitted here due to copyright concerns. The screenshot showed the view of a planet from a space shuttle cockpit (see also Ehninger et al., 2021; all items are available here: DOI 10.17605/OSF.IO/ZVP4B).

"Star Wars" item, at least for scoring the maximum number of points.

These two example items show that the requirements for solving an item (category) could vary in a competency test and that the content of the items could also differ. However, we did not know why one item (category) was more difficult than the other. Was the "Eurovision Song Contest" item more difficult because it contained more text, or was the "Eurovision Song Contest" item perhaps easier because the students both listened to music and watched a video? These questions can be answered by examining the characteristics of the items and relating them to the difficulty of the items.

Item difficulty and item characteristics

Item characteristics can be defined as the characteristics of an item associated with higher or lower demands on test takers, thereby influencing the solution probability (Hartig and Jude, 2007, p. 31). They are relevant for competence research primarily because the "competence" construct is defined by its context-specificity (whereas, e.g., intelligence is defined as generalized, context-independent cognitive dispositions that can only be learned to a limited extent; see, e.g., Hartig and Klieme, 2006; Hartig, 2008). From this context-specificity, one can derive the fundamental interest in the characteristics of a situation (i.e., the item in a test situation) in which competent performance manifests itself. Particular attention is paid to the characteristics of a situation that make competent performance easier or more difficult. This is because only "knowledge of the situational characteristics that influence successful performance enables a deeper understanding of the processes that underlie successful performance and thus a better understanding of the competence in question" (Hartig and Jude, 2007, p. 31; translation by the authors).

Nevertheless, there are also other arguments regarding the relevance of item characteristics: One interesting aspect is that they can be used to define levels of competence (e.g., Hartig, 2007). If different item difficulties can be explained empirically by a certain set of item characteristics, the levels of a competency can be described by means of the characteristics in question. These competency-level descriptions are then empirically validated and are also generalizable beyond the concrete test items used (Hartig and Jude, 2007).

Another aspect concerns the validity of a test. In their influential paper, Borsboom et al. (2004) argue for a reconceptualization of test validity: "A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes" (p. 1061). Therefore, validation research must be directed "at the processes that convey the effect of the measured attribute on the test scores" (p. 1061). Against this background, the formulation of item characteristics can be understood as hypotheses about the processes that cause variation in a competency test. Hence, from a test-theoretical point of view, the prediction of item difficulty by item characteristics can be regarded as a confirmation of the validity of the measurement instrument (see also Hartig, 2007).

Furthermore, if empirically validated item characteristics are provided, they can be used to design new test items (Nold and Rossa, 2007). It would then be possible to create specific "requirement profiles" for the items that are supposed to be developed, which would consist of different combinations and degrees of the item characteristics. Model-guided item development, in this sense, makes it possible to determine *a priori* which items should be easier or more difficult and the reasons for

TABLE 1 Coding scheme for the sample item "star wars."

Points	Description	Sample answers
0	Tautological justification or no reason	"Yes, because of the atmosphere that exists in space. The composer presented this very well." (VP_661)
1	Participants refer only to the musical atmosphere. If musical attributes are mentioned (or even a causal relationship is established between them and the atmosphere), this is done by referring to "basic" and superficial characteristics of the music (e.g., "bright notes," "long tones," "loud," "soft," "instruments that create tension").	"I think so because it sounds exciting and unusual, which, in my opinion, corresponds well with the atmosphere in outer space." (VP_714)
2	Participants relate the generated atmosphere to musical attributes. If instruments (e.g., "quiet strings") are mentioned, the answer is given two points.	"Yes, I find it very well done. The sound layers depict the infinite vastness of the universe … the synthesizers give the piece a futuristic character … single high notes to illustrate the stars." (VP_589)
3	Participants relate the generated atmosphere to musical attributes. A detailed description is provided (e.g., the musical form and the way the instruments are played).	"I find the composition convincing because the long notes (played by the violin) generate a feeling of width and yet (because of the high notes) sound quite exciting and dramatic, especially at the beginning. The fast (xylophone?) notes that go up and down the scale have a bright sound and are reminiscent of stars. The flourish at the beginning could suggest that a scenery of spectacular surroundings is just revealing itself to the audience." (VP_610)

This is a simplified and condensed version of the coding scheme, which was also published in Ehninger et al. (2021).

these differences. Accordingly, items can be developed explicitly for a certain competence profile or competence level.

Prenzel et al. (2002, p. 125) proposed categorizing item characteristics into three domains: formal task characteristics, cognitive demands in solving the tasks, and the characteristics of the knowledge base required for solving the tasks (similar categorizations can be found in, e.g., Nold and Rossa, 2007, and Hartig and Klieme, 2006). Knigge (2010) used this systemization for a music-specific item analysis of a competency test for musical perception (KoMus test). He systematized the item characteristics as follows: (1) formal item characteristics, (2) cognitive demands on auditory perception and musical memory, and (3) necessary activation of expertise.

(1) *Formal item* characteristics include the item format (closed vs. open), the formalities of the item content (e.g., picture stimulus vs. auditory stimulus), and the nature of the item stem (e.g., long vs. short question phrases). The influence of this item characteristic domain has been demonstrated in studies on the assessment of language and mathematical/scientific competencies (e.g., Prenzel et al., 2002; Cohors-Fresenborg et al., 2004; Beck and Klieme, 2007). With regard to musical competence, an influence of such general, non-music-specific characteristics also seems plausible, which was also confirmed by Knigge (2010) and Jordan (2014).

(2) There are several research results from other disciplines regarding the requisite cognitive processes for processing an item in language or mathematical tests (e.g., Hartig and Klieme, 2006; Nold and Rossa, 2007); however, these results are not directly transferable to musical competence. With regard to a competence test for musical perception, Knigge (2010) identified two cognitive demand domains of relevance to item processing related to auditory perception and musical memory. While these requirements are closely connected, they can also occur independently of each other.

(3) Finally, item characteristics can be characterized by the activation of subject-specific knowledge (e.g., Prenzel et al., 2002). In relation to the musical knowledge required to solve an auditory perception item, Knigge (2010) identified five item characteristics: knowledge of musical notation, knowledge of music theory, knowledge of music history, knowledge of musical styles and genres, and knowledge of the cultural and social contexts of music.

The categorization of item characteristics presented above was empirically validated by Knigge (2010) and Jordan (2014) who showed that the difficulty of an item was mainly influenced by the cognitive demands on auditory perception and necessary subject knowledge. For the entire set of item characteristics, a very strong prediction of item difficulty could be demonstrated for the KoMus competency test, with the explained variance being between 55% and 83% (regression analyses were conducted for all four subdimensions of the KoMus test; Jordan, 2014, pp. 136–139).

Research goal

The aim of this paper was to explore the question of why certain items of the MARKO test for music-related argumentation were more difficult than others. Based on previous studies on school-related achievement tests, we assumed that the differences in item difficulty were related to different item characteristics or combinations of

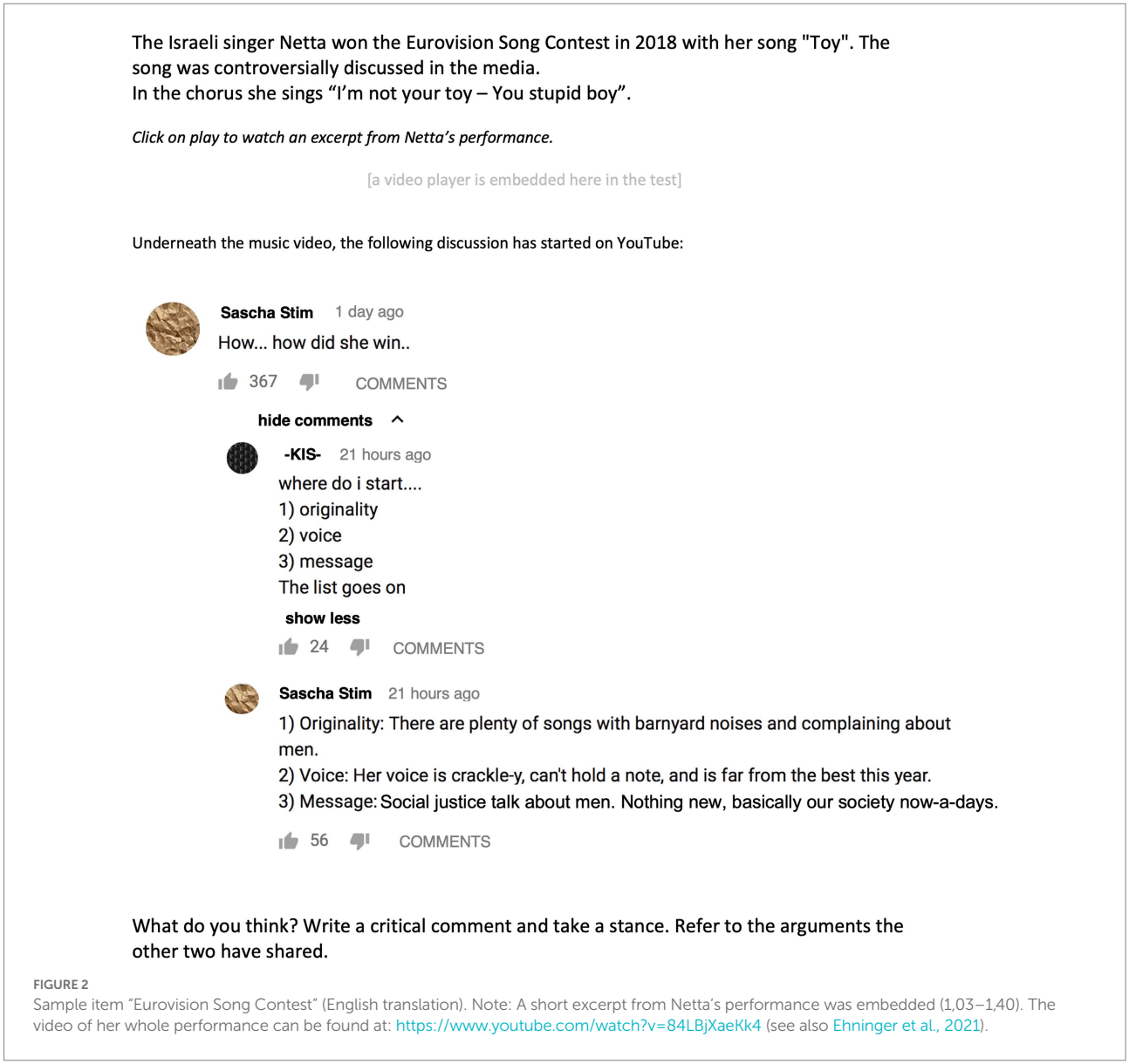


FIGURE 2
Sample item “Eurovision Song Contest” (English translation). Note: A short excerpt from Netta’s performance was embedded (1,03–1,40). The video of her whole performance can be found at: https://www.youtube.com/watch?v=84LBjXaeKk4 (see also Ehninger et al., 2021).

characteristics. Therefore, our aim was to identify the item characteristics relevant to the MARKO test and quantify their specific influence. In doing so, we hoped to gain a better understanding of the specific competence needed to solve the competency test items as well as examine the validity of the test.

Materials and methods

The methodological approach in this paper can be divided into three steps: First, the item characteristics were identified and categorized (“Identification and Categorization of Item Characteristics”). Second, the whole item pool (i.e., competency test) was coded according to the identified and categorized item characteristics (“Coding Item Characteristics”). Finally, multiple linear regression analyses were conducted with the item characteristics as predictors of item difficulty (“Multiple Regression Analyses”). The difficulty parameters of the test items were obtained in a prior study ($N = 440$; students from upper secondary schools and universities) employing IRT scaling (partial credit model; Ehninger et al., 2021).

Identification and categorization of item characteristics

We chose a combined deductive–inductive approach to identify item characteristics specific to the MARKO test:

- We adapted findings from previous research on item characteristics from music and other subjects (e.g., Knigge, 2010; Jordan, 2014).

TABLE 2 Coding scheme for the item "Eurovision song contest" (English translation).

Points	Coding scheme	Sample answers
0	The answer paraphrases parts of the YouTube discussion and/or refers to personal taste.	"I do not like the song either and agree with Sascha's comment. I also think that she does not hold the pitch very well, and I think that the crackling is a little ridiculous." (P2_11)
1	The answer takes into account the entire YouTube discussion but is paraphrasing it for the most part. The answer might include a new argument that has not come up in the YouTube discussion.	"The singer addresses a very important and current topic: social equality. However, I find it is not really appropriately communicated. The lyrics are presented with humor and thus they do not mean anything." (VP_142)
2	The answer contains at least two new arguments. Different perspectives are evaluated.	"Women's empowerment is a current topic of great importance. It is good that artists are setting an example. Sometimes, the lyrics are one-dimensional because women also 'play' with women. But often, it is the other way around and has been the case for centuries due to the unfair distribution of power, where women are neglected. Maybe she should have sung 'I'm not a toy, for no one' or something like that, which emphasizes the idea of equality. She represents a strong image of women, which is definitely socially critical. Because of the 'crackling,' as Sascha calls it, the song is unusual and different and differs from the social norm that influences the masses, as Sascha and 367 other people show. Have fun with your followers and mainstream boredom." (VP_89)

- We used the MARKO coding schemes and the MARKO competency model (Ehninger et al., 2021).
- We took theoretical assumptions on musical perception (see "Domain 1: cognitive requirements") and argumentative competence into account (e.g., Heller and Morek, 2015; see "Domain 1: cognitive requirements").
- We conducted in-depth analyses of the individual items.

Against this background, we conducted several coding sessions in which the applicability of the identified item characteristics was tested for the entire item pool (the MARKO test consists of 25 polytomous items). The coding sessions were conducted in a circular procedure that was carried out several times until interrater-reliability was acceptable. In those sessions, item characteristics were first coded by two independent raters for every single item category of the test. In a second step, interrater-reliability was calculated and ratings with low interrater-reliability were reviewed. Next, item characteristics were revised, and new item characteristics were added if necessary. Finally, all item categories were rated again. On this basis, several item characteristics were identified for the item pool of the MARKO test, resulting in three domains of item characteristics: *(1) cognitive requirements, (2) knowledge,* and *(3) formal item features.*

Domain 1: Cognitive requirements

The first domain of the identified item characteristics dealt with the cognitive requirements that a participant had to cope with when solving an item. Many cognitive requirements were described in the coding schemes for every item (Tables 1, 2). Three characteristics were identified as dealing with cognitive requirements: *(a) reference to perceived musical attributes, (b) cross-sentence argumentation,* and *(c) dialogical argumentation* (Table 3) .

A *reference to perceived musical attributes* was required for many items and was specified in the coding schemes. Research on music-related argumentation has shown that references to the musical attributes of a musical piece are a cognitive operation that is essential when engaging in music-related argumentation (Rolle, 2013; Knörzer et al., 2016). For example, the coding scheme of the "Star Wars" item (Table 1) specified that the participants only had to refer to "basic and superficial characteristics of the music (e.g., 'bright notes', 'long tones' […])" in order to score one point. To achieve two or more points for the item, the participants had to refer to more specific musical attributes such as musical instruments or the musical form. This item characteristic was also identified in an assessment test on music-related perception (Knigge, 2010; Jordan et al., 2012) and can be framed inside cognitive research in music psychology. From a cognitive psychology perspective, musical perception can be described as the active (re)construction of auditory events with the help of specific techniques and using existing knowledge that is strongly culturally influenced (Morrison and Demorest, 2009; for an overview of findings on musical perception in cognitive neuroscience see Koelsch, 2019). In general, we assume that if a MARKO item demands more complex musical perception, this will lead to an increase in item difficulty.

The item characteristic *cross-sentence argumentation* points to linguistic requirements for producing an answer to an item. The needed discourse competence has been modeled as a dimension of the overarching communicative competence (Canale and Swain, 1980). When people engage in argumentation, they do not "communicate with each other by simply producing words and sentences but by orienting to and accomplishing discursive activities above the sentence-level" (Heller and Morek, 2015, p. 181). In considering the item "Eurovision Song Contest" and its coding scheme (Figure 2; Table 2), it became clear that to score

two points for the item, the reasoning of the participant had to be consistent across several sentences.

In the MARKO test, several items were designed to assess the *dialogical* dimension of *argumentation* acknowledging that argumentation must not only be seen as a relationship between sentences but is a social practice (Eemeren et al., 2014, chapter 10). For this reason, in several items, the participants were confronted with opinions of others. An example of cognitive requirement was evident in the "Eurovision Song Contest" item (Figure 2) where the participants had to consider another perspective—an item characteristic called *dialogical argumentation*.

Domain 2: Knowledge

The second domain of item characteristics was entitled "knowledge," which included the item characteristics *(d) cultural and social context of music* and *(e) familiarity of musical genre* (Table 3). Similar item characteristics (knowledge of music history, knowledge of musical styles and genres, and knowledge of cultural and social contexts of music) were investigated and empirically validated by Knigge (2010) and Jordan (2014).

In accordance with Rolle's (2013) theoretical competency model, some items of the MARKO test included information about the *cultural and social contexts of music*. The YouTube discussion around the item "Eurovision Song Contest" (Table 2) referenced "women's empowerment" and "social justice." To understand these references, the participants had to know about the respective discourses and be familiar with them.

The second item characteristic in this domain was *familiarity of musical genre*. The test items included music from various musical genres, such as classical music, pop, musical theater, and hip-hop. The degree to which the participants were familiar with different musical genres varied considerably. For several musical pieces presented in the test, the students provided information on their familiarity with a specific kind of music. This item characteristic captured whether the participants were familiar with the type of music presented in the item. Here, we hypothesized that a person who is familiar with a music genre has more knowledge about this genre and is, therefore, more likely to be able to solve a respective item. Therefore, this item characteristic should lessen the difficulty.

Domain 3: Formal item features

The third domain of item characteristics involved *formal item features* and included item characteristics dealing with the content of the item: *(f) text length, (g) linguistic demands,* and *(h) visuals*.

For the item characteristic *text length*, it became clear that a comparison of the two sample items illustrated earlier (Figures 1, 2) led to significant differences in the amount of text that the participants had to read in order to solve the item. Text length was also identified by Knigge (2010, p. 209) as a difficulty-increasing item characteristic.

The item characteristic *linguistic demands* referenced the vocabulary and grammatical structure used in an item. Nold and Rossa (2007) and Knigge (2010, p. 209) also identified linguistic demands as a difficulty-increasing item characteristic. While all items included the music that the participants were listening to, some items also contained a video or picture. This formal item feature was represented by the item characteristic *visuals*.

Coding item characteristics

Following the identification of the item characteristics, all polytomous item categories were coded. This coding process is exemplified in Figure 2 through the "Eurovision Song Contest" item. This item had two item categories because the participants' answers were rated with 0, 1, or 2 points (Table 2). If a person received one point for the item, they solved item category one, and if they received two points, they solved item category two. While the item characteristics of *Domain 2 (knowledge and familiarity;* Table 4) and the *formal item features (Domain 3*; Table 5) were the same for item categories one and two, *the cognitive requirements (Domain 1;* Table 3) differed between the item categories.

Table 6 shows the item characteristics for both item categories. While there was no need to refer to complex *musical attributes (a)*, consistency in reasoning *across several sentences (b)* was required

TABLE 3 Domain 1 of the identified item characteristics (predictors classified as the cognitive requirements for solving the item).

Domain 1: Cognitive requirements

Predictor	Code	Description
(a) Reference to musical attributes	0	To solve the item, only a reference to salient musical attributes (e.g., "loud," "soft," "long tones") or no musical attribute is necessary.
	1	To solve the item, a reference has to be made to musical attributes that are more complex than salient musical attributes.
	2	To solve the item, several musical attributes have to be named precisely.
(b) Cross-sentence argumentation	0	No elaborate reasoning is needed to solve the item.
	1	Reasoning has to be consistent across several sentences.
(c) Dialogical argumentation	0	There is no need to discuss different perspectives or opinions in the answer.
	1	To solve the item, participants have to take into account different opinions and perspectives on the presented musical piece.

TABLE 4 Domain 2 of the identified item characteristics (predictors classified as "knowledge").

Domain 2: Knowledge		
Predictor	**Code**	**Description**
(d) Cultural and social context of music	0	The social and cultural context of music is not relevant to the item.
	1	The social and cultural context of music is a central element of the item.
(e) Familiarity of musical genre	0	Participants are unfamiliar or somewhat familiar with the type of music heard in the task.
	1	Participants are very familiar with the type of music heard in the task.

TABLE 5 Domain 3 of the identified item characteristics (predictors classified as formal task features).

Domain 3: Formal item features		
Predictor	**Code**	**Description**
(f) Text length	0	Item contains little text
	1	Item contains a lot of text
(g) Linguistic demands	0	Vocabulary: use of high-frequency words Grammar: simple syntactic structures (parataxis, avoidance of complex structures)
	1	Vocabulary: less frequent words, extended vocabulary Grammar: more complex structures
(h) Visuals	0	Item does not include visuals (video/picture).
	1	Item does include visuals (video/picture).

TABLE 6 Coded item characteristics for the item "Eurovision song contest."

Domain	**1 Cognitive requirement**			**2 Knowledge**		**3 Formal item features**		
Item characteristic	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***
Item category 1	*0*	*1*	*0*	*1*	*1*	*1*	*1*	*1*
Item category 2	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*

for both item categories. *Dialogical argumentation (c)* was not needed to score one point for the "Eurovision Song Contest" item. However, it was required for receiving two points since other people's opinions had to be referenced.

Multiple regression analyses

Multiple regression analyses were conducted in the final step. Here, the item characteristics were used to predict item difficulty. The item difficulty parameters had been estimated with IRT scaling in a prior study (weighted likelihood estimation; Ehninger et al., 2021), where the collected test data were modeled as a partial credit model, and threshold parameters τ were estimated. This presented item difficulty for each item category, with a higher τ value indicating a more difficult item category.

In the multiple regression analyses, the dummy coded item characteristics were used to predict item difficulty τ. In the equation below, τ_i stands for the item difficulty parameter τ of item i. β_0 represents the regression constant and β_c the regression weight for the item characteristic c. Finally, x_{1i} is the code for an item characteristic (1 if the characteristic was present in the item, 0 if it was not).

$$\tau_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \cdots + \beta_c \cdot x_{ci}$$

The difficulty τ of each item category was modeled as the weighted sum of the item characteristics present in a given item category. The regression weights β_c represented the magnitude of influence of an item characteristic on item difficulty. Thus, an item with the characteristic c was β_c more difficult than an item without this item characteristic.

It was assumed that Domain 1 (cognitive requirements) would have a greater impact on item difficulty than the characteristics of the two other domains. Thus, three regression models were estimated. The first model was estimated with predictors from Domain 1 (*cognitive requirements*), the second with predictors from Domain 1 and Domain 2 (*cognitive requirements* and *knowledge*), and the third with the predictors from all three domains. The three models were then compared to one another, and the analyses were conducted in *R* (version 4.1.2). We also checked several assumptions of our data, such as homoscedasticity and multicollinearity. Both the beta coefficients and the collinearity statistics had to be acceptable (VIF < 10; variance inflation factor). In addition, we analyzed the standard errors of the regression coefficients and the part and partial correlations of each predictor variable.

Results

All the item characteristics were dummy coded in preparation for the multiple linear regression analyses. The two-factor variable "reference to musical attributes" had to be converted into two dummy variables ("reference to musical attributes 1" and

"reference to musical attributes 2"). All the item characteristics were rated by two raters, who agreed to a great extent ($\kappa[0.71,1]$).

Next, block-wise multiple linear regression analyses were conducted. The first model included predictors from Domain 1 (*cognitive requirements*); the second model included item characteristics from Domain 1 and Domain 2 (*cognitive requirements* and *knowledge*); and the third model yielded all item characteristics from all three domains (*cognitive requirements*, *knowledge*, and *formal task features*).

Table 7 shows the results of the regression analyses. All four predictors in Model 1 were significant. The reference to musical attributes was the strongest predictor $(\beta = 0.46, \beta = 0.51)$, followed by cross-sentence argumentation $(\beta = 0.37)$ and dialogical argumentation $(\beta = 0.20)$, $R_{adj}^2 = 0.71$. The predictors added in Model 2 and Model 3, however, lay above the significance level $(p > 0.05)$ and roughly explained the same observed variance in item difficulty as in Model 1 $\left(R_{adj}^2 = 0.71\right)$. A model comparison confirmed that Model 2 did not explain more variance than Model 1, $F(2,46) = 1.24, p = 0.30$, and that Model 3 did not explain more variance than Model 2, $F(3,43) = 0.73, p = 0.54$. Therefore, Model 1 suited our data best and met the assumption of non-multicollinearity ($VIF[1.08,2.04]$). Figure 3 shows four residual plots illustrating the model specification, the normal distribution of the residuals, the

TABLE 7 Regression results with the criterion of item difficulty (thresholds).

Predictor		*b*	*SE*	*beta*	*beta* 95% CI [LL, UL]	*p*	Fit
Model 1: Cognitive requirements							
(Intercept)		−0.60	0.19			<0.01	
(a)	Musical attributes 1	1.80	0.31	0.46	[0.30, 0.61]	<0.001	
	Musical attributes 2	2.55	0.51	0.51	[0.31, 0.72]	<0.001	
(b)	Cross-sentence argumentation	1.45	0.42	0.37	[0.15, 0.58]	<0.01	
(c)	Dialogical argumentation	1.54	0.69	0.20	[0.02, 0.38]	0.03	
							$R_{adj}^2 = 0.71$**
Model 2: Cognitive requirements and knowledge							
(Intercept)		−0.48	0.20			0.02	
(a)	Musical attributes 1	1.83	0.31	0.46	[0.30, 0.62]	<0.001	
	Musical attributes 2	2.62	0.51	0.53	[0.32, 0.73]	<0.001	
(b)	Cross-sentence argumentation	1.42	0.43	0.36	[0.14, 0.58]	<0.01	
(c)	Dialogical argumentation	1.59	0.78	0.21	[0.00, 0.41]	<0.05	
(d)	Context of music	0.18	0.49	0.04	[−0.16, 0.24]	0.72	
(e)	Familiarity	−0.49	0.31	−0.13	[−0.29, 0.04]	0.12	
							$R_{adj}^2 = 0.71$**
Model 3: Cognitive requirements and knowledge and formal item features							
(Intercept)		−0.73	0.27			<0.01	
(a)	Musical attributes 1	1.90	0.32	0.48	[0.32, 0.64]	<0.001	
	Musical attributes 2	2.73	0.52	0.55	[0.34, 0.76]	<0.001	
(b)	Cross-sentence argumentation	1.43	0.43	0.36	[0.14, 0.58]	<0.01	
(c)	Dialogical argumentation	1.47	0.81	0.19	[−0.02, 0.40]	0.08	
(d)	Context of music	−0.20	0.58	−0.04	[−0.28, 0.20]	0.74	
(e)	Familiarity	−0.49	0.32	−0.13	[−0.30, 0.04]	0.14	
(f)	Text length	0.24	0.45	0.06	[−0.15, 0.26]	0.59	
(g)	Linguistic demands	0.35	0.33	0.09	[−0.09, 0.27]	0.31	
(h)	Visuals	0.39	0.35	0.10	[−0.08, 0.27]	0.27	
							$R_{adj}^2 = 0.71$**

b represents unstandardized regression weights; beta indicates the standardized regression weights; LL and UL indicate the lower and upper limits of a confidence interval, respectively; R^2 represents the adjusted determination coefficient; and **indicates $p < 0.01$.

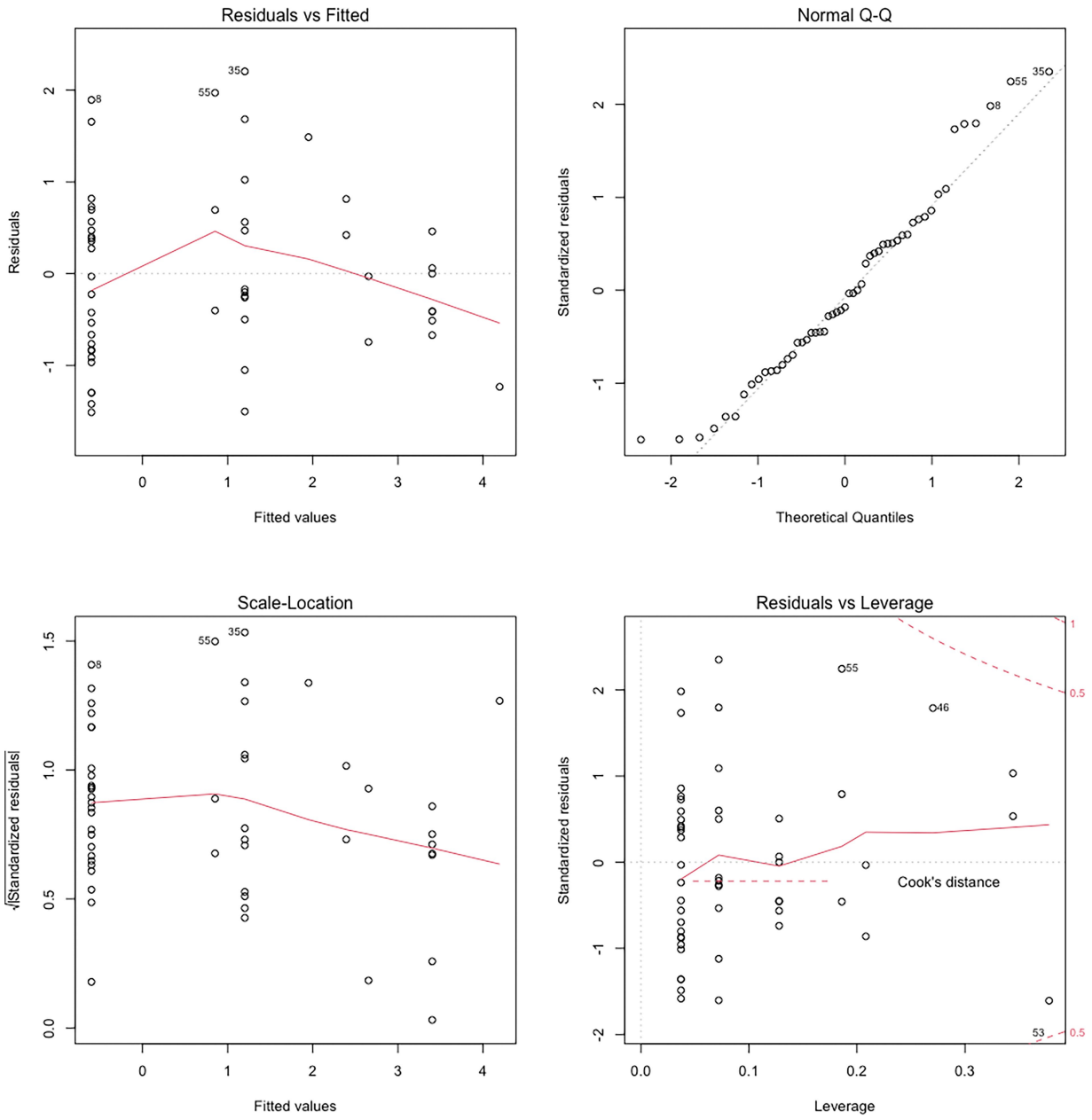


FIGURE 3
Four residual plots for regression Model 1 (cognitive requirements). The "Residuals vs. Fitted" plot (upper left) shows the fitted and unstandardized residual values. The "Normal Q-Q" plot illustrates the correct model specification and shows that the residuals are normally distributed. The assumption of homoscedasticity is shown in the "Scale-Location" diagram on the lower left. The outliers and influential values are shown on the plot "Residuals vs. Leverage" on the lower right (see also Luhmann, 2020, pp. 238–255; Field et al., 2012, pp. 266–276).

homoscedasticity assumption, and the identification of outliers and influential values.

Conclusion

Our findings show that the differences in item difficulty were predicted by the identified item characteristics. An important result was the categorization of the item characteristics into three domains: *cognitive requirements*, *knowledge*, and *formal item features*. A comparison of three regression models confirmed that Model 1, which held four predictors of the domain "cognitive requirements," best fit the study data ($R_{adj}^2 = 0.71$). The two regression models comprising predictors of the domains "knowledge" and "formal item features" failed to explain more variance. The strongest predictor in the final model was "reference to musical attributes" ($\beta = 0.46, \beta = 0.51$), followed by "cross-sentence argumentation" ($\beta = 0.37$) and "dialogical argumentation" ($\beta = 0.20$).

An interesting finding was that items containing visuals and longer or linguistically more complex texts were not more difficult. This is especially surprising since item characteristics related to reading skills have usually been found to increase item difficulty (see "item difficulty and item characteristics"). Therefore, we do not claim that reading skills are generally irrelevant for the MARKO test. On the contrary, we assume that linguistic skills are *particularly* important, which is reflected in two of the characteristics of the cognitive domain. Our analyses show that these specific features (cross-sentence argumentation and dialogical argumentation) were more important than the length or grammatical structure of the reading text. Thus, for individuals who were able to use complex and dialogical argumentation, it seemed to make no difference whether or not they had to read a great deal of text before completing an item. Technically speaking, we argue that it can be assumed that the length and complexity of an item text are relevant, in principle, but presumably, the linguistic features are confounded with each other so that only the strongest or most difficult characteristics could eventually be used as predictors.

Furthermore, we assumed that items containing music that was familiar to the participants were easier, but the respective predictor was not significant ($\beta = -0.49, p = 0.12$). However, it is important to note that we had little information about which musical pieces the students were familiar with. Therefore, further research needs to investigate the possible relation between familiarity with a musical genre and item difficulty.

Our research findings also have important implications for the music classroom and the question of how music-related argumentative competence can be fostered. The strongest predictor "reference to musical attributes" suggests that music-related perception is highly relevant when engaging in music-related argumentation. Thus, before being able to name a specific musical attribute, it first has to be perceived (see also Koelsch, 2019 and "domain 1: cognitive requirements"). The predictors "cross-sentence" and "dialogical argumentation" were both related to linguistic competence, pointing to the importance of linguistic skills when engaging in music-related argumentation. Further research needs to examine the interrelation between linguistic skills and music-related argumentative competence.

Although our findings seem promising, there are also some limitations of our methodological approach to music-related argumentation. Argumentation is an interactive event and an exchange of arguments with a real opponent can only be represented to a limited extent in a competency test. Although there were several items in the final MARKO test that imitated dialogical situations (such as the item "Eurovision Song Contest"), a competency test can never be as interactive as a conversation with an 'actual' person.

Our findings about the relations between item characteristics and item difficulty contribute to a better understanding of music-related argumentative competence in general and the validity of the competence test in particular. Since the final regression model only consists of item characteristics based on central assumptions hypothesized in the theoretical MARKO competency model (Rolle, 2013), this can be interpreted as proof of the construct validity of the MARKO test according to Borsboom et al. (2004). More specifically, our analyses support our assumption that the item characteristics ("attributes" in Borsboom et al., 2004's terminology) not only exist, but variations in the item characteristics causally produce variation in the competency test outcome. Furthermore, they can provide valuable information for scale anchoring in future studies (Hartig et al., 2012). The identified item characteristics can be important in developing further items measuring music-related argumentative competence, making it possible to determine beforehand which tasks are easier or more difficult and, therefore, can be developed for a specific requirement.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/t5cb8/?view_only=185c322c6dbf4a1d92aa88cd30b3afae.

Author contributions

All authors contributed to the design of the study and to the identification of the item characteristics of the study ("Identification and Categorization of Item Characteristics"). JE drafted the manuscript; JE and JK wrote sections of the manuscript. JE and JK conducted the statistical analyses; JE finalized statistical analyses. All authors revised this manuscript critically and made improvements on it. All authors approve the final version of the manuscript.

Funding

This project was supported by a grant from the Research Council of Norway (grant no. 320141).

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.

References

Beck, B., and Klieme, E. (Eds.). (2007). *Sprachliche Kompetenzen–Konzepte und Messung*: DESI-Studie (Deutsch Englisch Schülerleistungen International). Beltz.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bossen, A. (2017). "Sprache als Gegenstand der Musikpädagogischen Forschung und des musikdidaktischen Diskurses im Kontext einer Sprachbildung im Fach," in *Sprache im Musikunterricht. Ausgewählte Aspekte Sprachbewussten Handelns im Kontext von Inklusion*. eds. A. Bossen and B. Jank (Potsdam: Universitätsverlag Potsdam), 21–54.

Canale, M., and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Appl. Linguis.* 1, 1–47. doi: 10.1093/applin/1.1.1

Cohors-Fresenborg, E., Sjuts, J., and Sommer, N. (2004). "Komplexität von Denkvorgängen und Formalisierung von Wissen," in *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000*. ed. M. Neubrand (Wiesbaden: VS Verlag), 109–144.

Eemeren, F. H. V., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., and Wagemans, J. H. M. (2014). *Handbook of Argumentation Theory*. Dordrecht: Springer.

Ehninger, J. (2022). *Musikbezogenes Argumentieren. Testentwicklung und Kompetenzmodellierung* [Doctoral dissertation, University of Cologne]. KUPS. Available at: https://kups.ub.uni-koeln.de/61290/ (Accessed November 5, 2022).

Ehninger, J., Knigge, J., Schurig, M., and Rolle, C. (2021). A new measurement instrument for music-related argumentative competence: the MARKO competency test and competency model. *Front. Educ.* 6, 1–10. doi: 10.3389/feduc.2021.668538

Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*. Los Angeles: Sage Publications.

Hartig, J. (2007). "Skalierung und Definition von Kompetenzniveaus," in *Sprachliche Kompetenzen–Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. eds. B. Beck and E. Klieme (Weinheim: Beltz), 83–99.

Hartig, J. (2008). "Psychometric models for the assessment of competencies," in *Assessment of Competencies in Educational Settings*. eds. J. Hartig, E. Klieme and D. Leutner (Göttingen: Hogrefe & Huber), 69–90.

Hartig, J., Frey, A., Nold, G., and Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educ. Psychol. Meas.* 72, 665–686. doi: 10.1177/0013164411430707

Hartig, J., and Jude, N. (2007). "Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle," in *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung*. eds. J. Hartig and E. Klieme (Berlin: BMBF), 17–36.

Hartig, J., and Klieme, E. (2006). "Kompetenz und Kompetenzdiagnostik," in *Leistung und Leistungsdiagnostik*. ed. K. Schweizer (Heidelberg: Springer Medizin), 127–143.

Heller, V., and Morek, M. (2015). Academic discourse as situated practice: an introduction. *Linguist. Educ.* 31, 174–186. doi: 10.1016/j.linged.2014.01.008

Jordan, A.-K. (2014). "Empirische Validierung eines Kompetenzmodells für das Fach Musik–Teilkompetenz," in *Wahrnehmen und Kontextualisieren von Musik* (Münster: Waxmann)

Jordan, A.-K., Knigge, J., Lehmann, A. C., Niessen, A., and Lehmann-Wermser, A. (2012). Entwicklung und Validierung eines Kompetenzmodells im Fach Musik: Wahrnehmen und Kontextualisieren von Musik [development and validation of a competence model in music instruction. Perception and contextualization of music]. *Z. Pädagogik* 58, 500–521. doi: 10.25656/01:10392

Knigge, J. (2010). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz "Musik Wahrnehmen und Kontextualisieren"* [Doctoral dissertation, Universität Bremen]. Available at: http://nbn-resolving.de/urn:nbn:de:gbv:46-diss000120066 (Accessed November 5, 2022).

Knörzer, L., Stark, R., Park, B., and Rolle, C. (2016). "I like reggae and bob Marley is already dead": an empirical study on music-related argumentation. *Psychol. Music* 44, 1158–1174. doi: 10.1177/0305735615614095

Koelsch, S. (2019). "Neural basis of music perception: melody, harmony, and timbre," in *The Oxford Handbook of Music and the Brain*. eds. M. H. Thaut and D. A. Hodges (Oxford Academic), 187–211.

Kuhn, D. (2005). *Education for Thinking*. Cambridge: Harvard University Press.

Kultusministerkonferenz. (2005). Beschlüsse der Kultusministerkonferenz: Einheitliche Prüfungsanforderungen in der Abiturprüfung Musik. Available at: http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1989/1989_12_01-EPA-Musik.pdf

Lazarou, D., Sutherland, R., and Erduran, S. (2016). Argumentation in science education as a systemic activity: an activity-theoretical perspective. *Int. J. Educ. Res.* 79, 150–166. doi: 10.1016/j.ijer.2016.07.008

Luhmann, M. (2020). *R für Einsteiger: Einführung in die Statistik-Software für die Sozialwissenschaften*. Weinheim: Beltz.

Morek, M., and Heller, V. (2012). Bildungssprache–Kommunikative, epistemische, soziale und interaktive Aspekte ihres Gebrauchs. *Z. Angew. Linguisik* 57, 67–101. doi: 10.1515/zfal-2012-0011

Morek, M., Heller, V., and Quasthoff, U. (2017). "Erklären und Argumentieren. Modellierungen und empirische Befunde zu Strukturen und Varianzen," in *Begründen–Erklären–Argumentieren*. eds. I. Meißner and E. L. Wyss (Tübingen: Stauffenburg), 11–46.

Morrison, S. J., and Demorest, S. M. (2009). Cultural constraints on music perception and cognition. *Prog. Brain Res.* 178, 67–77. doi: 10.1016/S0079-6123(09)17805-6

Nold, G., and Rossa, H. (2007). "Hörverstehen," in *Sprachliche Kompetenzen-Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. eds. B. Beck and E. Klieme (Weinheim: Beltz), 178–196.

Prenzel, M., Häußler, P., Rost, J., and Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft* 30, 120–135. doi: 10.25656/01:7682

Quasthoff, U., Heller, V., and Morek, M. (2020a). "Diskurskompetenz und diskursive Partizipation als Schlüssel zur Teilhabe an Bildungsprozessen: Grundlegende Konzepte und Untersuchungslinien," in *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen*. eds. U. Quasthoff, V. Heller and M. Morek (Berlin: De Gruyter), 13–34.

Quasthoff, U., Wild, E., Domenech, M., Hollmann, J., Kluger, C., Krah, A., et al. (2020b). "Familiale Ressourcen für den Erwerb von Argumentationskompetenz," in *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen*. eds. U. Quasthoff, V. Heller and M. Morek (Berlin: De Gruyter), 79–106.

Rapanta, C. (2018). *Argumentation Strategies in the Classroom* . Wilmington, DE: Vernon Press.

Rolle, C. (2013). Argumentation skills in the music classroom: a quest for theory. In *Artistry*. eds. A. de Vugt and I. Malmberg Innsbruck: Helbling, 37–150.

Utdanningsdirektoratet. (2020). *Curriculum for Music (MUS01-02)*. Available at: https://www.udir.no/lk20/mus01-02?lang=eng (Accessed November 5, 2022).

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 23 February 2023
DOI 10.3389/feduc.2023.1049680

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Zhe Li,
Osaka University, Japan
Michael Schurig,
Technical University Dortmund, Germany

*CORRESPONDENCE
Ali Khoshnood
✉ a.khoshnood@pnu.ac.ir

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 20 September 2022
ACCEPTED 09 February 2023
PUBLISHED 23 February 2023

CITATION
Kafipour R and Khoshnood A (2023) Effect of feedback through dynamic assessment on EFL field-dependent and field-independent learners' speaking skill development. *Front. Educ.* 8:1049680. doi: 10.3389/feduc.2023.1049680

COPYRIGHT
© 2023 Kafipour and Khoshnood. 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.

Effect of feedback through dynamic assessment on EFL field-dependent and field-independent learners' speaking skill development

Reza Kafipour[1] and Ali Khoshnood[2]*

[1]Department of English Language, School of Paramedical Sciences, Shiraz University of Medical Sciences, Shiraz, Iran, [2]Faculty of Linguistics and Foreign Languages, Payame Noor University, Tehran, Iran

Introduction: Testing and assessment tools evaluate students' performance in a foreign language. Moreover, the ultimate goal of tests is to reinforce learning and motivate students. At the same time, instructors can gather information about learners' current level of knowledge through assessment to revise and enhance their teaching. This study aimed to investigate the effect of Dynamic Assessment on Iranian English as a Foreign Language (EFL) learners' speaking skills by considering language learners' cognitive styles (field dependence and field independence).

Methods: For this purpose, 60 Iranian intermediate-level EFL female learners were selected through convenience sampling from three language institutes with similar teaching methods in Shiraz, Iran. The current study has a quasi-experimental design since randomization was impossible. First, the authors used the Nelson Proficiency test and interview to determine the participants' proficiency level and speaking ability, respectively. Next, they took the group embedded figures test (GEFT) to determine the participants' type of cognitive style (field dependence or field independence). Next, the participants were randomly assigned to two experimental (FD and FI learners with the dynamic assessment) and two control groups. Paired and independent-sample t-test were applied to analyze the data.

Results and discussion: Results revealed that although dynamic assessment was effective for both experimental groups, the Field-dependent group with dynamic assessment outperformed the other. Thus, it can be concluded that in addition to the dynamic assessment, language learners' cognitive style can also play a vital role in increasing the assessment effectiveness. This type of assessment attracts instructors' attention to learners' potential to help the language learners gradually improve their performance. In addition, language institutes can introduce this new way of assessment in their advertisements and attract more students, leading to higher income and publicity for them.

KEYWORDS

dynamic assessment, field-dependent learners, field-independent learners, speaking skill, EFL

1. Introduction

While the need for learning English among English as a Foreign Language (EFL) learners increases globally, "there is a growing demand for standardized English language proficiency assessments as an objective measure to gauge each student's level of English language development" (Wolf and Butler, 2017, p. 3). Direct or indirect assessment is a topic of research in different studies as it can facilitate teaching and learning (Khoram et al., 2020; Saritas Akyol and Karakaya, 2021; Susilawati et al., 2022; Yusuf and Fajari, 2022). To address the increased need for appropriate measurement, assessment professionals and educators proposed dynamic assessment (DA) as a supplement for standardized testing, not a replacement for it (Lidz and Gindis, 2003). They believed that DA helps identify learners' differences and use these differences to enhance teaching. In this regard, DA emerged as a reaction to the traditional or static tests, which considers language learning as the outcome of the interaction between learners and instructors and considers teaching a part of the assessment. In contrast, there is no attempt to change the examinee's performance in the static assessment.

Assessing the language learners' abilities in second language evaluation has undergone different changes and development in the form, type, structure, and objectives behind. Consequently, testing an EFL learner's proficiency in speaking has been the primary concern of teachers who want to build practical criteria that can accurately assess oral reproduction because the evaluation is mostly subjective and many aspects of speaking such as pronunciation, pitch, tone, stress, intonation, sentence structures, and many other factors should be considered, so it is very important to have a practical, standard, and valid assessment procedure. Luoma (2004) believes that creating an instrument to evaluate concepts of accuracy and fluency and the learner's mastery of a spoken language's sound system and speech features is a difficult task. Test validation is another concern of the test developers to see if it serves as a reliable indicator of the level of student acquisition. And the other challenge is how to elaborate and describe pronunciation and its relevant standards (Luoma, 2004).

Hence, testing is needed to accompany teaching since it enables the teachers to change their teaching methods effectively to help different groups of students or individuals learn from their weaknesses by providing the details of their performances after the test (Heaton, 1989). Looking at different eras of language teaching in the past revealed that the emergence of each new approach and method in language teaching was followed by the appearance of a different language testing (Birjandi and Najafi Sarem, 2012). In addition, theorists and language teaching methodologists have developed language testing models, each appropriate for one language teaching model existing at a particular time.

Traditionally, the assessment was considered a way to gather information about learners' current level of knowledge. For this reason, researchers have called such assessments static assessments, which dominated language testing for many years (Heaton, 1989; Teo, 2012). Indeed, these static assessments aimed to determine if the learners have met expected achievement or not which is reflected in summative and formative assessment. Summative assessment measures how much a learner has learned after completion of the instruction while formative assessment measures how a learner is learning during instruction which is closer to dynamic assessment. Yet, the limitation of static assessment was that it might not stimulate learners to become independent knowledge constructors and problem solvers. Testing and teaching interact in DA to make learning successful. In traditional testing, examiners cannot intervene in the testing process, whereas, in DA, examiners are actively engaged in intervention and improvement of the examinees' cognitive outcome. According to Poehner (2008), instructors combine assessment and teaching as a single activity in DA. In contrast, they are distinguished from each other in traditional testing as different activities. He continued that in DA, the teacher tries to help language learners to complete a learning task.

On the other hand, Poehner (2005) believes that DA rejects the idea that any relation or interaction between examiners and examinees may negatively affect the reliability of the assessment. He claimed that DA assessment disregards the learners' performance in completing a specific task and tries to identify how much and what assistance the learners need. Moreover, according to holistic diagnostic feedback intervention, providing individualized feedback to the learners needs the instructor's expertise and skills. An experienced instructor in this issue can illicit and understand the students' weak points and strong points through examination, so the examination is not just for identification of achievement level; most importantly, it is used to provide better learning for the students. For this reason, intervention between examiner and examinee (learners and teachers) contributes to the adjustment of the learners' cognitive and metacognitive processes. And the students' self-regulation and motivation will be enhanced through feedback-driven strategies and skills (Von der Boom and Jang, 2018). The Sociocultural Theory of the Russian psychologist Vygotsky (1978), who initially proposed and formalized the approach as the zone of proximal development (ZPD), supports DA. Vygotsky (1978) stated that assessment, apart from revealing what led to the learners' poor performance, should provide solutions to remove problems and enhance the performance. According to ZPD, a more knowledgeable person can enhance the student's learning by guiding them through a task slightly higher than his/her ability level. So, whatever the student becomes more competent, the teacher gradually stops helping until the student can perform the task independently and completely.

Accordingly, when learners interact with others with higher knowledge and expertise, their learning will be positively influenced. ZPD confirms what people can do with the help of a more knowledgeable person is much more than what they can do individually. Dynamic assessment is a blend of instruction and assessment that coincide in an educational setting. Instructors, who use DA assessment in classrooms, try to support and help the learners by asking questions and providing hints or prompts as mediation. Mediation refers to the assistance provided by the instructor to help the learners find the answer and learn simultaneously (Malmeer and Zoghi, 2014). The instructors' interaction with learners is called reciprocity. Thus, reciprocity and mediation as two crucial elements in DA can show the learners'

Abbreviations: DA, dynamic assessment; EFL, English as a Foreign Language; GEFT, Group Embedded Figures Test; FD, field dependence; FI, field independence; ZDP, zone of proximal development; IELTS, International English Language Testing System; CLT, communicative language teaching; SPSS, Statistical Package for the Social Sciences.

progress and development in a specific area (Poehner and Lantolf, 2005).

As a recent approach, researchers have been interested in figuring out the effectiveness of DA on different language skills, especially speaking skills, which is the most significant one for EFL learners in communication with foreigners. However, in reality, many EFL learners are worried about their problems in oral production, and they always ask their instructors to help them improve their speaking skills (Rahmawati and Ertin, 2014).

However, besides the type of assessments, other factors can also fluctuate the effectiveness of this new approach, such as cognitive styles. Hence, it is acceptable to take into account field independence and field dependence cognitive styles while investigating the effectiveness of DA on EFL learners' speaking skill development as justified in the next paragraph. There are different studies which focused on learning styles, including visual, auditory, read/write, and kinesthetic and questioned their effectiveness in teaching and learning (Rohrer and Pashler, 2012; Newton and Salvi, 2020); however, some other studies showed them effective (Dunn and Dunn, 1993; Bates, 1994; Cassidy and Eachus, 2000; Birzer, 2003; Dunn and Griggs, 2004; Cassidy, 2010), so there are challenges and controversies in this issue which necessitates further evaluation as they claimed that learning styles are culture- and context-specific. Moreover, the current study is going to shed light on the effectiveness of another aspect of learning styles which is FD/FI rather than the traditional and early classification of the learning styles.

Field dependence/independence are among the learning styles that may enhance students' learning power and foster intellectual growth. The field-dependence/independence cognitive properties have continuously drawn researchers' attention. In the mid-70s, many researchers concluded that field-dependence/independence might have a crucial role in second/foreign language learning (Tucker et al., 1976). FI individuals tend to view the world objectively and make decisions based on an internal synthesis of relevant factors. FI learners are independent thinkers who focus on details separate from the context. These learners are characterized by their analytical approach and abilities to problem-solving.

On the other hand, FD learners focus on the overall meaning and the whole field. They are more relational, and they need more external reinforcements to keep them motivated. FI learners prefer formal learning contexts that respond to their competitive learning style (Witkin and Goodenough, 1981), while field-dependent learners, who are socially oriented and readily distracted, learn from the environments based on their experiences. As a result, they are less competitive compared with field-independent learners (Wooldridge, 1995).

Considering the EFL context, many Iranian EFL teachers have mentioned that many students are active and try to speak in their classrooms. However, their performance on the test cannot show their actual ability and vice versa. Therefore, it is vital to integrate testing and teaching and not judge the learners based on one round of performance. Consequently, it is crucial to create practical tests for EFL learners since they learn something from the test. Thus, dynamic assessment as an innovative type of assessment for EFL learners can be beneficial as it involves both assessment and instruction.

Some researchers have studied only the field-dependent and field-independent cognitive styles and tried to explore if the participants of their studies were field-dependent or field-independent (Cárdenas-Claros, 2005; Chapelle and Fraiser, 2009; Motahari and Norouzi, 2015; Mahevelati, 2019). Although the researchers have conducted several studies to show the contributive effect of dynamic assessment and even FD/FI on EFL learners' speaking skills, there is still room to highlight the learners' cognitive styles. Thus, the present study is designed to investigate the effect of dynamic assessment on EFL field-dependent and field-independent learners' speaking skill development. Moreover, some studies found FD and FI as effective to be considered in teaching and learning and some other studies found them ineffective. That is why the current study is going to evaluate and consider if they are found effective and make changes in the findings in our academic setting as FD and FI are context-specific and are officially called contextual factors (Kolb, 2015). As dynamic assessment is an interactive assessment which involves both teacher and learner, so the teachers' awareness of the learners' learning styles, here FD/FI, might play a role in teaching, learning, and assessment procedure. The primary purpose of this study is to examine the impact of dynamic assessment on EFL field-dependent and field-independent learners' speaking skills. Thus, this study seeks to answer the following questions:

1. Does dynamic assessment affect EFL field-dependent learners' speaking skills development?
2. Does dynamic assessment affect EFL field independent learners' speaking skills development?
3. Does dynamic assessment affect the development of EFL learners' speaking skills differently based on EFL learners' cognitive styles?

1.1. Theoretical framework

Dynamic assessment deals with identification of the individual differences and their implications in instruction and assessment. Dynamic assessment emphasizes on the processes rather than learning products. This type of assessment reflects what Vygotsky (1978) stated, "it is only in movement that a body shows what it is" (Gauvain, 2001, p. 35). For example, moving pictures imply different meanings and understandings compared to still pictures. DA has both psychoeducational and sociocultural importance; therefore, it emerged when product-oriented and static assessment failed to provide satisfactory results, along with the demand for culture-bound instruments which can consider contextual differences, such as socioeconomic, educational, and individual differences in language acquisition (Lidz, 1987; Haywood and Tzuriel, 1992; Lidz and Elliott, 2000).

Therefore, this study follows the theoretical framework proposed by Vygotsky (1978), whose sociocultural theory emphasizes the role of social interaction. ZPD has been considered as a guide for interaction in second language classrooms (Davin, 2013) and it refers to the difference between "individuals' actual ability and their potential for performing a task with assistance of a more capable individual." In another word, what people can do with the help of a more knowledgeable person is much more than what they can do individually.

2. Literature review

To date, multitudes of studies have been conducted to assess and explore the efficiency of DA on different aspects of language learning. Among the four language major skills, it seems speaking and oral productions of learners and components effective in speaking as vocabulary, in particular, have somehow received due empirical attention (Hayran, 2020; Uni, 2022). But Researchers such as O'Sullivan (2000), Poehner (2005), Hill and Sabet (2009), Davin (2013), Rahmawati and Ertin (2014), Karim and Haq (2014), Ahmadi Safa et al. (2016), Ebadi and Asakereh (2017), Hidri (2018), Minakova (2019), Safdari and Fathi (2020) have focused on the application of DA on learners' speaking abilities. However, the present study finds room to address the effect of dynamic assessment on EFL field-dependent and field-independent learners' speaking skill development, that is considering one more factor which might affect speaking and its assessment.

Hill and Sabet (2009) found that DA can enhance language learners' speaking skills while being an optimal means to assess the development of speaking skills. However, Ebadi and Asakereh (2017) argued that they overlooked learners' reciprocity and mediational patterns. These patterns are effective in obtaining reliable results. These studies show why the topic is challenging and should be investigated in different contexts.

Ableeva (2010) examined the impact of dynamic and traditional assessment on students' French listening skills. She concluded that DA, "due to its reliance on mediated dialogue, illuminates the sources of poor performance that are usually hidden during traditional assessments, which are non-dynamic in nature" (Ableeva, 2010, p. iv). Furthermore, she pointed out that DA can detect which areas learners need further improvement. However, this study, similar to most other studies, did not consider cognitive styles and field dependence, and field independence. Thus, the current study tried to cover the ignored aspects.

Teo (2012) examined the impact of dynamic assessment on Taiwanese EFL learners' reading skills. He applied DA to assess Taiwanese EFL college students' reading skills and taught them via mediation. His study indicated that suitable dynamic assessment procedures were beneficial in promoting learners' reading skills. However, this study just focused on reading skills, so there is no information about its effect on speaking skills as a productive skill in which many students have significant problems. That is why, the current study focused on speaking skills.

Malmeer and Zoghi (2014) attempted to determine the effect of dynamic assessment on Iranian EFL learners' grammar performance. They had 80 students as participants assigned into two groups of 40 (teenagers and adults). The results showed a significant difference between the pretest and posttest mean scores of the grammar test. Their study confirmed that adult EFL learners outperformed teenage EFL learners. As grammar is not taught directly anymore, it would be more effective to focus on the effect of assessment on different skills through which grammar is also used and practiced in future studies. That is why the current research focused on the impact of assessment on speaking skills.

Hidri (2014) initially examined the traditional assessment of listening skills prevalent among sixty Tunisian university EFL students. He argued that the current static assessment suffers from limitations and therefore proposed DA in that educational context to explore the relevance and effect of DA on the views of both the test-takers and raters. His study maintained that the tertiary learners could learn better when they join others in learning activities, which helps them overcome the test items' difficulty. He concluded that assessing the learners in a dynamic progress test can help "locate the areas of weaknesses in the language program or in the learners' cognitive and metacognitive strategies" (p. 15). This study also did not consider cognitive styles, field dependence, and field independence affecting the learners' performance.

Constant Leung pointed out two approaches to DA: interventionist and interactionist. He maintained that "Interventionist DA tends to involve quantifiable preprogrammed assistance and is oriented toward quantifiable psychometric measurement" (Leung, 2007, p. 260). In this approach, standardized interventions can measure learners' or groups of learners' ability in using "predetermined guidance, feedback, and support." Interactionist DA, however, "eschews measurement and is interested in the qualitative assessment of a person's learning potential" (p. 261).

Ahmadi Safa et al. (2016) investigated the influence of interventionist DA, interactionist DA, and non-DA on Iranian English language learners' speaking skills. They explored that interactionist DA improved the learners' speaking skills compared to interventionist DA and non-DA. However, Ebadi and Asakereh (2017) claimed that quantitative studies could not thoroughly reveal the learners' cognitive styles. They believed that qualitative studies best show learners' cognitive development via interpretation. To fill this gap, the current study focused on quasi-experimental design instead of qualitative approach.

Hidri (2018) examined the progress of the speaking skills of the EFL students of Persian Gulf countries. First, he argued that many teachers are constrained in class since male and female learners cannot communicate and interact due to sex segregation. Next, book designers consider the sociocultural context of Arabian countries in developing textbooks, and many daily conversations and topics are omitted as inappropriate materials according to Arabic traditions and religion. Although this study investigated speaking skills similar to the current study, it focused on cultural constraints and did not consider field dependence and independence. So, the current study covered this ignored aspect.

Siwathaworn and Wudthayagorn (2018) explored the impact of DA on Thai EFL university students who were found to be low proficient in speaking English. Their results showed that DA had promising potential and helped learners improve their speaking skills significantly in different ways. This study and similar studies act as the basis of hypotheses in the current study which point to the positive effects of dynamic assessment on speaking skills.

O'Sullivan (2000) has identified physical/physiological, psychological, and experiential characteristics as three factors affecting language learners' speaking test performance. The first group comprises unique measurements for examinees' physical illness or disabilities. Second, test takers' interests, emotional stage, motivation, and learning strategies. Finally, the third factor includes extrinsic impacts like former education, examination preparedness, examination experience, and communication experience. Siwathaworn and Wudthayagorn (2018) believed that DA could advocate all three sets of characteristics, and this is the reason why DA can assist learners in improving their speaking skills through speaking tests. This study fully supports

the design of the current study. Thus the present study seeks to see if the dynamic assessment can affect learners' speaking skills, as highlighted by Siwathaworn and Wudhayagorn.

3. Methodology

3.1. Design of the study

The present study employed a quasi-experimental design with a non-dynamic pretest and posttest design. After conducting the pretest, the participants took the group embedded figures test (GEFT) to be designated as Field Dependent or Field Independent learners. Next, they were randomly assigned to two experimental and two control groups. Finally, a posttest was given to the learners in the last session lasting 15 min.

3.2. Participants

The study participants included Persian intermediate-level EFL female learners from 3 language institutes (Parsian language institute, Goftar language institute, and Boostan language institute) in Shiraz. Randomization requires much time and financial support; that is why it was not feasible. Instead, the participants were selected through convenience sampling and then homogenized using the Nelson language Proficiency Test (Brown et al., 1993). To ensure homogeneity of the participants, the authors selected 60 participants, equally distributed between three institutes, among 120 test-takers whose score was one standard deviation below and one standard deviation above the mean, as the study samples. A standard deviation close to zero indicates that data points are close to the mean, whereas a high or low standard deviation indicates data points are respectively above or below the mean. That is why the scores having 1 SD above and below the mean were chosen to ensure the selection of intermediate learners and high and low scores were excluded. The participants' age range was 19–32 years. Fifteen learners were undergraduate students at university, 10 were high school students, and 7 had diploma. Moreover, those who had extremely high or low scores were excluded. Then, they were randomly assigned to experimental and control groups. The researchers decided to work on the intermediate level due to two reasons. First, based on their experience, most language learners concerned about finding a way to improve their speaking skills had intermediate levels. Second, more intermediate language learners were available in the institutes where the researchers decided to carry out their study.

3.3. Materials

The Top-Notch 3A student book (Saslow and Ascher, 2015) was used as the course material. Some researchers evaluated the *Top-Notch* series (Rezaiee et al., 2012; Alemi and Mesbah, 2013; Davari and Moini, 2016). They found that the series provides many interactions opportunities for EFL learners accompanied by positive and unbiased visual images. Many institutes in Shiraz, Iran, used this book as it consists of conversations and vocabularies which are practical for people who want to learn how to communicate with foreigners.

3.4. Instruments

First, the Nelson 350A Language Proficiency test (Fowler and Coe, 1976) was administered to the students to specify their level of proficiency and ensure the homogeneity of the sample. This test has a 50-item multiple-choice section with one close comprehension passage along with vocabulary, grammar, and pronunciation sections. It is a highly valid test whose validity and reliability have been estimated several times by Iranian researchers (Shahivand and Pazhakh, 2012). Next, the researchers applied the IELTS speaking skills test and rating scale to measure the students' speaking skill level at the beginning and end of the semester. It has three phases which last for 15 min. The first phase includes short questions and answers to make the candidate comfortable and familiar with the candidate. In the second phase, the candidate speaks on a specific topic for 2–3 min, and in the third phase, a two-way discussion about the subject between candidate and interviewer happens. The rating scale covers fluency, coherence, lexical resources, and pronunciation. Different researchers checked the validity and reliability of the test (Karim and Haq, 2014). Then, the authors administered the group embedded figures test (GEFT) developed by Witkin et al. (1971) to classify the participants into Field dependent and Field Independent learners. It has strong validity and reliability (Witkin et al., 1971). Pearson correlation coefficient (test-retest method) showed acceptable reliability for this test ($r = 0.82$). Moreover, Cronbach's alpha also showed a reliability coefficient of 0.87 which is quite acceptable.

The GEFT has three parts. The first part has seven items, and every one of the following two parts has nine questions. Part one is just for practice, so the number of simple figures correctly selected in parts two and three determines a participant's total score. Part one has a 2-min time limit followed by a 5-min time limit for parts two and three. Participants are to trace the simple figure embedded in the complex one. Raw scores range from 0 to 18, upper than 11.4 are identified as FI and the lower 11.4 as FD (Witkin et al., 1971).

3.5. Data collection procedure

The participants presented a short talk (10–15 min) to determine their speaking skills according to the IELTS speaking test rating scale. Considering IELTS speaking test rating scale, all 60 students had a band score of around three, so all 60 students were identified as homogenous in terms of their speaking skills and were included in the study.

Subsequently, these 60 students sat for the GEFT to find out which type of cognitive style each one owned. According to the GEFT scores, the students were divided into field-dependent and field-independent groups. Later, the researchers divided the participants into different experimental and control groups randomly. As dynamic assessment needs careful attention and is more practical for a small number of participants, 60 participants were divided into two experimental and two control groups. In addition, each experimental and control group involving 15

students was divided into classes of five students. Three of the classes were run by the dynamic assessment on Field dependent learners, while another three classes were run using the dynamic assessment on Field Independent learners, another three classes used the conventional method, communicative language teaching (CLT), on Field Dependent learners, and the last three classes used the conventional method (CLT) with Field Independent learners.

Then, the teachers applied DA for 12 sessions (12 weeks) providing the students with flexible mediation in a dialogue between the teacher and the learner. Each session lasted for 90 min. However, the control group students received traditional and communicative language teaching, although the learners were divided into field-dependent and field-independent control groups. They received the same material. CLT provides the students with real student-student and student-teacher communication, but the teacher will not provide feedback as it is done in dynamic assessment.

In the end, the participants took a non-dynamic posttest to see the effect of the treatment sessions on them. The pre and posttests were codified based on the IELTS rating scale to measure the students' speaking skills. To ensure the reliability of the scores, the authors asked two raters to score the participants' speaking skills to measure the inter-rater reliability to avoid the subjectivity of scoring as much as possible.

3.6. Data analysis procedure

The collected data through non-dynamic interviews were analyzed using SPSS software version 18. First, to understand whether there exists a significant difference between the Field dependent and Field independent groups, the authors ran an independent *t*-test as we have just one dependent variable and one independent one. Subsequently, the authors ran one independent *t*-test between each experimental group and its control group to ensure the effectiveness of the treatment. Finally, Pearson Correlation analysis was carried out in pre and posttest between the two raters' scores to ensure the reliability of their given scores.

4. Results

Pearson Correlation analysis showed agreement between two raters in pre-test (r = 0.82, sig. = 0.000) and post-test (r = 0.84, sig. = 000) in speaking skills test. So the rating of the scorer is considered reliable and acceptable.

4.1. Research question 1: *Does dynamic assessment affect EFL field-dependent learners' speaking skills development?*

The following table shows descriptive statistics for the students' performance in pretest and posttests in both dynamic and conventional assessment groups.

To answer the first question, pretest mean scores of both groups were analyzed to ensure the groups homogeneity in terms of speaking skills before treatment. As indicated in Table 1, the pretest mean score of the experimental group (M = 2.85, SD = 0.18) was slightly higher than the control group (M = 2.77, SD = 0.17) in which conventional assessment was conducted. An independent sample t-test was conducted to see if this difference is statistically significant. The results (t = 4.37, sig. = 0.102, $p > 0.05$) showed that this difference was not significant. It shows there was no significant difference between the two groups before treatment in terms of their speaking skills. When ensured about the homogeneity of the participants in both groups, the authors analyzed the post-test mean scores.

TABLE 1 Descriptive statistics for pretest and posttest mean scores of Field Dependent learners with dynamic assessment and conventional assessment.

Field dependent	*N*	Mean	SD
Pretest dynamic assessment	15	2.85	0.18
Posttest dynamic assessment	15	4.03	0.23
Pretest Conventional assessment	15	2.77	0.17
Posttest conventional assessment	15	3.67	0.16

According to Table 1, Field dependent learners in the experimental group that received dynamic assessment (M = 4.03, SD = 0.23) outperformed Field Dependent learners in the control group that received conventional assessment (M = 3.67, SD = 0.16). To ensure this finding is statistically significant, the authors ran an independent sample t-test. Based on the results (t = 5.50, sig. = 0.000, Cohen's d = 0.88), the difference between the mean scores of Field dependent learners who received dynamic and conventional assessment was significant ($P < 0.05$) with a large effect size and a statistical power of 95% calculated through Statistical Analysis System (SAS). In other words, Field Dependent learners with dynamic assessment significantly outperformed those who received a conventional assessment, which can lead the researchers to claim that the development of learners' speaking skills in the experimental group was due to the treatment.

4.2. Research question 2: *Does dynamic assessment affect EFL Field Independent learners' speaking skills development?*

First, the difference between pretest mean scores of field-independent learners with dynamic assessment and conventional assessment was evaluated using an independent sample t-test to ensure homogeneity of both groups before treatment.

According to Table 2, although the pretest mean score of the field-independent learners in the conventional assessment group (M = 2.79, SD = 0.14) was slightly higher than those in the dynamic assessment group (M = 2.70, SD = 0.19) before treatment, independent sample t-test did not show this difference as significant (t = 2.61, sig. = 0.109, $P > 0.05$). So, it can be concluded that both groups were homogenous, having similar speaking skills before treatment.

Comparing posttest scores in Table 3 shows that Field Independent learners in the experimental group who received dynamic assessment (M = 3.80, SD = 0.15) outperformed the Field-independent learners in the control group that received

TABLE 2 Descriptive statistics for pretest and posttest mean scores of Field Independent learners with dynamic and conventional assessment.

Field dependent	*N*	Mean	SD
Pretest dynamic assessment	15	2.70	0.19
Posttest dynamic assessment	15	3.80	0.15
Pretest conventional assessment	15	2.79	0.14
Posttest conventional assessment	15	3.68	0.17

TABLE 3 Descriptive and independent sample *t*-test for posttest mean scores of field independent and field dependent learners with dynamic assessment.

Dynamic assessment	*N*	Mean	SD	*t*	Sig.
Field independent	15	3.80	0.15	2.85	0.008
Field dependent	15	4.03	0.13		

conventional assessment (M = 3.68, SD = 0.17). To ensure this finding is statistically significant, the authors ran an independent sample t-test. The results (t = 2.14, sig. = 0.04, Cohen's d = 0.83) illustrated that Field Independent learners with dynamic assessment significantly outperformed the control group that received conventional assessment on the posttest ($P < 0.05$), which can be seen as a piece of evidence for confirming this issue that the development of learners in the experimental group was due to the treatment. Large effect size and a statistical power of 92% calculated through statistical analysis system (SAS) confirmed the results.

4.3. Research question 3: *Does dynamic assessment affect the development of EFL learners' speaking skills differently based on EFL learners' different types of cognitive styles?*

As Table 3 illustrates, there was a difference in the posttest scores for Field Dependent learners (M = 4.03, SD = 0.13) and Field Independent learners who received dynamic assessment (M = 3.80, SD = 0.15). Field-dependent learners outperformed Field-independent learners. An independent sample t-test was applied to test if this finding is significant. According to the results, the Field-dependent learners with dynamic assessment treatment significantly outperformed the Field- independent learners with dynamic assessment (t = 2.85, $P < 0.05$, Cohen's d = 0.90) with a large effect size and a statistical power of 88% calculated through SAS. These results suggest that although dynamic assessment is generally an effective way to help students improve their speaking performance, it is more beneficial for Field-dependent learners.

5. Discussion

This study investigated the effectiveness of dynamic assessment on Iranian EFL Field Dependent and Field Independent intermediate learners' speaking skills development. The two experimental instructional settings involved dynamic assessment for EFL Field Dependent learners and EFL Field independent learners separately. The first research question aimed to clarify whether the dynamic assessment is practical for improving field-dependent EFL learners' speaking skills. Results revealed the effectiveness of dynamic assessment on Persian EFL Field Dependent learners' speaking performance. Therefore, the answer to the first question was affirmative, leading to the efficiency of dynamic assessment on Persian EFL Field Dependent learners' speaking skill improvement.

The second research question was to determine if the dynamic assessment can help EFL intermediate learners with field independence cognitive style improve their speaking skills. The results obtained from their pretest and posttest illustrated the usefulness of dynamic assessment on Persian EFL Field Independent learners' speaking performance. Therefore, the answer to the second question was also affirmative, indicating the effectiveness of dynamic assessment on Persian EFL Field Independent learners' speaking skill improvement.

These findings align with Siwathaworn and Wudthayagorn (2018), who claimed that dynamic assessment affects tertiary EFL students' speaking skills. They tried to help the students improve their speaking skills in the elicitation limitation test task, where they were encouraged to repeat sentences. The results showed the positive effect of DA on the students' speaking skills. However, Davin and Donato (2013) criticized that only selected learners will find an opportunity to react actively to teacher mediation when DA is practiced in the classroom, so this method "limits the cognitive engagement of a majority of students in benefiting from the teacher's mediation." They further pointed out that "due to time constraints and the large number of students in a classroom, classroom DA alone is effective for those students who actively participate, but it is not sufficient to promote and monitor the language development of every student in a classroom" (p. 6).

Minakova (2019) carried out an experimental study to determine the impact of DA in language development through standardized test preparation. Despite the growing consensus on the fruitfulness of DA, she highlights an important limitation in her study, as "it does not offer solutions for teachers who work with large groups of students. The mediation program implemented in the present study was based on the individual meetings with the mediator, and its outcomes are more relevant to private IELTS tutors" (p. 206).

Another researcher, Teo (2012), also revealed the usefulness of dynamic assessment on Taiwanese EFL learners' reading skills. In another study, Poehner (2008) assessed the speaking skills of advanced French undergraduate learners. The results of his research agreed with this study which refers to the effectiveness of the dynamic assessment. His findings indicated that mediation, one of the significant parts of dynamic assessment, helped the learners better comprehend two tenses of *imparfait* and *passé compose* in French.

Ableeva (2010) was also another researcher who illustrated the effectiveness of the dynamic assessment. This researcher investigated the L2 listening comprehension ability of French university learners. These findings corroborate the results of many previous studies that confirmed the effectiveness of dynamic assessment in various instructional contexts (Lantolf and Poehner, 2011; Shrestha and Coffin, 2012; Teo, 2012; Nazari and Mansouri, 2014; Sadighi et al., 2018; Safdari and Fathi, 2020).

Finally, the findings revealed that dynamic assessment was more beneficial for Field Dependent learners, although Field-independent learners were not detached from the effectiveness of this treatment. Rassaei (2014) and Hoffman (1997) believe that this is because Field-dependent learners generally need guidance and assistance from the instructor and intend to interact with people. Furthermore, they add that field-dependent individuals take a holistic approach while under the influence of their surrounding context. In other words, Field-dependent learners are more successful in situations that need social sensitivity and empathizing with others. Consequently, dynamic assessment through interaction can be beneficial for them.

Furthermore, dynamic assessment in the mediation form of interaction provides a situation where the mediating agent, like the teacher, engages in a task with a learner and offers as much mediation as required to support the learner's performance in an activity (Davin, 2013).

The results are aligned with the studies done by Rassaei, 2014, Niroomand and Rostampour, 2014, and Wapner and Demick (2014). They claimed that the degree of the effectiveness of different treatments on second language learners' performance depends on their different cognitive learning styles (field dependence and field independence). So, these studies show that this learning styles as FD/FI still matters and should be considered to see if they affect the performance. Generally, depending on the situation and the kind of treatment, in some studies, Field-independent learners outperform field-dependent learners or vice versa. Therefore, in this study, as the dynamic assessment was more interaction and the teachers acted like assistants for language learners, Field-dependent learners with dynamic assessment treatment outperformed the other group. Therefore, the leaning style, in this study FD/FI should be controlled and considered in EFL in the future at least in the academic setting where the current study was conducted.

The findings of this study can benefit foreign language teachers, testers, and learners since foreign language teachers and testers will become aware of the effectiveness of a new way of assessment called dynamic assessment, which involves both assessment and instruction simultaneously. Thus, it can be beneficial in language teaching and help the learners improve their speaking skills. Moreover, the instructors and testers will notice whether EFL learners' cognitive styles interfere with dynamic assessment's effect on their speaking skills. New suggestions can be proposed with this method for improving the EFL learners' speaking skills.

6. Conclusion

The reported results in this study maintain several implications to provide more effective teaching and learning perceptions. The results can make EFL teachers aware of dynamic assessment that involves instruction and assessment simultaneously. In other words, the instructors offer assistance to students while assessing them simultaneously. The teachers gain a clearer understanding of the language learners' future by paying attention to language learners' responses to the mediations.

Indeed, teachers become familiar with the benefits of dynamic assessment on EFL learners' performance. This type of assessment attracts their attention to learners' potential and leads them to help the language learners gradually improve their performance. On the other hand, language learners figure out their potential development and promote their language skills. In addition, language institutes can introduce this new way of assessment in their advertisements and attract more students, leading to higher income and publicity for them.

Minakova's (2019) study corroborates both present and past studies, and she argues that her findings have crucial implications for educators. She furthers that "providing mediation during assessment allows them to uncover learners' latent abilities instead of simply documenting their current achievements. In other words, DA explores how one's performance is modifiable and what kind of mediation is needed to promote development within the learners' ZPD" (p. 186). To sum up, the current study showed the effect of dynamic assessment in developing the learners' speaking skills. Moreover, it showed that learners cognitive learning styles affect their performance when using dynamic assessment so cognitive styles as a type of individual difference among learners should be considered.

Data availability statement

The original contributions presented in this study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

Both authors declared to have equally contributed in writing the manuscript and read and approved the final version of submitted manuscript.

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.

References

Ableeva, R. (2010). *Dynamic assessment of listening comprehension in second language learning*, Ph.D. thesis. State College, PA: The Pennsylvania State University.

Ahmadi Safa, M., Donyaei, S., and Malek Mohamadi, R. (2016). An investigation into the effect of interactionist versus interventionist models of dynamic assessment on Iranian EFL learners' speaking skill proficiency. *J. Teach. English Lang.* 9, 153–172.

Alemi, M., and Mesbah, Z. (2013). Textbook evaluation based on the ACTFL standards: The case of top notch series. *Iran. EFL J.* 9, 162–179.

Bates, T. (1994). Career development for higher flyer. *Manag. Dev. Rev.* 7, 20–24. doi: 10.1108/09622519410771763

Birjandi, P., and Najafi Sarem, S. (2012). Dynamic assessment: An evolution of the current trends in language assessment and testing. *Theory Pract. Lang. Stud.* 2, 747–753. doi: 10.4304/tpls.2.4.747-753

Birzer, M. L. (2003). The theory of andragogy applied to police training. *Policing* 26, 29–42. doi: 10.1108/13639510310460288

Brown, J. A., Fishco, V. V., and Hanna, G. (1993). *Nelson-Denny reading test: Manual for scoring and interpretation, forms G & H*. Rolling Meadows, IL: Riverside Publishing.

Cárdenas-Claros, M. (2005). *Field dependence/field independence: How do students perform in CALL based listening activities?* Ames, IA: Iowa State University.

Cassidy, S. (2010). Learning styles: An overview of theories, models, and measures. *Educ. Psychol.* 24, 419–444. doi: 10.1080/0144341042000228834

Cassidy, S., and Eachus, P. (2000). Learning style, academic belief systems, self-report student proficiency and academic achievement in higher education. *Educ. Psychol.* 20, 307–322. doi: 10.1080/713663740

Chapelle, C., and Fraiser, T. (2009). Individual learner differences in call: The field independence/ dependence (FID) construct. *CALICO J.* 26, 246–266. doi: 10.1558/cj.v26i2.246-266

Davari, S., and Moini, M. R. (2016). The representation of social actors in top notch textbook series: A critical discourse analysis perspective. *Int. J. Foreign Lang. Teach. Res.* 4, 69–82.

Davin, K. J. (2013). Integration of dynamic assessment and instructional conversations to promote development and improve assessment in language classroom. *Lang. Teach. Res.* 17, 1–38. doi: 10.1177/1362168813482934

Davin, K. J., and Donato, R. (2013). Student collaboration and teacher-directed classroom dynamic assessment: A complementary pairing. *Foreign Lang. Ann.* 46, 5–22. doi: 10.1111/flan.12012

Dunn, R., and Dunn, K. (1993). *Teaching secondary students through their individual learning styles: Practical approach for grades 7-12*. Boston, MA: Allyn and Bacon.

Dunn, R., and Griggs, S. A. (Eds.) (2004). *Synthesis of the Dunn and Dunn learning-style model research: Who, what, when, where, and so what?* New York, NY: St. John's University's Center for the Study of Learning and Teaching Styles.

Ebadi, S., and Asakereh, A. (2017). Developing EFL learners' speaking skills through dynamic assessment: A case of a beginner and an advanced learner. *Cogent Educ.* 4:1419796. doi: 10.1080/2331186X.2017.1419796

Fowler, W. S., and Coe, N. (1976). *English language tests*. London: Butler & Tanner LTD.

Gauvain, M. (2001). *The social context of cognitive development*. New York, NY: Guilford.

Hayran, Z. (2020). Examining the speaking self-efficacy of pre-service teachers concerning different variables. *Eurasian J. Educ. Res.* 90, 1–18. doi: 10.14689/ejer.2020.90.1

Haywood, H. C., and Tzuriel, D. (1992). *Interactive assessment*. New York, NY: Springer-Verlag. doi: 10.1007/978-1-4612-4392-2

Heaton, J. B. (1989). *Writing English language test: Longman handbooks for language teachers*. London: Longman Inc.

Hidri, S. (2014). Developing and evaluating a dynamic assessment of listening comprehension in an EFL context. *Lang. Test. Asia* 4:4. doi: 10.1186/2229-0443-4-4

Hidri, S. (2018). "Assessing spoken language ability: A many-facet Rasch analysis," in *Revisiting the assessment of second language abilities: From theory to practice*, ed. S. Hidri (Midtown Manhattan, NY: Springer International Publishing), 23–48. doi: 10.1007/978-3-319-62884-4

Hill, K., and Sabet, M. (2009). Dynamic speaking assessments. *TESOL Q.* 43, 537–545. doi: 10.1002/j.1545-7249.2009.tb00251.x

Hoffman, S. (1997). Field Dependence/Independence in second language acquisition and implications for educators and instructional designers. *Foreign Lang. Ann.* 30, 222–234. doi: 10.1111/j.1944-9720.1997.tb02344.x

Karim, S., and Haq, N. (2014). An assessment of IELTS speaking test. *Int. J. Eval. Res. Educ.* 3, 152–157.

Khoram, A., Bazvand, A. D., and Sarhad, J. S. (2020). Error feedback in second language speaking: Investigating the impact of modalities of error feedback on intermediate EFL students' speaking ability. *Eurasian J. Appl. Linguist.* 6, 63–80. doi: 10.32601/ejal.710205

Kolb, D. A. (2015). *Experiential learning: Experience as the source of learning and development*, 2nd Edn. Upper Saddle River, NJ: Pearson Education.

Lantolf, J. P., and Poehner, M. E. (2011).). Dynamic assessment in the classroom: Vygostkian praxis for L2 development. *Lang. Teach. Res.* 15, 11–33. doi: 10.1177/1362168810383328

Leung, C. (2007). Dynamic assessment: Assessment for and as teaching? *Lang. Assess. Q.* 4, 257–278. doi: 10.1080/15434300701481127

Lidz, C. S. (1987). "Historical perspectives," in *Dynamic assessment: An interactional approach to evaluating learning potential*, ed. C. S. Lidz (New York, NY: Guilford), 3–34.

Lidz, C. S., and Elliott, J. (Eds.) (2000). *Dynamic assessment: Prevailing models and applications*. Amsterdam: Elsevier Science.

Lidz, C. S., and Gindis, B. (2003). "Dynamic assessment of the evolving cognitive functions in children," in *Vygotsky's educational theory in cultural context*, eds A. Kozulin, B. Gindis, V. S. Ageyev, and S. M. Miller (Cambridge: Cambridge University Press). doi: 10.1017/CBO9780511840975.007

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Mahevelati, E. (2019). Field-dependent/independent cognitive style preferences of EFL learners in an implicit learning task from information processing perspective: A qualitative investigation. *Int. J. Res. Stud. Lang. Learn.* 8, 39–60. doi: 10.5861/ijrsll.2019.3009

Malmeer, E., and Zoghi, M. (2014). Dynamic assessment of grammar with different age groups. *Theory Pract. Lang. Stud.* 14, 1707–1713.

Minakova, V. (2019). Dynamic assessment of IELTS speaking: A learning-oriented approach to test preparation. *Lang. Sociocult. Theory* 6, 184–212. doi: 10.1558/lst.36658

Motahari, M. S., and Norouzi, M. (2015). The difference between field independent and field dependent cognitive styles regarding translation quality. *Theory Pract. Lang. Stud.* 5, 1799–2591. doi: 10.17507/tpls.0511.23

Nazari, B., and Mansouri, S. (2014). Dynamic assessment versus static assessment: A study of reading comprehension ability in Iranian EFL learners. *J. Lang. Linguist. Stud.* 10, 134–156.

Newton, P. M., and Salvi, A. (2020). How common is belief in the learning styles neuromyth, and does it matter? A pragmatic systematic review. *Front. Educ.* 5:602451. doi: 10.3389/feduc.2020.602451

Niroomand, S. M., and Rostampour, M. (2014). The impact of field dependence/independence cognitive styles and gender differences on lexical knowledge: The case of Iranian academic EFL learners. *Theory Pract. Lang. Stud.* 4, 2173–2179. doi: 10.4304/tpls.4.10.2173-2179

O'Sullivan, B. (2000). *Towards a model of performance in oral language testing*. Reading: University of Reading.

Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French*. State College, PA: The Pennsylvania State University.

Poehner, M. E. (2008). *Dynamic assessment: A Vygostkian approach to understanding and promoting second language development*. Berlin: Springer.

Poehner, M. E., and Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Lang. Teach. Res.* 9, 233–265. doi: 10.1191/1362168805lr166oa

Rahmawati, Y., and Ertin, E. (2014). Developing assessment for speaking. *Indones. J. English Educ.* 1, 199–210. doi: 10.15408/ijee.v1i2.1345

Rassaei, E. (2014). Recasts, field dependence/independence cognitive style, and L2 development. *Lang. Teach. Res.* 19, 1–21. doi: 10.1177/1362168814541713

Rezaiee, A. A., Kouhpaeenejad, M. H., and Mohammadi, A. (2012). Iranian EFL learners' perspectives on new interchange series and top-notch series: A comparative study. *Proc. Soc. Behav. Sci.* 70, 827–840. doi: 10.1016/j.sbspro.2013.01.128

Rohrer, D., and Pashler, H. (2012). Learning styles: Where's the evidence? *Med. Educ.* 46, 634–635. doi: 10.1111/j.1365-2923.2012.04273.x

Sadighi, F., Jamasbi, F., and Ramezani, S. (2018). The impact of using dynamic assessment on Iranian's writing literacy. *Theory Pract. Lang. Stud.* 8, 1246–1251. doi: 10.17507/tpls.0809.21

Safdari, M., and Fathi, J. (2020). Investigating the role of dynamic assessment on speaking accuracy and fluency of pre-intermediate EFL learners. *Cogent Educ.* 7:1818924. doi: 10.1080/2331186X.2020.1818924

Saritas Akyol, S., and Karakaya, I. (2021). Investigating the consistency between students' and teachers' ratings for the assessment of problem-solving skills with many-facet rasch measurement model. *Eurasian J. Educ. Res.* 91, 281–300. doi: 10.14689/ejer.2021.91.13

Saslow, J., and Ascher, A. (2015). *Top notch 3 A*. New York, NY: Pearson Longman.

Shahivand, Z., and Pazhakh, A. (2012). The effects of test facets on the construct validity of tests in Irania EFL students. *High. Educ. Soc. Sci.* 2, 16–20.

Shrestha, P., and Coffin, C. (2012). Dynamic assessment, tutor mediation and academic writing development. *Assess. Writ.* 17, 55–70. doi: 10.1016/j.asw.2011.11.003

Siwathaworn, P., and Wudthayagorn, J. (2018). The impact of dynamic assessment on tertiary EFL students' speaking skills. *Asian J. Appl. Linguist.* 5, 142–155.

Susilawati, E., Lubis, H., Kesuma, S., and Pratama, I. (2022). Antecedents of student character in higher education: The role of the automated short essay scoring (ASES) digital technology-based assessment model. *Eurasian J. Educ. Res.* 98, 203–220.

Teo, A. K. (2012). Effects of dynamic assessment on college EFL learners' reading skill. *J. Asia TEFL* 9, 57–94.

Tucker, G., Hamayan, E., and Genesee, F. H. (1976). Affective, cognitive and social factors in second language acquisition. *Can. Modern Lang. Rev.* 32, 214–226.

Uni, K. (2022). Benefits of Arabic vocabulary for teaching Malay to Persian-speaking university students. *Eurasian J. Appl. Linguist.* 8, 133–142.

Von der Boom, E. H., and Jang, E. E. (2018). The effects of holistic diagnostic feedback intervention on improving struggling readers' reading skills. *J. Teach. Learn.* 12, 54–69. doi: 10.22329/jtl.v12i2.5105

Vygotsky, L. (1978). *Mind in society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Wapner, S., and Demick, J. (2014). *Field dependence-independence: Bio-psycho-social factors across the life span*. New York, NY: Psychology Press. doi: 10.4324/9781315807218

Witkin, H. A., and Goodenough, D. R. (1981). *Cognitive styles: Essence and origin*. New York, NY: International University Press.

Witkin, H. A., Oltman, P., Raskin, E., and Karp, S. (1971). *A manual for the embedded figures test*. Palo Alto, CA: Consulting Psychologist Press.

Wolf, M. K., and Butler, Y. G. (2017). *English language proficiency assessments for young learners*. New York, NY: Routledge. doi: doi.org/10.4324/9781315674391-1

Wooldridge, B. (1995). "Increasing the effectiveness of university/college instruction: Integrating the result s of learning style research into course design and delivery," in *The importance of learning styles: Understanding the implications for learning, course design, and education*, eds R. R. Sims and S. J. Sims (Westport, CT: Greenwood Press), 49–67.

Yusuf, F. A., and Fajari, L. E. W. (2022). Key success factors of various quality assessment institutions and quality of higher education services: A meta-analysis study. *Eurasian J. Educ. Res.* 98, 184–202.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 30 March 2023
DOI 10.3389/feduc.2023.1115259

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland,
New Zealand

REVIEWED BY
Zi Yan,
The Education University of Hong Kong,
Hong Kong SAR, China
Elena Tikhonova,
Peoples' Friendship University of Russia,
Russia

*CORRESPONDENCE
Ana Remesal
✉ aremesal@ub.edu

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 03 December 2022
ACCEPTED 23 February 2023
PUBLISHED 30 March 2023

CITATION
Remesal A and Estrada FG (2023) Synchronous self-assessment: First experience for higher education instructors.
Front. Educ. 8:1115259.
doi: 10.3389/feduc.2023.1115259

COPYRIGHT
© 2023 Remesal and Estrada. 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.

Synchronous self-assessment: First experience for higher education instructors

Ana Remesal* and Flor G. Estrada

Departamento de Cognición, Desarrollo y Psicología de la Educación, Facultad de Psicología, Universidad de Barcelona, Barcelona, Spain

We present a qualitative study of four cases of university instructors (teacher educators) implementing synchronous self-assessment (SSA). SSA consists of an innovative assessment strategy during written exam situations, which highlights the students' voice and agency, giving it greater weight in the power balance traditionally established between instructors and students in classroom assessment practices. In this article, we focus on the effects and pedagogical potential of this assessment strategy from instructors' point of view. In our study, three instructors were novels in implementing this strategy; the fourth instructor had several years of experience with it. The four instructors agreed on basic design features for an end-of-semester exam offered in four groups of first-year students of the same shared program at a Bachelor's degree for Kindergarten Educator and Primary School Teacher. The instructors were individually interviewed after the assessment session in their course and the exams were gathered for analysis. Content and discursive analysis was carried out on the data. Results show substantial differences in the evaluative artefacts (instructors' exams) in terms of cognitive demand and formative assessment potential, and point to noticeable needs for professional development in pursuit of assessment literacy in Higher Education.

KEYWORDS

self-assessment, higher education, assessment literacy, teachers' conceptions of assessment, innovation, exam design

1. Introduction

Learning assessment is still today one of the most significant challenges for instructors, regardless of educational level. In an international context of curricular renewal focused on the development of competencies, new challenges arise for teaching and assessment. In this article, we want to present an innovative proposal that leads us to analyse the potential for improvement of particular exam situations as a whole (Hernández Nodarse, 2007), unlike other previous proposals that refer to the nature of specific assessment activities. (e.g., Villarroel et al., 2021).

Multiple-choice standardised exams are frequent in higher education, especially in non-humanistic areas (e.g., López Espinosa et al., 2014; Roméu and Díaz Quiñones, 2015; Herrero and Medina, 2019; Imbulpitiya et al., 2021). Interest in test design is relatively recent in the university context. Up to now, literature on classroom assessment does not recognise a test's unitary value as an interactive classroom experience (Hernández Nodarse, 2007) but analyse assessment tasks in a non-contextual and isolated way. In our approach, however, exams are precisely taken as a unitary interactive experience in the teaching-learning process, as we will explain later.

Xu and Brown's (2016) most recent review of teacher education for assessment literacy is the first to highlight the importance of instructors' beliefs and conceptions about assessment itself, as well as their emotional experience. The 13 key points highlighted by Popham 2 years earlier (2014) did not yet include them. University instructors' assessment literacy still needs to be improved. Many instructors lack specific initial training for assessing unless their field of research, or their individual motivation, leads them to explore the teaching and learning processes (López Espinosa et al., 2014). Thus, we face a deficient field of knowledge as a starting point.

In this article, we will first present the essence of our proposed evaluative strategy. Second, we report the results of a quadruple case study in the context of higher education; finally, we will provide a set of theoretical reflections and proposals for open lines of research.

1.1. What is synchronous self-assessment?

Synchronous Self-Assessment (SSA) is an innovative classroom assessment strategy (Remesal et al., 2019; Remesal, 2021) that allows the students to make crucial and impact-full decisions within their assessment process, recognising and valuing their learning, thus emphasising their agency. Ideally, this strategy comes to life in written exam situations. Unlike other assessment situations or activities, the exam has particular characteristics that make it ideal for SSA. First, it is an explicit assessment situation, where all the participants are aware of a series of rules that aim at exposing the learning generated or elaborated during a specific time. Second, it happens under a certain time pressure in which students know that their best version of the learning effort made is actually at stake. Therefore, the quality of synchronicity does not refer to technological aspects on this occasion. It indicates the simultaneity of the hetero –or external-assessment processes as led by the instructor and the self –or internal-assessment processes led by the students in the shared classroom interactive space and concerning the same assessment activities.

The students' active role comes to the foreground through two decisions they must take. First, they must select a series of activities from a total to solve. The number of activities to choose and solve will depend on the educational level and the time available for completion. Secondly, they must also choose a weighted grading for their solved activities (equal scores, maximum or minimum difference). The grading options may also vary regarding the educational level and duration of the exam. Unlike the proposals about self-grading (Crowell, 2015), SSA provokes a qualitative comparison of personal performance in a short series of solved assessment activities immediately after resolution. SSA relates to recent conceptual proposals such as the evaluative judgement (Tai et al., 2018). Since this concept is not associated with any specific pedagogical measure, we propose that SSA would offer a concrete way to educate towards this evaluative judgement, as students make a value judgement about their performance in a short series of activities in a comparative manner, applying personal quality criteria that they may construct throughout the learning process.

SSA rebalances the power relation between instructor and student in the assessment situation, with the student assuming much more responsibility in strategic aspects such as "By which activities am I going to demonstrate the learning I have achieved?" and "How will my performance be valued?." With this innovative proposal, students actively participate in their learning assessment. SSA launches deep metacognitive processes that potentially lead the students to a new and greater awareness of everything learned. Thus, facilitating internal self-assessment in a natural way (Nicol, 2021) as a subjective phenomenon that accompanies the entire learning process and, of course, its evaluation (Yan and Brown, 2017; Yan, 2020), opening up new learning opportunities (Yan and Boud, 2021) and, eventually, promoting self-awareness and emotional self-management. Thus, we propose that the exam situation, as a moment of purposeful assessment, explicitly shared between the participants, with the goal of external demonstration of the maximum learning achieved, and all this under a certain time pressure (Remesal et al., 2022), is an ideal opportunity to encourage the student's agency.

1.2. The multidimensional model of classroom assessment practices

According to the multidimensional model of classroom assessment practices (MMCAP) (Coll et al., 2012), we distinguish five different moments or segments of interactivity in classrooms related to learning assessment: (1) preparation, (2) assessment *de facto* or data collection, (3) correction, evaluation or grading, (4) feedback, and (5) posthoc pedagogical enhancement. Each of these segments constitutes links in a chain with particular actions, roles and contingent compromises by the protagonists of the educational process, teacher and learner. A whole set of these five links constitute an *assessment situation*. Figure 1 presents this model. A preliminary phase of instructional design is necessary. It encompasses all five steps, and more than those, the bigger picture of all the assessment situations within a course (a term, a unit, etc., whatever pedagogical unit we may refer to), which Coll et al. define as assessment *program*.

A series of design decisions must be taken concerning the whole assessment situation. For example: What is the object of assessment? What are the learning goals relative to this assessment occasion? How are we assessing? With what activities, instruments, and resources? What purposes do we follow? When is this assessment situation taking place? Referring to which evaluation criteria and thresholds will performance be evaluated? What rules shall regulate students' participation? What kind of feedback shall students receive? Subsequently, the instructor implements this evaluative situation in its different parts:

- In the *(1) preparation segment*, the instructor aims to facilitate the students' best possible performance. Typically, contents will be recapitulated in the preparation segment, and assessment rules, learning strategies, and coping strategies will be shared. The location in time of the preparation segment might occur some minutes before the data collection segment, or a week before, for instance.
- Next, there is *(2) the data collection segment.* In this central segment, traditionally called 'assessment' *de facto*, students and teachers are both aware of the purpose of data collection, namely, to gather evidence of learning.
- Afterwards, the instructor evaluates the collected data in *(3) the correction segment*, applying the previously established assessment criteria.
- In fourth place, the teacher offers *(4) feedback to the students*, according to the previous decision on how, what, when, etc.,

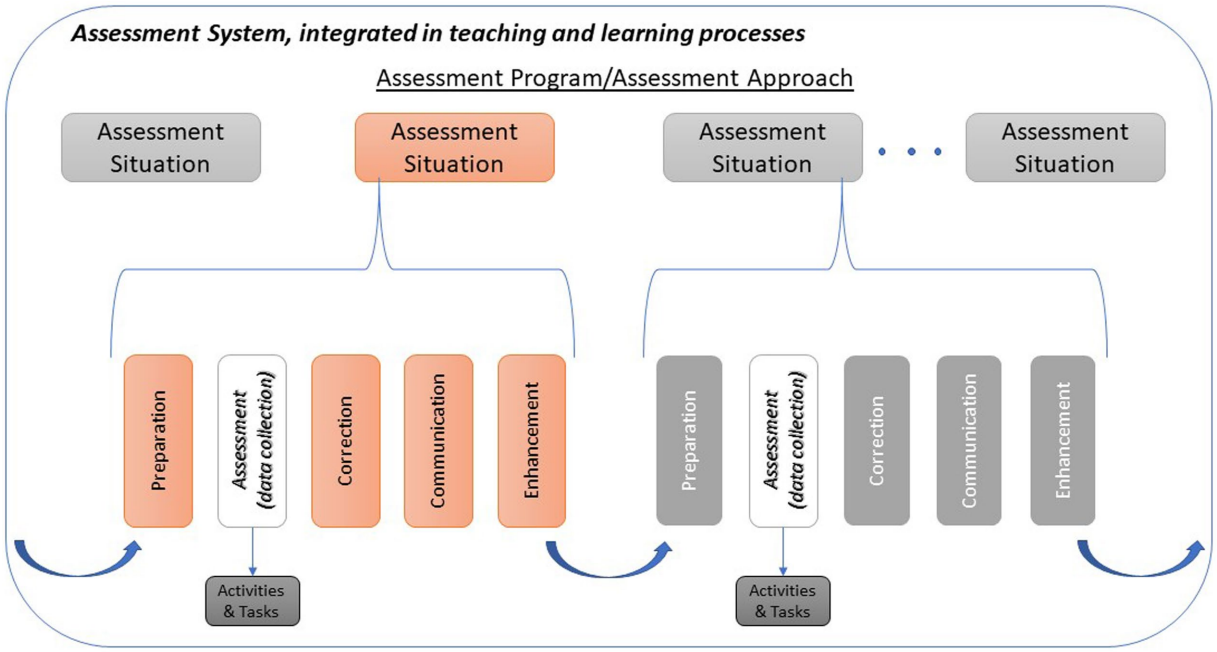


FIGURE 1
Multidimensional Model of Classroom Assessment Practices (adapted from Coll et al., 2012).

- Eventually, teachers will make certain decisions as a consequence of the learning results in *(5) the posthoc enhancement segment*. Whether these decisions are pedagogical or formative (like adjusting teaching resources, looking for extra activities) or summative (like merely handing out a set of grades and moving on with a course) depends on many factors, starting with the teacher's conceptions of assessment, but also related to contextual or systemic constraints (Remesal, 2011).

This multidimensional model of classroom assessment proposes a basic scheme to analyse and understand educational interaction for assessing learning. It allows researchers and practitioners to reflect on it and eventually change whatever needs improvement. As many variations as we can think of, they all could be located in this basic scheme. For example, students' co-evaluation would be a variant of the correction segment, in which students are called in. Also, students' participation in a possible co-construction of assessment activities or evaluation criteria would be a variant of the preparation segment.

A written exam would be typically an instrument of the data collection phase; it might be part of a broader evaluation program, and it -quite naturally-would include elements of greater detail or lower level, such as assessment *activities* and *tasks*. Let us clarify the difference between these two latter elements. An *assessment activity* presents a global action request to the student, a statement or an utterance graphically identifiable, with a beginning and an ending, as separated from other action requests, clearly identifiable within the exam. At the same time, an assessment activity can contain from only one to a variable number of assessment tasks, which suppose specific and unitary cognitive demands for the student (Remesal, 2006; Remesal et al., 2022). For example, in a typical reading comprehension assessment activity the student receives a text with several associated questions. The text followed by the questions would make up the assessment *activity*. However, each of the individual questions would constitute an assessment *task* that requires a detailed, independent response from the student. In turn, they offer unitary opportunities for good or bad performance. Sometimes several assessment tasks can be linked to each other so that the quality of a first response compromises subsequent responses. In any case, the assessment tasks will be contrasted one by one with the assessment and grading criteria.

1.3. Synchronous self-assessment within the MMCAP

How does synchronous self-assessment relate to the whole picture of the MMCAP? Figure 2 presents how SSA fits into this model. This strategy is best implemented during a written exam, that is, during the data collection phase. However, like any other assessment activity, it would permeate the remaining interactional segments. First, design decisions must be taken as to what form the exam should take (number of activities and tasks, their features, assessment criteria, post-hoc decisions).

The key to synchronous self-assessment lies, indeed, in the design of the exam. Throughout educational history, there is abundant literature against exams (e.g., Yu and Suen, 2005); however, we contend that there is still a high pedagogical potential in such artefacts, if analysed carefully. The exam itself is a *unitary pedagogical artefact* which demands complex design strategies, especially if aiming at assessment *for* and *as* learning (Yan and Boud, 2021). Recent experiences in diverse disciplinary areas like Law (Beca et al., 2019) and Computer Science (Rusak and Yan, 2021) bring evidence of higher education instructors' current worries concerning exam design. Following current approaches to assessment for learning and assessment as learning, an exam should present complex,

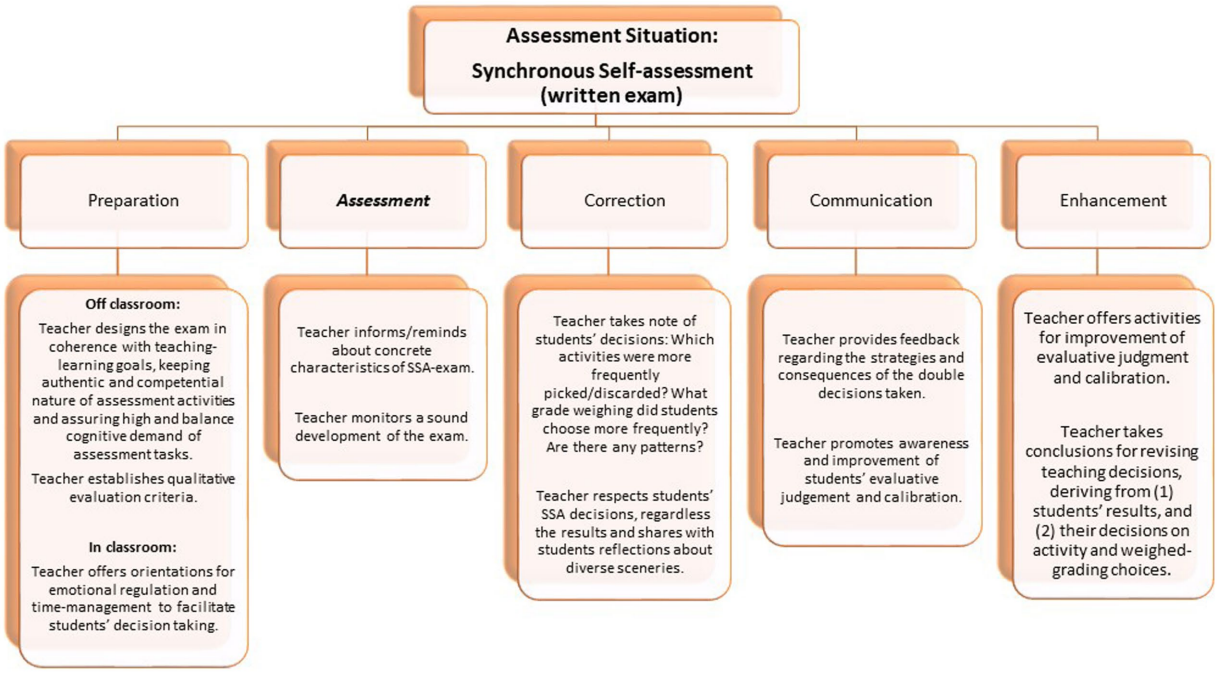


FIGURE 2
Synchronous Self-assessment within the multidimensional model of classroom assessment.

contextualised, realistic, argumentative activities and tasks. In other words, assessment activities must meet, as far as possible, the expectations of the so-called competency or authentic assessment (Villarroel et al., 2021). Authentic assessment requires complex assessment tasks. Ideally, the cognitive demand of these exams should be at least in a middle to high level.

Despite critical voices against 'exams' and in favour of other alternative assessment activities and instruments, we defend that the exam does not necessarily constrain the authenticity of each activity or task. Each exam has a *microstructure*, referring to the characteristics of each of the assessment activities and tasks included in it. Secondly, it has a *macrostructure*, referring to how these activities and tasks are distributed in time and space and how they are coherently related (or not) to each other. Moreover, finally, it has a set of *rules of interaction* that determine what each of the participants can or cannot do or is expected to do, during the development of the exam itself (for example, consulting sources or not, using the calculator, answering individually or in a group, giving spoken or written answers).

In the specific case of SSA, and more important, when the strategy is applied for the first time for students and teachers, we propose to design exams with two differentiated parts. The first part would be common and compulsory for all students to solve; the second part would offer elective activities for the students to carry out the SSA *per se*. Dividing the exam into these two parts follows a twofold objective:

- From the *student's* perspective, the common and compulsory part guarantees sufficient cognitive activation *before* undergoing SSA. Activities in this first part, hence, ideally should be designed to bring to the surface the learning needed to tackle the second part of the exam. So, in a certain way, the first common part constitutes a preparatory segment to the SSA realisation.
- From the *instructor's* side, the first common part establishes the minimum standards that all students should demonstrate, contributing both to the formative but also to assessment's accreditation and accountability purpose.

This novel strategy of synchronous self-assessment permeates all five segments of classroom interaction related to learning assessment. In Figure 2, we present the key aspects that researchers and practitioners should consider when implementing SSA: from preparation to pedagogical enhancement, if we intend to push SSA to the limit of its potential, all interactional segments ought to be considered.

We are currently engaged in a research plan for the medium-long term. Our first exploratory effort of this new strategy focused on the student. Some results have already been published concerning students' emotional experience and metacognitive engagement and management (Remesal et al., 2019; Remesal, 2021; Remesal et al., 2021). As these first studies demonstrate, SSA offers benefits in terms of a significant increase in confidence or sense of control before solving the exam. It also raises awareness and strategic resources management to increase performance. The time has arrived to look at the teaching figure. Some first results regarding the conceptions of assessment in connection with SSA have already been presented (Estrada, 2021): the positive adoption of SSA seems to be more likely amongst teachers with a richer formative conception of assessment in terms of Remesal's model of conceptions (Remesal, 2011; Brown and Remesal, 2017), that is, solid formative beliefs affecting all four dimensions (affection on teaching, on learning, on accreditation and accountability), whilst teachers with a summative conception or even just a weaker formative conception (affection of just two dimensions) are less prone to implement SSA to its whole extent.

1.4. Research goals

After exposing the conceptual basis of this novel assessment strategy, in this present study we pursue to deepen our knowledge of SSA's implications for the educational practice. Our concrete goals are:

1. To identify the characteristics of the exams designed by the participating instructors (one expert and three novels regarding SSA).
2. To explore the instructors' reflections associated with a SSA experience in their course.
3. To identify possible training needs for an adoption plan of the evaluation strategy of SSA, as well as possible lines of research that are open to us for improving assessment literacy.

2. Materials and methods

2.1. Participants

We carried out a quadruple case study (Yin, 2014) with four university instructors who implemented the SSA strategy in four groups of first-year university students at a Teacher Education Degree (two groups of Primary Education and two groups of Early Childhood Education). Table 1 presents the demographic characteristics of the four participants. Each instructor attended a group of 30 to 40 students. The selection of these four instructors responds to an intentional sampling strategy: (1) the four instructors are part of a larger teaching team and share the same course plan of Educational Psychology, (2) the four instructors also have extensive teaching experience in these grades and share the same basic formative assessment approach, at least intentionally, and (3) finally, one of the instructors could be considered an expert in SSA, whilst the other three are novels.

2.2. Data generation

This study collected data of diverse nature and origins: classroom natural artefacts and interviews. First, the participating instructors delivered their end-of-semester written exams, as they had personally designed these assessment artefacts. The instructors agreed on basic guidelines for designing and carrying out these exams:

- *Micro-structure*: the activities should aim to have particular features: they should preferably be complex activities with a high cognitive level, open argumentative resolution based on professional actions or contexts, and as authentic as possible (Villarroel et al., 2021).
- *Macro-structure*: the exam would split into two parts: a first part, common and mandatory for all students, and a second part, with elective (SSA) activities. In the first part, students would have to solve between two and four activities. In this second part, students first would select three activities out of five offered and secondly, they would choose between a triple evaluation variant: *equitable* -all solved activities would weigh the same potential maximum value-, *maximal difference* -one solved activity would weigh for 50%, one for 30%, and the third one would account for 20% of the final grade, and *minimal difference* -two solved activities would weigh for 40%, and the last one for 20% of the final grade.
- *Interactional norms*: Students would be informed about the specific norms and innovative strategy just starting the exam. The exam would last 120 min, with parts 1 and 2 explicitly separated. Throughout the exam, students could consult doubts with the instructor.

TABLE 1 Demographic description of participants.

Instructor	Age (range)	Teaching experience at HE (range)	Experience with SSA
Instructor #1	Over 45y.	Over 20y.	No
Instructor #2	40–45y.	5–10y.	No
Instructor #3	Over 45y.	5–10y.	No
Instructor #4	40–45y.	15–20y.	Yes

After the students sat the exam and instructors had time to revise and grade students' responses, we conducted a semi-structured individual interview with each instructor, recorded on audio and transcribed. This interview sought to collect information about different evaluative decisions made by the instructors and their evaluation of the experience implementing the SSA strategy. The interview script was elaborated on the basis of the MMCAP.

2.3. Analysis procedure

Both authors participated equally in the analysis process; a third analyst's collaboration is acknowledged at the end of the text. To perform the analysis, we proceeded in three steps:

- First, for the analysis of the exams, the revised Bloom's taxonomy was our reference (Anderson and Krathwohl, 2001; Remesal, 2006; Villarroel et al., 2021). We started determining the quantity and internal relation of assessment activities and tasks (macrostructure). Secondly, we identified the level of cognitive demand of the assessment tasks (microstructure), considering three basic levels (*low* –to remember, to identify-, *medium* –to understand, to apply-and *high* –to evaluate, to create). Each of these levels received accordingly 1 (low), 2 (medium) or 3 (high) points as *cognitive demand value*.
- For the interviews, we conducted a qualitative analysis of the instructors' discourse (Braun and Clarke, 2006). The five phases, or interactional segments, of the MMCAP (Coll et al., 2012) were the point of departure of a recursive analysis procedure which advanced in a series of loops of individual analysis -by each of the three analysts-and later contrast for discussion of discrepancies until reaching a consensus. During this back-and-forth procedure of deductive and inductive analysis, emerging themes relative to the specificity of SSA were particularly in focus.
- In the final analysis step, we contrasted both data, artefacts and interview discourse, in order to identify points of coherence (or lack thereof) in each case.

3. Results

In order to offer a complex but also clear and panoramic picture of the results, we organise them in the following way: first, we will present the essential characteristics (micro-and macrostructure, and cognitive demand) of all the artefacts (four written exams as designed by the participating instructors); second, we will dedicate particular sections to each case addressing at once the results from the interviews and the contrasting discursive analysis of the artefacts; thirdly, we present results referring to the specificities of SSA and instructors' evaluation of this innovative experience.

3.1. Exam features: Macrostructure, microstructure and cognitive demand

Tables 2, 3 and Figures 3, 4 present the results corresponding to the exploratory, descriptive analysis of the designed exams. The cognitive demand of an assessment activity gives us clues about the performance expectations that an instructor places on students. In order to understand the results in their context, it is important to remember that all four participating instructors are members of a teaching team. As a team, they adopted a basic agreement that affected the design and development of the exam in all groups equally, as indicated in a previous section. However, as results show, differences can be seen between the analysed exams. The absolute quantity of assessment tasks that the instructors assigned to each exam part differs, oscillating in a range from 18 to 29, taking the exam as a whole, but ranging from 7 to 10 and between 11 and 20, when looking at first and second part separately (see Table 2):

Table 3 completes the look into the four assessment artefacts designed by the participants, showing the average cognitive demand. Here we find two exams with (a) relatively low and (b) balanced cognitive demand between part 1 and part 2 of the exam (instructors

TABLE 2 Quantity of assessment tasks in exams and cognitive demand.

	Exam designer			
Cognitive demand	Instr.1	Instr.2	Instr.3	Instr.4
Activities part 1 / Activities part 2	2 / 5	3 / 5	5 / 4	2 / 5
Total of assessment tasks	18	19	29	26
Common section	7	6	10	6
Low (remember, identify)	4	5	5	2
Medium (apply, analyse)	2	0	3	1
High (create, evaluate)	1	1	2	3
Elective section	11	13	19	20
Low (remember, identify)	8	11	14	1
Medium (apply, analyse)	0	0	5	10
High (create, evaluate)	3	2	0	9

TABLE 3 Level of cognitive demand [1-low, 2-medium, 3-high]. Mean (standard deviation).

	Instr.1	Instr.2	Instr.3	Instr.4
Common section	1.57 (0.78)	1.33 (0.81)	1.70 (0.82)	2.16 (0.98)
Elective section	1.54 (0.93)	1.30 (0.75)	1.26 (0.45)	2.40 (0.59)

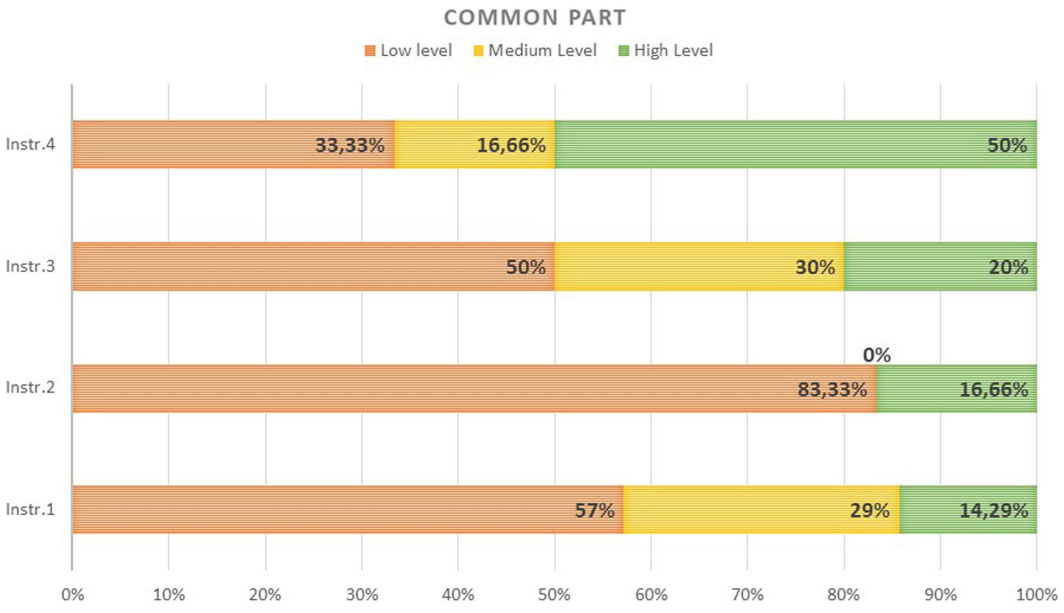


FIGURE 3
Exams' common part. Cognitive demand.

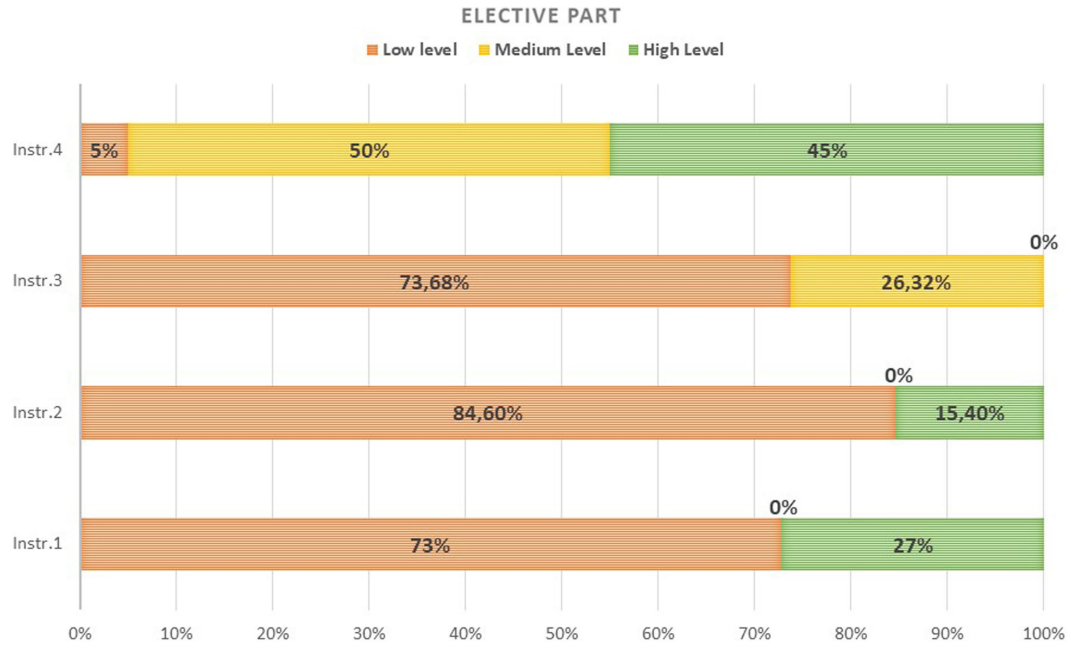


FIGURE 4
Exams' elective part. Cognitive demand.

#1 and #2). The third artefact presented (a) low and (b) unbalanced cognitive demand with a higher demand in the first-common part (instructor #3), and the fourth artefact posed (a) medium to high cognitive demand and (b) unbalanced cognitive demand with a higher demand at the SSA-part (instructor #4).

Finally, Figures 3, 4 show these same descriptive results in percentual terms for a richer comparison. Instructor #4, who had previous experience implementing SSA, presented an exam with much higher cognitive demand to her students, in contrast with the other three instructors who implemented the innovative strategy for the first time.

3.2. Discursive analysis: Looking into details

Identifying cognitive demand is just a first look at the designed exams. The discursive analysis of the assessment activities in the exams as they were presented to the students allows us to deepen our understanding of these evaluative artefacts, shared in the following subsections. We will first report about the three novel instructors, followed by the expert one. For each case, we report results of the analysis of the assessment artefacts (exams), first and second part, and provide some excerpts from the interviews to either support to or contrast against the exams. The enunciates of the assessment activities extracted from the exams are framed to distinguish them from interview excerpts.

3.2.1. What do we learn from case #1?

As we have already seen, more than half of the tasks of the first part of exam #1 present a low-level cognitive demand ($M = 1.57$; $SD = 0.78$). It is mainly in the third activity when Instructor #1's caution in this first experience with SSA is most evident. This activity would initially set a high cognitive level (assess and propose improvement). Nevertheless, the cognitive demand is curtailed by the secondary instruction, exposing in advance the key action to be carried out (italics added), thus reducing the task to a mere follow-up of direct orders:

> Identify the errors in the following conceptual map and correct them-*introducing the information missing in the relationship between concepts and modifying the one that is erroneous by the information appropriate to the same map.*

In the second, elective part of the exam, all five activities (out of which the students choose three to solve) revolve around a single case narrative that brings its own contextual boundaries. Each of these activities presents a disparate level of cognitive demand ($M = 1.54$; $SD = 0.93$). The critical concepts sought in each activity are marked in bold by the instructor herself as a cognitive aid. A careful reading reveals that four of the five activities do not use the case as an actual trigger for a creative and genuine response but rather as an excuse or a frame for simple identification-remembrance. For example (bold in original):

> Define meaning in learning and identify examples of three conditions for **meaning attribution in the text**.

Unlike the first four activities, the last one does not anticipate response hints. On the contrary, it challenges students to construct their scheme for constructing their responses. The question best collects high-level cognitive skills such as analysing, comparing, and evaluating. However, its cognitive demand is very disproportionate compared to the previous four activities. It is, altogether, a very unbalanced exam, and we can expect that very few students would choose this fifth activity (bold in the original):

> Comment on the text to explain, in your opinion, **the most relevant part of Ana's learning case** using the contents of the **three thematic blocks worked through the course**.

Comparing part 1 and part 2, we find a balanced exam, with no difference in cognitive demand ($d = 0.03$). In the interview, Instructor #1 acknowledges having designed her exam with a strong accounting purpose since she expects all students to be able to demonstrate a minimum, declarative and defining standard of knowledge in the first part common to all students. In contrast, the second part, where the SSA conditions are applied, would be the space to demonstrate knowledge in use.

> "This criterion guided me as a basis, that is, I tried to ensure in the first part some conceptual knowledge that seemed relevant to me, and in the second part, well, the use of conceptual knowledge in one case, (…) how to put this knowledge into practice." (Instr.#1)

She also adds the following idea of minimal account giving as a leading thread for the exam design:

> "Very nuclear [knowledge] in terms of how we had worked on it and in terms of the importance we gave to these ideas at work and also very nuclear in the sense that if the students do not know how to answer this, well… then we have a problem" (Instr.1)

3.2.2. What do we learn from case #2?

The first part of the exam designed by Instructor #2 presents mostly tasks of low cognitive demand ($M = 1.33$; SD = 0.81) but also includes one task of high cognitive level, exemplification and argumentation (italics added):

> Define self-regulation. *Why is it important teaching it at early childhood education? Give a justified example of how you would work towards self-regulation in your classroom.*

However, the second part of the exam, with SSA, presents two types of activities with a notable difference between them with regards to cognitive demand ($M = 1.30$; SD = 0.75). The first two activities refer to a single concept of the programme and ask to analyse and explain equally. These are activities that we consider to be parallel. Nevertheless, the third and fourth activities only request the identification of various concepts presented in the case narrative. The fifth activity, suddenly, raises cognitive demand (bold in original):

> Analyse and explain how the three conditions of **meaningful learning** are presented in Carmen's case to learn Anatomy.
>
> Analyse and explain how the three conditions of the **attribution of meaning** are presented in Carmen's case to learn Anatomy.
>
> What **goals and motivational orientation** does Carmen present in her medical studies?
>
> What kind of **learning approach, goals and motivational orientation** does Carlos suggest to Carmen?
>
> Evaluate the **distance** between Carmen's prior knowledge and the anatomy instructor's teaching and explain the **effects** of working at that distance on the student's motivation.

The internal contrast of both exam parts presents no difference, hence it is a balanced exam (d = 0.03) with regards to the cognitive demand. Instructor #2 states in the interview his goal that the assessment activities of the elective SSA part maintained a consistent level of difficulty. In this way, he underlines that the student's choice of activities be based on the mastery of thematic knowledge and not on possible unbalanced diverse demands. Hence, after the analysis of the five activities, we can raise our doubts about the accomplishment of this instructor's goal regarding the balanced cognitive demand. Here too, the accountability function of the exam can be appreciated:

> "I tried to ensure that they were balanced in the different issues and that they were, that they were somewhat equitable, (…) that there was not a very simple or a very complicated one (…) and I did the most complex part based on the case, if you want to excel or have a good grade, you have to go a little further". (Instr.2)

3.2.3. What do we learn from case #3?

The exam designed by Instructor #3 presents in overall a higher cognitive demand than the two previous ones, particularly in the first part. In this third case we also find an increased amount of assessment tasks embedded in the activities, both in part 1 and part 2. An additional difference with respect to the other two novel instructors lays in the lack of balance between the common and the elective part of the exam concerning cognitive demand. Instructor #3 designs an exam with a significantly greater cognitive demand ($d = 0.66$) in the common part ($M = 1.70$; SD = 0.82) and less cognitive demand in the SSA part ($M = 1.26$; SD = 0.45) to be solved by students. Additionally, she inverts the order of presentation, so that students have to respond first to the elective section (easier, according to her intended design) and in second place they respond to the common section (more difficult, from her point of view). In the elective part, she presents the students, for example, two sets of parallel activities that only differ by the alluded learning content. All of these activities are of low cognitive demand (reminder or presentation of a conceptual network) and are complementary to each other in terms of the content evaluated:

> Explain the differences between collaborative and cooperative work. Briefly describe the three interpsychological mechanisms involved in the construction of knowledge amongst peers.
>
> Explain the differences between collaborative and cooperative work. Briefly describe the three dimensions of analysis of collaboration for learning.
>
> Considering the two mechanisms of educational influence that operate in classroom interactivity situations, explain the construction process of shared meanings.
>
> Considering the two mechanisms of educational influence that operate in situations of classroom interactivity, explain the process of progressive transfer of control.

In the interview with Instructor #3 we learn about another crucial difference between her and her colleagues. Regarding the macrostructure, for instructors #1 and #2 the first part of the exam supposes the verification of the lowest common denominator of knowledge, hence, the basic learning goals, and the optional part entails a higher performance expectation. In contrast, for Instructor #3, the opposite happens: the optional part includes basic activities for the instructor and the first part -for compulsory response-supposes the opportunity for individual excellence. This inversion introduced by Instructor #3 also affects the rules of interaction of the exam since her students are exposed to SSA conditions when starting the exam time, unlike all the other groups. This change of rules are evident in the following excerpt:

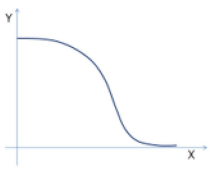
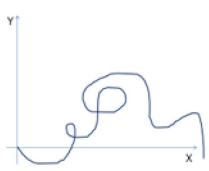
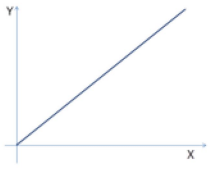
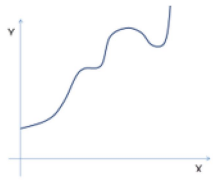
> "60% were the questions that they could choose, so they had five questions, which is what we agreed on in the teaching team, and I tried to ensure that they were balanced in the different topics (…), that there was not one that was very simple or one that was very complicated, but that they were all at a medium level, of difficulty, this is the part that they had to choose. And the part that they couldn't choose was a case, there are three questions about the case, which had a higher degree of difficulty. The truth is that I thought, 'well, here [in the first part], they have the advantage that they can choose' and well, it would be the most affordable part of the exam and the most complex part I did base on the case" (Instr.3)

3.2.4. What do we learn from case #4?

Instructor #4, with at least five-year experience in implementing SSA, designs the exam with the highest cognitive demand, comparing all four cases. Both, part 1-common part- and part 2-SSA-have over medium values (part 1, $M = 2.16$; $SD = 0.98$; part 2, $M = 2.40$; $SD = 0.59$). There is a small difference between both parts in their cognitive demand ($d = 0.29$), prioritising the elective part, thus emphasising students' agency. We present an example of one activity of each of the exam parts. In the first one, a direct question requires the student to elaborate a new argument, incorporating some 'safety net' questions (Van Den Heuvel-Panhuizen, 2005) to prevent the student's mental blockage in a single option:

> Is it possible to meaningfully learn something eventually wrong? If yes, what could be the consequences? If not, what do we conclude? Justify your answer and provide an example.

In the second example, from the SSA part, students must confront and combine knowledge and skills from different subject areas, such as psychology and mathematics, evaluating and appreciating the feasibility of four given options:

> Look at the following images. The abscissa axis (X) represents time in all graphs, and the ordinate axis (Y) represents knowledge. If the learning process could be represented in such a simple way (something in itself impossible, as we know, because too many variables are involved), which of the following graphs would best represent a meaningful learning process? Why?
>
> GRÁFICO 1
>
> 
>
>
> GRÁFICO 2
>
> 
>
>
> GRÁFICO 3
>
> 
>
>
> GRÁFICO 4
>
> 
>

Instructor #4 declares in the interview that her exam aims to offer students the challenge of solving tasks of a complex nature. In both parts of the exam, assessment tasks present medium or high cognitive demand, since the activities require analysis, argumentation, and/or creative exemplification. The prompts do not request direct declarative memory responses or explicit definitions. Instead, they expect students to apply conceptual knowledge of those definitions or 'assemble' argumentative responses. This fourth exam is consistent with Instructor #4's discourse:

> "In the first part, where everyone must demonstrate basic learning, I try to offer questions that go to the core of the concepts and that help the students reflect on that core [of concepts]. And in the synchronous strategy part, I try to propose them activities that are also of a competence challenge, more applied and creative". (Instr.4)

3.3. Instructors' evaluation of the experience: Assessment literacy needs detected

One of the keys to the proper development of the SSA strategy lies in the design of the exam. As an evaluative artefact, it presents a specific microstructure, macrostructure and interactional rules that manage the student and instructor's agency space. The analysis of the presented exams, as designed by the instructors individually, show a great difference in terms of cognitive demand to the student. Comparing Instructor #4 with her colleagues, all of them novels in SSA, we find significant differences in both exam sections. Table 4 shows the varying size effects of these differences.

Through the discursive analysis of the interviews, we identified various emerging themes revealing some inconsistencies, which point to likely training needs to improve assessment literacy and assessment practices. We present below some extracts around the following aspects: emotional reaction to the experience, understanding of self-assessment and, in particular, the SSA, effects on the correction phase and the interpretation phase and use of learning outcomes for subsequent teaching decisions.

3.3.1. Emotional reactions to the first experience with SSA

Three of the instructors were novel in the implementation of the SSA strategy. In their interview, all of them manifested a positive acceptance of the challenge of this project, although they also added objections of two kinds, first of all, emotional objections. In the instructors' responses, we find concerns dealing with their own emotional response as instructors and with an empathic identification with the evaluated student.

TABLE 4 Comparing cognitive demand of the four exams (Cohen's *d*).

	Part of the exam	
	Common	Elective
Instr.4 versus Instr.1	0.66	1.10
Instr.4 versus Instr.2	0.92	1.63
Instr.4 versus Instr.3	0.51	2.17

Instructor #1, for example, expressed concerns about increased anxiety on the student's part:

> "[I really wanted] to do it, to see what would happen, because the idea of choice specifically, that they could reflect on what score they would give each of their answers… I found it interesting, wanted to do it to see what would happen" (…) "In relation to the format, in fact, that is, deep down my concern as a teacher myself was originally that it would not affect them, aah that they would not be overwhelmed by time [pressure], by the fact of having to take their decisions". (Instr.1)

Instructor #2, on the other hand, predicted positive reactions on the students' part, since the increase in a personal agency in decision-making would also favour them:

> "I thought it was a good chance to promote sense and meaning itself, in choosing the activities. Something that has always interested me; I had always thought of, the fact that the students could take some decision or choice on the activities, but I had never considered about the grading of the questions, I thought that this was more on the teacher, right? I think this was the most innovative part." (Instr.2)

Instructor #3, finally, kept a double perspective, empathic towards the student but also expressing some concern from her teaching position, anticipating an increase in the workload associated with the grading, from her point of view:

> "First, I thought that the students would welcome it as something positive, a good idea for the students that would be accepted with pleasure, and that it would benefit them. And then I thought of the teaching part that I might have difficulties when correcting, because each activity would have a different weigh depending on the student's choice, which complicated the issue of evaluation for me, indeed". (Instr.3)

The emotional concern anticipated by Instructor #1 was not confirmed. Students had a positive reception in all cases, as shown in previous publications (Remesal et al., 2019, 2021):

> "They were calm and they answered and they had enough time to do it, so there was not this effect that I was worried about … at all." (Instr.1)

> "(…) the students are super happy, they even asked for the next exam to have the same [form], they see it as an advantage." (Instr.3)

> "In general, students are often surprised by the new rules of the exam, but they receive it well, (…) I think that above all they are happy when they hear the rules of the exam. In my experience, they've been always positive." (Instr.4)

3.3.2. SSA as a particular modality of student self-assessment

A central theme in this exploratory study is the instructors' conception of student self-assessment to understand to what extent SSA can be perceived and appreciated as beneficial. In this sense, we highlight the reflection of Instructor #1, which reveals a conception of retrospective, let us say, traditional self-assessment. Students' self-assessment most frequently occurs as a separate metacognitive behaviour, after and outside the learning process and the demonstration/performance of learning. This conception of self-assessment of learning as an ex-post-facto action is widespread in the literature of the field, as well as in educational practice (Andrade, 2019):

> "A process where the student understands what her learning process has been and how she also evaluates the results of this process, so that's it. In the process, both the work that [the student] has done is included, as well as reflecting on other elements that may affect other aspects of the educational process, it is also reflecting on like "well, I have done this, it's ok for me" or "I might need some help in this new topic", but well, above all this, it's how the student evaluates his process and the results he has obtained". (Instr.1)

In contrast, for Instructor #2, self-assessment implies a continuous metacognitive demand. Quite a challenge if beginning students have not yet established the habit:

> "for me it consists of a process of reflection on how the [learning] process has gone… and how the process is going, right? that has a great impact on the students throughout the process, on how they are learning, what difficulties they are facing, what progress they are having and well, I try to make them reflect during the process… on how they're progressing… that, on their learning". (Intr.2)

Instructor #3 refers to self-assessment as a process of shared dialogue between instructor and student (Sutton, 2009):

> "Self-assessment has to be a joint exercise, the student must have all the information to be able to evaluate himself, but then there has to be --not a grade from the instructor, but a conversation, a dialogue, an exchange". (Instr.3)

Finally, for Instructor #4, the only instructor with previous experience in SSA in this study, the particular contribution would be in the multidimensionality of the self-assessment processes that underlie the moment of the exam since the students face the required decisions:

> "Through this strategy, the student develops more meta-knowledge, more awareness of what they really know or do not know, or to what degree they know it… also more awareness of their cognitive and emotional resources at the time of facing the evaluation, or like knowing or recognising their nervousness, know how to manage it, know how to manage time, know how to organise themselves" (Instr.4)

Instructors new to the SSA strategy, however, were cautious in assessing the potential benefits regarding the assessment of actual complex competencies:

> "I imagine that at the time of the exam they do a quick assessment of what they know and what they do not know, so what I meant is

> self-assessment as a process or self-assessment as an evaluation of the results of whether this grade seems good or not good to me, or whether it is appropriate or not. Another aspect that would precisely have to do more with the learning results, not with the qualification, yeah, I kind of think so, that at that moment that helps them think "what do I know better" (…) so I think so, that, in this sense, I do think that it contributes a little more to the assessment of competencies because it precisely contributes, I understand, to this nuance of self-evaluation of what one quickly thinks what one knows and what not, and with what, what he knows and what he doesn't and with what he feels/ and-or I don't know, I don't know, with what he feels safer when answering". (Instr.1)

> "I think it goes a little step further because it asks them to make a choice right in the middle of an exam… that was not expected, let's say, eh, and it goes a little step [further]". (Instr.2)

> "We would be working on the competence of personal autonomy and learning to learn, so it would be a competence development (…) that they have to develop and that, normally, because these competences are not worked through or evaluated, but with this (SSA) it allows them to reflect on their learning process, on their knowledge and it would be closely related to the concept of self-assessment before, although it would not be a self-assessment with a numerical grade but rather a self-assessment of their own knowledge". (Instr.3)

3.3.3. SSA and the multidimensional model of classroom assessment practices

As we assume, SSA would indispensably affect all five evaluative segments of the interactional process of classroom assessment established by the MMCAP (Coll et al., 2012). However, according to the interviews of the three instructors who implemented it for the first time, the strategy only had a noticeable impact on two of these segments: correcting or marking the students' answers and making sense of the choices to take consequent decisions.

3.3.3.1. Impact on the marking phase

All three novices in SSA seemed to apply assessment criteria based on dichotomous absolutes (correct-incorrect answer), and they gave evidence of little reflection on the possible impact of the SSA:

> "The truth is that it wasn't difficult, I mean when I corrected, I saw what was right and what was wrong and then depending on the student's choice of their score, I adjusted the score, that's it, easy, it does not really add any complexity this [SSA] do you know what? I marked their answers as usual". (Instr.1)

> "The marking of each answer is based on a series of verifiers that I prepare beforehand to make the correction, so that it be a fair correction for everybody… for all the answers. So, in that sense, it has not changed, well I had to think of more indicators because there were two more activities". (Instr.2)

> "I organised myself, the rubric and stuff, I was able to organise myself relatively well, yes, it is a little more challenging, because you have to be doing more calculations, but it was OK in the end". (Instr.3)

Compared to the three novices in SSA, we have Instructor #4's reflections on her more detailed correction procedure, based on more qualitative aspects, to which she adds ethical-moral factors:

> "I make like a scheme of the correction criteria and what I do is collecting the concepts I hope the students introduce in their arguments and what kind of examples I am going to consider valid or non-valid. So, starting from there, I always value the example equally or even more than any conceptual definition, the definition of the concept is always implicit to me, so I always give the example as knowledge in use a more weighted score. (…) later with the SSA the student says "I want this activity to be worth 3 points, 2 or 1", so I apply that factor to what I have corrected, but it does not affect it in any other way (…) This is very important, precisely, in the sense that the instructor should not be affected by the students' choices on weighing scheme. If I was affected, I could be including another type of evaluation that could end up being unfair to the student". (Instr.4)

3.3.3.2. Impact on consecutive decision taking: potential for enhancement

SSA puts on the instructor the need to make pedagogical sense of the students' new choices, sometimes unexpected ones. Previously published results on the students' experience tell us many different reasons hidden behind students' decisions (Remesal, 2020). Some of those reasons are related to the students' conceptions of learning assessment and their traditionally passive role in the process, which might hinder them from assuming a more active role. Other reasons have to do with specific circumstances during the exam development (e.g., emotional or cognitive blockades). Other reasons are related to personal preferences or interests or to the self-awareness of personal skills. In this regard, we have also identified notable differences between novice instructors (#1, #2 and #3) and Instructor #4, with previous experience in SSA. To begin with, in the interviews, novice instructors shared representations of students as mainly passive subjects, lacking agency and strategic or reflective decision-making capacity, or they attribute decisions to non-controlled temperamental factors. For example:

> "It seemed to me, at first, I thought "man, this is a bit weird", but I really suppose that they distributed a bit… randomly, I don't know (…) those who choose maximal difference… I also had the feeling that they had chosen a bit by chance, a bit miscalculating". (Instr.1)

> "There were students who strategically decided to assign the greater points to the answer they thought was best formulated, and then the other two types of choices, minimal difference and equitable, were like sort of balanced… and in these cases some students had kind of more doubts when they didn't know what they had answered correctly or wrong, they strategically decided to distribute the scores equally, right?… the student who has a more strategic vision of his performance can choose the most daring options, let's say the maximal difference for example… but

> then those who are more insecure or hesitant, right? They have chosen more options… of more equal distribution or either minimal difference". (Instr.2)

> "(5 seconds silence) I think that students' choice may depend on various factors, it depends on their knowledge, their learning processes and then there is another factor which is personality, there are people who are much more risk-tempted and people who are less risk-bound, so to say, so I get the feeling that students who are more conservative, who are not as risky, take equal distribution. But then, there are students who are more risk-bound, I guess, and then they pick maximal difference, and then the last option, minimal difference, that one has been taken by fewer students, as far as I remember, because this one does involve a process of reflection". (Instr.3)

Instructor #4, in contrast, stated:

> "Two or three students choosing an equitable grades distribution, for example, may have very different motives behind their choice. So, for me as an instructor, the fact of learning about that diversity of motivations behind the same decision also helps me learn many things from the students, about the complexity of the learning process (…) and also helps to understand misunderstandings". (Instr.4)

The positive reflection of Instructor #4 on SSA generates more introspective and self-critical reflective processes of her teaching or management of the assessment instruments to capture students' knowledge and learning process. Even assessment activities and tasks become an object of her reflection in such a way that assessment becomes truly formative, leading to regulate not only learning processes but also teaching processes (Yan and Boud, 2021):

> "It adds this layer of complexity to the entire process of reflection on learning, what I still have to improve is the activities themselves, because I reckon that sometimes I'm perhaps not very explicit when it comes to proposing activities (…) So this SSA strategy forces students to read the entire exam first, take a panoramic perspective, reflect on each activity, (…) so that panoramic reading implies a different activation of knowledge. It is not sequential, (…) but it is a horizontal and panoramic reading (…) having to decide forces you to having to read all [the activities] in a row, take a step back, evaluate the situation, put all your knowledge in active working memory and strategically see better what you can solve, also always with the aim of giving your best performance. I believe that this is also a matter of justice or fairness, right? To put the student before his best possible self, well give him the option to show his best selfie, sort of, not the students portrait the instructor is looking for, from his definition of the goals and assessment criteria in the course, what we decide would be a good student, but the best possible self-portrait by the student himself". (Instr.4)

4. Discussion of results

In this article, we present the results of a first approximation to SSA from the teachers' perspective. As an innovative proposal, it is challenging to discuss the results in a regular way due to the absence of previous specific literature to contrast with. Therefore, we propose a series of reflections on new questions and challenges that arise for teachers and, consequently, for teacher educators as a result of implementing SSA as a means for assessment for learning.

To begin with, we have been able to detect areas in severe need of improvement regarding assessment literacy: exam design (with and without SSA) is far from satisfying and incoherent with the assessment for learning discourse (Hernández Nodarse, 2007; Hernández Nodarse et al., 2018). Crucial questions arise; for example, what is the link between the experience in implementing SSA and the level of cognitive demand of the proposed exam designed by the instructors? Does Instructor #4 pose a more demanding exam due to the fact (or as a consequence) of having more experience in the implementation of SSA? Or are they independent phenomena? Or perhaps linked to other factors?

The emotional impact of assessment on the teacher has not yet been sufficiently studied. Brown et al.' chapter (2018) is one of the few publications, if not the only one by date, on this important issue. Much work has been done on the side of students' emotions related to assessment (Schutz and Pekrun, 2007), but previous works on the teacher's side (e.g., Schutz and Zembylas, 2009; Sutton et al., 2009), do not specifically consider *assessment* amongst the emotion-loaded phenomena in the teaching profession. One of the latest advances in instruments design for the study of teacher emotions, still does not pay attention to this particular chapter of the teaching profession (Hong et al., 2016). We wonder, thus, if the emotional concerns exposed by these three novel teachers are related to a lack of confidence in students' capability of assuming more agency and responsibility in the assessment of their learning. Or else, their emotional concerns could relate to the loss of (teacher's) power provoked by the SSA strategy.

In all three cases of novices in SSA to a greater or lesser extent, we identified notable inconsistencies between their discourse and the designed exam, as well as between the basic agreements of norms and macro-and microstructure, on the one hand, and the final design and development, on the other. Despite all the participating instructors being members of the same teaching team, with a shared pedagogical approach and teaching programme (Fulton and Britton, 2011), the evaluative artefacts (exams in this case) designed by each of these instructors differ notably. The contrasted exams share nothing more than the superficial structure and the conceptual contents object of evaluation, which are part of the course programme. It becomes evident once again that educational assessment needs continuous reflection and constant and real teamwork; otherwise, its validity and even the fairness of treatment received by the students are firmly at stake (Buckley-Walker and Lipscombe, 2021).

The results also highlight the need for training for university instructors, even those with long teaching experience, regarding the ability to identify the cognitive demand of assessment activities and tasks. For many years, the training towards assessment literacy has focused on what we could call 'superior' or 'meso/macro' assessment levels (Popham, 2009), such as using alternative and complementary instruments (for example, the implementation of rubrics). However, at a 'lower' or 'micro' level, deep reflection on the very nature of assessment activities and tasks is still pending (Bonner, 2017). We will only be able to respond to the call of Villarroel et al. (2021) to implement competency and authentic assessment if we are concretely aware

of the shortcomings of our current praxis. We contend that the very participation in this case study led the four instructors, and particularly those novels to SSA, to reflect on their conceptions and practices, however, it was only a first step.

Our results also show that instructors' conceptions about 'ordinary' or 'traditional' students' self-assessment affect the chances of acceptance and effective implementation of SSA. Undoubtedly, much remains to be explored from the view of the most general conceptions about assessment, be they linked to accountability or pedagogical regulation (Remesal, 2011; Estrada, 2021). It might be attractive to link SSA to summative or accrediting purposes since it occurs in a traditional summative practice -such as a written exam. However, we know that both assessment functions, accreditation and pedagogical or formative regulation, are always related and in constant tension in any educational system (Taras, 2009; Remesal, 2011; Black and Wiliam, 2018). Therefore, each assessment activity contains the potential to positively inform both summative and formative decisions that affect students' learning (Lau, 2016). We want to underline that our exam definition does not find a limit in paper-and-pencil format or an individual resolution (Remesal et al., 2022). In our understanding, the key and the specificity of the 'exam' lies in the explicitness of the shared purpose (for both teacher and student) to expose an optimal performance under a time-constrained condition. This broad definition of exam actually accepts a great variety of forms in the classroom. It would be worth exploring alternative conditions for implementing SSA, such as spoken or enactive resolution, or as a group experience. In short, it would be convenient to explore the potential of exam situations from an essentially formative point of view.

In terms of practical implications of SSA, we see two additional advantages in SSA in comparison with more traditional evaluation practices, which we also submit for consideration. In the first place, the correction or marking phase is usually the most tedious for the instructor in traditional conditions since it is repetitive when all the students must solve the same activities in a traditional "one-size-fits-all" exam. Under SSA conditions, however, this marking phase, due to students' particular choices, becomes more diverse and colourful and thus less monotonous, facilitating a higher level of attention and constant reflection on the teachers' side. With SSA, each exam can be unique due to the double decision of each student. The information that the instructor may collect is more complex and qualitative, both about the individual student and the group class as a whole. In turn, the specific students' choice of assessment activities provides rich information about the activities *per se* (their intelligibility, the level of challenge they pose, the interest they raise, and the understanding of the referred contents). Secondly, students' decisions about the weighting of answers can give clues about their self-competence or perceived difficulties during the exam. Every instructor faces the challenge of interpreting this information, so the assessment situation is also an opportunity to make decisions that effectively improve teaching. In an ideal case of an essential trust established between instructor and students, the students themselves could share the reasons for their decisions with the instructor, which further enriches all the formative potential of SSA.

As a matter of fact, our proposal of SSA is closely related to other recent concepts, such as evaluative judgement. (Boud et al., 2018; Tai et al., 2018; Panadero et al., 2019). We contend that SSA is indeed a practical proposal for developing evaluative judgement in the long run, since it opens up occasions for reflection on personal excellence criteria. When using SSA to actively develop students' evaluative judgement, it can be introduced gradually, for example, one type of choice at a time (activity or weighted grading). This same gradual strategy could be applied in teachers' professional development for enhancing their assessment literacy, offering the chance to reflect upon changes more deeply.

5. Conclusions and open roads

The quadruple case study we have presented has obvious limitations: too many factors remain inevitably unattended, and specific context features might not apply to other situations. However, the results are robust enough to deserve our attention and raise essential questions for future research and praxis revision. From the multidimensional model of classroom assessment practices (MMCAP), that allows us to examine classroom assessment in a more comprehensive and qualitative way, we propose the following list of challenges for future research and practice improvement: How does SSA contribute to the development of evaluative judgement and students' self-efficacy? How can the instructor intervene in this direction? Regarding pedagogical potential, when is it best to inform students of the particular conditions of an SSA experience? How may this affect students' study strategies and learning approaches? How to design challenging exam situations that suppose real new learning opportunities? How can instructors be helped to manage the power rebalance with the student? How to promote a fair evaluation for all students? How can students' agency be raised? How can a teacher take pedagogical advantage of students' choices -both at an individual and group level? How is the instructor affected emotionally in this process? How could an instructor interpret the results of SSA globally as well as individually? What formative decisions can SSA promote? How can this strategy cater for the diversity of students? How can it contribute to educational excellence? Is SSA a suitable strategy in response to the new claim of personalised educational practices?

Finally, this study has been carried out at the university level, but we must also regard implementing this assessment strategy at earlier educational levels. In this case, we could ponder the appropriate adaptations depending on the students' developmental stage, or the curricular area. Many questions remain open; from these pages, we urge the educational community to accept the challenge of SSA. This first small case study is a tiny step into a new long road.

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. The patients/participants provided their written informed consent to participate in this study.

Author contributions

AR carried out the data collection prior to the incorporation of FE to the research team. Once FE started the research collaboration, her contribution to the analysis and elaboration of the text is remarkable. AR and FE are equally in charge of the final result. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors want to acknowledge and thank the contribution of C. L. Corrial (Pontificia Universidad Católica de Valparaiso, Chile) to the analysis of the data implied in this study. This research was funded by the University of Barcelona, ref. REDICE182200.

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.

References

Anderson, L. W., and Krathwohl, D. R. (Eds.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: NY: White Plains Longman.

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Front. Educ.* 4:87. doi: 10.3389/feduc.2019.00087

Beca, J. P., Castillo, E., Cid, A., Darritchon, E., and Lagos, S. (2019). Diseño de un examen de grado por competencias en la carrera de Derecho. *Rev. Pedagog. Univ. y Didáct. del Derecho* 6, 101–130. doi: 10.5354/0719-5885.2019.53747

Black, P., and Wiliam, D. (2018). Classroom assessment and pedagogy. *Assess. Educ. Princ. policy Pract.* 25, 551–575. doi: 10.1080/0969594X.20181441807

Bonner, S. (2017). "Using learning theory and validity theory to improve classroom assessment research, design, and implementation" in *Paper presented at the Annual Meeting of the National Council on Measurement in Education*. New York.

Boud, D., Ajjawi, R., Dawson, P., and Tai, J. (2018). *Developing Evaluative Judgement in Higher Education: Assessment for Knowing and Producing Quality Work*. Routledge, England.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

Brown, G. T., Gebril, A., Michaelides, M. P., and Remesal, A. (2018). "Assessment as an emotional practice: emotional challenges faced by L2 instructors within assessment" in *Emotions in Second Language Teaching*. ed. J. D. Martínez Agudo (Cham: Springer), 205–222.

Brown, G. T., and Remesal, A. (2017). Teachers' conceptions of assessment: comparing two inventories with Ecuadorian teachers. *Stud. Educ. Eval.* 55, 68–74. doi: 10.1016/j.stueduc.2017.07.003

Buckley-Walker, K., and Lipscombe, K. (2021). Validity and the design of classroom assessment in instructor teams. The. *Aust. Educ. Res.* 20pp 49, 425–444. doi: 10.1007/s13384-021-00437-9

Coll, C., Mauri, T., and Rochera, M. (2012). La práctica de la evaluación como contexto para aprender a ser un aprendiz competente. *Profesorado, revista de currículum y formación del profesorado* 16, 49–59. http://www.ugr.es/~recfpro/rev161ART4.pdf, https://revistaseug.ugr.es/index.php/profesorado/article/view/19769

Crowell, T. L. (2015). Student self grading: perception vs. reality. *Am. J. Educ. Res.* 3, 450–455. doi: 10.12691/education-3-4-10

Estrada, F.G. (2021). *Revisión y modificación de concepciones del profesorado sobre la evaluación a través de la autoevaluación sincrónica*. Paper presented at XV-Congreso Internacional de Educación e Innovación. Florence, Italy. December, 13-15th, 2021.

Fulton, K., and Britton, T. (2011). *STEM Instructors in Professional Learning Communities: From Good Instructors to Great Teaching*. Washington, DC: National Commission on Teaching and America's Future.

Hernández Nodarse, M. (2007). Perfeccionando los exámenes escritos: reflexiones y sugerencias metodológicas. *Rev. Iberoam. de Educ.* 41, 1–25.

Hernández Nodarse, M., Gregori, M., and Tomalá, D. (2018). ¿Cómo puede transformarse verdaderamente la evaluación del aprendizaje practicada por los profesores? Una experiencia pedagógica por la mejora del profesorado. *Revista Ciencias Pedagógicas e Innovación* 6, 88–106. doi: 10.26423/rcpi.v6i1.221

Herrero, P. G., and Medina, A. R. (2019). La exigencia cognitiva en los exámenes tipo test en contexto universitario y su relación con los enfoques de aprendizaje, la autorregulación, los métodos docentes y el rendimiento académico. *Eur. J. Investig. Health, Psychol. Educ.* 9, 177–187. doi: 10.30552/ejihpe.v9i3.333

Hong, J., Nie, Y., Heddy, B., Monobe, G., Ruan, J., You, S., et al. (2016). Revising and validating achievement emotions questionnaire – teachers (AEQ-T). *Int. J. Educ. Psychol.* 5, 80–107. doi: 10.17583/ijep.2016.1395

Imbulpitiya, A., Whalley, J., and Senapathi, M. (2021). Examining the exams: bloom and database modelling and design. In Australasian Computing Education Conference (pp. 21–29). doi: 10.1145/3441636.3442301

Lau, A. M. S. (2016). 'Formative good, summative bad?'–a review of the dichotomy in assessment literature. *J. Furth. High. Educ.* 40, 509–525. doi: 10.1080/0309877X.2014.984600

López Espinosa, G. J., Quintana, R., Rodríguez Cruz, O., Gómez López, L., Pérez de Armas, A., and Aparicio, G. (2014). El profesor principal y su preparación para diseñar instrumentos de evaluación escritos. *Edumecentro* 6, 94–109. Available at: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2077-28742014000200007&lng=es

Nicol, D. (2021). The power of internal feedback: exploiting natural comparison processes. *Assessment & Evaluation in Higher Education* 46, 756–778. doi: 10.1080/02602938.2020.1823314

Panadero, E., Broadbent, J., Boud, D., and Lodge, J. M. (2019). Using formative assessment to influence self-and co-regulated learning: the role of evaluative judgement. *Eur. J. Psychol. Educ.* 34, 535–557. doi: 10.1007/s10212-018-0407-8

Popham, W. J. (2009). Assessment literacy for instructors: faddish or fundamental? *Theory Pract.* 48, 4–11. doi: 10.1080/00405840802577536

Remesal, A. (2006). *Los Problemas En La Evaluación Del Area De Matemáticas En La Educación Obligatoria: Perspectiva De Profesores y Alumnos*. Tesis Doctoral presentada en la Universidad de Barcelona. Director: Dr.César Coll Salvador.

Remesal, A. (2011). Primary and secondary instructors' conceptions of assessment: a qualitative study. *Teach. Instructor Educ.* 27, 472–482. doi: 10.1016/j.tate.2010.09.017

Remesal, A. (2020). *Autoevaluación sincrónica. Conferencia presentada en el Primer Encuentro Virtual Internacional: Enseñanza, Aprendizaje y Evaluación en la Educación Superior*. Universidad del Desarrollo, Santiago de Chile.

Remesal, A. (2021). "Synchronous self-assessment: assessment from the other side of the Mirror" in *Assessment as Learning: Maximising Opportunities for Student Learning and Achievement*. eds. Z. Yan and L. Yang (England: Routledge)

Remesal, A., Colomina, R., Nadal, E., Vidosa, H., and Antón, A. M. (2021). "Diseño de exámenes para autoevaluación sincrónica: emociones implicadas" in *Paper Presented at XI CIDUI 2020+1. Más Allá De Las Competencias: Nuevos Retos En La Sociedad Digital* (Barcelona). Universitat de Barcelona.

Remesal, A., Estrada, F. G., and Corrial, C. L. (2022). "Exams for the purpose of meaningful learning: new chances with synchronous self-assessment" in *Design and Measurement Strategies for Meaningful Learning*. ed. J. L. Gómez-Ramos (ISI-Global), 192–211. Pensilvania.

Remesal, A., Khanbeiki, A., Attareivani, S., and Parham, Z. (2019). *In Sito, Synchronous Self-Assessment: A New Research Strategy for Accessing Individual SA Processes*. Paper presented at EARLI 2019. Aachen, Germany, 12th-16th, August, 2019.

Roméu, M. R., and Díaz Quiñones, J. A. (2015). Valoración metodológica de la confección de temarios de exámenes finales de Medicina y Estomatología. *Educ. Méd. Super.* 29, 522–531. Available at: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21412015000300011&lng=es&tlng=es

Rusak, G., and Yan, L. (2021). Unique exams: designing assessments for integrity and fairness. In Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (pp. 1170–1176).

Schutz, P. A., and Pekrun, R. E. (Eds.) (2007). *Emotion in Education* Elsevier.

Schutz, P. A., and Zembylas, M. (Eds.) (2009). *Advances in Teacher Emotion Research: The Impact on Teachers' Lives* Springer NY.

Sutton, P. (2009). Towards dialogic feedback. *Critical Reflective Pract. Educ.* 1, 1–10. Cambrigde.

Sutton, R. E., Mudrey-Camino, R., and Knight, C. C. (2009). Teachers' emotions and classroom effectiveness: implications from recent research. *Theory Pract.* 48, 130–137. doi: 10.1080/00405840902776418

Tai, J., Ajjawi, R., Boud, D., Dawson, P., and Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *High. Educ.* 76, 467–481. doi: 10.1007/s10734-017-0220-3

Taras, M. (2009). Summative assessment: the missing link for formative assessment. *J. Furth. High. Educ.* 33, 57–69. doi: 10.1080/03098770802638671

Van Den Heuvel-Panhuizen, M. (2005). The role of contexts in assessment problems in mathematics. *For the learning of mathematics* 25, 2–23.

Villarroel, V., Bruna, D., Brown, G. T., and Bustos, C. (2021). Changing the quality of instructors' written tests by implementing an authentic assessment instructors' training program. *Int. J. Instr.* 14, 987–1000. doi: 10.29333/iji.2021.14256a

Xu, Y., and Brown, G. T. (2016). Instructor assessment literacy in practice: a reconceptualisation. *Teach. Instructor Educ.* 58, 149–162. doi: 10.1016/j.tate.2016.05.010

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assess. Eval. High. Educ.* 45, 224–238. doi: 10.1080/02602938.2019.1629390

Yan, Z., and Boud, D. (2021). "Conceptualising assessment as learning" in *Assessment as Learning. Maximising Opportunities for Student Learning and Achievement*. eds. Z. Yan and L. Yang (England: Routledge), 11–24.

Yan, Z., and Brown, G. T. L. (2017). A cyclical self-assessment process: towards a model of how students engage in self-assessment. *Assess. Eval. High. Educ.* 42, 1247–1262. doi: 10.1080/02602938.2016.1260091

Yin, R. K. (2014). *Case Study Research: Design and Methods. 5th* SAGE, California

Yu, L., and Suen, H. K. (2005). Historical and contemporary exam-driven education fever in China. *KEDI J. Educ. Policy* 2, 17–33.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 06 April 2023
DOI 10.3389/feduc.2023.1154592

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Stefan Daniel Keller,
University of Teacher Education Zuerich, Switzerland
Fernando José Sadio-Ramos,
Instituto Politécnico de Coimbra, Portugal

*CORRESPONDENCE
David Baidoo-Anu
✉ baidooanu.david@queensu.ca

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 30 January 2023
ACCEPTED 10 March 2023
PUBLISHED 06 April 2023

CITATION
Kissi P, Baidoo-Anu D, Anane E and Annan-Brew RK (2023) Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence.
Front. Educ. 8:1154592.
doi: 10.3389/feduc.2023.1154592

COPYRIGHT
© 2023 Kissi, Baidoo-Anu, Anane and Annan-Brew. 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.

Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence

Prosper Kissi[1], David Baidoo-Anu[2]*, Eric Anane[3] and Ruth Keziah Annan-Brew[4]

[1]Directorate of Academic Planning and Quality Assurance, University of Cape Coast, Cape Coast, Ghana, [2]Faculty of Education, Queen's University, Kingston, ON, Canada, [3]Institute of Education, University of Cape Coast, Cape Coast, Ghana, [4]Department of Education and Psychology, University of Cape Coast, Cape Coast, Ghana

This study explored the relationship between multiple choice test construction competence and the quality of multiple-choice tests among senior high school teachers in Ghana. In all, 157 teachers were selected from four senior high schools in the Kwahu-South District. Participants responded to self-designed questionnaire developed to assess teachers' multiple-choice items construction competencies. A three-factor structure emanated from the exploratory factor analysis on teachers' multiple choice test construction competence—content validity, item "options" handling, and test items assembling. Teachers in this study perceived more competence in ensuring content validity, followed by test item assembling, and handling of "options" (that is, alternatives) of the test items. The study also found serious problems with copies of multiple-choice items teachers have constructed for the students. Findings from this study provide unique and compelling evidence regarding teachers' perceived test construction competence and analysis of their multiple-choice tests. Implications for policy and practice are discussed.

KEYWORDS

assessment, test construction, Multiple-choice test, teachers, Ghana, senior high schools

Introduction

Classroom assessment plays an instrumental role in supporting and improving teaching and learning (Black and Wiliam, 1998, 2010). As part of the tools used in classroom assessment, teacher-made tests play a crucial role in the assessment process. That is, teacher-made tests aid in pre-assessment (the assessment of what students already know before teaching), formative assessment (the assessment of student performance incorporated into the act of teaching), and summative assessment (the assessment of student learning at the end of some instructional period) of students' learning outcomes (Gareis and Grant, 2015), which, in turn, informs relevant educational decisions. The need for teachers to understand and use teacher-made tests to improve students' learning is increasingly becoming important in the field of education (Guskey, 2003; Guskey and Jung, 2013). Teachers must be proficient and competent in the area

of assessment as they have traditionally been in the areas of curriculum and instruction (Gareis and Grant, 2015). Extant literature on test construction competencies and the quality of teacher-made tests showed that test construction competencies are related to the quality of test items (Marso and Pigge, 1989; Dosumu, 2002; Magno, 2003; Agu et al., 2013; Kinyua and Okunya, 2014). Thus, a teacher's competence in constructing test items is directly related to achieving good quality test instruments (Chau, as cited in Hamafyelto et al., 2015). When classroom teachers have limited test construction skills, the quality of the tests they construct is reduced. Tests that are poor in quality negatively affect the assessment validity (Amedahe and Asamoah-Gyimah, 2016). School teachers and administrators are not able to provide support and educational opportunities that meet each student's needs when the assessment tools constructed by teachers are low in quality (Agu et al., 2013). In other words, the lack of or low degree of validity of the test leads to undependable inferences about student learning (Gareis and Grant, 2015; Amedahe and Asamoah-Gyimah, 2016). Based on this, educational decisions such as the selection of students for educational opportunities would be wrongfully made.

The Ghanaian context

In Ghana, questionnaires as a self-report measure have been a common instrument that has been used to investigate test construction competencies or practices of classroom teachers (see Oduro-Okyireh, 2008; Anhwere, 2009; Wiredu, 2013; Armah, 2018). Accordingly, Wiredu (2013) suggested investigating teachers' responses to questionnaire items by directly examining samples of tests developed by the teachers for construction flaws. This will help provide qualitative information concerning the quality of the teacher-made tests. Previous studies on teachers' test construction competencies did not involve the analysis of samples of teacher-made tests in understanding the relationship between teachers' responses to a questionnaire on test construction knowledge and test construction practices (Oduro-Okyireh, 2008; Anhwere, 2009). To go beyond just relying on the responses of teachers on self-report measures, Oduro-Okyireh (2008) recommended that research should be conducted to understand teachers' actual test construction competencies (the quality of teacher-made tests). One of the measurement theories that call for the need to use item analysis (quantitative and qualitative methods) to evaluate the quality of teacher-made tests is the classical true-score theory.

It is imperative to note that most norm-referenced achievement tests are commonly designed to differentiate examinees with regard to their competence in the measured areas (Nitko, 2001). That is, the test is designed to yield a broad range of scores, maximizing discrimination among all examinees taking the test. This is based on the crucial assumption that psychological differences exist and can be detected through a well-designed measurement process (Furr and Bacharach, 2014). The well-designed measurement process is a question of the quality of the test constructed to detect individual differences in a given psychological construct such as achievement in mathematics. Therefore, constructing a test of good quality largely depends on an individual's ability to quantify the differences among people (Furr and Bacharach, 2014). For example, in educational settings, the onus rests on the teacher's ability (competence) to construct a measuring instrument that would help detect students who have gained mastery in a given content area and those who have not. However, test-related factors (format and construction flaws) which are attributed to the test construction competence of the classroom teachers affect how well their tests can detect high achievers from low achievers in a given subject area. Item analysis procedures, based on the assumptions of the classical true-score theory, create an avenue to validate teachers' responses to any self-report measure used in assessing their test construction competence.

Quantitative item analysis is a numerical method for analyzing test items' difficulty and discrimination indices employing student-response alternatives or options (Kubiszyn and Borich, 2013). The indices from the quantitative item analysis communicate the presence of problem items and errors that minimize the tests' utility in separating high achievers from low achievers. Accordingly, performing qualitative item analysis helps reveal more specific problems that contribute to unacceptable difficulty and discrimination indices (Nitko, 2001).

Given that the educational system in Ghana is examination-oriented (Baidoo-Anu and Ennu Baidoo, 2022), teachers are expected to develop competencies in test construction to be able to construct sound and quality tests (tests that are useful for measuring differences in students' achievement in a given subject area). However, with the large class size in Ghanaian classrooms, teachers are mostly forced to rely on multiple-choice tests to assess their students (Kissi, 2020). Despite the predominant use of multiple-choice items in Ghanaian classrooms, attention has not been given to teachers' multiple-choice test construction competence and the quality of the multiple-choice tests they construct. Accordingly, the crux of our study is to explore senior high school teachers' perceived multiple-choice test construction competence and the quality of multiple-choice tests they construct. Based on the objective of our study, the following research questions were developed to guide the study.

1. What is the perceived multiple-choice test construction competence of teachers in senior high schools in Ghana?
2. What are the characteristics of the multiple-choice test items constructed by the teachers based on the following criteria: difficulty index and discrimination index?
3. What are the common types of error associated with teacher-made multiple-choice tests among senior high school teachers in Ghana?

Literature review

Classical true-score theory

The theory conceptualizes any observed score on a test as the composite of two hypothetical components–a true score and a random error component. Mathematically, this is expressed in the form "X = T + E", where X represents the observed test score; T is the individual's true score; E is the random error component (Crocker and Algina, 2008). Thus, the theory is a simple mathematical model that describes how measurement errors can influence observed scores (Allen and Yen, 2002). The theory states that for every observed score, there is a true score or true underlying ability that can be observed accurately if there were no measurement errors (Allen and Yen, 2002). The observed score refers to a value that is obtained from the

measurement of some characteristic of an individual. A true score is a theoretical idea that refers to the average score taken over repeated independent testing with the same test or alternative forms. It is also the real or actual level of performance on the psychological attribute being measured by a test (Furr and Bacharach, 2014). Apart from the influence of true scores, probable factors that affect observed scores are described by the theory as errors of measurement (Furr and Bacharach, 2014). True scores and error scores are unobservable theoretical constructs while observed scores are observable in nature (Nitko, 2001).

Regarding the definition of observed scores in the theory, in a situation where there are no errors of measurement in observed scores, one can greatly and confidently depend on observed scores for relevant decisions. This is because, from repeated independent testing, all observed scores reflect the true ability of the candidate who is assessed. Also, supposing that a core mathematics achievement test is administered to a group of students who differ in ability and their observed scores are without measurement errors, the teacher would place his or her confidence in the assessment results because differences (variability) in the students' test scores accurately reflect the differences in their true levels of knowledge in mathematics.

Nevertheless, the existence of errors of measurement results in deviations of observed scores from the true score(s) (Bhattacherjee, 2012), and this minimizes one's confidence and dependability on the assessment results. How much confidence one can place in test results is a question of two main concepts pertaining to the quality of assessment procedures: (a) reliability and (b) validity. Therefore, the classical true-score theory provides an understanding of factors (measurement errors) that influence observed scores' reliability and validity. Examples of such factors include ambiguous items, poor instructions on a test, fatigue, and guesswork because of item difficulty (Crocker and Algina, 2008; Amedahe and Asamoah-Gyimah, 2016). To contribute to the reliability and validity of assessment results by ensuring test quality, the theory ties a good test to the test construction competence of the test constructor. Accordingly, the theory emphasizes some procedures and principles for test construction and evaluation to aid in the effective control and reduction of the impact of measurement errors related to a given test.

Principles, guidelines, or suggestions for constructing and improving the quality of multiple-choice tests

Errors associated with multiple-choice tests negatively affect the reliability and validity of the entire assessment results. To help improve the quality of the multiple-choice test, some principles, guidelines, or suggestions have been given by researchers, professionals, and experts in the educational assessment of students and psychological testing. In constructing multiple-choice tests, it is quintessential to follow the general principles of test construction and specific item format test construction principles. The outlined general test construction principles and specific principles for the construction of multiple-choice tests are organized as indicated by Nitko (2001), Joshua (2005), Kubiszyn and Borich (2013), and Etsey (as cited in Amedahe and Asamoah-Gyimah, 2016).

General principles for test construction

a. Begin writing items far enough in advance so as to have time to revise them.
b. Align the content of the test with instructional objectives.
c. Include items or questions with varying difficulty levels.
d. Match test items to the vocabulary level of the students.
e. Be sure that the item deals with an important aspect of the content area.
f. Write or prepare more items than are actually needed.
g. Be sure that the problem posed is clear and unambiguous.
h. Be sure that each item is independent of all other items. That is, the answer to one item should not be required as a condition for answering the next item. A hint to one answer should not be embedded in another item.
i. Be sure the item has one correct or best answer on which all experts would agree.
j. Prevent unintended clues to the answer in the statement or question. Grammatical inconsistencies such as "a" or "an" give clues to the correct answer to those students who are not well prepared.
k. Give specific instructions on the test. For example, instructions should be given as to how students are required to answer the questions.
l. Give the appropriate time limit for the completion of the test.
m. Appropriately assemble the test items. For example, use a font size that students can see and read, properly space the items, and arrange test items according to difficulty level (that is, from low to high); number the items one after the other without interruption, and appropriately assign page numbers.
n. Use an appropriate number of items to test students' achievement.
o. Review items for construction errors.
p. Evaluate the test items for clarity, practicality, efficiency, and fairness.

Specific principles for constructing multiple-choice test

a. Present the stem as a direct question.
b. Present a definite, explicit, and singular question or problem in the stem.
c. Eliminate excessive verbiage or irrelevant information from the stem.
d. Include in the stem any word(s) that might otherwise be repeated in each alternative.
e. Use negatively stated stems carefully (by underlining and/or capitalizing or bolding the negative word in the stem).
f. Make alternatives grammatically parallel with each other and consistent with the stem.
g. Make alternatives mutually exclusive or independent of each other.
h. Avoid the use of "none of the above" as an option when an item is of the best answer type.
i. Avoid the use of "all of the above" as part of the options to the stem of an item.
j. Make alternatives approximately equal in length.
k. Present alternatives in a logical order (for example, chronological, most to least, or alphabetical) when possible.
l. Keep all parts of an item (stem and its options) on the same page.

m. Arrange the alternatives in a vertical manner.
n. Use plausible distractors/options/alternatives.

Though the classical true-score theory has been described as a weak theory, its application to examine the quality of test items was of particular interest as it helps to understand the question, "Why is there a need for teachers to be competent in applying the principles of test construction?" It also endorses the use of quantitative methods of evaluating the quality of test items based on test scores and complements such evaluation with qualitative item analysis.

Assessment competence and assessment practice

The construction of tests for assessment is an aspect of classroom assessment practices that requires some level of assessment competence. Kissi, (2020) defined assessment competence as "an acquired, modifiable, and unobservable but demonstrable ability which is an integration of an individual's knowledge, skills, attitudes, and values in/on assessment" (p. 70). In light of this, assessment competence refers to an individual's ability to use or demonstrate the knowledge and skills acquired through assessment training in order to assess students' learning (Kissi, 2020). In contrast, assessment practice is the process of acquiring, analyzing, and interpreting data regarding student learning. It entails making crucial decisions on the student and the procedures involved in imparting knowledge to the learner (Nitko, 2001). Assessment competence in the view of Kissi, (2020):

> *answers the question: how well do classroom teachers employ their ability (which is an integration of their knowledge, skills, attitudes, and values in/on assessment) to successfully carry out those activities to match expected standards or to ensure improvement in their assessment activities? Since assessment competence in itself cannot directly be observed, such a construct can be inferred from what teachers do in terms of how well they go about their assessment practices (p. 71).*

Assessment competence and multiple-choice test construction competence

In the view of Gareis and Grant (2015), classroom teachers should be able to apply adequate knowledge and skills in assessment as they have usually been doing when it comes to activities involved in the transfer of knowledge to students. According to Nitko (2001), because the activities involved in the assessment are relevant to making relevant educational decisions, teachers have to be competent in choosing and using assessment tools. As stipulated in the standards for teacher competence in the educational assessment of students, for teachers to function effectively in assessment, they should be competent in assessment. According to the American Federation of Teachers (AFT), National Council of Measurement in Education (NCME), and National Education Association (NEA) requirements (as cited in Nitko, 2001), teachers should be capable of selecting, creating, administering, scoring, interpreting, and utilizing assessment data for pertinent educational decisions following legal and ethical norms (Kissi, 2020). From the foregoing, it can be seen that one of the standards for teachers' assessment competencies is their capacity to design and create tests. This standard indicates the test construction skills they need to have (Kissi, 2020).

One factor that directly affects the test quality, in Chau's view (as cited in Hamafyelto et al., 2015), is the proficiency of classroom teachers in constructing assessment tools. To gather information to improve teaching and learning, it is possible to identify students' areas of weakness and instructional issues given that the assessment tools used are of good quality (Nitko, 2001). According to McMillan (2000), understanding how general, fundamental assessment guidelines and ideas may be applied to improve student learning and teacher effectiveness is what is most important about assessment.

As one of the assessment competence criteria, test construction competence requires teachers to be adept at adhering to specific principles while creating assessment instruments or procedures that are suitable for instructional decisions. AFT, NCME, and NEA (as cited in Nitko, 2001) indicate that instructors who are proficient in this area will have the following conceptual and application skills in (a) planning the construction of assessment tools that help to inform decisions about students and instructional procedures; (b) selecting an appropriate technique which meets the intent of their instruction; (c) adhering to appropriate principles for developing and using assessment methods or techniques in their teaching, and avoiding common mistakes in student assessment; (d) using student data to examine the quality of each assessment technique used. In order to effectively assess pupils or students in accordance with the instructional objectives presented in class, teachers must select item format(s) that is or are suitable to the intent of their instruction. In Ghanaian senior high schools, the predominant item formats used in constructing end-of-term examination questions or tests are the essay and the multiple-choice item formats (Kissi, 2020). Hence, understanding teachers' multiple-choice test items' construction is needed within Ghana's educational system.

What is a multiple-choice item?

A multiple-choice item is an item that is made up of one or more introductory sentences followed by a list of two or more suggested responses (Nitko, 2001). The student is required to choose the correct answer from among the responses the teacher gives (Nitko, 2001). The part of the item that asks the question is called the stem. Instead of asking a question, it may set the task a student must perform or state the problem a student must solve. The list of suggested responses to the stem is called options. The options are also known as alternatives, responses, or choices (Morrow et al., 2000; Nitko, 2001). Usually, only one of the options is the correct or best answer to the question or problem the teacher poses. This is called the keyed answer, keyed alternative, or simply the key. The remaining incorrect options are called distractors or foils (Nitko, 2001; Joshua, 2005). To ensure that the assessment task neither prevents nor inhibits a student's ability to demonstrate attainment of the learning target, care should be taken to follow the guidelines for constructing multiple-choice tests. For instance, avoiding ambiguous and imprecise items, inappropriate and unfamiliar vocabulary, and poorly worded directions. After the first draft of the items, the items should be reviewed and edited. Moreover, the marking scheme should be prepared in conjunction with drafting the items (Etsey, as cited in Amedahe and Asamoah-Gyimah, 2016).

Quality of assessment procedures

The quality of assessment procedures is of great concern when it comes to the assessment of student learning. Ghanaian classroom teachers (trained and untrained) from the basic level to the university level construct, administer, and score classroom achievement tests regardless of whether they have had training in measurement and evaluation or not (Anhwere, 2009). When classroom teachers encounter some difficulties and/or do not possess adequate skills in test construction, the quality of the tests they construct is questionable. This is because, according to Chau (as cited in Hamafyelto et al., 2015), a teacher's test construction competence is directly related to ensuring the quality of a test. Poor test quality negatively affects the validity of assessment results (Amedahe and Asamoah-Gyimah, 2016). From the aforesaid, by implication, when teacher-made tests are low in quality, school administrators and teachers will not be able to make available support and educational opportunities that each student needs (Agu et al., 2013). In other words, a lack of or low degree of validity of test results leads to undependable inferences about student learning (Gareis and Grant, 2015; Amedahe and Asamoah-Gyimah, 2016) based on which educational decisions such as promotion and selection of students for educational opportunities would be wrongfully made. To avoid or minimize the negative effects of assessment procedures that are low in quality, the onus rests on classroom teachers to ensure the quality of the assessment procedures they employ. However, in Ghana, since classroom teachers hardly engage in quantitative item analysis as a way of assessing the utility of their multiple-choice test items, this study was relevant in (a) measuring their perceived competence in test construction; (b) evaluating their perceived competence in terms of the difficulty and discrimination indices of their multiple-choice test items; (c) employing qualitative item analysis to examine what test-related errors affected some of the observed indices.

Examining test construction competence through quantitative and qualitative item analysis

Test tryout, administration, and quantitative evaluation of the test

Quantitative evaluation (or item analysis) is a numerical method for analyzing test items employing student response alternatives or options (Kubiszyn and Borich, 2013). Before one would be able to conduct quantitative item analysis, the test should be administered to a sample with similar characteristics as the actual group who will be taking the final test (Shillingburg, 2016). This is called a test tryout. According to Cohen and Swerdlik (2010), for classroom teachers, test tryouts (pilot work) need not be part of the process of developing their tests for classroom use. However, the classroom teacher can engage in quantitative evaluation of test items after a test has been administered. The technique will enable them to assess the quality or utility of the items. It does so by identifying distractors or response options that are not doing what they are supposed to be doing. Quantitative evaluation of test items is ideally suited for examining the usefulness of multiple-choice formats (Kubiszyn and Borich, 2013).

Qualitative evaluation of the test

This method is used to review items on printed copies for test construction errors (Kubiszyn and Borich, 2013). Again, it is required for assessing the worth of the test before it is produced in large numbers to be administered (Amedahe and Asamoah-Gyimah, 2016). It is also done after a test is administered following quantitative item analysis and its purpose is to find out qualitative information about what led to unacceptable indices from quantitative item analysis. Hence, qualitative evaluation (or item analysis) is a non-numerical method for analyzing test items not employing student responses but considering content validity, clarity, practicality, efficiency, and fairness (Amedahe and Asamoah-Gyimah, 2016). Content validity, as one of the qualitative evaluation criteria, answers the questions: Are the items representative samples of the instructional objectives covered in a class? Does the test genuinely reflect the level of difficulty of the materials covered in a class? If the answer is "Yes," then content-related validity evidence is established (Amedahe and Asamoah-Gyimah, 2016). Clarity as another measure of evaluating the worth of the test refers to how the items are constructed and phrased while simultaneously judging them against the ability levels of the students. That is, the test material should be clear to students as to what is being measured and what they are required to do in attending to the questions (Nitko, 2001).

Practicality is concerned with the adequacy of the necessary materials and the appropriateness of time allocated for the completion of the test (Brown, 2004). The efficiency of a test seeks information as to whether the way the test is presented is the best to assess the desired knowledge, skill, or attitude of examinees in relation to instructional objectives (Amedahe and Asamoah-Gyimah, 2016). Conversely, fairness refers to the freedom of a test from any kind of bias. The test should be judged as appropriate for all qualified examinees irrespective of race, religion, gender, or age. The test should not disadvantage any examinee or group of examinees on any basis other than the examinee's lack of knowledge and skills the test is intended to measure (Nitko, 2001). Since the study focused on examining the characteristics of the teacher-made end-of-term multiple-choice tests, after the quantitative item analysis, the qualitative item analysis was used to identify possible test-related factors that affected the psychometric properties of the tests (in terms of difficulty and discrimination indices). Though qualitative evaluation is wide in scope, for the purpose of the study, it was operationalized as the deviations observed with respect to the principles of test construction using the multiple-choice test error analysis checklist (see Appendix A).

Methods

The study employed descriptive research design to understand senior high school teachers' multiple-choice test construction competence and the quality of multiple-choice test items they constructed. The study was done in two phases. The first phase was to obtain information on the multiple-choice test construction competence of the teachers. The second phase was to help validate the perceived multiple-choice test construction competence of the teachers through quantitative and qualitative item analysis.

Participants

Participants' selection was done in two phases.

Phase one

We selected 157 teachers from four senior high schools in the Kwahu-South District in the Eastern Region of Ghana. These participants were form one, form two, and form three teachers distributed across seven subject teaching areas (Financial Accounting, Cost Accounting, Business Management, Economics, English Language, Integrated Science, and Core Mathematics). The 157 participants responded to the self-designed questionnaire developed to assess teachers' multiple-choice test construction competence.

Phase two

The 157 participants who responded to the self-designed questionnaire were asked if they were willing to provide (a) copies of their latest end-of-semester self-constructed and administered multiple-choice test, (b) marking scheme, and (c) students' responses on the administered end-of-semester multiple-choice test items. Out of 157 participants, 47 teachers (across all the subject areas) provided these documents for further analysis. Accordingly, the 47 teachers were selected for the item analysis of the multiple-choice items they have constructed. Out of the 47 teachers, 68.09% had a first degree with education, 23.40% had a first degree without education, 4.26% of the participants had a master of philosophy, and 4.26% had completed a master of education programme. It is evident from the results that most of the participants were first-degree holders with a background in education. In the pursuit of a first degree, master of education, and master of philosophy, one is introduced to courses related to educational assessment of students' learning outcomes. From the cumulative percent, most of the participants (76.60%) possessed basic competence in the assessment of students (Kissi, 2020).

Instruments

The two instruments used for the data collection exercise were questionnaires and document examination. A 20-item self-designed instrument titled *Teachers' Multiple-Choice Test Construction Competence Questionnaire (TTCCQ-MC)* was used to assess teachers' multiple-choice test construction competence. The instrument was developed based on a comprehensive literature review on test construction competence. The instrument is made up of two sections namely "Section A" and "Section B." Section A is made up of items that help to obtain demographic information on teachers and Section B is made up of items that help to assess teachers' multiple-choice test construction competence. The scale of measurement that is used for the items under Section B is a 4-point Likert-type scale on a continuum of strongly disagree (SD), disagree (D), agree (A), and strongly agree (SA). The content validity of the research instrument was established by making sure that it objectively, fairly, and comprehensively covered the domain that it purports to cover. The instrument which was used for the study was initially made-up of 23 items. However, the items were reduced to 20 after experts' and teachers' judgment and pre-testing of the instrument among 130 teachers in a different district with similar characteristics as the study area. Concerning the experts' and teachers' judgment of the items on the questionnaire, items that were ambiguous and difficult to understand were rephrased so that the respondents could easily read and understand.

Document examination in this study covered students' responses on multiple-choice test items for end-of-semester administered teacher-made tests, copies of the marking schemes, and end-of-semester teacher-made tests. Using the marking schemes and students' responses on multiple-choice test items administered by the research participants, quantitative item analysis was performed to assess the characteristics of the multiple-choice test items for each of the classroom teachers. The assessment criteria used in assessing the characteristics of the items are based on the following item analysis descriptive statistics indices: (a) difficulty index (p-value), and (b) discrimination index (DI).

Based on the literature reviewed, the criteria suggested by Allen and Yen (1979) in terms of acceptable difficulty indices ranging from 0.30 to 0.70 and Kubiszyn and Borich's (2013) recommendation of at least a positive discrimination index for norm-reference tests, the following criteria were used in determining the characteristics of the teacher-made test items:

1. An item is judged as a good item if it is within the range of 0.30 to 0.70 and has a positive discrimination index.
2. An item is a problem item if it is within the range of 0.30 to 0.70 but has a zero discrimination index.
3. An item is a problem item if it is within the range of 0.30 to 0.70 but has a negative discrimination index.
4. An item is a problem item if it falls outside the range of 0.30 to 0.70 but has a positive discrimination index.
5. An item is a problem item if it falls outside the range of 0.30 to 0.70 and has a zero discrimination index.
6. An item is a problem item if it falls outside the range of 0.30 to 0.70 and has a negative discrimination index.

In addition, with regard to a qualitative evaluation of the teacher-made tests for format and construction flaws, the participants' end-of-semester administered Business Management and Core Mathematics multiple-choice tests were assessed for errors using the "Multiple-Choice Test Error Analysis Checklist" (see Appendix A).

Data analysis

Research question one sought to explore and describe the multiple-choice test construction competence of teachers in assessing students' learning outcomes at the senior high school level in the Kwahu-South District. The scoring of items based on the 4-point Likert scale of measurement was strongly agree = 4, agree = 3, disagree = 2, and strongly disagree = 1. After the scoring, exploratory factor analysis was performed to determine the factor structure of the multiple-choice test construction competence of the teachers, and the factors were ranked based on their respective explained variance and mean. The standard deviations associated with each mean were also provided. Concerning the use of mean, a criterion score (CS) of 2.50 (that is, $[1+2+3+4]/4 = 2.50$) using the item's mean was established to determine the level of the respondents' agreement or disagreement towards their perceived test construction competence. An item mean score of 2.50 or above indicates teachers' positive attitudes, while a

mean below 2.50 indicates teachers' negative attitudes which are embedded in each indicator of how well they employ their competence in constructing multiple-choice tests. After obtaining the difficulty and the discrimination indices for each set of items constructed by the teachers, means, standard deviations, and sum were used to analyze data collected on research question two. In addition, the mean of the problem items and the good items were compared using MedCalc's comparison of means calculator after meeting the assumptions that permit such analysis. Concerning research question three, "common format and construction flaws" is a categorical variable; therefore, frequency count was reported.

Results

Research question one

To answer research question one, there was a need to understand the structural patterns from teachers' responses to the TTCCQ-MC; thus, exploratory factor analysis was conducted using principal component analysis (PCA) with orthogonal rotation (varimax). Prior to starting the factor analysis, data were checked to ensure appropriateness for factor analysis. Exploratory factor analysis was conducted using 20 items that assess teachers' multiple-choice test construction competence. In testing the assumptions for PCA, the determinant of the correlation matrix as an indicator of multicollinearity was 0.015, which was substantially greater than the minimum recommended value of 0.00001. This meant that multicollinearity was not a problem in conducting PCA. The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = 0.70, and all KMO values for individual items were > 0.50, which was above the acceptable limit of 0.50 (Field, 2018). This meant that the sample size was adequate for PCA. Bartlett's test of sphericity was significant (χ^2 (190) = 619.939, $p < 0.001$). This indicated that correlations between items were good for PCA.

After satisfying the assumptions for PCA, an initial analysis was run to obtain the eigenvalue for each component in the data. Seven components had eigenvalues over Kaiser's criterion of one and in combination explained 60.65% of the total variance. The scree plot (see Figure 1) showed a point of inflexion that would justify retaining three components. Given the sample size of 157 and 20 items, the Kaiser's criterion on seven components, and convergence of the scree plot on three components, parallel analysis (PA) was conducted in addition to examine the appropriate number of components to maintain. Hayton et al. (2004) have pointed out that PA helps to identify the meaningful number of emerging factors from the set of items that are to be maintained. The PA also endorsed maintaining three factors. The results from PCA and reliability analysis endorsed 19-item TTCCQ-MC. That is, considering the absolute cut-off value of 0.40 for factor loadings, one of the items did not load on any of the factors since their loadings were below the cut-off value. Therefore, the final 19-item questionnaire with an overall reliability coefficient of 0.73 was considered valid for assessing teachers' multiple-choice test construction competence.

Based on the exploratory factor analysis, a three-factor structure emanated to help us understand teachers' multiple-choice construction competence. The first factor was termed test item assembling. Seven items were loaded into this factor, explaining 13.43% of the variance. These teachers prioritized ensuring proper spacing of test items for easy reading, keeping all parts of an item (stem and its options) on the same page, making sure options are approximately equal in length, and appropriately assigning page numbers to the test with clear specific instructions on the test. Factor two was named test content validity. Six items were loaded into this factor, explaining 12.50% of the variance. The items focused on teachers' priority in making sure that test items are matched to instructional objectives (intended outcomes of the appropriate difficulty level), preparing the marking scheme while constructing the items, and ensuring that each item deals with an important aspect of the content area and pose clear and unambiguous items. The third and final factor was named item "options" handling. Six items were loaded into this factor, explaining 11.77% of the variance. These teachers focused on ensuring that item options (i.e., alternatives) are approximately equal in length, options are presented in some logical order (e.g., chronological, most to least, or alphabetical) when possible, options are made independent of each other, and they also avoided the use of "none of the above" as an option when an item is of the best answer type. Exploratory factor analysis of teachers' multiple-choice test construction competence is presented in Table 1.

Rankings based on the percentage of explained variance indicated that the teachers perceived much more competence in assembling test items (13.43%) followed by competence in ensuring content validity (12.50%). Items options handling was the least perceived competence (11.77%) by teachers in our study. The results as presented in Table 2 confirm the preceding observations.

As seen in Table 2, comparing the mean of means for each component [competence in test item assembling (MM = 3.37, SD = 0.37), competence in ensuring content validity (MM = 3.24, SD = 0.41), and competence in handling items' options (MM = 2.80, SD = 0.48)] to the criterion score of 2.50, it can be said that, generally, for each component, most of the teachers perceived their competence as high. However, based on the rankings, it can be said that most of the research participants found it very easy to exhibit competence in assembling test items (MM = 3.37, SD = 0.37, R = 1st), easy to demonstrate competence in achieving content validity (MM = 3.24, SD = 0.41, R = 2nd), and quite difficult to demonstrate competence in handling the items' alternative (MM = 2.80, SD = 0.48, R = 3rd).

Research question two

The result of the characteristics of the multiple-choice items developed by the research participants is presented in Table 3.

As shown in Table 3, based on quantitative items analysis statistics, items that met both acceptable criteria for the discrimination index and difficulty index were judged as good items. Items that did not meet the set criteria were judged as problem items. The result showed that out of the total number of 2,325 items, 2,306 were deemed valid for item analysis (that is, multiple-choice items with four options). This means that 19 items were excluded from the items analysis. With respect to the set criteria for assessing the characteristics of the items, out of the total of 2,306 items, 1,199 items were described as good items, and 1,107 items were identified as problem items. This means that most of the test items constructed by the teachers are described as good items per their respective difficulty and discrimination indices. However, the 1,107 items identified as problem items might

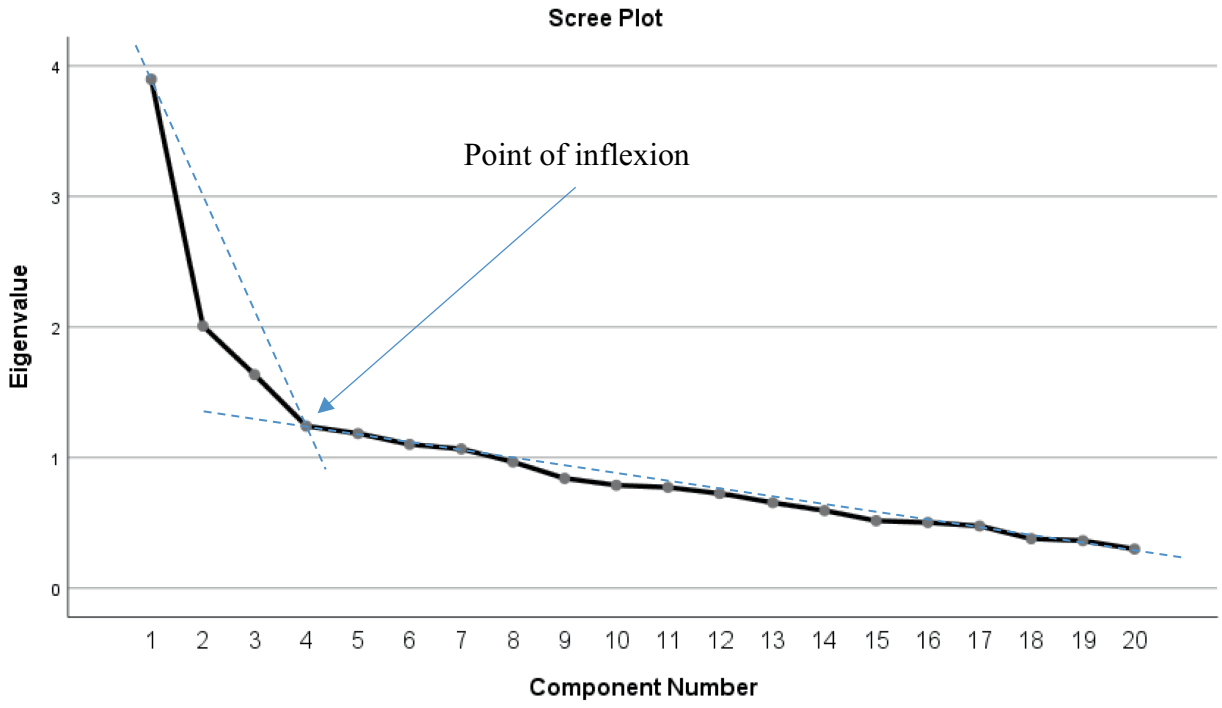


FIGURE 1
Scree plot for factor extraction.

have posed serious consequences for students who responded to these items.

Further analysis using "MedCalc's Comparison of means calculator" suggests that the average value for the number of good items produced by the classroom teachers (M = 25.51, SD = 8.51) was not statistically greater than the average value for the number of problems items produced (M = 23.55, SD = 8.98), t (92) = 0.03, p = 0.28, 2-tailed. Accordingly, it can be said that with respect to test characteristics, in general, the test items for assessing students' achievement lacked a suitable level of psychometric properties. This is attributable to the fact that the total number of good items produced by the teachers was not statistically different from the total number of problem items. Table 4 presents the result on problem items based on unacceptable difficulty indices that are less than 0.30, difficulty indices that are greater than 0.70, and discrimination indices that are less than or equal to 0.00.

Table 4 shows that out of the total number of 2,306 valid items for item analysis, 664 had difficulty indices less than 0.30 (difficult items) and 295 had difficulty indices greater than 0.70 (easy items). This means that most of the items were difficult. Further, in sum, the unacceptable number of items according to Allen and Yen's (1979) item evaluation criteria for item difficulty is 959 (that is, 664 + 295). On the other hand, out of the 2,306 valid items, 395 items had unacceptable discrimination indices less than or equal to zero based on Kubiszyn and Borich's (2013) recommendation that one can seriously consider any item with a positive discrimination index for the norm-referenced test(s). This means that most of the items had unacceptable difficulty indices as compared to the discrimination indices.

Research question three

The literature review on the use of quantitative item analysis in assessing items' characteristics revealed that the presence of problem items calls for qualitative evaluation of the multiple-choice test. Thus, research question three was formulated to help identify the multiple-choice format and item construction errors associated with teacher-made multiple-choice tests. In addressing this research question, the participants' end-of-semester administered Business Management (BM) and Core Mathematics (CM) multiple-choice tests were assessed for errors using the "Multiple-Choice Test Error Analysis Checklist" (see Appendix A). In all, 12 achievement tests (BM, 4; CM, 8) were qualitatively examined. The results are presented in Table 5.

As can be seen from Table 5, with specific reference to format errors, 11 out of 12 tests (BM, 3 out of 4; CM, 8 out of 8) were identified to have a detectable pattern of correct answers. Also, 6 out of 12 tests had items with a font size that some of the students could find more difficult to see and read (BM, 0 out of 4; CM, 6 out of 8). Therefore, it could be said that most of the tests were identified with the problem of a detectable pattern of correct answers as compared to the use of font size that students could find difficult to see and read.

To examine construction flaws associated with the tests, problem items were qualitatively examined. From Table 5, each of the following errors was observed with the problem items across 9 out of the 12 tests: (a) clues to the correct answer, (b) instruction-related issues (no and/or incomplete instruction), and (c) time for completion of items not indicated on the test. These observed errors are followed by other errors such as the use of implausible distractors (that is, 8 out of 12 tests) and ambiguous items/more

TABLE 1 Exploratory factor analysis of the teachers' multiple-choice test construction competence.

Description		Factors		
		1	2	3
Percentage of variance explained (after rotation)		13.432	12.502	11.766
Initial eigenvalue		3.898	2.007	1.635
Parallel Analysis (Random eigenvalues)		1.687	1.556	1.463
Q/N	When constructing multiple-choice tests, I:	1 Test Items Assembling	2 Content Validity	3 Item Options Handling
1	match test items to instructional objectives (intended outcomes of the appropriate difficulty level)		0.699	
2	make sure each item deals with an important aspect of the content area		0.757	
3	prepare the marking scheme while constructing the items		0.528	
4	pose clear and unambiguous items		0.454	
5	give specific instructions on the test	0.481		
6	include in the stem any word(s) that might otherwise be repeated in each option			0.569
7	make the options grammatically consistent with the stem			0.544
8	make options independent of each other			0.604
9	avoid the use of "none of the above" as an option when an item is of the best answer type			0.529
10	make options approximately equal in length			0.717
11	present options in some logical order (e.g., chronological, most to least, or alphabetical) when possible			0.535
12	include questions of varying difficulty		0.492	
13	match items to the vocabulary level of the students	-	-	-
14	give appropriate time for completion of the test		0.464	
15	use the appropriate number of test items	0.632		
16	number the test items one after the other	0.584		
17	appropriately assign page numbers to the test	0.475		
18	properly space the test items for easy reading	0.683		
19	keep all parts of an item (stem and its options) on the same page	0.502		
20	review test items for construction errors	0.571		

Extraction Method: Principal Component Analysis.

than one correct answer (that is, 7 out of 12 tests). On the contrary, 1 out of the 12 tests was identified with clueing and linking items (that is, BM, 1 out of 4; CM, 0 out of 8). Thus, the result suggests that most of the tests examined with reference to construction errors associated with problem items had the following issues: (a) clues to the correct answer, (b) instruction-related issues (no or incomplete instruction), (c) time for completion of items not indicated on the test, (d) implausible distractors, and (e) ambiguous items/more than one correct answer as opposed to clueing and linking items.

Discussion

The purpose of this study was to explore senior high school teachers perceived multiple-choice test construction competence and the quality of their multiple-choice tests. To understand the teachers' multiple-choice test construction competence in the Kwahu-South District, a three-factor structure emanated from the factor analysis - content validity, item "options" handing, and test item assembling. Findings showed that, generally, most of the teachers judged themselves as competent in constructing multiple-choice tests. In other words, they perceived themselves as possessing competence in achieving content validity, handling the options to the item's stems, and assembling the test. This observation could be related to the fact that most of the respondents had a background in education.

In the educational assessment of student learning outcomes, ensuring content validity and appropriately handling options of the item stems are more relevant competence areas as compared to competence in assembling test items. However, the teachers perceived more competence in test item assembling than ensuring content validity of the test and using appropriate "options or alternatives" of the test items. For example, the teachers perceived more competence in the proper spacing of test items for easy reading, keeping all parts of an item (stem and its options) on the same page rather than ensuring that test items are matched to instructional objectives

TABLE 2 Ranks of the teachers' multiple-choice test construction competence.

Component	Number of items	Mean of means (MM)	Std. deviation of MM	Ranks (R)
Competence in test item assembling	7	3.37	0.37	1st
Competence in Ensuring content validity	6	3.24	0.41	2nd
Competence in handling Items' options	6	2.80	0.48	3rd

TABLE 3 Characteristics of the multiple-choice items developed by the teachers.

Description	Items constructed by the teachers	Valid items for item analysis	Problem items	Good items
Sum (total)	2325.00	2306.00	1107.00	1199.00
Mean	49.47	49.06	23.55	25.51
Std. Deviation	14.15	14.14	8.98	8.51

TABLE 4 Summary of Items based on unacceptable difficulty and discrimination indices.

Description	Sum	Mean	Std. Deviation
Difficulty indices less than 0.30	664	14.13	6.98
Difficulty Indices greater than 0.70	295	6.28	4.56
Discrimination indices less than or equal to 0.00	395	8.40	5.74

(intended outcomes of the appropriate difficulty level), making the options independent of each other, and crafting options that are grammatically consistent with the stem to avoid clues to the correct answer.

Findings from this study revealed that the teachers perceived themselves as competent in constructing multiple-choice tests. However, findings from the quantitative evaluation of the items revealed that there were serious problems with copies of the multiple-choice tests the teachers constructed for assessing their students. Thus, most of the teachers in this study perceived themselves as competent multiple-choice test constructors; however, an analysis of the sample of their actual test items showed otherwise. This study confirms the recommendation by Ary et al. (2010) that direct observation of the behavior of a random sample of respondents is a brilliant strategy to validate their responses to self-report measures. The problems observed through the direct analysis of items were unacceptable difficulty and discrimination indices. In relation to item difficulty, from Nitko's (2001) point of view, teachers should ensure that the test they construct contains items that are not too difficult or too easy for their students. However, many of the items were described as problem items with respect to the high and low difficulty indices. Consequently, the quality of the assessment results used in grading the students was questionable. The findings from the quantitative item analysis support prior work that found that teachers often have inadequate prerequisite skills to construct quality multiple-choice items that effectively assess the learning achievements of students (Rivera, 2011; Agu et al., 2013; Kinyua and Okunya, 2014; Hamafyelto et al., 2015; Tshabalala et al., 2015). To address the issue, Nitko (2001) calls on classroom teachers to develop competence in tailoring test items to each of the student's ability levels. This is necessary as the reliability of an assessment is affected when test difficulty is not matched to the ability of the students involved (Amedahe and Asamoah-Gyimah, 2016).

Concerning the discrimination indices, Furr and Bacharach (2014) have stated that it is the responsibility of the classroom teacher to construct test items that effectively discriminate those who have mastered a given content area from those who have not. Where deficiencies exist, in norm-referencing, these items should not be considered in terms of the total number of items that make up students' composite scores in a given achievement test (Nitko, 2001; Crocker and Algina, 2008; Kubiszyn and Borich, 2013). However, these items were considered in arriving at the composite scores based on which grades were assigned.

According to Hambleton and Jones (1993), classical true-score theory item analysis procedures have the potential to provide invaluable information concerning construction flaws such as implausible distractors and double negatives. Therefore, informed by this assertion, research question three was established to identify the associated multiple-choice format and item construction errors that contributed toward the poor difficulty and discrimination indices through qualitative evaluation of the tests. The qualitative analysis of teachers-constructed multiple-choice items revealed fundamental flaws in the items' write-up and format errors which might have explained the problem items as identified with the teacher-made multiple-choice tests.

Generally, findings in relation to research question three reveal that most of the tests were identified with the problem of a detectable pattern of correct answers as compared to the use of font size that students could find difficult to see and read. Moreover, most of the tests examined with reference to construction errors were associated with problem items that had the following issues: (a) clues to the correct answer, (b) instruction-related issues (no or incomplete instruction), (c) time for completion of items not indicated on the test, (d) implausible distractors, and (e) ambiguous items/more than one correct answer as opposed to clueing and linking items. This supports Rivera's finding that classroom teachers do not possess adequate skills in constructing test items (Rivera, 2011).

Researchers have stated that the presence of format and construction errors reduces the quality of assessment results (Morrow et al., 2000; Nitko, 2001; Joshua, 2005; Kubiszyn and Borich, 2013; Amedahe and Asamoah-Gyimah, 2016). Therefore, problem items identified with the tests can pose serious consequences for students who responded to these items because these examinations in Ghana are high-stake (Amoako, 2019; Baidoo-Anu et al., 2022; Baidoo-Anu, 2022, Baidoo-Anu and Ennu Baidoo, 2022). Test results are used to make high stake decisions about students, especially determining their progress in the educational system. The findings from this study are

TABLE 5 Format and construction errors identified with the business management and core mathematics tests.

Type of errors	BM	CM	Total
Test format errors	Freq.	Freq.	Freq.
Alternatives not presented in some logical order	4/4	3/8	7/12
A detectable pattern of correct answers	3/4	8/8	11/12
The horizontal arrangement of options	3/4	4/8	7/12
Options of items appearing in different columns/pages	3/4	5/8	8/12
Page numbers not assigned	4/4	6/8	10/12
Poor arrangement of items/spacing of test items	2/4	4/8	6/12
Use of font size that is difficult to see and read	0/4	6/8	6/12
Item construction errors	Freq.	Freq.	Freq.
Ambiguous items/more than one correct answer	2/4	5/8	7/12
The central theme, task, or problem is not presented in the stem	3/4	0/8	3/12
Clues to the correct answer	4/4	5/8	9/12
Heterogeneous options	2/4	0/8	2/12
Grammatical, punctuation, and spelling errors	4/4	0/8	4/12
Implausible distractors	2/4	6/8	8/12
Instructional-related issues (no/ incomplete instruction)	4/4	5/8	9/12
Clueing and linking items	3/4	0/8	1/12
No answer	1/4`	3/8	4/12
Wrong key to the item	3/4	4/8	5/12
Not emphasizing (e.g., bolding, underlining or capitalizing) negative word in the stem	3/4	0/8	3/12
Time for completion of items not indicated on the test	4/4	5/8	9/12
Wrong answer	1/4	4/8	5/12

Key: Freq. = Frequency; / = out of.

consistent with several studies that found flaws in the multiple-choice items of teachers (Downing, 2004; Tarrant and Ware, 2008; DiBattista and Kurzawa, 2011; Wiredu, 2013). For example, Downing, 2004 argued that too high or low item difficulty disadvantaged some students. While Downing (2004, 2005) was not explicit on the type of students who are affected, Tarrant and Ware (2008) were more explicit and argued that flaws in high-stakes multiple-choice questions did not only disadvantage borderline students, rather high-achieving students were more likely than borderline students to be penalized by flawed items.

In sum, the direct assessment procedure helped to validate teachers' responses to the self-report measure used in the assessment of their competence in constructing multiple-choice. Both quantitative and qualitative item analyses were employed to validate the self-reported competence of the teachers. These methods revealed that though the teachers reported high levels of competence in constructing multiple-choice tests, the validation of their perceived competence using quantitative item analysis revealed that generally across all the seven subject areas, the number of problem items raise a concern about what they perceived about themselves and what their competence produced. Burton et al. (1991) have indicated that good multiple-choice test items are more demanding and take a lot of time to craft as compared to other types of test items. Given that multiple-choice test construction has different stages with each stage playing a significant role in test quality, teachers' less competence in any of the stages has the potential to mar the quality of tests (Agu et al., 2013). Thus, there is a need to ensure classroom teachers are practically exposed to item writing skills, especially ensuring content validity and crafting options to a multiple-choice item stem with good quality. According to Rivera (2011), classroom teachers can master the writing of test items through practice. Maba (2017) has also indicated that competence as an ability is modifiable and new experiences can be integrated. For instance, faculty members' (teachers) competence in developing multiple-choice test items with acceptable difficulty and discrimination indices improved significantly through training in constructing multiple-choice tests (Abdulghani et al., 2015). Consequently, new experiences gained by teachers as a result of exposure to constant training and practice in ensuring the quality of multiple-choice tests can lead to the integration and modification of their multiple-choice test construction competence.

Implication for policy and practice

Findings from this study provide unique and compelling evidence in the Ghanaian context regarding teachers' perceived test construction competence and analysis of teachers-constructed multiple-choice tests. Examinations results in Ghana are used to make high stake decisions regarding schools, teachers, and students (Baidoo-Anu and Ennu Baidoo, 2022). For instance, exam results determine students' progress from one grade to the other. Failure to pass these exams has dire consequences sometimes, including being retained in their present grade until they have passed the exams. This delays their progress and costs the family an extra year or more of associated schooling costs. According to the Ghana Ministry of Education (2018), more than 12% of senior high school students are retained in each grade level. Unfortunately, multiple-choice test items are the predominant type of items that are used during almost every examination in Ghana largely due to large class sizes. Thus, poor multiple-choice test constructions do not only affect students but also their families and the country's quality of education. This is because teachers' decisions made from these low-quality multiple-choice items

may lack valid evidence and may not represent the actual achievements of students. This implies that educational stakeholders will not be able to adequately provide support and educational opportunities that meet each student's needs. Therefore, Ghana Education Service should priorities providing in-service professional development training opportunities for teachers to develop the prerequisite skills needed to construct quality multiple-choice items. Professional development training of this nature is not a one-day workshop but demands ongoing long-term support and resources for teachers. Moreover, evidence (course outline) showed that teacher education programs have test construction as part of the topics in educational assessment courses; however, teaching this course is more theoretical and does not provide the opportunity to practically engage pre-service teachers. Hence, we recommend that teacher education programs in Ghana could also incorporate practical lessons or training in their curriculum to help pre-service teachers develop competence in test construction with specific emphasis on achieving content validity and effective handling of multiple-choice item stem options.

We want to highlight that findings from this study were shared with teachers and district education directors, especially those who participated in the study. The common problems identified including recommendations were also shared with them.

Limitations and suggestions for future research

The study employed 157 teachers to respond to the question and 47 teachers for the sample multiple-choice test analysis. Moreover, the sample multiple-choice test analysis was carried out on mathematics and business management test. Performing qualitative evaluation in other subject areas could have revealed more specific problems in all subject areas that contributed to unacceptable difficulty and discrimination indices. However, such general evaluation was not feasible in terms of easy access to subject area experts in English, Financial Accounting, Economics, Cost Accounting, and Integrated Science to help in qualitatively examining tests for construction flaws such as ambiguities, more than one answer, and clues to correct answers. Consequently, the conclusions based on the relatively small sample of teachers do not present a holistic view of the test construction competence of the entire population of teachers considered for the study. Given the significant nature of this study, future research could expand the scope and sample to allow the generalization of the findings across the country.

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 University of Cape Coast-Ghana. The patients/participants provided their written informed consent to participate in this study.

Author contributions

PK conceived of the presented idea. PK, DB-A, EA, and RA-B contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The work was supported by the School of Graduate Studies and Graduate Students' Association of Ghana, University of Cape Coast, through the Samuel and Emelia Brew-Butler -SGS/GRASAG, UCC Research Grant.

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.

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., et al. (2015). Faculty development programs improve the quality of multiple choice questions items' writing. *Sci. Rep.* 5:9556. doi: 10.1038/srep09556

Agu, N. N., Onyekuba, C., and Anyichie, C. A. (2013). Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: need for a test construction skill inventory. *Educ. Res. Rev.* 8, 431–439. doi: 10.5897/ERR12.219

Allen, M. J., and Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press Inc.

Allen, M. J., and Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Amedahe, F. K., and Asamoah-Gyimah, K. (2016). *Introduction to measurement and evaluation* (*7th ed.*). Cape Coast: Hampton Press.

Amoako, I. (2019). What's at stake in high stakes testing in Ghana: Implication for curriculum implementation in basic schools. *Int. J. Soc. Sci. Educ. Stud.* 5:72. doi: 10.23918/ijsses.v5i3p72

Anhwere, Y. M. (2009). Assessment practices of teacher training college tutors in Ghana. Available at: https://ir.ucc.edu.gh/xmlui/bitstream/handle/123456789/1690/ANHWERE%202009.pdf?sequence=1&isAllowed=y (Accessed May 23, 2020).

Armah, C. (2018). Test construction and administration practices among lecturers and staff of examinations unit of the university of Cape Coast in Ghana. Available at: https://ir.ucc.edu.gh/xmlui/handle/123456789/3868 (Accessed November 10, 2022).

Ary, D., Jacobs, C. L., Sorensen, C., and Razavieh, A. (2010). *Introduction to research methods in education* (*8th ed.*). Belmont, CA: Wadsworth.

Baidoo-Anu, D. (2022). Between-school streaming: unpacking the experiences of secondary school teachers and students in category C schools in Ghana. *Int. J. Educ. Res. Open* 3, 100188–100189. doi: 10.1016/j.ijedro.2022.100188

Baidoo-Anu, D., and Ennu Baidoo, I. (2022). Performance-based accountability: exploring Ghanaian teachers perception of the influence of large-scale testing on teaching and learning. *Educ. Inq.* 1-18, 1–18. doi: 10.1080/20004508.2022.2110673

Baidoo-Anu, D., Gyamerah, K., and Chanimbe, T. (2022). Secondary school categorization in Ghana: silent plights of students and implications for equitable learning. *J. Hum. Behav. Soc. Environ.* 33, 348–365. doi: 10.1080/10911359.2022.2061665

Bhattacherjee, A. (2012). Social science research: principles, methods, and practices (2nd ed.). Available at: http://scholarcommons.usf.edu/oa-tesxtbook/3 (Accessed May 5, 2020).

Black, P. J., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract.* 5, 7–74. doi: 10.1080/0969595980050102

Black, P., and Wiliam, D. (2010). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan* 92, 81–90. doi: 10.1177/003172171009200119

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., and Wood, B. (1991). How to prepare better multiple-choice test items: Guidelines for university faculty. Available at: https://testing.byu.edu/handbooks/betteritems.pdf

Cohen, R. J., and Swerdlik, M. J. (2010). *Psychological testing and assessment: An introduction to tests and measurement* (*7th ed.*). New York: McGaw-Hill.

Crocker, L., and Algina, J. (2008). *Introduction to classical & modern test theory*. Mason, OH: Cengage Learning.

DiBattista, D., and Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *CJSoTL* 2:4. doi: 10.5206/cjsotl-rcacea.2011.2.4

Dosumu, C. T. (2002). *Issues in teacher-made tests*. Ibadan: Olatunji and Sons Publishers.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Med. Educ.* 38, 1006–1012. doi: 10.1111/j.1365-2929.2004.01932.x

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Adv. Health Sci. Educ.* 10, 133–143. doi: 10.1007/s10459-004-4019-5

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (*5th ed.*). London: Sage.

Furr, R. M., and Bacharach, V. R. (2014). *Psychometrics: An introduction* (*2nd ed.*). London: Sage.

Gareis, R. C., and Grant, W. L. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning* (*2nd ed.*). New York: Routledge.

Ghana Ministry of Education (2018). The education strategic plan 2018–2030. Available at: https://www.globalpartnership.org/sites/default/files/2019-05-education-strategic-plan-2018-2030.pdf (Accessed May 20, 2022).

Guskey, T. R. (2003). How classroom assessments improve learning. *Educ. Leadersh.* 60, 6–11.

Guskey, T. R., and Jung, L. A. (2013). *Answer to essential questions about standards, assessments, grading, & reporting*. Thousand Oaks, CA: Corwin.

Hamafyelto, R. S., Hamman-Tukur, A., and Hamafyelto, S. S. (2015). Assessing teacher competence in test construction and content validity of teacher-made examination questions in commerce in Borno State, Nigeria. *J. Educ.* 5, 123–128. doi: 10.5923/j.edu.20150505.01

Hambleton, R. K., and Jones, R. W. (1993). An NCME instructional module on: comparison of classical test theory and item response theory and their applications to test development. *Educ. Meas. Issues Pract.* 12, 38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor retention decisions in exploratory analysis: a tutorial on parallel analysis. *Organ. Res. Methods* 7, 191–205. doi: 10.1177/1094428104263675

Joshua, M. T. (2005) *Fundamentals of test and measurement in education*. Calabar: University of Calabar Press.

Kinyua, K., and Okunya, L. O. (2014). Validity and reliability of teacher-made tests: case study of year 11 physics in Nyahururu District of Kenya. *Afr. Educ. Res. J.* 2, 61–71.

Kissi, P. (2020). *Multiple-choice construction competencies and items' quality: Evidence from selected senior high school subject teachers in Kwahu-South District*. Available at: http://hdl.handle.net/123456789/4641 (Accessed January 5, 2023).

Kubiszyn, T., and Borich, G. D. (2013). *Educational testing and measurement: Classroom application and practice* (*10th ed.*). Hoboken, NJ: Wiley.

Maba, W. (2017). Teachers' sustainable professional development through classroom action research. *Int. J. Res. Soc. Sci.* 7, 718–732.

Magno, C. (2003). The profile of teacher-made test construction of the professors of University of Perpetual Help Laguna. *UPHL Institut. J.* 1, 48–55.

Marso, R. N., and Pigge, F. L. (1989). "The status of classroom teachers' test construction proficiencies: assessments by teachers, principals, and supervisors validated by analyses of actual teacher-made tests." in *Paper presented at the annual meeting of the National Council on measurement in education, San Francisco, California*. (ERIC Document Reproduction Service No. ED306283).

McMillan, J. H. (2000). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin publishing company.

Morrow, J. R., Jr., Jackson, A. W., Disch, J. G., and Mood, D. P. (2000). *Measurement and evaluation in human performance* (*2nd ed.*). Champaign, IL: Human Kinetics.

Nitko, A. J. (2001). *Educational assessment of students* (*3rd ed.*). Upper Saddle River, New Jersey: Prentice-Hall.

Oduro-Okyireh, G. (2008). Testing practices of senior secondary school teachers in the Ashanti region. Available at: https://ir.ucc.edu.gh/xmlui/handle/123456789/1324 (Accessed April 3, 2020).

Rivera, J. E. (2011). Test item construction and validation: developing a statewide assessment for agricultural science education. *Career Tech. Educ. Res.* 36, 69–80. doi: 10.5328/cter36.2.69

Shillingburg, W. (2016). Understanding validity and reliability in classroom, school-wide, or district-wide assessments to be used in teacher/principal evaluations. Available at: https://cms.azed.gov/home/GetDocumentFile?id=57f6d9b3aadebf0a04b2691a (Accessed April 8, 2020).

Tarrant, M., and Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Med. Educ.* 42, 198–206. doi: 10.1111/j.1365-2923.2007.02957.x

Tshabalala, T., Mapolisa, T., Gazimbe, P., and Ncube, A. C. (2015). Establishing the effectiveness of teacher-made tests in Nkayi district primary schools. *Nova J. Humanit. Soc. Sci.* 4, 1–6.

Wiredu, S. G. (2013). Assessment practices of tutors in the nurses' training colleges in the Western and Central regions of Ghana. Available at: https://ir.ucc.edu.gh/xmlui/bitstream/handle/123456789/2685/WIREDU%202013.pdf?sequence=1&isAllowed=y (Accessed May 6, 2020).

10.3389/feduc.2023.1154592

Appendix A

Multiple-Choice Test Error Analysis Checklist Instruction: Record once if each error has occurred several times or once for each test.

Q/N	Errors in constructing multiple - choice test	Number of occurrences across tests	Total
	Test format errors		
1.	Alternatives not presented in some logical order		
2.	A detectable pattern of correct answers		
3.	The horizontal arrangement of options		
4.	Options of items appearing in different columns/pages		
5.	Page numbers not assigned		
6.	Poor arrangement of items/spacing of test items		
7.	Use of font size difficult to see and read		
	Item construction errors		
8.	Ambiguous items/More than one correct answer		
9.	The central theme, task, or problem is not presented in the stem		
10.	Clues to the correct answer		
11.	Clueing and linking items		
12.	Grammatical, punctuation, and spelling errors		
13.	Heterogeneous options		
14.	Implausible distractors		
15.	Instructional-related issues (no/incomplete instruction)		
16.	No answer		
17.	Not emphasizing (e.g., bolding, underlining or capitalizing) negative word in the stem		
18.	Time for completion of items not indicated on the test		
19.	Use of "all of the above"		
20.	Wrong answer		
21.	Wrong key to the item		
22.	Wrong usage of "none of the above"		

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 15 June 2023
DOI 10.3389/feduc.2023.1198709

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Zhe Li,
Osaka University, Japan
Dmitry Ryumin,
St. Petersburg Institute for Informatics and Automation (RAS), Russia

*CORRESPONDENCE
Shuwei Xue
✉ xueshuwei@ouc.edu.cn
Shifa Chen
✉ chenshifa99@ouc.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 02 April 2023
ACCEPTED 30 May 2023
PUBLISHED 15 June 2023

CITATION
Xue S, Xue X, Son YJ, Jiang Y, Zhou H and Chen S (2023) A data-driven multidimensional assessment model for English listening and speaking courses in higher education. *Front. Educ.* 8:1198709. doi: 10.3389/feduc.2023.1198709

COPYRIGHT
© 2023 Xue, Xue, Son, Jiang, Zhou and Chen. 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.

A data-driven multidimensional assessment model for English listening and speaking courses in higher education

Shuwei Xue[1*†], Xin Xue[2], Ye Jun Son[1†], Yaxuan Jiang[1], Hang Zhou[1] and Shifa Chen[1*]

[1]College of Foreign Languages, Ocean University of China, Qingdao, China, [2]School of Tourism Sciences, Beijing International Studies University, Beijing, China

Based on multiple assessment approach, this study used factor analysis and neural network modeling methods to build a data-driven multidimensional assessment model for English listening and speaking courses in higher education. We found that: (1) Peer assessment, student self-assessment, previous academic records, and teacher assessment were the four effective assessors of the multi-dimensional assessment of English listening and speaking courses; (2) The multidimensional assessment model based on the four effective assessors can predict the final academic performance of students in English listening and speaking courses, with previous academic records contributing the most, followed by peer assessment, teacher assessment, and student self-assessment. Therefore, a multidimensional assessment model for English listening and speaking courses in higher education was proposed: the academic performance of students (on a percentage basis) should be composed of 29% previous academic records, 28% peer assessment, 26% teacher assessment, and 17% student self-assessment. This model can guide teachers to intervene with students who need help in a timely manner, based on various assessors, thereby effectively improving their academic performance.

KEYWORDS

assessment methods, English listening and speaking courses, factor analysis, neural network modeling, peer assensment, self assessement, teacher assessment

1. Introduction

Assessment is one of the most complex cognitive behavior in the cognitive domain of educational goals (Bloom, 1956), and requires a rational and in-depth assessment of the essence of things. Currently, educational assessment is mostly static assessment or standardized testing (normative/standardized assessment; Haywood and Lidz, 2006). The tools and processes used in such assessments are standardized, and individual abilities are represented by statistical numbers (Sternberg and Grigorenko, 2002). The advantage of static assessment lies in its design objectivity, precision, and structuralism. However, it only provides test scores, focuses only on students' existing abilities, and the teacher is the only assessor. This assessment method is one-sided, easily leading students into a repetition of ineffective rote memorization tactics, seriously undermining students' learning interest and confidence (Thanh Pham and Renshaw, 2015). The teaching assessment of listening and speaking courses in universities, which are often the first to incorporate new teaching methods such as digital technologies, must evolve from traditional static assessment methods to more comprehensive and accurate ones. Therefore, it

is necessary to reform and improve the teaching assessment system of these courses.

The multiple assessment approach (Maki, 2002), based on the theory of multiple intelligences and constructivism, emphasizes the diversity of assessment methods, content, and subjects (Linn, 1994; Messick, 1994; Brennan and Johnson, 1995; Flake et al., 1997; Lane and Stone, 2006; Lane, 2013). Among them, the diversification of assessment subjects refers to the assessment of students by teachers (teacher assessment), student peers (peer assessment), and students themselves (student self-assessment). This is beneficial for expanding the sources of assessment information and potentially improving the reliability and validity of the assessment (Sadler, 1989; Shepard, 2000; Hattie and Timperley, 2007; Li et al., 2019; Ghafoori et al., 2021). The diversification of subjects aims to break the single-subject assessment model, allow more people to participate in the assessment activities, and transform assessment from one-way to multi-directional, to construct an assessment model that combines student self-assessment, peer assessment, and teacher assessment.

Student self-assessment is the learner's value judgment of his or her own knowledge and ability level (Bailey, 1996); peer assessment is the value judgment of a student's ability level, course participation, and effort by classmates (Topping et al., 2010); and teacher assessment is the value judgment of a student's learning situation made by the teacher. Educationalist Rogers believes that true learning can only occur when learners have a clear understanding of learning goals and assessment criteria (Rogers, 1969). Students as the main assessors embody the idea that "assessment is a learning tool" (Sitthiworachart and Joy, 2003), which is conducive to enhancing students' metacognitive and self-regulation abilities (Nicol, 2010), promoting teachers and students to discover each other's strengths and weaknesses, and timely improving temporary shortcomings. However, when using these three sources of assessment, we must be cautious about potential biases, such as reliability, grading, social response bias, response style, and trust/respect (Dunning et al., 2004; van Gennip et al., 2009; Van Gennip et al., 2010; Brown et al., 2015; Panadero, 2016; Meissel et al., 2017). To ensure optimal conditions for accuracy and avoid known pitfalls, both teachers and students should strive to be as objective as possible when evaluating performance. This highlights the importance of having a comprehensive assessment system rather than relying on a one-sided system.

Research on the diversification of assessment subjects has mainly focused on exploring the relationship between the three types of assessment mentioned above (To and Panadero, 2019; Xie and Guo, 2022). Studies showed that the relationships between the assessment subjects are weak (Boud and Falchikov, 1989; Falchikov and Boud, 1989; Falchikov and Goldfinch, 2000; Chang et al., 2012; Brown and Harris, 2013; Double et al., 2020; Yan et al., 2022). For instance, the results of student self-assessment and peer assessment were not consistent with the assessments given by teachers (Goldfinch and Raeside, 1990; Kwan and Leung, 1996; Tsai et al., 2002). Some studies have found that 39% of students overestimate their performance (Sullivan and Hall, 1997), while other studies have found that students' self-assessment scores were significantly lower than the scores given by teachers(Cassidy, 2007; Lew et al., 2009; Matsuno, 2009). The results of these studies were influenced by the type of task and the individual characteristics of the learners, and these factors need to be considered when interpreting the results. This allows for the possibility of conducting a factor analysis to differentiate assessments from the various assessors involved.

In addition to the factors related to the assessment subjects, students' previous academic performance is also a key factor that influences their current academic performance due to the cumulative effect of learning (Plant et al., 2005; Brown et al., 2008). As a student progresses through their education, the knowledge and skills they acquire build upon each other. Thus, if a student struggles in a prerequisite course or fails to master certain concepts, it can hinder their ability to succeed in subsequent courses. Additionally, a student's previous academic performance can affect their confidence and motivation, which can in turn impact their current academic performance (Ciarrochi et al., 2007). There are situations where previous performance has a stronger influence than other predictors, creating what is known as an autoregressive relationship (Biesanz, 2012). Overall, previous academic performance can serve as an assessor of future academic success, highlighting the importance of consistent effort and dedication in one's education. Therefore, in this study, we also included previous academic performance in the construction of the multidimensional assessment model.

With the increase of assessors in the assessment system, it is a more meaningful research problem to mine the association of various assessors in the data to provide decision-making guidance for education. Currently, in the field of computer science, the representative methods for data dimensionality reduction and modeling are factor analysis (Kim et al., 1978) and machine learning-based predictive modeling (Alpaydin, 2016). This study used the two methods to reduce dimensionality and model different sources of assessment information, and discovered patterns in complex data.

Factor analysis is a multivariate statistical method that can screen out the most influential factors from numerous items and use these factors to explain the most observed facts, thus revealing the essential connections between things (Tweedie and Harald Baayen, 1998). Factor analysis has a long history of successful application in corresponding education research (Cudeck and MacCallum, 2007). For instance, in China scholars analyzed various items that affect students' comprehensive quality using factor analysis, calculated each student's comprehensive score, and compared it with traditional assessment methods. They found that this method can make up for the shortcomings of relying solely on Grade Point Average (GPA) (Chang and Lu, 2010).

The application of machine learning-based predictive modeling methods in assessment studies has also become increasingly widespread. Among them, the neural network model, inspired by the structure of the human brain neuron, can simultaneously include multiple predictive variables in the model and calculate the contribution of variables to the model (Lecun et al., 2015). It is a multilayer perceptron. During the training phase, the connections between layers are assigned different weights. The hidden layer(s) also performs a kind of dimensionality reduction (like PCA) which helps to learn the most relevant of the many (correlated) features. It can implicitly detect all possible (linear or nonlinear) interactions between predictors which is advantageous over general linear regression models when dealing with complex stimulus–response environments (e.g., Tu, 1996). Scholars found that compared to the regression methods, the deep learning-based models were more effective in predicting students' performance (Okubo et al., 2017; Kim et al., 2018). Online English teaching assistance system, using decision tree

algorithms and neural network models, was also implemented, which improved the efficiency of teaching (Fancsali et al., 2018; Zheng et al., 2019; Sun et al., 2021). However, the hidden layer(s) likes a black box, as the common factors inside cannot be directly observed. Therefore, it is necessary to combine with other methods to "unbox" the intermediate stage.

Factor analysis and machine learning methods have different approaches, but they can be combined to improve the interpretability of the model. For instance, factor analysis can be used as a pre-processing step to simplify the input items before feeding them into a machine learning model. This reduces model complexity and clearly extracts factors and their constituent components. Machine learning methods can improve the predictive accuracy, aiding in interpreting the factor analysis results. The combination of the two methods has been used in various fields (Nefeslioglu et al., 2008; Marzouk and Elkadi, 2016), including education (Suleiman et al., 2019). Hence, this study attempts to combine factor analysis and machine learning methods in assessing English listening and speaking courses in universities to investigate the feasibility of using a multidimensional assessment model in these courses.

2. The present study

As mentioned, this study aims to explore effective assessment methods for English listening and speaking courses in higher education and construct a multidimensional assessment model. Specifically, it has two main research questions: (1) What are the key factors of the multidimensional assessment model? (2) Can the multidimensional assessment model constructed based on these key factors predict students' academic performance, and what is the significance of each factor in the model?

To address the above issues, we collected assessment data from various sources, including peer ratings of learners' language abilities and classroom performance, self-ratings by learners, teacher ratings, and previous academic records. With the help of computational science methods, specific assessment factors were extracted from complex data, and the assessment factors were tested to see if they could successfully predict students' current academic performance (See Figure 1 for an illustration). We hypothesized that: (1) factor analysis can distinguish different sources of assessment data, which can be summarized into four common factors: previous academic records, peer assessment, and teacher assessment, and student self-assessment; (2) the neural network model can use these four common factors to establish a prediction model for students' academic performance.

3. Methods

3.1. Participants

Sixty-two undergraduate students majoring in English from a university in China were recruited for this study. Among them, there were 55 female and 7 male students, with a *mean* age of 19.37 years ($SD_{age} = 0.71$), ranging from 18 to 21 years old. All participants were native Chinese speakers with English as their second language, with a *mean* age of acquisition (AOA) of 8.65 years ($SD_{AOA} = 1.56$). Prior to the experiment, all participants signed an informed consent form.

3.2. Tools

The research tools used in this study were mainly paper-and-pencil materials, including a background survey questionnaire, a student self-assessment scale, a peer assessment scale, a teacher assessment scale, and tests.

- The background survey questionnaire included three items: the English score in the college entrance examination (out of 150 points), the academic grade in the first semester of the listening and speaking course (out of 100 points), and the academic grade in the second semester of the listening and speaking course (out of 100 points).
- The student self-assessment scale required students to assess their own English proficiency, including listening ability, speaking ability, reading ability, and writing ability. It was a five-point Likert scale.
- The peer assessment scale required students to rate their classmates, except for themselves, based on their understanding of their classmates and their performance in the course. The scale included seven items: listening ability, speaking ability, reading ability, writing ability, class participation, cooperation and competitiveness awareness, and learning attitude and perseverance. It was also a five-point Likert scale.
- The teacher assessment scale required teachers to grade each student (out of 100 points) based on their comprehensive performance in the course.
- The tests included regular in-class tests and the final exam. There were seven regular in-class tests, with multiple-choice questions based on IELTS listening and a listening textbook. The average correct rate of the seven quizzes was used to represent the students' in-class test score (in percentage). The final exam comprised of a combination of randomly selected textbook exercises and TOEFL listening questions. The students' current academic performance was being evaluated based on the final exam score (out of 100 points).

3.3. Data collection

This study was conducted in the English Listening and Speaking course for undergraduate English majors. The course lasted for 16 weeks, with two class hours per week, taught offline by one teacher. The textbook used in the course was the *Viewing, Listening and Speaking, Student Book*, authored by Zhang E., Deng Y., and Xu W., published by Shanghai Foreign Language Education Press in January 2020, with an International Standard Book Number of 978–7–5,446-6,080-8.

In the course small group presentation sessions were designed, with each group consisting of 3–4 students, who chose a topic to prepare and present together, thereby enhancing the teacher's understanding of the students and the students' understanding of each other. Starting from the seventh week, in-class tests were randomly

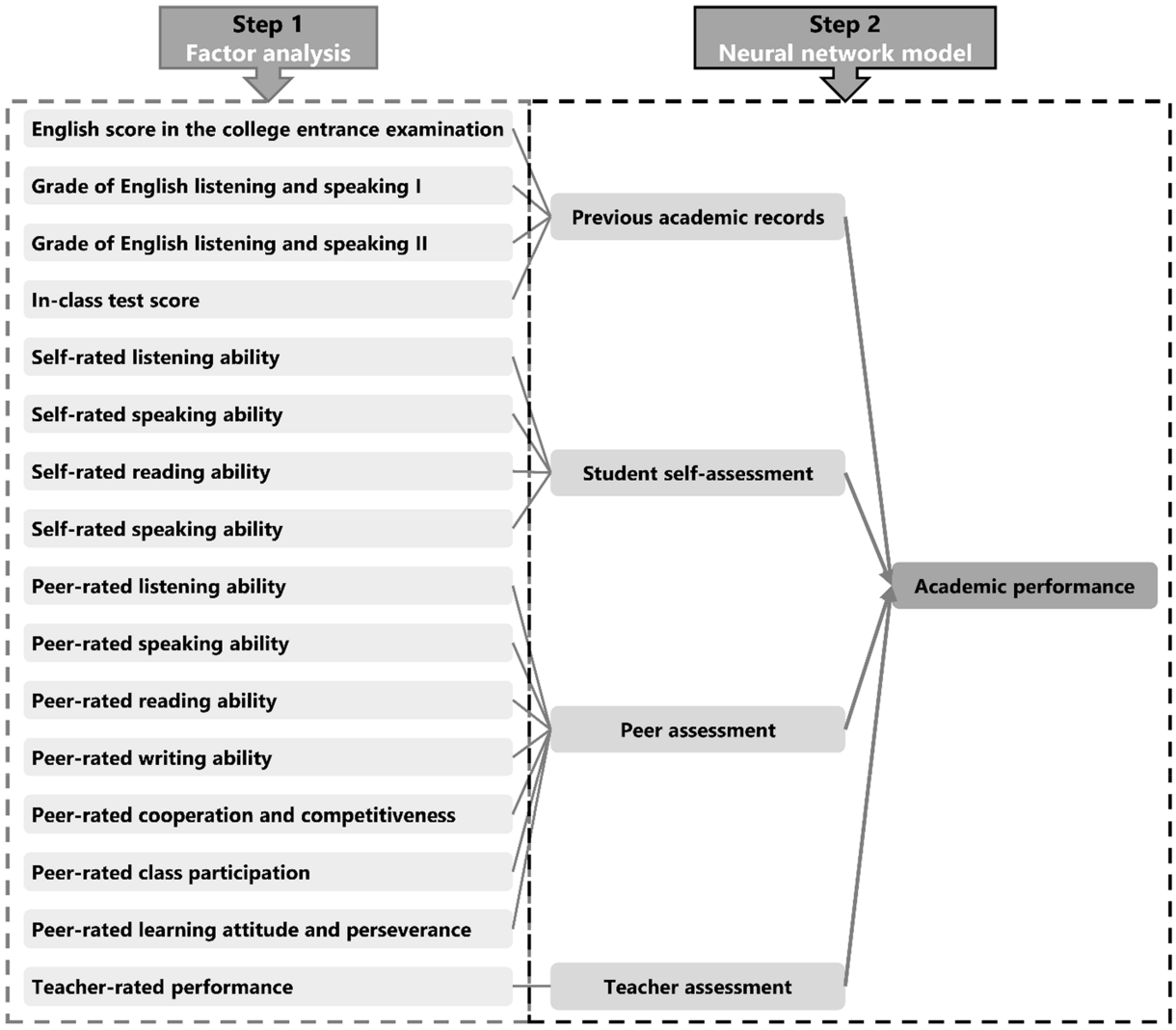


FIGURE 1
Conceptual model of multidimensional assessment. Items from four assessors were collected. Factor analysis was used to identify the key factors, and neural network model was used to construct the multidimensional assessment model based on the key factors.

arranged before class, totaling seven times. At the end of the course, students completed the background survey questionnaire, self-assessment scale, and peer assessment scale. Finally, a final exam was administered, and the teacher evaluated the students' test scores and rate each student based on his/her classroom performance.

3.4. Data analysis

3.4.1. Factor analysis

First, the data of 16 items were analyzed using Kaiser-Meyer-Olkin (KMO) test (Kaiser, 1974) and Bartlett's sphericity test (Stone et al., 2008) in JMP 14 Pro software (SAS Institute Inc., Cary, NC) to determine if the data was the factorability of the data for factor analysis (a KMO value of less than 0.60 indicates unsuitability for factor analysis, and if the null hypothesis of Bartlett's sphericity test is accepted, factor analysis cannot be performed). Second, items with loading of greater than or equal to 0.30 were determined to be statistically significant. Third, maximum likelihood method and oblimin rotation technique based on a correlation matrix were used to extract the factors and determine the number of factors. Fourth, common factors were extracted, and the factors were named to determine whether they reflected students' self-assessment, peer assessment, teacher assessment, and previous academic records, respectively.

We also conducted a confirmatory factor analysis using seven items due to the small sample size, which is generally recommended to have at least 10 people per item for factor analysis (Costello and Osborne, 2005). By shrinking the items to seven (Marsh et al., 1998), our sample had approximately 9 people per item. The selected seven items were specifically focused on listening and speaking courses, including academic grades in the first and the second semester, self-assessed listening and speaking ability, peer-assessed listening and speaking ability, and teacher assessment.

These items were chosen because they better represented the four hypothesized factors.

3.4.2. Neural network modeling

Using neural network modeling method in JMP 14 Pro software (SAS Institute Inc., Cary, NC), with the common factors extracted by factor analysis as the predictors (the standardized *mean* values of the items included in the common factors; e.g., Suhr, 2005, 2006), a predictive model for academic performance was constructed to explore the key assessment predictors affecting academic performance. The specific parameters of the model were as follows: the neural network model had three layers (input layer, hidden layer, and output layer), with three nodes in the hidden layer and a hyperbolic tangent (TanH) activation function. The model learning rate was set at 0.1, the number of boosting models was 10, and the number of tours was 10. In order to address the issue of overfitting, cross-validation was employed in the study. K-fold cross-validation was deemed more suitable when dealing with small sample sizes (Refaeilzadeh et al., 2009). This method divides the data into K subsets, and each of the subsets is used to test the model fit on the remaining data, resulting in K models. The best-performing model, based on test statistics, is selected as the final model. 10-fold cross-validation is typically recommended as it provides the least biased accuracy estimation (Kohavi, 1995).

The feature importance of each predictor in the model (feature importance) was calculated using the dependent resampled inputs method, with values ranging from 0 to 1. A value greater than 0.10 was considered a key factor affecting the outcome variable (Saltelli, 2002; Strobl et al., 2009). To avoid grouping errors in the cross-validation dataset, the 10-fold cross-validation process was repeated 100 times (iterations), and the model fit and the feature importance of predictors reported were the *means* of these 100 iterations (Were et al., 2015).

4. Results

4.1. Results of the factor analysis

The results of the factorability test for factor analysis based on 16 items showed that the KMO value was 0.75, and the Bartlett's sphericity test was significant ($\chi^2 = 533.56$, $df = 120$, $p < 0.001$), indicating the validity of conducting factor analysis on the data.

Four factors were extracted using the maximum likelihood method and oblimin rotation technique on a correlation matrix. The extraction was based on the eigenvalues (>1) and the "elbow" on the scree plot (refer to Figure 2) where the item's load on the common factor reached 0.30. As shown in Table 1, the results of the factor analysis were ideal, with a cumulative explained variance of 74.22%.

The results of the factorability test for factor analysis based on 7 selected items showed that the KMO value was 0.63, and the Bartlett's sphericity test was significant ($\chi^2 = 156.78$, $df = 21$, $p < 0.001$), indicating the validity of conducting factor analysis on the data. The four factors identified by the confirmatory factor analysis were shown in Table 2, and the cumulative explained variance of 84.99. Since this analysis was a supplementary analysis to validate the findings based on 16 items, we will continue to use the 16 items in the neural nets model.

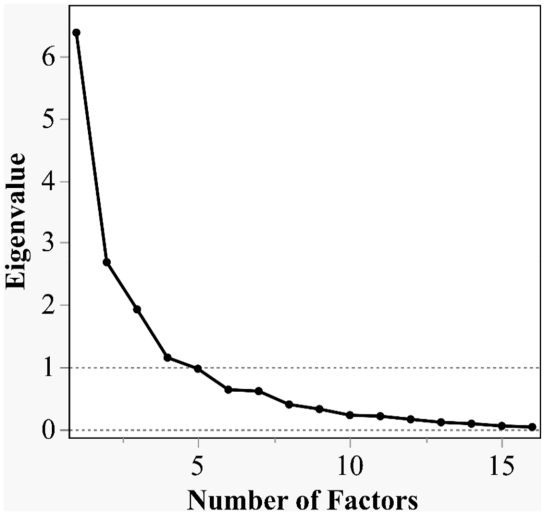


FIGURE 2
The scree plot from the factor analysis of 16 items. The "elbow" was at the fourth point.

4.2. Results of the neural network modeling analysis

As factor analysis cannot directly provide a predictive model for student academic performance, we need to use the neural network model method to build this predictive model on this basis. As shown in Figure 3, using the four standardized common factors obtained from factor analysis (the *mean* of the items contained in the standardized common factors) as predictor variables, the model fits of the predictive model for the academic performance in the current academic performance of the participants were acceptable (Mean $r^2_{training} = 0.84$, SD $r^2_{training} = 0.02$; Mean $r^2_{test} = 0.78$, SD $r^2_{test} = 0.14$).

We further calculated the feature importance of the four common factors in the model (see Figure 4). The results showed that all assessors played a critical role in predicting the academic performance in the current academic performance of the participants (Mean $_{feature\ importance} > 0.10$). The feature importance of the four assessors was as follows: previous academic records (Mean = 0.38; SD = 0.09), peer assessment (Mean = 0.36; SD = 0.08), student self-assessment (Mean = 0.22; SD = 0.09), and teacher assessment (Mean = 0.33; SD = 0.09). Among the four assessors, previous academic records, peer assessment, and teacher assessment had a greater contribution than student self-assessment.

4.3. The multidimensional assessment model in a percentage system

Based on the above results, we preliminarily constructed a multidimensional assessment model for English listening and speaking courses in higher education institutions (as shown in Figure 5).

The model was composed of a set of 16 assessment items. Based on the results of factor analysis, the four largest common factors that had the most impact were selected. The *mean* of the

TABLE 1 Loading of the 16 items of the multidimensional assessments in the factor analysis.

Items	Factor 1	Factor 2	Factor 3	Factor 4
Peer assessment				
Cooperation and competitiveness	0.97			
Listening ability	0.91			
Speaking ability	0.88			
Reading ability	0.85			
Writing ability	0.85			
Class participation	0.83			
Learning attitude and perseverance	0.57			
Self-assessment				
Listening ability		0.81		
Reading ability		0.78		
Speaking ability		0.78		
Writing ability		0.77		
Previous academic records				
Grade of the first semester's listening and speaking course			0.83	
Grade of the second semester's listening and speaking course			0.73	
In-class test score			0.51	
English score in the college entrance examination			0.41	
Teacher assessment				
Student's performance				0.63
Variance	5.58	3.10	2.61	0.59
Communicative Percent (%)	34.85	54.20	70.53	74.22

Only the loadings with absolute values greater than 0.30 are displayed.

items included in the common factors was standardized as the predictive variable of the neural network model, and a predictive model of student academic performance was constructed. Since the contribution of each predictive variable in the original model included both the main effect of the predictive variable and the interaction effect with other variables, the sum of the feature importance was greater than 100%. In order to make the maximum predicted value of academic performance 100 points, we converted the model to a percentage system. While this process was deemed necessary for our study because most courses in China use the centesimal system, it may not be necessary for other studies. The results showed that students' academic performance (in percentage) should be composed of

TABLE 2 Loading of the 7 items of the multidimensional assessments in the factor analysis.

Items	Factor 1	Factor 2	Factor 3	Factor 4
Peer assessment				
Listening ability	0.98			
Speaking ability	0.87			
Self-assessment				
Speaking ability		0.99		
Listening ability		0.67		
Previous academic records				
Grade of the second semester's listening and speaking course			0.89	
Grade of the first semester's listening and speaking course			0.69	
Teacher assessment				
Student's performance				0.98
Variance	1.87	1.54	1.54	1.00
Communicative Percent (%)	26.71	48.70	70.64	84.99

Only the loadings with absolute values greater than 0.30 are displayed.

29% for previous grades, 28% for peer assessment, 26% for teacher assessment, and 17% for student self-assessment.

5. Discussion

5.1. The effective components of the multidimensional assessment model based on factor analysis

The four common factors reveal that the seven assessments from classmates have high loading on factor 1, which reflects the results of peer assessment. Therefore, factor 1 can be named "peer assessment." The four assessments from student themselves have high loading on factor 2, which reflects the results of student self-assessment. Therefore, factor 2 can be named "self-assessment." The four items from the first two semesters' listening and writing grades, the in-class test score and the English college entrance exam score, have high loading on factor 3, reflecting the early academic performance of the participants. Therefore, factor 3 can be named "previous academic records." The teacher's assessment has high loading on factor 4, which reflects the rating results of the teacher. Therefore, factor 4 can be named "teacher assessment." The result indicates that the assessment items from different sources are relatively independent and have a certain level of discriminant validity. The multiple assessments of students' English listening and speaking courses can be composed of these four factors.

5.2. A predictive model for student academic performance based on neural network model method

The results of the predictive model built using the neural network method showed that the four assessors (previous academic records, peer assessment, teacher assessment, and self-assessment) could predict student academic performance and all of them were key factors in predicting academic performance (Mean $_{\text{feature importance}} > 0.10$). The order of their feature importance was previous academic records, peer assessment, teacher assessment, and student self-assessment.

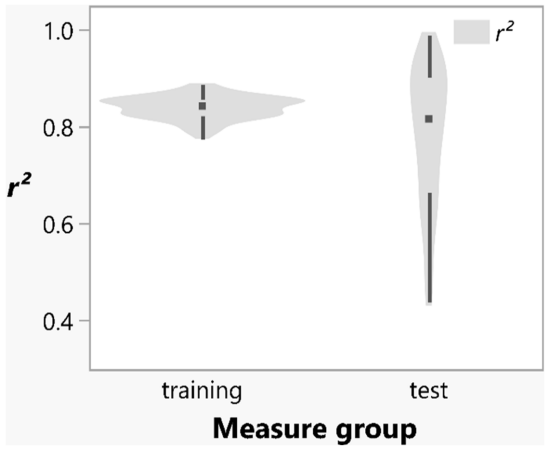


FIGURE 3
Distribution of the model fits of the neural network model. This figure shows the *mean* r^2 values obtained from 100 iterations of the training and test groups, and the error bars represent the standard deviations of 100 iterations.

The results of the present study indicate the previous academic grades and regular in-class test score had the strongest explanatory power (this may not necessarily hold true for other studies). Both of them were based on paper-and-pencil tests that were similar in form and content to the final exam of the current semester and were familiar to students. Research has shown that previous academic achievement can have a positive impact on learning strategies and motivation through the mediating effect of positive academic emotions (Elias and MacDonald, 2007; Vettori et al., 2020). When students have good previous academic performance, they experience positive emotions such as happiness, pride, and relaxation, which can motivate them to use cognitive strategies more flexibly, which in turn can have a positive impact on their subsequent academic performance.

Peer assessment was based on students' mutual understanding and can more objectively and comprehensively reflect students' abilities and daily performance (Shen et al., 2020). This study confirms previous findings that peer assessment scores have high reliability and are significantly correlated with students' final grades (Li et al., 2016). For the evaluated students, the timely and rich feedback provided by peer assessment helps to avoid deepening confusion and accumulating mistakes. For teachers,

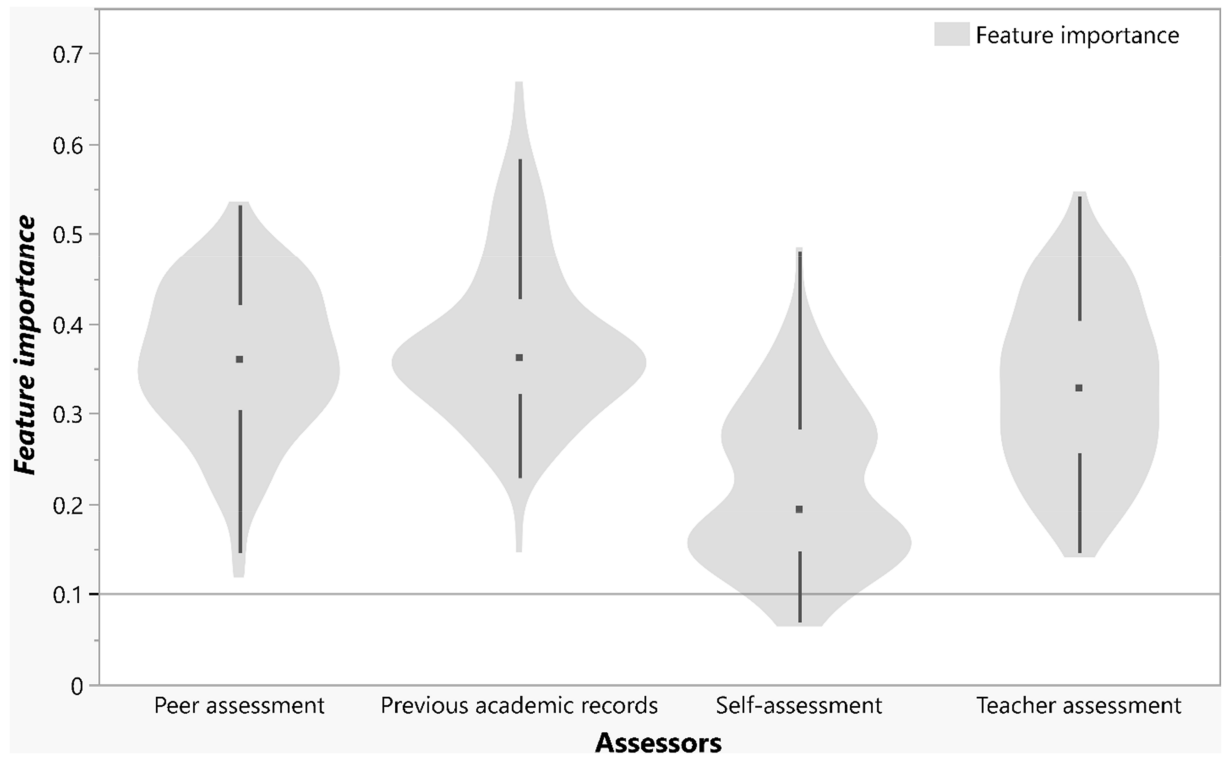


FIGURE 4
Distribution of the feature importance of the assessors in the neural network model. This figure shows the *mean* feature importance values obtained from 100 iterations, and the error bars represent the standard errors of 100 iterations.

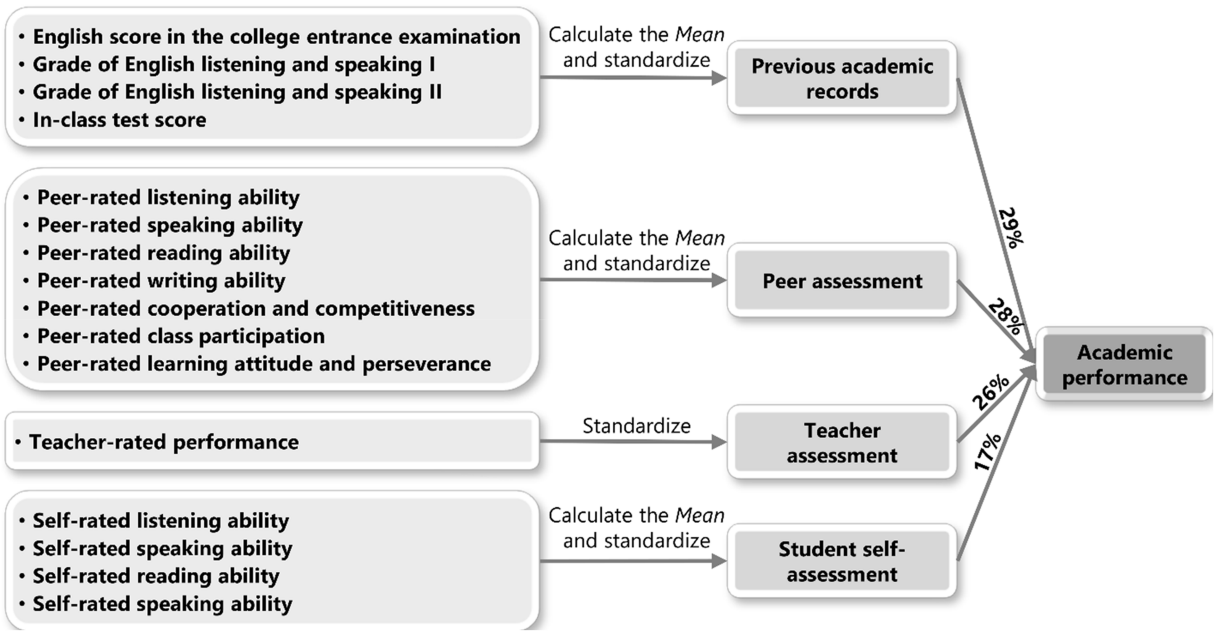


FIGURE 5
Distribution of the feature importance of the assessors in the neural network model. This figure shows the *mean* feature importance values obtained from 100 iterations, and the error bars represent the standard errors of 100 iterations.

peer assessment can to some extent replace teacher assessment, thereby reducing teachers' workload.

Teacher assessment is also an important predictor of academic performance. Students who have positive and supportive relationships with their teachers are more likely to achieve higher levels of success than those with more conflicted relationships (Aultman et al., 2009). Teachers who report interacting with students more frequently may be better equipped to connect their subject matter to students' interests. This, in turn, can help teachers to make the subject matter more relatable and engaging for the students, leading to better learning outcomes (Panadero et al., 2017). Based on one semester of communication, teachers may know students well, so they could successfully predict their performance.

Student self-assessment is also a good assessment index for predicting academic performance (Puustinen and Pulkkinen, 2010; Yan and Carless, 2022; Yan et al., 2023), but in this study, its predictive power was the lowest. This may be because individuals find it difficult to make accurate self-assessments of their abilities, for example, self-assessment of abilities such as humor, grammar, and logical reasoning can be easily influenced by other factors (Ferraro, 2010; Park and Santos-Pinto, 2010). Especially in a culture like China, where interdependence is emphasized, the habit of modesty may lead individuals to show self-depreciation when self-evaluating in order to obtain more social approval (Fay et al., 2012).

It is worth noting that the order of the four common factors in factor analysis and the order of the assessors' feature importance in the neural network model were not consistent. The four common factors extracted by factor analysis were ranked: peer assessment, student self-assessment, previous academic records, and teacher assessment; while the order of the assessors' feature importance in the neural network model was: previous academic records, peer assessment, teacher assessment, and student self-assessment. The reason for this discrepancy is that factor analysis adds up the loadings of the items contained in the common factor and ranks them according to the total amount. The more items, the higher the ranking. However, the neural network model takes the average and standardized score data of each item in the common factor and inputs it into the model for prediction, so the results obtained may be slightly different.

5.3. Data-driven multidimensional assessment model

According to the data-driven multiple assessment model we constructed, the students' final academic performance in the English Listening and Speaking III course in college can be roughly summarized as follows: Academic performance of the Listening and Speaking III = 29% × Previous academic records (standardized average scores of the college entrance examination English test, Listening and Speaking I, Listening and Speaking II, and in-class tests) + 28% × Peer assessment (standardized average scores of peers' assessment of listening, speaking, reading, and writing abilities, class participation, cooperation and competitiveness, and learning attitude and perseverance) + 26% × Teacher assessment (standardized teacher ratings) + 17% × Self-Assessment (standardized average scores of students' self-assessment of listening, speaking, reading, and writing abilities).

This model provides a new solution for course assessment. Practically, teachers can establish their own course assessment methods and assign course scores to students based on the model. Using only the final exam score to evaluate English listening and speaking courses in higher education is not adequate. Learners cannot receive accurate and timely feedback during the learning process, and

teachers cannot provide personalized advice for each student. This is not conducive to language learners' learning. Introducing the theory of multiple assessments into the educational assessment system can promote the theoretical construction and practical development of the assessment system in higher education. Based on the results of this study, we can try to incorporate different assessment subjects into educational assessments, such as allowing students to participate in assessments, and having students themselves and peers rate learners' language abilities, and presentation skills and classroom performance. During the teaching process, teachers should actively collect data on students' previous academic records, self-assessment, peer assessment, and teacher assessment to establish a more comprehensive assessment for each student. Before the final exam, predicting students' learning performance can give more attention to students who may have lower grades, and ensure that each student can achieve satisfactory results in the final exam. Excellent performance in this semester will also have a positive impact on future semesters, forming a virtuous circle.

6. Conclusions and outlook

In this study, factor analysis and neural network models were used to explore the relationships between multiple assessments and academic performance in English listening and speaking courses in higher education. The results showed that factor analysis could sort out assessments from different sources, and the four factors were from the students themselves, their peers, teachers, and previous academic records, respectively. This demonstrated the independence of multiple assessments in practical applications. These four assessors were further incorporated into a predictive model for academic performance, and all of them were found to be important variables for predicting the current academic performance. Therefore, a data-driven multidimensional assessment model for English listening and speaking courses in higher education was constructed. This study actively responded to the demand for interdisciplinary research methods, integrated assessment, teaching, and computer science and technology based on multiple assessment theory, verified the effectiveness of multiple assessments, and provided a reference for the reform of English educational assessment in universities.

However, this study is a preliminary exploration of multiple assessment theory in educational practice, and there are still many shortcomings that need to be addressed through further research. This is mainly reflected in the fact that multiple assessments strive for holistic assessment, emphasizing the diversification of assessment methods, content, subjects, etc. The first limitation is that this study only focused on the diversification of assessment subjects, considering assessments from students, peers, and teachers, but the diversification of assessment methods and contents still requires further research. The second limitation is that due to the small-class setting in China and the avoidance of teacher variances, we only have a small sample size, which may cause the possible lack of statistical power. The third limitation is that we only used the eigenvalue and scree plot to determine the number of factors, which is a poor basis. The forth limitation is that we used a non-refined method to determine the factor scores. It is possible for future studies to refine our proposed method with a larger number of participants.

Moreover, our study focuses on courses that seek to establish a comprehensive assessment system for developing interpersonal abilities, such as speaking or listening skills. For courses that aim to provide fundamental knowledge or skills, such as programming, mathematics, or surgical skills, a comprehensive assessment system may not be urgent.

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 The study was conducted in accordance with the Ocean University of China's policies on Research Ethics for studies involving human participants. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SX: conceptualization and funding acquisition. YJ and HZ: investigation. SX and XX: formal analysis. SX and YS: writing–original draft preparation. SX and SC: writing–review and editing. SC: resources. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Postdoctoral Science Foundation of China [Grant no. 2021M703043], Shandong Provincial Education Science Planning Project [Grant no. 2021WYB005], Shandong Provincial Natural Science Foundation [Grant no. ZR2022QC261], Fundamental Research Funds for the Central Universities of China [Grant no. 202213007], and Chunhui Program of the Ministry of Education of China [Grant no. HZKY20220461].

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.

References

Alpaydin, E. (2016). *Machine learning: The new AI*. Cambridge, MA: MIT Press.

Aultman, L. P., Williams-Johnson, M. R., and Schutz, P. A. (2009). Boundary dilemmas in teacher–student relationships: struggling with "the line.". *Teach. Teach. Educ.* 25, 636–646. doi: 10.1016/J.TATE.2008.10.002

Bailey, K. M. (1996). Working for washback: a review of the washback concept in language testing. *Lang. Test.* 13, 257–279. doi: 10.1177/026553229601300303

Biesanz, J. C. (2012). "Autoregressive longitudinal models" in *Handbook of structural equation modeling*. ed. R. H. Hoyle (New York: The Guilford Press), 459–471.

Bloom, B. S.. (1956). *Taxonomy of educational objectives: Cognitive domain*. Harlow, UK: Longmans.

Boud, D., and Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *High. Educ.* 18, 529–549. doi: 10.1007/BF00138746/METRICS

Brennan, R. L., and Johnson, E. G. (1995). Generalizability of performance assessments. *Educ. Meas. Issues Pract.* 14, 9–12. doi: 10.1111/J.1745-3992.1995.TB00882.X

Brown, G. T. L., Andrade, H. L., and Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assess. Educ.* 22, 444–457. doi: 10.1080/0969594X.2014.996523

Brown, G. T. L., and Harris, L. R. (2013). "Student self-assessment" in *SAGE handbook of research on classroom assessment*. ed. J. H. McMillan (Thousand Oaks, CA: Sage), 367–394.

Brown, S. D., Tramayne, S., Hoxha, D., Telander, K., Fan, X., and Lent, R. W. (2008). Social cognitive predictors of college students' academic performance and persistence: a meta-analytic path analysis. *J. Vocat. Behav.* 72, 298–308. doi: 10.1016/J.JVB.2007.09.003

Cassidy, S. (2007). Assessing 'inexperienced' students' ability to self-assess: exploring links with learning style and academic personal control. *Assess. Eval. High. Educ.* 32, 313–330. doi: 10.1080/02602930600896704

Chang, H., and Lu, J. (2010). Multivariate Statistical Analysis Application in the Evaluation of Student's Synthesis Diathesis. *J. Appl. Stat. Manag.* 29, 754–760. doi: 10.13860/j.cnki.sltj.2010.04.008

Chang, C. C., Tseng, K. H., and Lou, S. J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a web-based portfolio assessment environment for high school students. *Comput. Educ.* 58, 303–320. doi: 10.1016/J.COMPEDU.2011.08.005

Ciarrochi, J., Heaven, P. C. L., and Davies, F. (2007). The impact of hope, self-esteem, and attributional style on adolescents' school grades and emotional well-being: a longitudinal study. *J. Res. Pers.* 41, 1161–1178. doi: 10.1016/J.JRP.2007.02.001

Costello, A. B., and Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Pract. Assess. Res. Eval.* 10:7. doi: 10.7275/jyj1-4868

Cudeck, R., and MacCallum, R. C. (2007). "Factor analysis at 100: historical developments and future directions" in *Factor analysis at 100: Historical developments and future directions*. eds. R. Cudeck and R. C. MacCallum (Mahwah, NJ: Lawrence Erlbaum Associates Publishers)

Double, K. S., McGrane, J. A., and Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: a Meta-analysis of control group studies. *Educ. Psychol. Rev.* 32, 481–509. doi: 10.1007/S10648-019-09510-3/FIGURES/4

Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment: implications for health, education, and the workplace. *Psychol. Sci. Public Interest* 5, 69–106. doi: 10.1111/J.1529-1006.2004.00018.X

Elias, S. M., and MacDonald, S. (2007). Using past performance, proxy efficacy, and academic self-efficacy to predict college performance. *J. Appl. Soc. Psychol.* 37, 2518–2531. doi: 10.1111/J.1559-1816.2007.00268.X

Falchikov, N., and Boud, D. (1989). Student self-assessment in higher education: a meta-analysis. *Rev. Educ. Res.* 59, 395–430. doi: 10.3102/00346543059004395

Falchikov, N., and Goldfinch, J. (2000). Student peer assessment in higher education: a Meta-analysis comparing peer and teacher Marks. *Rev. Educ. Res.* 70:287. doi: 10.2307/1170785

Fancsali, S. E., Zheng, G., Tan, Y., Ritter, S., Berman, S. R., and Galyardt, A. (2018). *Using embedded formative assessment to predict state summative test scores*. ACM International Conference Proceeding Series, pp. 161–170.

Fay, A. J., Jordan, A. H., and Ehrlinger, J. (2012). How social norms promote misleading social feedback and inaccurate self-assessment. *Soc. Personal. Psychol. Compass* 6, 206–216. doi: 10.1111/J.1751-9004.2011.00420.X

Ferraro, P. J. (2010). Know thyself: competence and self-awareness. *Atl. Econ. J.* 38, 183–196. doi: 10.1007/S11293-010-9226-2/TABLES/2

Flake, B. S., Hambleton, R. K., and Jaeger, R. M. (1997). A new standard-setting method for performance assessments: the dominant profile judgment method and some field-test results. *Educ. Psychol. Meas.* 57, 400–411. doi: 10.1177/0013164497057003002

Ghafoori, M., Birjandi, M., and Izadpanah, P. (2021). Self-assessment, peer assessment, teacher assessment and their comparative effect on EFL learners' second language writing strategy development. *J. Engl. Lang. Teach. Learn. Univ. Tabriz* 13, 201–216. doi: 10.22034/ELT.2021.48543.2456

Goldfinch, J., and Raeside, R. (1990). Development of a peer assessment technique for obtaining individual marks on a group project. *Assess. Eval. High. Educ.* 15, 210–231. doi: 10.1080/0260293900150304

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Haywood, H. C., and Lidz, C. S. (2006). *Dynamic assessment in practice: Clinical and educational applications*. Cambridge: Cambridge University Press.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika* 39, 31–36. doi: 10.1007/BF02291575/METRICS

Kim, J.-O., Ahtola, O., Spector, P. E., and Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it*. Thousand Oaks, CA: Sage.

Kim, B. H., Vizitei, E., and Ganapathi, V.. (2018). *GritNet: student performance prediction with deep learning*. Proceedings of the 11th international conference on educational data mining, EDM. Available at: https://arxiv.org/abs/1804.07405v1.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Appears in the International Joint Conference on Arti Cial Intelligence (IJCAI), pp. 1137–1145. Available at: http://robotics.stanford.edu/~ronnyk

Kwan, K. P., and Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assess. Eval. High. Educ.* 21, 205–214. doi: 10.1080/0260293960210301

Lane, S. (2013). "Performance assessment in education" in *APA Handbook of Testing and Assessment in Psychology*. eds. K. F. Geisinger, B. A. Bracken, J. F. Carlson, N. R. Kuncel, S. P. Reise and M. C. Rodriguez (Washington, DC: American Psychological Association), 329–339.

Lane, S., and Stone, C. A. (2006). "Setting performance standards" in *Performance assessment*. ed. R. L. Brennan. *4th* ed (Westport, Connecticut: Praeger), 387–431.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lew, M. D. N., Alwis, W. A. M., and Schmidt, H. G. (2009). Accuracy of students' self-assessment and their beliefs about its utility. *Assess. Eval. High. Educ.* 35, 135–156. doi: 10.1080/02602930802687737

Li, M., Liu, Y., and Zhou, Q. (2016). The analysis of the reliability and characteristics of peer assessment. *E-Educ. Res.* 9, 48–54. doi: 10.13811/j.cnki.eer.2016.09.008

Li, H., Xiong, Y., Hunter, C. V., Guo, X., and Tywoniw, R. (2019). Does peer assessment promote student learning? A meta-analysis. *Assess. Eval. High. Educ.* 45, 193–211. doi: 10.1080/02602938.2019.1620679

Linn, R. L. (1994). Performance assessment: policy promises and technical measurement standards. *Educ. Res.* 23:4. doi: 10.2307/1177043

Maki, P.. (2002). *Using multiple assessment methods to explore student learning and development inside and outside of the classroom*. NASPA's Net Results.

Marsh, H. W., Hau, K. T., Balla, J. R., and Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivar. Behav. Res.* 33, 181–220. doi: 10.1207/S15327906MBR3302_1

Marzouk, M., and Elkadi, M. (2016). Estimating water treatment plants costs using factor analysis and artificial neural networks. *J. Clean. Prod.* 112, 4540–4549. doi: 10.1016/J.JCLEPRO.2015.09.015

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Lang. Test.* 26, 075–100. doi: 10.1177/0265532208097337

Meissel, K., Meyer, F., Yao, E. S., and Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educ.* 65, 48–60. doi: 10.1016/J.TATE.2017.02.021

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educ. Res.* 23:13. doi: 10.2307/1176219

Nefeslioglu, H. A., Gokceoglu, C., and Sonmez, H. (2008). An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. *Eng. Geol.* 97, 171–191. doi: 10.1016/J.ENGGEO.2008.01.004

Nicol, D. (2010). From monologue to dialogue: improving written feedback processes in mass higher education. *Assess. Eval. High. Educ.* 35, 501–517. doi: 10.1080/02602931003786559

Okubo, F., Shimada, A., Yamashita, T., and Ogata, H. (2017). *A neural network approach for students' performance prediction*. ACM International Conference Proceeding Series, 598–599.

Panadero, E. (2016). "Is it safe? Social, interpersonal, and human effects of peer assessment: a review and future directions" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (Abingdon: Routledge), 247–266.

Panadero, E., Jonsson, A., and Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: four meta-analyses. *Educ. Res. Rev.* 22, 74–98. doi: 10.1016/J.EDUREV.2017.08.004

Park, Y. J., and Santos-Pinto, L. (2010). Overconfidence in tournaments: evidence from the field. *Theor. Decis.* 69, 143–166. doi: 10.1007/S11238-010-9200-0/METRICS

Plant, E. A., Ericsson, K. A., Hill, L., and Asberg, K. (2005). Why study time does not predict grade point average across college students: implications of deliberate practice for academic performance. *Contemp. Educ. Psychol.* 30, 96–116. doi: 10.1016/J.CEDPSYCH.2004.06.001

Puustinen, M., and Pulkkinen, L. (2010). Models of self-regulated learning: a review. *Scand. J. Educ. Res.* 45, 269–286. doi: 10.1080/00313830120074206

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). *Cross-validation. Encyclopedia of database systems*. Berlin: Springer.

Rogers, C. R. (1969). *Freedom to learn: A view of what education might become*. New York: CE Merrill.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instr. Sci.* 18, 119–144. doi: 10.1007/BF00117714/METRICS

Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Anal.* 22, 579–590. doi: 10.1111/0272-4332.00040

Shen, B., Bai, B., and Xue, W. (2020). The effects of peer assessment on learner autonomy: an empirical study in a Chinese college English writing class. *Stud. Educ. Eval.* 64:100821. doi: 10.1016/J.STUEDUC.2019.100821

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educ. Res.* 29, 4–14. doi: 10.3102/0013189X029007004/ASSET/0013189X029007004.FP.PNG_V03

Sitthiworachart, J., and Joy, M. (2003). *Web-based peer assessment in learning computer programming. Proceedings 3rd IEEE International Conference on Advanced Learning Technologies, ICALT*, 180–184.

Sternberg, R. J., and Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, UK: Cambridge University Press.

Stone, H., Peet, M., Bhadeshia, H. K. D. H., Withers, P., Babu, S. S., and Specht, E. D. (2008). *Properties of sufficiency and statistical tests*. Proceedings of the Royal Society of London. Series a - mathematical and physical sciences, pp. 1009–1027.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 14, 323–348. doi: 10.1037/a0016973

Suhr, D. D. (2005). *Principal component analysis vs. exploratory factor analysis*. SUGI 30 proceedings.

Suhr, D. D. (2006). *Exploratory or confirmatory factor analysis?* SUGI 31 proceedings.

Suleiman, S., Lawal, A., Usman, U., Usman Gulumbe, S., and Bui Muhammad, A. (2019). Student's academic performance prediction using factor analysis based neural network. *Int. J. Data Sci. Anal.* 5, 61–66. doi: 10.11648/j.ijdsa.20190504.12

Sullivan, K., and Hall, C. (1997). Introducing students to self-assessment. *Assess. Eval. High. Educ.* 22, 289–305. doi: 10.1080/0260293970220303

Sun, Z., Anbarasan, M., and Praveen Kumar, D. (2021). Design of online intelligent English teaching platform based on artificial intelligence techniques. *Comput. Intell.* 37, 1166–1180. doi: 10.1111/COIN.12351

Thanh Pham, T. H., and Renshaw, P. (2015). Formative assessment in Confucian heritage culture classrooms: activity theory analysis of tensions, contradictions and hybrid practices. *Assess. Eval. High. Educ.* 40, 45–59. doi: 10.1080/02602938.2014.886325

To, J., and Panadero, E. (2019). Peer assessment effects on the self-assessment process of first-year undergraduates. *Assess. Eval. High. Educ.* 44, 920–932. doi: 10.1080/02602938.2018.1548559

Topping, K. J., Smith, E. F., Swanson, I., and Elliot, A. (2010). Formative peer assessment of academic writing between postgraduate students. *Assess. Eval. High. Educ.* 25, 149–169. doi: 10.1080/713611428

Tsai, C. C., Lin, S. S. J., and Yuan, S. M. (2002). Developing science activities through a networked peer assessment system. *Comput. Educ.* 38, 241–252. doi: 10.1016/S0360-1315(01)00069-0

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* 49, 1225–1231. doi: 10.1016/S0895-4356(96)00002-9

Tweedie, F. J., and Harald Baayen, R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Comput. Hum.* 32, 323–352. doi: 10.1023/A:1001749303137/METRICS

van Gennip, N. A. E., Segers, M. S. R., and Tillema, H. H. (2009). Peer assessment for learning from a social perspective: the influence of interpersonal variables and structural features. *Educ. Res. Rev.* 4, 41–54. doi: 10.1016/J.EDUREV.2008.11.002

Van Gennip, N. A. E., Segers, M. S. R., and Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learn. Instr.* 20, 280–290. doi: 10.1016/J.LEARNINSTRUC.2009.08.010

Vettori, G., Vezzani, C., Pinto, G., and Bigozzi, L. (2020). The predictive role of prior achievements and conceptions of learning in university success: evidence from a retrospective longitudinal study in the Italian context. *High. Educ. Res. Dev.* 40, 1564–1577. doi: 10.1080/07294360.2020.1817875

Were, K., Bui, D. T., Dick, Ø. B., and Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol. Indic.* 52, 394–403. doi: 10.1016/J.ECOLIND.2014.12.028

Xie, X., and Guo, J. (2022). Influence of teacher-and-peer support on positive academic emotions in EFL learning: the mediating role of mindfulness. *Asia Pac. Educ. Res.* 2:665. doi: 10.1007/s40299-022-00665-2

Yan, Z., and Carless, D. (2022). Self-assessment is about more than self: the enabling role of feedback literacy. *Assess. Eval. High. Educ.* 47, 1116–1128. doi: 10.1080/02602938.2021.2001431

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., and Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: a meta-analysis. *Educ. Res. Rev.* 37:100484. doi: 10.1016/J.EDUREV.2022.100484

Yan, Z., Wang, X., Boud, D., and Lao, H. (2023). The effect of self-assessment on academic performance and the role of explicitness: a meta-analysis. *Assess. Eval. High. Educ.* 48, 1–15. doi: 10.1080/02602938.2021.2012644

Zheng, G., Fancsali, S. E., Ritter, S., and Berman, S. R. (2019). Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *J. Learn. Anal.* 6, 153–174. doi: 10.18608/jla.2019.62.11

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 16 June 2023
DOI 10.3389/feduc.2023.1198011

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Ana Remesal,
University of Barcelona, Spain
Serafina Pastore,
University of Bari Aldo Moro, Italy

*CORRESPONDENCE
Morgane Senden
✉ morgane.senden@uclouvain.be

RECEIVED 31 March 2023
ACCEPTED 25 May 2023
PUBLISHED 16 June 2023

CITATION
Senden M, De Jaeger D and Coertjens L (2023) Safe and sound: examining the effect of a training targeting psychological safety and trust in peer assessment.
Front. Educ. 8:1198011.
doi: 10.3389/feduc.2023.1198011

COPYRIGHT
© 2023 Senden, De Jaeger and Coertjens. 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.

Safe and sound: examining the effect of a training targeting psychological safety and trust in peer assessment

Morgane Senden[1]*, Dominique De Jaeger[2] and Liesje Coertjens[1]

[1]Psychological Sciences Research Institute (IPSY), Université Catholique de Louvain, Louvain-la-Neuve, Belgium, [2]Faculty of Motor Sciences, Institute of Neuroscience (IoNS), Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Introduction: Peer feedback can be very beneficial for student learning in higher education, yet students may feel uncomfortable providing and receiving peer feedback: they may for example not feel safe in the group or have little trust in their peers' abilities to provide feedback. Surprisingly, only few studies have investigated how students' feelings of discomfort can be reduced. To fill this gap, we created a 1-h training session using active learning methods. The training focuses on enhancing students' perceptions of psychological safety, trust in their abilities and in their peers' abilities to provide feedback.

Methods: The efficacy of this training was tested using a quasi-experiment with pre- and post-test design. Third-year bachelor students in physical education participated in a peer feedback activity to fulfill the requirement of an obligatory course. In 2019–2020, 47 students participated in a peer assessment activity without specific training on psychological safety and trust (control group), while in 2021–2022, 42 students received specific training before peer assessment (experimental group).

Results: Analyses include a comparison of the control and experimental groups with regard to (1) the evolution of their perceptions (psychological safety, trust in their abilities, and trust in their peers' abilities) for pre-to post-test, (2) the quality of the feedback they provided to their peers (3) and the improvement of students' work between the draft submitted for the peer activity and the final version submitted to the professor.

Discussion: Results do not support the training's efficacy, yet suggest pathways for future research.

KEYWORDS

peer assessment, psychological safety, trust, training, higher education, interpersonal factors, quasi-experiment, intervention

1. Introduction

Peer assessment is a practice that can contribute to achieving various goals of higher education. In addition to being an important tool for higher education students' learning, the use of peer assessment is in line with a more participatory culture of learning (Kollar and Fischer, 2010). Involving students in the assessment pushes them to take responsibility for their learning and to develop important skills such as self-regulatory practices (Carless et al., 2011; Planas Lladó et al., 2014). These practices are necessary to attain an important goal of higher education: enable students to become independent lifelong learners (Tai et al., 2018).

Topping (1998, p. 250) defined peer assessment as "an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status." Several recent meta-analyses have shown that peer assessment not only has a positive impact on learning compared to the absence of any sort of assessment but also that it impacts student learning more positively than teacher assessment does (Double et al., 2020; Li et al., 2020; Wisniewski et al., 2020).

While peer assessment is an activity with a lot of potential for students' learning, it is also an activity full of complexities, which involves interpersonal and intrapersonal factors (Panadero et al., 2023). Even though most students perceive the educational value of peer assessment (Mulder et al., 2014), they also express a series of concerns. Students may feel uncomfortable in both the assessee and the assessor's roles. As assessor, they may not feel competent or they may find it difficult to be objective, while as an assessee they may fear their peers will be biased or will judge them (Hanrahan and Isaacs, 2001; Mulder et al., 2014; Wilson et al., 2015).

Although it is known that interpersonal and intrapersonal factors involved in peer assessment are important to consider because they can affect peer assessment activities (Panadero et al., 2023), few studies have investigated if an intervention could positively affect them. Moreover, it is not clear if the impact of interpersonal and intrapersonal factors on peer assessment outcomes occurs through an impact on the quality of received and provided feedback. To fill this gap, this study aims to investigate if a training targeting psychological safety and trust can enhance students' perceptions, incites them to provide more elaborated feedback, and supports them to benefit more from the peer assessment. In the next sections, we will define the three factors this study focuses on, and detail research findings on their role in peer assessment and on how we could intervene.

1.1. The impact of psychological safety and trust in peer assessment

Trust in the self as an assessor, trust in the other as an assessor, and psychological safety will be the focus of this study. These three factors are strong predictors of peer assessment outcomes (van Gennip et al., 2010).

1.1.1. Psychological safety

The notion of psychological safety originally came from organizational psychology (Edmondson and Lei, 2014). In the working environment, where most of the research on psychological safety has been conducted, an emphasis is placed on the fact that psychological safety enables employees, teams, and organizations to learn (Edmondson and Lei, 2014). Therefore, it seems logical that the concept of psychological safety was subsequently studied in educational contexts. Psychological safety creates an environment where students feel comfortable discussing their performance and errors and asking for feedback which has a positive impact on their learning and their grades (Soares and Lopes, 2020). From there it becomes clear why psychological safety is a requirement for peer assessment. In the context of peer assessment, psychological safety is defined as "the extent to which students feel safe to give sincere feedback as an assessor and do not fear receiving inappropriate negative feedback or to respond to negative feedback" (Panadero et al., 2023, p. 5).

Students' perceptions of psychological safety are a direct predictor of students' conceptions of peer assessment and an indirect predictor of students' perceived learning (van Gennip et al., 2010). Moreover, students with higher perceptions of psychological safety adopt deeper learning approaches in peer assessment and hold more cohesive conceptions of learning, which mean that they do not only see learning as an accumulation of knowledge but as an association of new content with prior knowledge (Cheng and Tsai, 2012). However, psychological safety is not associated with the perceived educational value of peer assessment (Rotsaert et al., 2017).

1.1.2. Trust in the self as an assessor and trust in the other as an assessor

Trust in the self as an assessor is defined as the "belief about the ability to perform peer assessment as assessor," while trust in the other as an assessor is defined as the "confidence in a peer's capability to perform a fair and/or accurate peer assessment" (Panadero et al., 2023, p. 5). A student can trust his or her ability but not his or her peers' ability, or vice versa. It seems that most students tend to trust their abilities to assess their peers, while a smaller percentage of students also trusted their peers' abilities (Cheng and Tsai, 2012).

Like psychological safety, trust in the self as an assessor and trust in the other as an assessor both positively predict students' conceptions of peer assessment (van Gennip et al., 2010) and are associated with deeper learning approaches (Cheng and Tsai, 2012). However, only trust in the self is associated with more cohesive conceptions of learning. Moreover, students' perceptions of trust in the self as an assessor and trust in the other as an assessor both affect their implication in peer assessment (Zou et al., 2018) and are associated with a higher perceived educational value of peer assessment (Rotsaert et al., 2017).

1.2. Interventions targeting psychological safety and trust

The relationships between levels of trust and psychological safety and peer assessment processes or outcomes, even though only correlational, highlight the important role that these perceptions can play. This raises the question of how to ensure that students feel psychologically safe and have trust in their peers and the self when participating in peer assessment.

The main suggestion in literature to overcome students' concerns related to peer assessment is to use anonymity. Indeed, evidence suggests that students appreciate anonymity. They feel more comfortable during anonymous peer assessment, both as an assessor and as an assessee (Su, 2022), and they would rather be anonymous or use a nickname than use their real name (Yu and Liu, 2009). Moreover, anonymity can have positive impacts on students' perceptions, such as reducing peer pressure and fear of disapproval (Vanderhoven et al., 2015). Although the use of anonymity in peer assessment has some positive effects, to the best of our knowledge, no study shows that anonymity has a positive impact on psychological safety and trust specifically. Studies have either found an absence of impact (e.g., Rotsaert et al., 2018) or even a negative one: a study found that pupils viewed their assessors more positively (which includes a higher trust in them) when real names were used in the peer assessment than when assessors were anonymous or used nicknames (Yu and Wu,

2011). According to Panadero and Alqassab's (2019) review of the effects of anonymity in peer assessment, anonymity seems to positively affect students' perceptions related to the peer assessment activity, but negatively affects their perceptions related to interpersonal factors, such as trust and psychological safety.

To the best of our knowledge, besides anonymity, the only existing intervention that aims to increase students' perceptions of psychological safety and trust before or during peer assessment is the use of role-playing described by Ching and Hsu (2016). In their study, students had to choose a stakeholder's role during peer assessment and provided feedback to their peers from this stakeholder's perspective. Students reported a high level of psychological safety and trust during this kind of peer assessment and they provided feedback of high quality. However, as there were no control group or pre-test measures, it remains up to present unclear if students' level of trust and psychological safety would have been lower without the role-play. Moreover, psychological safety and trust were measured with different types of items than in other studies (e.g., van Gennip et al., 2010; Rotsaert et al., 2018), and may not reflect the same concepts. In addition, the peers' work that had to be assessed in Ching and Hsu (2016)'s study is well suited to be assessed while adopting a stakeholder's role, but for other types of student productions (e.g., mathematical problem solving), it could be less adapted and the use of role-play might not work in these cases. If role-play is not possible or relevant during peer assessment, it could be included in a training session that would take place before peer assessment.

There is evidence that including a training session is an effective way to ease some tensions related to peer assessment. Students who followed a training that aimed to control the peer pressure they can feel during a peer feedback activity saw more value in the peer assessment and experienced less peer pressure than the students who did not follow it (Li, 2017). To the best of our knowledge, no training has ever been designed to enhance students' perceptions of psychological safety or trust in the context of peer assessment. For psychological safety, which is important in various situations, inspiration can be found beyond this specific context. A study in the context of group work found no effect of a 50-min training session that aimed to increase students' perceptions of psychological safety before working in small groups (Dusenberry and Robinson, 2020). This absence of effect could be explained by two factors: the training was not context-specific and the consisted mostly of non-active learning methods. By overcoming these limits, effective training for psychological safety could be developed.

A training enhancing perceptions of trust and psychological safety could also positively affect students' learning and performances. Indeed, in Li (2017)'s study, students who had participated in the training obtained higher scores for their revised work compared to students who did not receive the training. The mechanisms by which students' perceptions of psychological safety and trust could affect learning gains following peer assessment are unclear, due to a lack of research linking interpersonal and intrapersonal factors and peer assessment learning outcomes (Panadero et al., 2023).

1.3. The present study

This study aims to investigate if students' perceptions of psychological safety, trust in the self as an assessor, and trust in the other as an assessor can be increased through a training targeting these factors. To this end, we designed a 1-h training session and provided it to a group of students before they participated in a peer assessment activity. This cohort of students was compared to a cohort of students who did not receive the training. We assumed that providing a training on these aspects would increase students' perceptions of psychological safety, trust in the self as an assessor, and trust in the other as an assessor.

Our first hypothesis is therefore:

> *H1*: Perceptions of psychological safety, trust in the self as an assessor, and trust in the other as an assessor increase more in the experimental group than in the control group.

Another aim of this study is to investigate the relationship between students' perceptions of psychological safety and trust and their learning gains following peer assessment. Indeed, there is a lack of research on how students' perceptions can hinder or facilitate learning through peer assessment (Panadero et al., 2023). It is known that intrapersonal and interpersonal factors can affect students' engagement in peer feedback (Zou et al., 2018), but not how this engagement will affect students' learning gains. Prior research has found that there is a relationship between the quality of feedback students provide and the improvement in their final work following peer assessment (Li et al., 2010). It could mean that an increase in trust and psychological safety would affect students' implication in the peer assessment, which would be reflected by the quality of feedback provided during the peer assessment and by an improvement in performance due to the peer assessment. Thus, our second and third hypotheses are the following:

> *H2*: Students in the experimental group provide feedback of higher quality to their peers than students in the control group.
>
> *H3*: Students' grades improve more after the peer assessment activity in the experimental than in the control group.

2. Materials and methods

2.1. Study design and participants

To test the effect of the training on psychological safety and trust, we designed an intervention study, using a quasi-experiment with pre and post-test design. The first cohort of students followed the course in 2019–2020. These students were the control group; they participated in the peer assessment as it was implemented in the course, without any modification. The second cohort of students followed the course in 2021–2022. These students were the experimental group; before participating in the peer assessment, they received specific training on psychological safety and trust.

The participants were third-year physical education students enrolled in a course on acrobatic sports didactics at a French-speaking Belgian university. In physical education, an important part of courses is practical, and students are used to (1) being active, (2) interacting with each other, and (3) receiving formative feedback from peers and professors. Fifty-one students followed the course in 2019–2020, and 56 students followed it in 2021–2022. Of these 107 students, 100 (94%)

agreed to participate in the study, but among them, 11 were absent during the training session or did not participate in the peer assessment activity (both mandatory for the course). Therefore, the sample consisted of 89 students, 47 in the first cohort (control group) and 42 in the second cohort (experimental group). Sixty-nine percent of the participants were men. On average, participating students were 21.5 years old (SD = 1.79). Most participants (89%) indicated having already participated in peer assessment before, yet the professor of the course confirmed that students have not received training prior to this. There was no difference in terms of gender, age, or prior experiences with peer assessment between the control and the experimental groups.

2.2. Procedure and intervention

The study took place in the context of a course on the didactics of acrobatic sports. This course was built on prerequisites that students have acquired in previous courses (e.g., gymnastics, biomechanics), but students were novices in didactics. In parallel, they followed courses on other sports' didactics. The course aimed to teach students how to develop relevant learning situations in acrobatic sports. To this end, students had to create an instruction sheet describing a learning exercise. This sheet consisted of a diagram illustrating the exercise, one or two instructions explaining how to perform it, and one or two criteria for success (an example is provided in Supplementary material). They also had to create a video of the exercise.

After completing the first draft of their sheet, students participated in the peer assessment activity on Moodle. Using a rubric created by the professor (second author), each student assessed the instruction sheet of seven of their peers (randomly assigned). The rubric is available in Supplementary material. For each criterion (diagram, video, instructions, success criteria, and overall coherence), students chose a level in the rubric and wrote qualitative feedback. After the peer assessment activity, students could improve their sheet before submitting it for the final assessment by the course's professor. A description of all design elements of the peer assessment is available as Supplementary material using Panadero et al. (2023)'s reporting instrument.

As a course requirement, students from both groups participated in a 1-h session that prepared them for the peer assessment: the procedure of the peer assessment activity was detailed, they learned how to give effective feedback based on Hattie and Timperley (2007)'s feedback model, and they had the opportunity to practice in class. Directly after this session, a researcher (first author) provided the training to the experimental group. This 1-h training aimed to increase students' perceptions of psychological safety, trust in the self as an assessor, and trust in the other as an assessor.

The training included three stages (see Figure 1). The first stage took place a week before the training session. To discover students' thoughts on peer assessment and adapt the training based on it, we asked them to answer a few questions (such as "What are the disadvantages of peer assessment?") using an interactive platform. The second stage included role-plays and discussions, two methods recommended to transform attitudes (Blanchard and Thacker, 2012). For the role-plays, students were split into groups of six and received a detailed roadmap with two role-play scenarios and instructions on how to play these scenarios and how to discuss them afterward. After performing and discussing the two role-plays, they stayed in sub-groups to summarize their discussions. The third and final stage was an open discussion with the entire class to summarize all the ideas. Based on the discussion, we created a mind map that students received a few days later. A detailed presentation of the training is available in the Supplementary material. The presence of a training session for the experimental group was the only difference between the groups, otherwise, the procedure was identical (see Figure 2). Students filled out questionnaires about their perceptions (pre-test and post-test) during the first and the last sessions of the course.

2.3. Measurements

2.3.1. Survey data

Students' perceptions (trust in the self as an assessor, trust in the other as an assessor, psychological safety, and the importance of anonymity) were measured with a French translation of the scales of Rotsaert et al. (2018). These scales have been shown to be reliable and their items operationalize well the concepts as defined in the introduction of this study. Items were measured using a seven-point Likert scale ranging from 1 (totally disagree) to 7 (totally agree). Macdonald's ω suggest good reliability (see Table 1).

2.3.2. Students' feedback

To operationalize the quality of feedback, we used the distinction between verification and elaboration feedback (Shute, 2008). Verification feedback only tells students if their productions are correct or not, while elaboration feedback contains additional information to help students to arrive at the correct answer. Effective feedback contains both verification and elaboration (Shute, 2008).

To compare feedback provided by students from the control group to those provided by students from the experimental group, feedback from all students was coded. First, we divided each feedback into units of analysis following Strijbos et al. (2006)'s procedure. This resulted in a database of 2,874 units of meaning. Each unit of meaning was then coded according to the criterion it refers to (examples are provided in Table 2) and to the feedback style (examples are provided in Table 3). To code feedback according to their style (verification or elaboration), we relied on the coding scheme developed by Alqassab et al. (2018) that we adapted to the context of the present study. According to this scheme, elaboration feedback can be one of five types: correction, confirmation, justification, question, or suggestion. Given the very small number of meaning units belonging to the "confirmation" category, this category was omitted. Some units of meaning did not fit into any of these categories (e.g., "Good luck for the next step!") and were classified as "other."

To test inter-rater agreement, a second coder coded a random subsample of 287 units of meaning (i.e., 10% of the total). As the verification and correction feedback were more frequent than the other types, resulting in unbalanced marginals, the Gwet's AC_1 is more reliable than the Cohen's κ (Black et al., 2016). The Gwet's AC_1 indicated a moderate to substantial agreement [Gwet' $AC_1 = 0.61$ (95% CI, 0.546–0.674), $p < 0.001$] (Gwet, 2014).

2.3.3. Students' grades

To investigate the effect of the training session on students' performance, the course professor assessed both the first draft and the

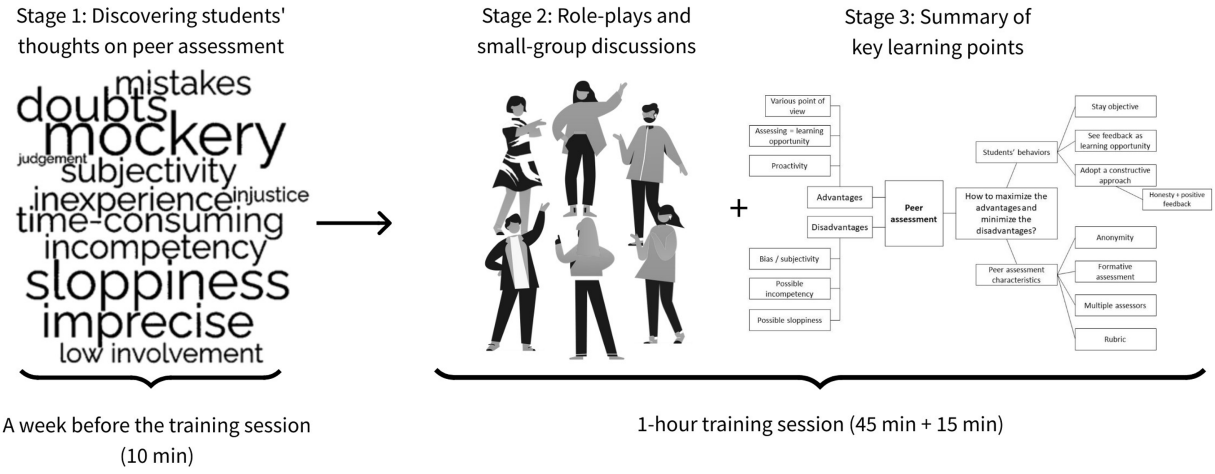


FIGURE 1
Training session.

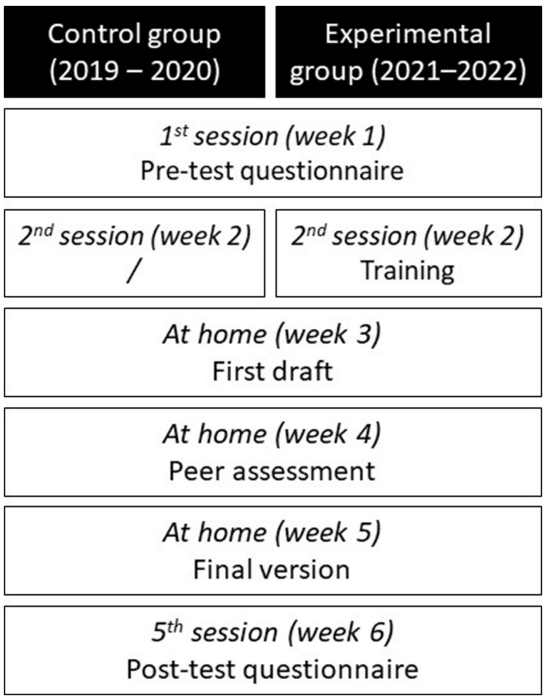


FIGURE 2
Study timeline.

final version of students' instruction sheets. The first draft was assessed for research purposes only, students did not receive their grades and it did not affect whether they passed the course or not. The criteria provided to the students were slightly different from those that the professor used, in order to guide them more in the peer assessment. For example, the video, which was only there to help the global comprehension of the exercise, was in students' criteria to ensure they would watch it. The criteria used by the professor to grade both versions were the following: "diagram," "relevance of the situation," "instructions," and "success criteria." Each criterion was scored out of 5, giving an overall score out of 20.

2.4. Data analysis

2.4.1. Survey data

There were some missing data for the questionnaire. In each cohort, some students were absent (e.g., due to illness) at the pre or post-test. We created an online version of the questionnaire and shared the link with absent students, but nevertheless, there was 17% of missing data at the pre-test and 7% at the post-test. There were no item-level missing data.

To handle these missing data, we used multiple imputations, which is considered the gold standard method when data are missing at random (Enders, 2010). The variables used for the imputations were the age, gender, and group (experimental or control), as well as the perceptions available (if the data were missing for the pre-test, we used post-test data and inversely). The analyses were conducted in R (version 4.1.0) with the "mice" package. Following von Hippel's (2020) two-stage calculation, we calculated that we needed at least 22 imputations. Therefore, we did 30 imputations. We used Rubin's rules (Rubin, 1987) to pool the results.

Once the data were imputed and pooled, linear mixed-effects models were conducted to test our hypothesis according to which an interaction effect would be present, with a higher increase in the experimental group.

2.4.2. Students' feedback

To compare the proportion of feedback provided by students in the control group and the experimental group, we calculated a score between 0 and 1 for each student regarding each type of feedback. This score was calculated by dividing the number of feedback elements provided by the total number of feedback units that could have been provided. A score of 0 means that the student did not provide any feedback of this type, while a score of 1 means that the student provided feedback of this type at every opportunity. For example, a student who assessed seven peers on six criteria and provided 34 verification feedback had a score of 0.81 (34/42) for verification feedback. To investigate the impact of the training on the type of feedback given, Student's *t*-tests and Mann–Whitney's *U* tests were conducted on Jamovi (version 2.2.5).

TABLE 1 Reliability of survey scales.

Scales	Number of items	Macdonald's ω			
		Control group		Experimental group	
		Pre-test	Post-test	Pre-test	Post-test
Psychological safety Example item: In this group, I can share my opinion without hesitation	4	0.91	0.91	0.86	0.90
Trust in the self as an assessor Example item: I feel I am able to assess my peers	7	0.92	0.85	0.90	0.93
Trust in the other as an assessor Example item: I think my peers are able to give me objective feedback	6	0.69	0.85	0.88	0.82

TABLE 2 Example of feedback for each criterion.

Criterion	Example
Diagram	For the shape of the diagram, there is just the fact that one pushes on the palm of the hand and not on the fingers
Video	Properly executed movement
Overall coherence	Overall coherence: Good, nothing to say, simple and clear
Instructions	The instructions are very clear and precise, we know at what moment to push the shoulders and to pull the heels
Success criteria	Another criterion for success that you could have used instead is to reach far with your hands on the jumping table
Global	Great work!

TABLE 3 Example of feedback for each type.

Type	Example
Verification	Success criteria: very good
Correction	I think it would be more correct if the arms (Figures 2, 3) and legs (Figure 3) were aligned
Justification	[To say "maintain muscular tension" is not an instruction] in the sense that in almost all exercises you have to maintain muscular tension
Question	Where do we need to jump? And how?
Suggestion	I would have written for example: "do not exceed 1/3 of the big mattress"

2.4.3. Students' grades

Given that students' grades did not follow a normal distribution, we conducted Wilcoxon tests to investigate if there was an improvement in students' performance. These analyses were conducted on Jamovi (version 2.2.5). Then, to test if the improvement was more important in the experimental group than in the control group, we calculated difference scores between the two versions (for the global version and each criterion) and conducted Mann–Whitney's *U* tests to compare the difference scores between the control and experimental groups.

Spearman correlations between all study variables are presented in Table 4.

3. Results

3.1. Impact of the training on students' perceptions

The mean scores and standard deviation of students' perceptions of psychological safety, trust in the self as an assessor, and trust in the other as an assessor are display in Table 5.

The level of psychological safety, trust in the self as an assessor, and trust in the other as an assessor were high at the pre-test. It means that students already felt safe and had positive perceptions of their own assessment abilities and those of peers before the training and the peer assessment activity.

3.1.1. Psychological safety

For psychological safety, the results show an absence of a significant effect. Contrary to our hypothesis, there was no significant interaction effect [$t(1432.829) = 1.181$, $p = 0.238$]. Moreover, the two groups were comparable [$t(14187.907) = -0.957$, $p = 0.338$] and, in both groups, the levels of psychological safety were not altered significantly from pre-test to post-test [$t(362.903) = 1.161$, $p = 0.246$].

3.1.2. Trust in the self as an assessor

Contrary to our hypothesis, there was no significant interaction effect for trust in the self as an assessor [$t(1377.802) = -1.164$, $p = 0.245$]. It is important to note that the two groups were not comparable. Levels of trust in the self as an assessor were higher in the experimental than in the control group [$t(8621.171) = -2.054$, $p = 0.040$]. In both groups, levels of trust in the self as an assessor did not change statistically from pre-test to post-test [$t(359.655) = 1.327$, $p = 0.185$].

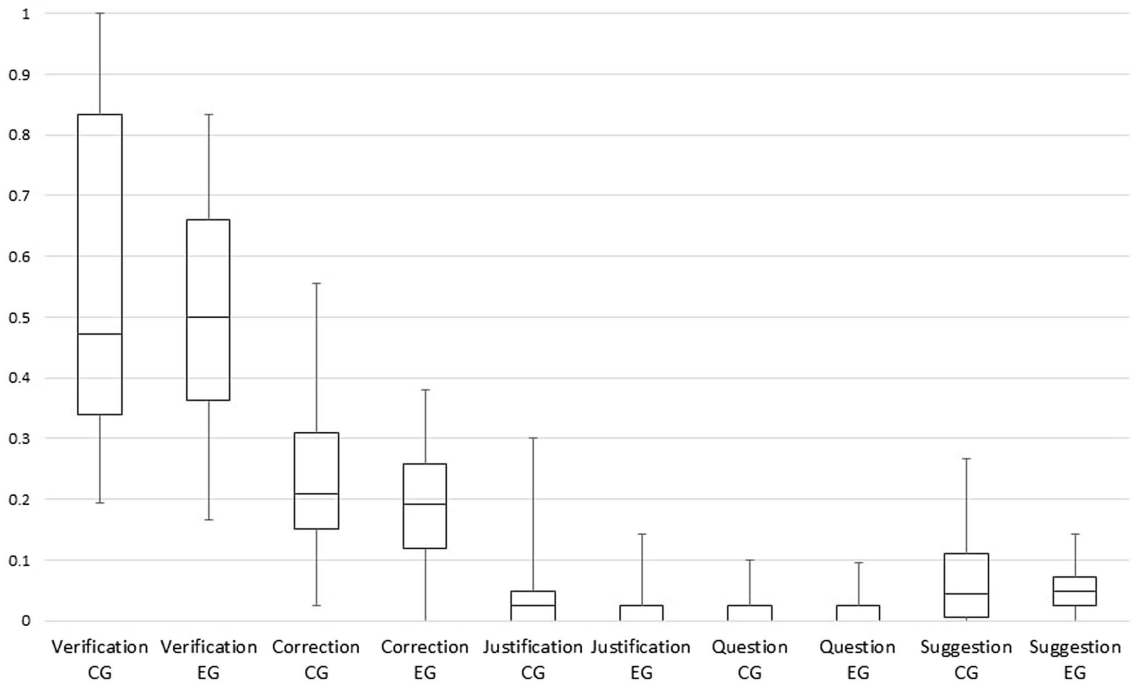


FIGURE 3
Box-plots of students' scores for each feedback type in the control group and the experimental group. CG, Control Group; EG, Experimental Group.

3.1.3. Trust in the other as an assessor

There was no significant interaction effect for trust in the other as an assessor [$t(980.787) = 1.282$, $p = 0.200$]. As for trust in the self as an assessor, the two groups were not comparable, but the difference was in the other direction: levels of trust in the other as an assessor were higher in the control than in the experimental group [$t(867.910) = -5.997$, $p < 0.001$]. In both groups, levels of trust in the other as an assessor did not change statistically from pre-test to post-test [$t(798.583) = -0.422$, $p = 0.673$].

3.2. Impact of the training on the feedback provided by students

As shown in Figure 3, the pattern of provided feedback is rather similar for both groups. Verification feedback is frequently provided, but with important variability among students. Correction feedback is the second most frequent feedback, again with quite some variability. The other types of feedback (justification, question, and suggestion) are rarely provided.

Scores for verification, justification, question, and suggestion did not follow a normal distribution. Therefore, some Mann–Whitney U tests were conducted to test for differences between the two groups. There was no significant effect for verification feedback ($U = 1{,}030$, $p = 0.472$), question feedback ($U = 1{,}072$, $p = 0.632$), and suggestion feedback ($U = 963$, $p = 0.218$). However, there was a significant effect for justification feedback ($U = 840$, $p = 0.020$), but this effect was in the opposite direction compared to our initial hypotheses: students in the control group provided more feedback of this type than students in the experimental group did. A t-test was conducted for correction feedback given that, for this type of feedback, the scores followed a normal distribution. The results were similar to those for justification feedback: students in the control group provided more correction feedback than students in the experimental group [$t(93) = 2.13$; $p = 0.036$].

3.3. Impact of the training on students' grades

First, the impact of the training on students' performance was investigated with the global grade. Figure 4 illustrates the evolution of students' grades for both groups.

We started by testing if, for both groups taken together, there was an improvement in performance using a Wilcoxon test. It was the case ($W = 546$, $p < 0.001$, $r = -0.491$), which confirmed that students' work in both groups improved after the peer assessment.

A Mann–Whitney U test was conducted to investigate if the improvement was more important in the experimental than in the control group. The effect was not significant ($U = 915$, $p = 0.548$), implying that we could not show that the grade improvement was different in the experimental ($M = 0.29$, SD = 0.74) and in the control group ($M = 0.40$, SD = 1.03).

Observing Figure 4, it seems like students from the control group performed better than those in the experimental group. Mann–Whitney's U tests were conducted to check if these differences were significant. Students from the control group received a higher grade for their first draft ($U = 588$, $p < 0.001$) and for their final version ($U = 559$, $p < 0.001$).

Secondly, the analyses were conducted for each criterion. Wilcoxon tests' results indicated that there was a significant grade improvement for the criteria "diagram" ($W = 347$, $p = 0.003$) and "relevance of the situation" ($W = 70$, $p = 0.006$), but not for the criteria "instructions" ($W = 365$, $p = 0.167$) and "success criteria" ($W = 354$, $p = 0.599$). The

10.3389/feduc.2023.1198011

TABLE 4 Correlation matrix.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Psychological safety pre-test															
2. Psychological safety post-test	0.76**														
3. Trust self pre-test	0.40**	0.37**													
4. Trust self post-test	0.53**	0.68**	0.63**												
5. Trust other pre-test	0.12	0.05	−0.06	−0.2											
6. Trust other post-test	0.29*	0.18	−0.10	0.13	0.65**										
7. Improvement global grade	−0.09	−0.18	−0.02	−0.12	0.06	−0.12									
8. Improvement grade "diagram"	−0.18	−0.04	−0.08	0.00	−0.08	−0.15	0.67**								
9. Improvement grade "relevance of the situation"	0.04	0.03	0.12	0.09	0.40**	0.20	0.19	0.01							
10. Improvement grade "instructions"	−0.05	−0.16	−0.12	−0.16	−0.04	−0.26	0.51**	0.15	−0.11						
11. Improvement grade "success criteria"	0.13	−0.12	0.13	−0.10	0.06	0.01	0.53**	0.01	0.04	0.11					
12. Verification feedback	0.00	−0.05	−0.05	−0.08	−0.03	0.04	−0.08	−0.04	−0.11	0.02	−0.14				
13. Correction feedback	−0.03	−0.07	−0.11	−0.15	−0.10	0.00	−0.06	−0.01	−0.19	−0.01	−0.03	0.57**			
14. Justification feedback	0.05	0.03	0.00	−0.02	0.18	0.24*	−0.02	0.03	−0.02	−0.06	0.04	0.36**	0.29**		
15. Question feedback	−0.12	0.03	0.02	0.10	−0.11	0.13	−0.24*	−0.14	−0.15	−0.07	−0.20	0.10	0.21*	0.11	
16. Suggestion feedback	0.04	0.00	−0.14	−0.12	−0.02	0.18	0.03	0.09	−0.17	0.09	0.02	0.30**	0.19	0.37**	0.19

*$p < 0.05$; **$p < 0.01$.

difference score between the two versions was calculated for the two criteria with a significant improvement and Mann–Whitney *U* tests were conducted to investigate if the improvement was different in both groups (see Table 6). There is only a significant difference for the criterion "relevance of the situation." This difference is in the opposite direction of our hypothesis: students in the control group improved more than students in the experimental group did.

4. Discussion

4.1. General discussion

As a reminder, this study aims to investigate if a training on psychological safety and trust can reduce students' concerns regarding peer assessment. Using an intervention design, a 1-h training session was provided to a group of students ($N = 42$), while a previous cohort served as a control group ($N = 47$). We examined whether the training session affected (1) students' perceptions of psychological safety, trust in the self as an assessor, and trust in the other as an assessor, (2) the type of feedback they provided to their peers, and (3) their performance improvement following the peer assessment.

Our first hypothesis was that students' perceptions of psychological safety, of trust in the self as an assessor, and of trust in the other as an assessor would increase more in the experimental than in the control group. The results did not support this hypothesis; in both groups, levels of students' perceptions did not evolve significantly from pre-test to post-test. Our training did not have a visible impact on students' perceptions. This result is not coherent with another

TABLE 5 Means and standard deviations of students' perceptions.

	Control group		Experimental group	
	Pre-test	Post-test	Pre-test	Post-test
Psychological safety	5.24 (0.92)	5.44 (1.00)	5 (1.22)	5.42 (1.14)
Trust in the self as an assessor	4.96 (0.68)	5.17 (0.78)	5.40 (0.85)	5.35 (0.93)
Trust in the other as an assessor	5 (0.62)	4.99 (0.85)	4.07 (1.04)	4.29 (1.09)

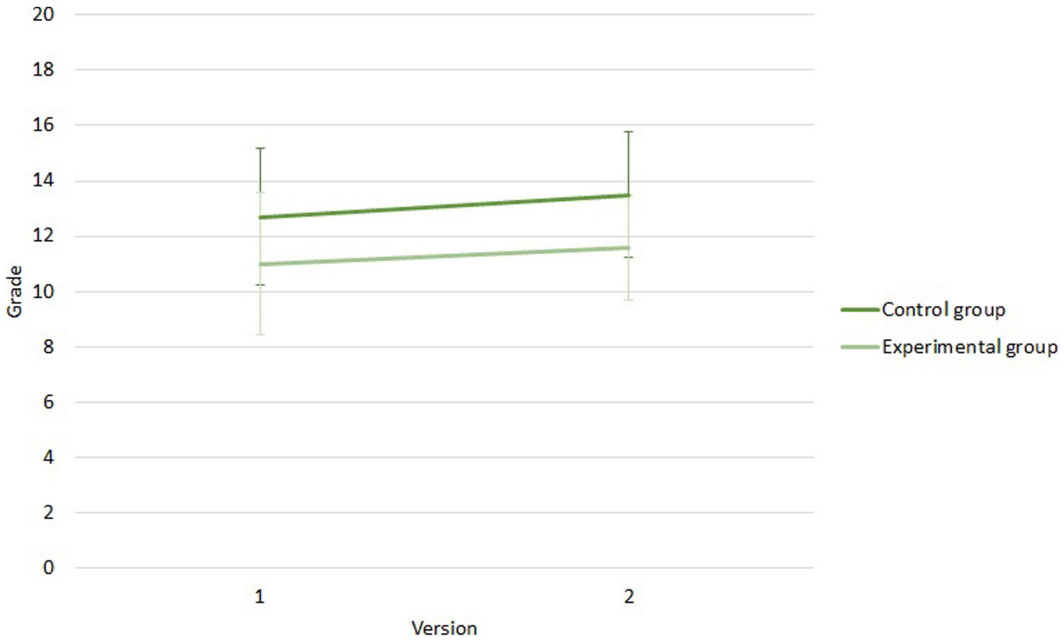


FIGURE 4
Evolution of students' grades (on/20).

TABLE 6 Means, standard deviations, and Mann–Whitney *U* tests' results for difference scores.

	M (SD) in the control group	*M* (SD) in the experimental group	Mann–Whitney U	Sig.	Effect size (*r*)
Difference score for the global grade	0.80 (2.07)	0.57 (1.48)	915	0.548	0.07
Difference score for "diagram"	0.42 (1.00)	0.24 (0.98)	873	0.324	0.12
Difference score for "relevance of the situation"	0.31 (0.66)	0.02 (0.35)	708	0.004	0.28

study that also aimed at making students feel at ease before peer assessment but targeting perceptions of peer pressure (Li, 2017).

This absence of effect may be explained by students' positive perceptions at the pre-test; mean scores were around five on a 7-point scale. Two factors could explain these positive perceptions. First, the majority of students in the sample (89%) had prior peer assessment experiences and such prior experiences have been found related to less negative attitudes toward peer assessment (Wen and Tsai, 2006). Second, some studies show that students' perceptions of peer assessment may vary according to students' majors (e.g., Zou et al., 2018). In this study, participants studied physical education, a major in which students know each other well. They are used to practicing sports together, both during and outside courses, and they excel at different sports, which allows them to help each other out a lot.

A consequence of the initial positive perceptions is that they could have been more difficult to affect. Indeed, most learnings follow a diminishing-returns curve: it begins slowly, increases exponentially, and then slows importantly when approaching mastery (Ritter and Schooler, 2001). This learning curve has been found for the acquisition of numerous intellectual and perceptual-motor skills (Rosenbaum et al., 2001). It implies that the most important improvement happens near the start, which suggests that it is at this time that a training will have its biggest impact. Even though in the present study we targeted attitudes and not skills, this reasoning may suggest that students' perceptions may have been too high at the pre-test for the training to have a visible impact.

There was one exception regarding the high scores at the pre-test: trust in the other in the experimental group showed

moderate scores at the pre-test. These were significantly lower compared to the levels of trust in the other as an assessor for the students in the control group. A possible explanation for these differences between groups could be the COVID-19 pandemic. The data collection for the control group ended just before the first lockdown in Belgium, while the data collection for the experimental group took place after on-and-off lockdowns for almost 2 years. The use of online teaching reduced contact with peers and had a deep impact on the way students could interact with each other (Alger and Eyckmans, 2022), which likely reduced the proximity students in physical education usually experience.

Our second hypothesis was that students in the experimental group would provide more elaborated feedback (including verification and an elaboration element such as correction, justification, question, or suggestion) than students in the control group. This hypothesis was not supported either; students in the experimental group did not provide any type of feedback in a higher proportion than students in the control group did. The absence of effects supporting our hypothesis is not surprising given the absence of visible effects of the training on students' perceptions. Our initial hypothesis was that the training would increase students' perceptions, which would incite them to provide more elaborated feedback. However, given that the training did not seem to affect students' perceptions, this assumption did not hold anymore.

In addition, significant effects in the opposite direction of our hypotheses were observed: students in the control group provided more correction and justification feedback than students in the experimental group. This effect may be explained by the initial performance differences between the two groups. The higher grades received by students from the control group suggest they had a better understanding of the course, which could explain why they seemed able to give more elaborated feedback than students in the experimental group. It is known that higher-performing students tend to provide more useful feedback to their peers (Wu and Schunn, 2023). This explanation is supported by the positive correlations between students' grades for their first draft and the proportion of question feedback provided.

Our third hypothesis was that students' performance would improve more in the experimental than in the control group. We started by looking at whether there was a performance improvement. As expected, in both groups, students' grades were higher for their final version than for their first draft. More precisely, students improved their grades for the criteria "diagram" and "relevance of the situation," but not for the criteria "instructions" and "success criteria." The particularity of these last two criteria is that students must master both higher-order aspects (e.g., find the most relevant instruction for the exercise) and lower-order aspects (e.g., write the instruction with adequate vocabulary and without spelling mistakes). If students only acted upon peer feedback on lower-order aspects – which prior research has shown that students tend to do (e.g., Aben et al., 2022; Van Meenen et al., 2023)—and left important problems with higher-order aspects, the improvement would not result in a higher grade from the professor.

The comparison of the performance improvement in the control and the experimental group did not support our hypothesis that students from the experimental group would improve more than those in the control group. The performance improvement was similar in both groups for the global grade. Here too, one could argue that it is due to the absence of a visible impact of the training on students' perceptions. Our initial hypothesis was that the training would increase the perceptions of students from the experimental group, which would allow them to gain more from the peer assessment than students from the control group. Given that the training did not seem to affect students' perceptions, the assumption that students from the experimental group would improve more did not hold anymore.

On the contrary, considering that students in the control group trusted their peers more, they are expected to benefit more from their feedback (van Gennip et al., 2010). We would therefore expect a larger improvement in the control than in the experimental group. This was the case, but only for the criterion "relevance of the situation" (a criterion for which an improvement is only possible if important changes were made). For this criterion, students who trust their peers may have considered their feedback more carefully and, therefore, improved their performance. An element that supports this is the moderate positive correlation between students' perceptions of trust in the peer as an assessor and the performance gain for the criteria "relevance of the situation."

4.2. Pathways for future research

All results taken together, we could not find evidence supporting the training's efficacy. There may be several reasons for this finding, which should be investigated in future research.

First, some limitations of the study may have prevented us to find an effect of the training. As the study was conducted in a natural setting, it was impractical and unethical to use random sampling (Cohen et al., 2007). Therefore, we used a quasi-experiment with pre-and post-test design with two cohorts of students. Although this design is relatively strong, having the two groups participating in the study in different years hindered the comparability between the control and the experimental groups, especially in this case with a pandemic occurring in between. Students in the experimental group participated in this study after experiencing several lockdowns during which the majority of their classes were held online. Online teaching due to the COVID-19 pandemic deeply impacted interactions between students (Alger and Eyckmans, 2022) and students' learning (Di Pietro, 2023), which could explain the differences between the two groups at pre-test. To ensure group comparability while remaining ethical, future studies could be conducted in a course with a large cohort of students and multiple peer assessment opportunities. In this context, a waiting list control group would be possible (Elliott and Brown, 2002); the training could be introduced at different times for different groups of students so that it would be possible to compare them while still allowing every student to benefit from the training before the outcomes of peer assessment could impact their grades.

Another limitation is linked to the population of the study. As explained before, the perceptions of psychological safety, trust in the self as an assessor, and trust in the other as an assessor of students participating in the study were already quite positive at the pre-test, which could have made it more difficult to affect them (Ritter and Schooler, 2001). Future studies could investigate the impact of the training on students who have negative perceptions of psychological safety and trust at baseline. These studies could be conducted with students for whom more concerns related to peer assessment could be expected, such as students who never experienced peer assessment

before (Wen and Tsai, 2006) or students who are studying humanities (Praver et al., 2011).

Second, it is possible that the training had both positive and negative impacts on students' perceptions, which canceled each other out. During the training, students had to opportunity to think about and discuss their concerns linked to peer assessment. If students feel safe and had trust in themselves and their peers before the training, which seems to be the case, the training may have challenged these positive perceptions at first, before positively affecting them. For trust in the self as an assessor more particularly, the training may have both allowed students to acquire skills in providing feedback and to realize that they may have overestimated their skills before. Indeed, according to the Dunning-Kruger effect, unskilled people overestimated themselves due to their lack of skills (Kruger and Dunning, 1999).

Third, the training could be more effective with a smaller students per instructor-ratio. In the study, the relatively large number of students made it difficult for the instructor to visit each group, while they were role-playing, which is important to keep students on task, answer their questions, and provide suggestions (Bolinger and Stanton, 2020). Moreover, to ensure student participation, group discussions were also conducted in small groups and, therefore, could not be facilitated by an instructor, contrary to what is recommended (Blanchard and Thacker, 2012). Future studies could investigate the efficacy of an improved version of the training, with more time for role-plays and discussions, diverse training methods, and more instructors to guide the students.

In addition, choices made to keep training time-and cost-efficient may have reduced its efficacy. Our objective was to create a training that could easily be implemented in any higher education course. Therefore, the training had been designated so that it was short enough to fit into the busy course schedules and so that a single instructor could train a group of about 50 students. At the same time, to overcome a limitation of the training on psychological safety developed by Dusenberry and Robinson (2020), our training almost exclusively contained active-learning methods, namely role-play and group discussion. Active-learning methods and interactions between learners are indispensable for learning, but they require time (Martin et al., 2014), which was limited in our 1-h session. Students only participated in two role-plays (one as active participants, and one as observers), while it is recommended to have three sets of role-plays (Blanchard and Thacker, 2012). Moreover, with a 1-h session, it was not possible to combine more than two training methods, although this combination enhances training effectiveness (Martin et al., 2014). Although short training is effective to impact peer pressure (Li, 2017), longer training may be required to impact psychological safety and trust, because as Hunt et al. (2021) argued it is difficult to foster psychological safety.

It is also possible that given this difficulty to foster psychological safety (Hunt et al., 2021), a training is not an effective way to affect it. The only other study we know of that tried to affect psychological safety through training did not find a significant effect either (Dusenberry and Robinson, 2020). Even when designing the training with this study's limitations in mind to overcome, we did not find a significant impact of the training on students' perceptions of psychological safety. This would suggest that, contrary to peer pressure (Li, 2017), psychological safety is a factor that cannot be effectively enhanced through a training, and that other kinds of interventions are needed.

To the best of our knowledge, the present study is the first that aims to incite trust in the self and the peer as an assessor through training. While the results of the present study indicate an absence of effect, more studies are warranted. Concluding that trust cannot be positively affected by a training appears premature.

To have an intervention that is both feasible and effective, a possibility could be to use guidance during peer assessment, eventually in combination with a 1 or 2-h training session prior to the peer assessment exercice. Prior research found that guidance can enhance the peer assessment process (e.g., Bloxham and Campbell, 2010; Harland et al., 2017; Misiejuk and Wasson, 2021). Three types of guidance could also enhance students' perceptions of psychological safety and trust. The first is the use of backward evaluation, the feedback that an assessee provides to his/her assessor about the quality of the received feedback (Misiejuk and Wasson, 2021). The second is the use of cover sheets, sheets that allow an assessee to ask specific questions and explain what type of feedback they would like to receive (Bloxham and Campbell, 2010). The third is the use of a rebuttal, a text that assessee must write to explain why they found the received feedback relevant or not and justify how they acted upon it (Harland et al., 2017). These sorts of guidance could help students realize they are not passive recipients of feedback, but that they can critically appraise the feedback they received and that they can improve themselves in providing feedback, and therefore, could enhance their perceptions. Future studies could investigate if, combined with a training session, these kinds of guidance have positive impacts on students' perceptions of psychological safety and trust.

Besides testing the effectiveness of the training we developed, a second aim of this study was to investigate the impact of an increase in students' perceptions of psychological safety and trust in the self and the other as assessor on peer assessment outcomes. Given that we could not find any impact of the training on students' perceptions, we have no evidence that an increase in psychological safety or trust would result in higher quality feedback or bigger performance improvement. However, there was a correlation between students' perceptions of trust in the peer as assessor and their performance gain for one of the criteria ("relevance of the situation"). Previous research found that trust in the other as an assessor is a predictor of students' learning following a peer assessment activity when learning is measured by asking students if they feel they have learned (van Gennip et al., 2010). Our result corroborates this positive relationship between trust in the other and learning, with the learning being measured by students' grade improvement in our study.

The relationship between perceptions of trust in the other and performance improvement did not seem to be explained by the quality of feedback given by students (there was no significant correlation in our study). However, the quality of provided feedback is not the only variable that could mediate the relationship between students' perceptions and their performance improvement, feedback uptake could also play an important role. Some studies show that peer feedback uptake tends to be low and that students make few revisions following peer assessment (e.g., Winstone et al., 2017; Berndt et al., 2018; Aben et al., 2022; Bouwer and Dirkx, 2023; Van Meenen et al., 2023). Possibly, this low feedback uptake is linked to low levels of psychological safety and trust, but empirical evidence is lacking on this. Future research on the relationship between students' feedback uptake on the one hand and their perceptions of psychological safety and trust on the other hand is warranted.

4.3. Conclusion

This study aimed to investigate the efficacy of a training session created to enhance students' perceptions of psychological safety and trust before a peer assessment activity. Findings suggested that students' perceptions are related to students' performance improvement, which confirms that it is important to develop interventions targeting these perceptions. Contrary to our hypotheses, we could not find evidence of the training's effectiveness on students' perceptions, students' performances, or the type of feedback provided by students. This absence of visible impact could be explained by limitations of the study, or by limitations of the training in itself. Future studies could investigate if an improved longer version of the training would be more effective, or if the training should be combined with other guidance during peer assessment.

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 Commission d'éthique IPSY, Université catholique de Louvain. The patients/participants provided their written informed consent to participate in this study.

Author contributions

MS, DJ, and LC conceptualized the research project, defined the research question, and developed the research design. MS and DJ implemented the intervention and collected the data, under the supervision of LC. MS performed the statistical analysis with the help of LC. MS wrote the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors would like to thank Maryam Alqassab for sharing her detailed coding scheme, Aurélie Bertrand helping us with the analyses, and Quentin Brouhier for coding feedback.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1198011/full#supplementary-material

References

Aben, J., Timmermans, A. C., Dingyloudi, F., Mascareño Lara, M., and Strijbos, J. (2022). What influences students' peer-feedback uptake? Relations between error tolerance, feedback tolerance, writing self-efficacy, perceived language skills and peer-feedback processing. *Learn. Individ. Differ.* 97:102175. doi: 10.1016/j.lindif.2022.102175

Alger, M., and Eyckmans, J. (2022). "I took physical lessons for granted": a case study exploring students' interpersonal interactions in online synchronous lessons during the outbreak of COVID-19. *System (Linköping)* 105:102716. doi: 10.1016/j.system.2021.102716

Alqassab, M., Strijbos, J., and Ufer, S. (2018). The impact of peer solution quality on peer-feedback provision on geometry proofs: evidence from eye-movement analysis. *Learn. Instr.* 58, 182–192. doi: 10.1016/j.learninstruc.2018.07.003

Berndt, M., Strijbos, J.-W., and Fischer, F. (2018). Effects of written peer-feedback content and sender's competence on perceptions, performance, and mindful cognitive processing. *Eur. J. Psychol. Educ.* 33, 31–49. doi: 10.1007/s10212-017-0343-z

Black, N., Mullan, B., and Sharpe, L. (2016). Computer-delivered interventions for reducing alcohol consumption: meta-analysis and meta-regression using behaviour change techniques and theory. *Health Psychol. Rev.* 10, 341–357. doi: 10.1080/17437199.2016.1168268

Blanchard, P. N., and Thacker, J. W. (2012). *Effective training: Systems, strategies, and practices. 5th* Edn. Upper Saddle River, NJ: Pearson/Prentice Hall.

Bloxham, S., and Campbell, L. (2010). Generating dialogue in assessment feedback: exploring the use of interactive cover sheets. *Assess. Eval. High. Educ.* 35, 291–300. doi: 10.1080/02602931003650045

Bolinger, A. R., and Stanton, J. V. (2020). *Role-play simulations*. Cheltenham: Elgar.

Bouwer, R., and Dirkx, K. (2023). The eye-mind of processing written feedback: unraveling how students read and use feedback for revision. *Learn. Instr.* 85:101745. doi: 10.1016/j.learninstruc.2023.101745

Carless, D., Salter, D., Yang, M., and Lam, J. (2011). Developing sustainable feedback practices. *Stud. High. Educ.* 36, 395–407. doi: 10.1080/03075071003642449

Cheng, K., and Tsai, C. (2012). Students' interpersonal perspectives on, conceptions of and approaches to learning in online peer assessment. *Australas. J. Educ. Technol.* 28, 599–618. doi: 10.14742/ajet.830

Ching, Y., and Hsu, Y. (2016). Learners' interpersonal beliefs and generated feedback in an online role-playing peer-feedback activity: an exploratory study. *Int. Rev. Res. Open Dist. Learn.* 17, 105–122. doi: 10.19173/irrodl.v17i2.2221

Cohen, L., Manion, L., and Morrison, K. (2007). *Research methods in education. 6th* Edn. London: Routledge/Taylor & Francis Group.

Di Pietro, G. (2023). The impact of covid-19 on student achievement: evidence from a recent meta-analysis. *Educ. Res. Rev.* 39:100530. doi: 10.1016/j.edurev.2023.100530

Double, K., McGrane, J., and Hopfenbeck, T. (2020). The impact of peer assessment on academic performance: a meta-analysis of control group studies. *Educ. Psychol. Rev.* 32, 481–509. doi: 10.1007/s10648-019-09510-3

Dusenberry, L., and Robinson, J. (2020). Building psychological safety through training interventions: manage the team, not just the project. *IEEE Trans. Prof. Commun.* 63, 207–226. doi: 10.1109/TPC.2020.3014483

Edmondson, A. C., and Lei, Z. (2014). Psychological safety: the history, renaissance, and future of an interpersonal construct. *Annu. Rev. Organ. Psych. Organ. Behav.* 1, 23–43. doi: 10.1146/annurev-orgpsych-031413-091305

Elliott, S. A., and Brown, J. S. L. (2002). What are we doing to waiting list controls? *Behav. Res. Ther.* 40, 1047–1052. doi: 10.1016/S0005-7967(01)00082-1

Enders, C. K. (2010). *Applied missing data analysis*. New York: The Guilford Press.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among multiple raters 4*). Gaithersburg, MD: Advanced Analytics, LLC.

Hanrahan, S. J., and Isaacs, G. (2001). Assessing self-and peer-assessment: the students' views. *High. Educ. Res. Dev.* 20, 53–70. doi: 10.1080/07294360123776

Harland, T., Wald, N., and Randhawa, H. (2017). Student peer review: enhancing formative feedback with a rebuttal. *Assess. Eval. High. Educ.* 42, 801–811. doi: 10.1080/02602938.2016.1194368

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Hunt, D., Bailey, J., Lennox, B., Crofts, M., and Vincent, C. (2021). Enhancing psychological safety in mental health services. *Int. J. Ment. Heal. Syst.* 15:33. doi: 10.1186/s13033-021-00439-1

Kollar, I., and Fischer, F. (2010). Peer assessment as collaborative learning: a cognitive perspective. *Learn. Instr.* 20, 344–348. doi: 10.1016/j.learninstruc.2009.08.005

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Li, L. (2017). The role of anonymity in peer assessment. *Assess. Eval. High. Educ.* 42, 645–656. doi: 10.1080/02602938.2016.1174766

Li, L., Liu, X., and Steckelberg, A. L. (2010). Assessor or assessee: how student learning improves by giving and receiving peer feedback. *Br. J. Educ. Technol.* 41, 525–536. doi: 10.1111/j.1467-8535.2009.00968.x

Li, H., Xiong, Y., Hunter, C. V., Guo, X., and Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assess. Eval. High. Educ.* 45, 193–211. doi: 10.1080/02602938.2019.1620679

Martin, B. O., Kolomitro, K., and Lam, T. C. M. (2014). Training methods: a review and analysis. *Hum. Resour. Dev. Rev.* 13, 11–35. doi: 10.1177/1534484313497947

Misiejuk, K., and Wasson, B. (2021). Backward evaluation in peer assessment: a scoping review. *Comput. Educ.* 175:104319. doi: 10.1016/j.compedu.2021.104319

Mulder, R. A., Pearce, J. M., and Baik, C. (2014). Peer review in higher education: student perceptions before and after participation. *Act. Learn. High. Educ.* 15, 157–171. doi: 10.1177/1469787414527391

Panadero, E., and Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assess. Eval. High. Educ.* 44, 1253–1278. doi: 10.1080/02602938.2019.1600186

Panadero, E., Alqassab, M., Fernández Ruiz, J., and Ocampo, J. C. (2023). A systematic review on peer assessment: Intrapersonal and interpersonal factors. *Assess. Evaluat. Higher Educ.* ahead-of-print, 1–23. doi: 10.1080/02602938.2023.2164884

Planas Lladó, A., Soley, L. F., Fraguell Sansbelló, R. M., Pujolras, G. A., Planella, J. P., Roura-Pascual, N., et al. (2014). Student perceptions of peer assessment: an interdisciplinary study. *Assess. Eval. High. Educ.* 39, 592–610. doi: 10.1080/02602938.2013.860077

Praver, M., Rouault, G., and Eidswick, J. (2011). Attitudes and affect toward peer evaluation in EFL reading circles. *Read. Matrix* 11, 89–101.

Ritter, F. E., and Schooler, L. J. (2001). "The learning curve," in *International encyclopedia of social and behavioral sciences*. eds. N. J. Smelser and P. B. Baltes (Amsterdam: Pergamon).

Rosenbaum, D. A., Carlson, R. A., and Gilmore, R. O. (2001). Acquisition of intellectual and perceptual-motor skills. *Annu. Rev. Psychol.* 52, 453–470. doi: 10.1146/annurev.psych.52.1.453

Rotsaert, T., Panadero, E., Estrada, E., and Schellens, T. (2017). How do students perceive the educational value of peer assessment in relation to its social nature? A survey study in flanders. *Stud. Educ. Eval.* 53, 29–40. doi: 10.1016/j.stueduc.2017.02.003

Rotsaert, T., Panadero, E., and Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *Eur. J. Psychol. Educ.* 33, 75–99. doi: 10.1007/s10212-017-0339-8

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795

Soares, A. E., and Lopes, M. P. (2020). Are your students safe to learn? The role of lecturer's authentic leadership in the creation of psychologically safe environments and their impact on academic performance. *Act. Learn. High. Educ.* 21, 65–78. doi: 10.1177/1469787417742023

Strijbos, J., Martens, R. L., Prins, F. J., and Jochems, W. M. G. (2006). Content analysis: what are they talking about? *Comput. Educ.* 46, 29–48. doi: 10.1016/j.compedu.2005.04.002

Su, W. (2022). Masked ball for all: how anonymity affects students' perceived comfort levels in peer feedback. *Assess. Evaluat. Higher Educ.* ahead-of-print, 1–11. doi: 10.1080/02602938.2022.2089348

Tai, J., Ajjawi, R., Boud, D., Dawson, P., and Panadero, E. (2018). Developing evaluative judgement: enabling students to make decisions about the quality of work. *High. Educ.* 76, 467–481. doi: 10.1007/s10734-017-0220-3

Topping, K. (1998). Peer assessment between students in colleges and universities. *Rev. Educ. Res.* 68, 249–276. doi: 10.3102/00346543068003249

van Gennip, N., Segers, M., and Tillema, H. (2010). Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learn. Instr.* 20, 280–290. doi: 10.1016/j.learninstruc.2009.08.010

Van Meenen, F., Masson, N., Catrysse, L., and Coertjens, L. (2023). Taking a closer look at how higher education students process and use (discrepant) peer feedback. *Learn. Instr.* 84:101711. doi: 10.1016/j.learninstruc.2022.101711

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., and Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Comput. Educ.* 81, 123–132. doi: 10.1016/j.compedu.2014.10.001

von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociol. Methods Res.* 49, 699–718. doi: 10.1177/0049124117747303

Wen, M. L., and Tsai, C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *High. Educ.* 51, 27–44. doi: 10.1007/s10734-004-6375-8

Wilson, M. J., Diao, M. M., and Huang, L. (2015). 'I'm not here to learn how to mark someone else's stuff': an investigation of an online peer-to-peer review workshop tool. *Assess. Eval. High. Educ.* 40, 15–32. doi: 10.1080/02602938.2014.881980

Winstone, N. E., Nash, R. A., Parker, M., and Rowntree, J. (2017). Supporting Learners' Agentic engagement with feedback: a systematic review and a taxonomy of Recipience processes. *Educ. Psychol.* 52, 17–37. doi: 10.1080/00461520.2016.1207538

Wisniewski, B., Zierer, K., and Hattie, J. (2020). The power of feedback revisited: a meta-analysis of educational feedback research. *Front. Psychol.* 10:3087. doi: 10.3389/fpsyg.2019.03087

Wu, Y., and Schunn, C. D. (2023). Assessor writing performance on peer feedback: exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback. *J. Educ. Psychol.* 115, 118–142. doi: 10.1037/edu0000768

Yu, F., and Liu, Y. (2009). Creating a psychologically safe online space for a student-generated questions learning activity via different identity revelation modes. *Br. J. Educ. Technol.* 40, 1109–1123. doi: 10.1111/j.1467-8535.2008.00905.x

Yu, F., and Wu, C. (2011). Different identity revelation modes in an online peer-assessment learning environment: effects on perceptions toward assessors, classroom climate and learning activities. *Comput. Educ.* 57, 2167–2177. doi: 10.1016/j.compedu.2011.05.012

Zou, Y., Schunn, C. D., Wang, Y., and Zhang, F. (2018). Student attitudes that predict participation in peer assessment. *Assess. Eval. High. Educ.* 43, 800–811. doi: 10.1080/02602938.2017.1409872

frontiers | Frontiers in Education

TYPE Systematic Review
PUBLISHED 06 July 2023
DOI 10.3389/feduc.2023.1217167

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Yueting Xu,
South China Normal University, China
Lenore Adie,
Australian Catholic University, Australia

*CORRESPONDENCE
Serafina Pastore
✉ serafina.pastore@uniba.it

RECEIVED 04 May 2023
ACCEPTED 08 June 2023
PUBLISHED 06 July 2023

CITATION
Pastore S (2023) Teacher assessment literacy: a systematic review. *Front. Educ.* 8:1217167. doi: 10.3389/feduc.2023.1217167

COPYRIGHT
© 2023 Pastore. 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.

Teacher assessment literacy: a systematic review

Serafina Pastore*

Department of Research and Humanities Innovation, University of Bari Aldo Moro, Bari, Italy

Teacher assessment literacy, generally defined as a set of knowledge and skills a teacher needs to effectively enact assessment in the classroom, has been a priority in the educational policy and educational research agenda for decades. For a long time, it has been identified with standardized measurement and classroom testing. The interest in this topic is related not only to the accountability pressures and the identification of assessment as a lever for school and system reform but also to the need for teachers to support student learning by developing and implementing responsive assessments within their classrooms. Considerable efforts have been made to prepare novice and expert teachers in understanding how to deal with aspects of assessment practice and how to use the assessment results. Although the research on teacher assessment literacy is quite wide, it continues to demonstrate how teachers struggle with assessment, especially when they are required to transfer new approaches and theories into the actual classroom context. This systematic review synthetizes the literature on teacher assessment literacy considering how it has been defined and studied over the last 10 years (2013–2022). Documenting and comparing the different expressions and definitions of assessment literacy used in the 42 selected studies, this systematic review offers a detailed overview of the changes that occurred in the conceptualizations of assessment literacy. Along with the analysis of the theoretical/conceptual frameworks and research methods used to investigate teacher assessment literacy, the scrutiny of its foundational components represents a useful base to orient pre- and in-service teacher education. Against the backdrop of strengths and weaknesses of this review, research priorities and practical implications of the findings are discussed.

KEYWORDS

teacher assessment literacy, assessment education, teacher education, educational assessment, systematic review

1. Introduction

In educational discourses, assessment has always been one of the hottest topics (Flórez Petour, 2015; Roberts et al., 2021). The crucial trait of educational reform (e.g., Assessment Reform Group in the UK, or No Child Left Behind in the US) (Ball, 2015; OECD, 2022), assessment represented a leading force in education, advocated both as a powerful tool of educational policy and school improvement (Hilton et al., 2013; Scheerens, 2016; Torres and Weiner, 2018; Mouraz et al., 2019) and as a fundamental component of teacher instructional practice (Black and Wiliam, 2018; Allal, 2020; Yan, 2021; Brown, 2022). Given its strategic nature, assessment has been identified as a core principle underlying curriculum in many educational systems around the world, as well as a key aspect of teacher professionalism and teaching quality (O'Neill and Adams, 2014; Darling-Hammond, 2016; Cochran-Smith, 2023).

The attention deserved for teacher assessment literacy, as well as the emphasis on assessment implications for educational policy and educational practice, has bolstered the interest of policy makers, teacher educators, and researchers in preparing teachers for assessment (Popham, 2009, 2018; DeLuca et al., 2016b; Stiggins, 2017). Although a large body of literature exists, the concept of assessment literacy remains, *per se*, a complex and contested concept, difficult to define. Furthermore, if persistent levels of assessment illiteracy continue to be pointed out by researchers within the pre-service and in-service contexts (Beziat and Coleman, 2015; DeLuca et al., 2016a; Christoforidou and Kyriakides, 2021; Atjonen et al., 2022), then how teachers adapt their assessment practice following institutional changes or policy requirements, or inedited instructional circumstances (such as those arouse during the COVID-19 pandemic period), represents a challenging research topic.

Assuming that a unique definition of assessment literacy is not tenable, given the recognized contextual, cultural, and social nature of assessment (Willis et al., 2013), this study offers a scrutiny of the current landscape on assessment literacy and provides, through a systematic review, an overview of current knowledge about teacher assessment literacy.

Against the backdrop of teacher professionalism and teacher competency concepts, this article, in the first section, illustrates the research questions and the method employed to identify and select articles included in the analysis. In the following, instead, the results are shaped and discussed considering the study limitations, as well as the implications for educational research and educational practice.

It is important to recognize that the concept at the heart of this article, teacher assessment literacy, is not without complications and that several words in which research writes about it exist. Thus, although words such as assessment knowledge and skills, assessment competence, assessment approaches, and assessment capabilities can be found (and they have been used in performing the systematic review), over this article, the expression teacher assessment literacy has been preferred.

2. Study aims

With the purpose of mapping out how teacher assessment literacy has been studied over the last 10 years, a systematic review has been performed. Responding to the following research questions, this article aims to provide new insights useful to inform evidence-based actions in the educational research, policy, and practice fields:

- How is teacher assessment literacy defined in research and what are its foundational constructs/components?
    - On which theories/approaches did teacher assessment literacy definitions rely on?
    - To what extent and in what ways had teacher assessment literacy been investigated?

3. Conceptual framework

This review of teacher assessment literacy definitions sets the stage for the current efforts to map the key components in the assessment domain and supports (pre- and in-service) teachers to incorporate them into their practice.

Before proceeding with the methodology of this review, a brief reflection on what is meant by assessment literacy, within the broad framework of teacher professionalism, is offered.

For nearly more than 60 years, several studies have scrutinized what teachers should essentially know and be able to do with assessment. These studies were part of the attempts made to re-conceptualize teacher work focusing on how "teachers acquire, generate, and learn to use knowledge in teaching" (Feiman-Nemser, 2008, p. 697). The growing interest in practical knowledge, along with the recognition of the role and significance of daily experience in teacher work, led not only to revise the ideas of teacher and teaching but also to question the modalities through which teachers acquire and transfer their professionality, as well as how they mediate, adjust, and preserve their professional expertise.

Aligned with the idea that (teacher) learning is practical, redundant, spiraliform, and context-embedded (i.e., classroom, school, and national school system), the competence-based approach (Blïomeke and Kaiser, 2017; Day, 2017) deeply affected the debate on teaching quality and teaching effectiveness (O'Neill and Adams, 2014; Darling-Hammond, 2016; Torres and Weiner, 2018; Cochran-Smith, 2023). Stressing the socio-cultural perspective, Shepard pointed out that "teachers need the opportunities to construct their own understanding in the context of their practice and in ways consistent with their identity as thoughtful professionals" (Shepard, 2017, p. XXII). Professional competence is not made by a list of fixed cognitive and affective components. Assuming as an *a priori*, the idea of a continuum improvement, teacher competence is, instead, a complex set of "woven-together assumptions and meaning about what is important to do and be" (Shepard, 2017, p. XXII). One of the most important implications of this approach is the recognition that different levels of teacher competence and development trajectories exist (Blïomeke and Kaiser, 2017; Day, 2017; Dall'Alba, 2018; Christoforidou and Kyriakides, 2021). Teachers tend to develop their profession over time, in a complex, iterative learning process, which is influenced by pre- and in-service teacher education paths and different variables (e.g., duration, content knowledge, training models, and opportunities), as well as by on-going teaching experience and contextual factors, such as collaboration with colleagues, school participation, and professional standards. Within the specific areas of educational assessment and teacher education, the idea that teachers' continuous learning of knowledge and abilities would be linked to the improvement of classroom instructional practices (and therefore, to the increase of student learning) has been touted by teacher educators, policy makers, and educational researches as a key aspect to raise education quality (Hilton et al., 2013; Torres and Weiner, 2018). If, on the one hand, these research efforts have reduced the original vague and ambiguous definitions of assessment literacy, then, on the other hand, they have tried to ensure an effective assessment preparation for teachers. However, while teacher assessment literacy has been

recognized as a key component of effective teaching and learning, research has continued to point out how teachers demonstrate low levels of assessment literacy and tend to perceive themselves as not confident in assessing student learning (DeLuca and Bellara, 2013; Poskitt, 2014; Stiggins, 2017). Different studies, in fact, have shown how, historically, assessment education received little consideration by research on initial teacher education (DeLuca, 2012). Other studies, instead, have tried to shed light on the effects of pre-service assessment education models on novice teachers (Brooks, 2021; Cochran-Smith et al., 2021). The studies and the professional development initiatives purposed to support teachers' assessment practice in their classrooms, instead, have identified more effective training models and strategies. Active, collaborative, engaging, and classroom-embedded learning models are generally recognized as more sustainable by teachers (Neuman and Cunningham, 2009). At the same time, long-term on-the-job training programs have been identified as having more influence on teachers' assessment conceptions change (Desimone, 2009). Despite the consensus that high-quality professional development could provide teachers with knowledge and skills useful to deal with innovation and challenges, some authors have pointed out that continuous development in the assessment field seems to be ineffective and time-consuming (O'Neill and Adams, 2014; Torres and Weiner, 2018). As a consequence, DeLuca et al. (2020) noticed that teachers continue to struggle with assessment practices, especially when they are required to transfer new approaches and theories into the actual classroom context. The current attempts to define teacher assessment literacy, as well as the growing research interest in teacher conceptions of assessment (Brown, 2004; Deneen and Brown, 2016), finally, offer an opportunity to better identify what counts to be an assessment literate teacher and to detect which critical features an education/training path should include (e.g., course content and pedagogies professional drivers) to effectively meet teachers' learning needs in the assessment domain.

4. The present research study

The next paragraphs describe how the systematic review has been realized; moreover, they present the main results corresponding to the research questions, which consider:

- Definitions of assessment literacy and its foundational components,
- Theories/approaches for teacher assessment literacy; and
- Research methods/methodologies used in the selected studies.

4.1. Procedure

Using the approach of Petticrew and Roberts (2006) for systematic reviews in social studies, research questions and search terms have been first defined. Then, databases have been selected and interrogated. After defining inclusion and exclusion criteria, extracted data have been categorized and summarized by theoretical/conceptual approaches to teacher assessment literacy, research aims research methods, and population and sample. Furthermore, selected studies have been evaluated in terms of scientific quality. Finally, a contrastive analysis of reviewed publications has been performed.

4.2. Search string design and databases

Although the term assessment literacy is very specific, terms such as assessment competence, assessment capability, and assessment approaches are often used synonymously in the literature and appear in a wide range of studies. Therefore, to retrieve as many relevant studies as possible, the following search string was designed:

> "teacher assessment literacy*" OR "teacher assessment competence*" OR "teacher assessment approach*" OR "teacher assessment literacy capability*" NOT "higher education" NOT "university" NOT "second language" NOT "language" NOT "efl"

This research string delimited the search to compulsory education. The relationships between teacher assessment literacy and second language as a teacher subject domain were not considered. These delimitations allowed us to better situate retrieved studies within the broader context of teacher professionalism and teacher education. To identify additional sources and ensure that influential work was not overlooked, snowballing strategies, such as tracking the reference lists of included sources, and checking the researchers' outputs were also applied (Alexander, 2020; Dekkers et al., 2022). In this way, 28 additional articles were identified.

The literature search in this study was performed across 5 databases: Eric, PsycINFO, PubMed, Scopus, and Web of Science.

4.3. Inclusion and exclusion criteria

The publications selection followed these set inclusion and exclusion criteria:

- The study was published in a scientific, international, peer-reviewed journal. To systematically pool together updated high-quality research studies, chapters in edited books, doctoral theses, conference papers, books, as well as working papers and reports were excluded. Moreover, although the inclusion of other languages allows us to consider more studies, and, as pointed out by Dekkers et al. (2022) to improve "the internal and external validity of findings in a review" (p. 205), only studies written in English were included.
- The study reported a definition of teacher assessment literacy and research work (i.e., quantitative, qualitative, and mixed approach) on this topic. Given the focus of this review (i.e., provide insight into how teacher assessment literacy is defined), however, theoretical articles and reviews were also included.
- The study was realized in the context of primary or secondary education.

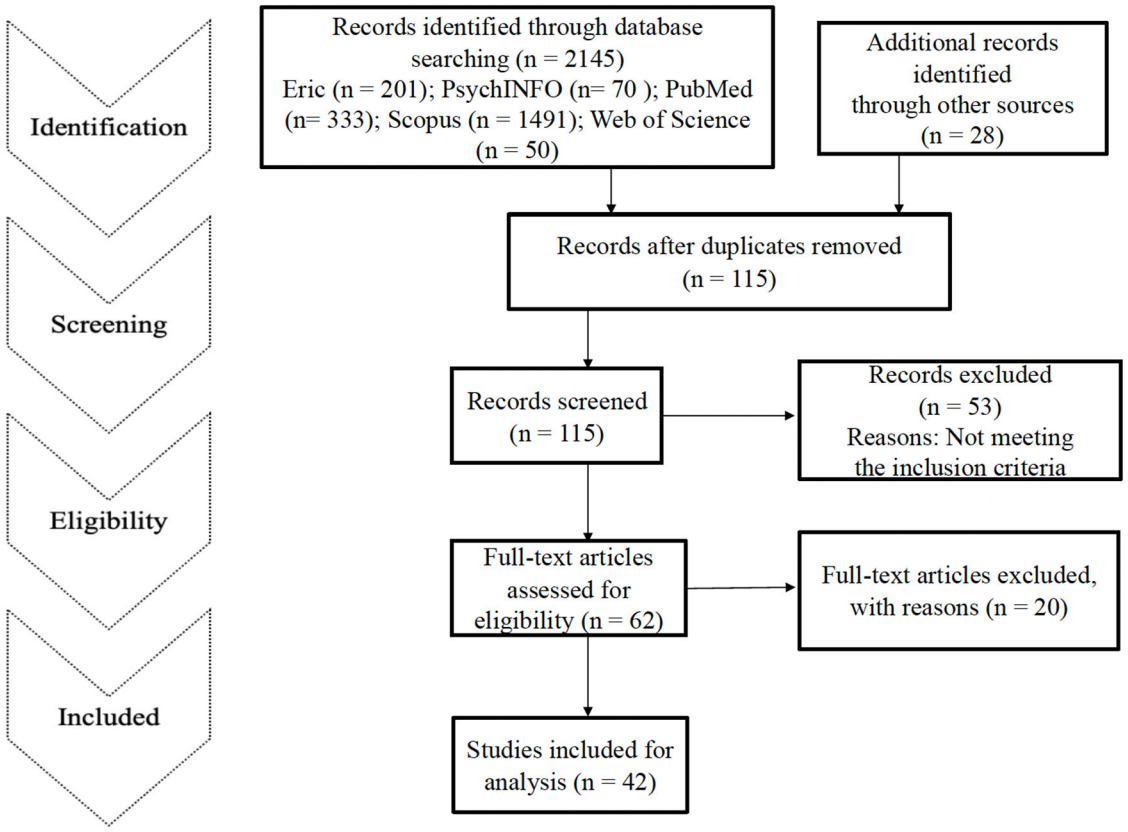


FIGURE 1
Flowchart of the search and screening process. Adapted from Moher et al. (2009).

- The study was published in the last 10 years (2013–2022). This last criterion restricted the selection to the most recent research studies realized in the field.

In case of doubtful or not complete satisfaction with all the inclusion criteria, the publication remained in the selection. Once the title and abstract scan phase was completed, the author and two research assistants assessed the relevance of each study and discussed inclusion decisions. When a publication did not match the inclusion criteria, it was removed.

4.4. Data extraction and data analysis

The studies included in the final set were categorized and summarized using a data extraction form which consisted of the following sections:

- General information of the article (i.e., authors, title, journal, year of publication, and *k*-words);
- Definition of teacher assessment literacy (i.e., principal components of assessment literacy);
- Study characteristics (i.e., country, theoretical or conceptual framework, research aims/questions, and research design);
- Participant characteristics (i.e., number of participants, educational setting, and sample size).

The data extraction form allowed a more effective comparison and cross-checking of data. Publications of the initial search were 2,173 (Figure 1). After removing duplicate and screening titles and abstracts, the full-text articles assessed on inclusion criteria were 115.

Moreover, 53 articles were further excluded because not directly focused on teacher assessment literacy. Thus, 62 studies were included in the coding. Of these, 20 articles were excluded: 17 because without a clear definition of teacher assessment literacy and 3 because of general commentaries.

5. Results

In total, 42 studies were considered for the analysis. In the following, the study characteristics are reported (Table 1).

In the teacher assessment literacy research domain, qualitative ($N = 18$) and quantitative research approaches ($N = 15$) are more frequent than mixed-method ones ($N = 9$). More specifically, the qualitative research studies panel included two comparative studies (one on assessment literacy measures and another on assessment education systems) (DeLuca et al., 2019a,c), five reviews (Fulmer et al., 2015; Xu and Brown, 2016; Looney et al., 2018; Oo et al.,

TABLE 1 Overview of study characteristics.

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
1.	Willis et al. (2013)	Assessment literacy is a dynamic context-dependent social practice that involves teachers articulating and negotiating classroom and cultural knowledge with one another and with learners, in the initiation, development and practice of assessment to achieve the learning goals of students	AU	Conceptualize assessment literacy	Bernstein's socio-cultural theory	No	No	No	No	No
2.	DeLuca and Bellara (2013)	Assessment literate teachers understand how to construct, administer, and score reliable assessments and communicate valid interpretations about student learning [. . .] Furthermore, assessment literacy involves integrating assessment practices, theories, and philosophies to support teaching and learning within a standards-based framework of education	US, CA	Analyze the alignment of teacher education policies, teacher standards for educational assessment, and learning expectations of pre-service assessment course syllabi across three alignment dimensions (i.e., content focus, depth of knowledge, and range of knowledge)	Pre-service assessment education	QL	No	K-12	No	No
3.	Gunn and Gilmore (2014)	The notion of assessment capability recognizes the important role of both teachers and learners in assessment for learning processes (Dixon and Haigh, 2009). It acknowledges that being capable involves being "motivated to access, interpret and use information from quality assessment in ways that affirm or further learning" (Absolum et al., 2009, p. 9)	NZ	Explore early childhood teacher education students' assessment conceptions, by examining the extent to which these conceptions engage with current policy expectations, and by considering the extent to which they can shed light on student teachers' developing assessment capability	ITE	MX	TE	CE	No	37

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
4.	Hill et al. (2014)	Assessment capability includes the elements of assessment literacy but goes further to include teachers and their students, being "able and motivated to access, interpret and use information from quality assessments in ways that affirm or further learning" (Absolum et al., 2009, p. 9)	NZ	Report on "Learning to Become Assessment Capable" project	ITE	MX	TE	PE, CE	No	11 Teacher educators 224, 214, 100 ITE students
5.	Cowie et al. (2014)	Assessment capability, also discussed as assessment literacy, includes the ability to develop assessments that transform learning goals into assessment activities that accurately reflect student understanding and achievement [. . .]. The ability to interpret assessment results and to use these to adapt instruction to address student learning needs, interests, and strengths, along with the capacity to effectively communicate valid interpretations about student learning, are key components of assessment literacy	NZ	Detect what student teachers had learned about assessment in their university courses and while on practicum, how their views had changed	ITE	QL	TE	Primary	No	3

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
6.	Kaden and Patterson (2014)	Assessment expertise and practice, also called assessment literacy, can be defined as the ability to design, select, interpret, and use assessments appropriately for educational decisions [. . .]. It refers to the necessary knowledge that an educator possesses and uses during her or his practice. Assessment literacy can be further described as the ability to develop assessments that transform learning goals into assessment activities that accurately reflect students' understanding and achievement (Mertler, 2009)	US	Understand how rural and urban candidates are changing assessment practices to support learning, what variables facilitate that change, and to what extent assessment practices differ between rural and urban candidates	No	MX	TE	ME, SE	Maths and Science	11
7.	Gotch and French (2014)	Assessment literacy can be defined as "an individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (Popham, 2011, p. 267)	US	Review available assessment literacy measures, and evaluate the psychometric evidence to support their adequacy for use in teacher evaluation	No	QT	No	No	No	No
8.	Poskitt (2014)	Willis et al.'s (2013) definition of AL	NZ	Demonstrate that another component is required to enhance our understanding of assessment literacies and professional learning: the need to include and be responsive to the wider community and political contexts; to transform the enactment of assessment for the betterment of student and teacher learning	Guskey's (2002) model evaluating professional learning Carless' (2005) theoretical framework on embedding professional learning changes in schools	MX	TP	No	No	No

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
9.	Hailaya et al. (2014)	Assessment literacy is defined as teachers' 'knowledge of and abilities to apply assessment concepts and techniques to inform decision making and guide practice' (Mertler, 2005, p. 16)	PH	Add to ALI's previous validation findings which were based on samples from the United States; ascertain its measurement properties and utility; and gauge its portability to other education systems such as the countries in the Asia Pacific region	Assessment Literacy Inventory (Mertler, 2005)	QT	No	PE, SE	No	582
10.	Beziat and Coleman (2015)	Assessment literacy "as the knowledge of means for assessing what students know and can do, how to interpret the results from these assessments, and how to apply these results to improve student learning and program effectiveness" (Webb, 2002, p. 1)	US	Pinpoint areas of strengths and weaknesses to improve our professional programs in more effectively preparing to teach candidates the knowledge and skills of classroom assessment	Teacher Assessment Literacy (Plake et al., 1993)	QT	TE	PE	No	26
11.	Clark (2015)	Willis et al.'s (2013) definition of AL	US	Analyze the preservice teachers thinking about their assessment decisions	Teaching culture Assessment literacy of preservice teachers Ecological aspects of an individual's agency (Biesta and Tedder, 2006)	QL	TE	SE	Social Studies	3
12.	Fulmer et al. (2015)	[. . .] assessment literacy refers to teachers' ability to apply their assessment knowledge in the revision, development, or implementation of assessment tasks in their classrooms	SG	Present a review of literature on teachers' assessment practices organized around distinguishable levels of contextual factors that may affect them	Kozma (2003) conceptual model	QL	NA	NA	NA	NA

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context						
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
13.	DeLuca et al. (2016a)	[. . .] assessment literacy has been defined as the skills and knowledge teachers require to measure and support student learning through assessment (Brookhart, 2011; Popham, 2013).	US, CA	Construct a reliable instrument reflective of contemporary assessment practices and contexts	Classroom Assessment Standards (Joint Committee on Standards for Education Evaluation (JCSEE), 2015)	MX	63% in-service teachers and 37% preservice teachers	K-12	NA	Panel of 24 North American educational assessment 404 Canadian teachers experts
14.	Deneen and Brown (2016)	Becoming assessment literate requires developing an understanding of theory and application of diverse assessment practices and skills to appropriately administer and interpret assessments at the classroom and jurisdictional levels [. . .] This is not, however, a values-free definition	US	Examine how and to what degree a graduate-level course designed to enhance assessment literacy among practicing and pre-service teachers interacts with and mediates conceptions of assessment	Assessment Education for Teacher Candidates Teacher Conceptions	MX	TE	NA	NA	32
15.	DeLuca et al. (2016b)	Teacher assessment literacy (i.e., teacher competency in educational assessment) is a professional requirement within the current accountability framework of public education across many parts of the world [. . .] Assessment literacy involves the ability to construct reliable assessments and then administer and score these assessments to facilitate valid instructional decisions anchored to state or provincial educational standards (Stiggins, 2002, 2004; Popham, 2004, 2013)	US, CA, AU, NZ, UK	Analyze assessment literacy standards from five English-speaking countries (i.e., Australia, Canada, NZ, UK, and USA) plus mainland Europe to understand shifts in the assessment landscape over time and across regions and analyze prominent assessment literacy measures developed after 1990	NO	QL	No	No	No	No

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context						
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
16.	Xu and Brown (2016)	Assessment literacy is dependent on a combination of cognitive traits, affective and belief systems, and socio-cultural and institutional influences, all of which are central to teacher education.		Reconceptualize teacher assessment literacy (AL) by connecting two fields of research: educational assessment and teacher education	Willis et al.'s (2013) definition of AL and DeLuca (2012) framework	QL	TE & TP	No	No	No
17.	Edwards (2017)	Assessment knowledge and skills	NZ	Develop an analytical rubric of Summative Assessment Literacy, and use the rubric to track the development of summative assessment literacy	Assessment Education	QL	TE	SE	Science	8
18.	Charteris and Dargusch (2018)	Assessment capability includes building awareness of and skill in the variety of assessment modes and ensuring understanding of validity, reliability, and task design in assessments	AU	Deploy the notion of practice architectures to consider the richness of higher education and schooling settings, and the challenge of supporting the skills, knowledge, and dispositions of emerging practitioners in assessment	Practice theory (Kemmis et al., 2014, 2017)	No	TE	PE	No	No
19.	Looney et al. (2018)	Assessment literacy is usually broadly defined, encompassing both assessment knowledge and skills related to the teacher practice (Stiggins, 1995; Popham, 2009) as well as the use and interpretation of evidence to inform instruction, generate feedback, guide student learning, and report student achievement	NZ	Define teacher assessor identity	Teacher assessment identity	QL	No	No	No	No

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
20.	Coombs et al. (2018)	Wills et al.'s, definition of AL	CA, US	Examine the relationship between teachers' approaches to assessment across a set of dimensions (including their conceptions of assessment purposes, processes, fairness, and measurement theory) and career stage	Teachers' approaches to assessment	QT	TE & PD	No	No	727
21.	DeLuca et al. (2018)	Teachers' approaches to assessment are shaped by several factors including their previous experiences with assessment (as students and teachers), their values and beliefs on what constitutes valid and useful evidence of student learning, their knowledge of assessment theory, and the (Popham, 2004; Harrison, 2005)	CA, US	Analyze classroom teachers' differential responses to contemporary assessment scenarios and explore their perceived skill related to specific assessment responsibilities relevant to accountability and standards-based frameworks of education	Assessment Education and Classroom Assessment Standards	QT	TE & PD	K-12	No	404
22.	Herppich et al. (2018)	[. . .] teachers' assessment competence as a measurable cognitive disposition that is acquired by dealing with assessment demands in relevant educational situations and that enables teachers to master these demands quantifiably in a range of similar situations in a relatively stable and relatively consistent way	DE, AU, LU	Defining the subject, scope, and contextual framing of the teacher assessment competence model.	Analytic conceptualization of professional competence	No	No	No	No	No
23	DeLuca et al. (2019a)	Willis et al.'s (2013) definition of AL	CA	Explore teachers' approaches to assessment by examining their differential responses to common classroom assessment scenarios.	Assessment Education Teacher assessment literacy Teacher conceptions	QT	TE	PE, SE	No	453
24	DeLuca et al. (2019b)	Teacher assessment approaches definition (ACAI)	CA	Examine the relationship between teacher candidates' mindsets about learning and their approach to classroom assessment	Intelligence Theory Approaches to the assessment framework	QT	TE	PE		396

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
25.	Gotch and McLean (2019)	Xu and Brown's (2016) framework of AL	US	Examine teachers' knowledge of assessment development, consistent with fulfilling the classroom summative and interim/benchmark roles within Herman (2016) comprehensive assessment system	Xu and Brown's (2016) framework of AL Herman's comprehensive assessment system (2016)	QT	PD	PE, SE	No	144
26.	Pastore and Andrade (2019)	Assessment literacy is an interrelated set of knowledge, skills, and dispositions that a teacher can use to design and implement a coherent and appropriate approach to assessment within the classroom context and the school system	NO	Gather experts' feedback on the clarity, completeness, and usefulness of a new definition of teacher assessment literacy that could inform the design of a new model	Holistic approach competence	QL	No	No	No	35, 27 educational assessment experts
27.	DeLuca et al. (2019c)	Assessment capability involves situated professional judgment, that is the ability to draw on learning and assessment theories and experiences to purposefully design, interpret, and use a range of assessment evidence in the service of student learning	CA, AU, NZ, E	Characterize the complex state of assessment education in four country regions by describing the influences across vertical and horizontal knowledge systems in Australia, Canada, NZ, and England	Bernstein's theory	No	TE	No	No	No
28.	DeLuca et al. (2020)	Contemporary conceptions of assessment competence have begun to recognize the socio-cultural nature of teacher assessment knowledge, beliefs, and practices as situated within teaching and learning contexts and as influenced by multiple factors including policy requirements, teacher professional development, learning environment, and teacher-student negotiations (Tierney, 2006; Willis et al., 2013)	CA, DE	Explore how beginning teachers in two distinct educational cultures—Germany and Canada—approach classroom assessment. Investigate the influence of educational culture as one mechanism behind the development of teacher assessment competence	Socio-cultural perspective of assessment competence	QT	TE	PE, SE	Science, English, Maths, SE Certification	206 (CA) 182 (DE)

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
29.	Barnes et al. (2020)	A shared characteristic of these emerging conceptualizations is an understanding that teachers' assessment literacy is: (a) situated within social, cultural, and policy contexts; (b) a multidimensional construct, including assessment knowledge, assessment philosophy, and affective dimensions; and (c) shaped by teachers' learning experiences, personal dispositions, and theoretical orientations to teaching and learning	US	Analyze teacher candidates' developing approaches to classroom assessment at three points during their preservice program: (a) before explicit instruction in assessment through a one-credit assessment course; (b) upon completion of the assessment course and (c) after completing a full-time, 10-week student teaching experience	Approaches to the assessment framework	QT	TE	SE	No	34 teacher candidates and 23 co-operating teachers
30	Schneider et al. (2020)	Willis et al.'s (2013) definition of AL	CA, DE	Examine the role and influence of personality on pre-service student teachers' approaches to classroom assessment across two country samples	Socio-cultural approach	QT	TE	PE, SE	Science, English, Maths, SE Cert.	206 (CA) 182 (DE)
31.	Adie et al. (2020)	Assessment literacy is broadly described as "the skills and knowledge teachers require to measure and support student learning through assessment" (DeLuca et al., 2016a, p. 248)	NZ	Investigate whether Standard 5 of APST captures the assessment knowledge, procedural knowledge and skills (also known as assessment literacy), and professional judgment as developmental skills responsive to context and reflexive action	Professional standards, teachers' assessment literacies, and models of expertise	QL	No	No	No	No
32.	Akhtar et al. (2021)	Assessment literacy is the set of knowledge, skills, and competence of an individual to distinguish between appropriate and inappropriate methods of assessment	PK	Investigate the extent to which the teachings of assessment and evaluation prepare the prospective teachers to be assessment literate in Pakistan	No	MX	TE	No	No	18 tutors and 344 prospective teachers

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
33.	Christoforidou and Kyriakides (2021)	Assessment literacy has been defined as an understanding of the principles of sound assessment (Stiggins, 2002; Popham, 2004), emphasizing specific assessment knowledge and skills that teachers must possess to be effective	CY	Examine the impact of two different approaches to TPD [i.e., the Competency-Based Approach (CBA) and the Dynamic Approach (DA)] on teacher assessment skills and students' learning outcomes	Measurement framework (EER) Xu and Brown's (2016) framework of AL	QT	TE	PE	No	178
34.	Doyle et al. (2021)	Xu and Brown's (2016) framework of AL	IE	Reimagine Teacher Assessment Identity (TAI) through the lens of Xu and Brown's (2016) conceptual framework of teacher assessment literacy in practice	Xu and Brown's (2016) framework of teacher assessment literacy	MX	TE	PE	BREHM program Religious Education CBA	96
35.	DeLuca et al. (2021)	Approaches to assessment as teachers' philosophical and theoretical orientations toward 12 assessment dimensions that shape how they practice classroom assessment within their sociocultural and policy contexts	US, CHN, CA	Provide initial insights into teachers' approaches to assessment in the U.S., China, and Canada	ACAI	QT	No	No	No	227 (US) 250 CHN) 233 (CA)
36.	Rogers et al. (2022)	[. . .] the concept of assessment literacy [. . .] involves the skills and knowledge necessary to "gather information about student learning to inform education-related decisions" (National Taskforce on Assessment Education for Teachers, 2016)	US	Drawing on frameworks of assessment literacy analyze key aspects of formative assessment	Assessment literacy and novice teacher learning.	QL	TE	PE, SE	Literacy	3
37.	Oo et al. (2022)	Teacher assessment literacy comprises the knowledge and skills to make highly contextualized, fair, consistent, and trustworthy assessment decisions to inform learning and teaching to effectively support both students' and teachers' professional learning (Alonzo et al., 2021, p. 58)	No	Examine extant ITE assessment programs	ITE and literature on assessment literacy	QL	TE	No	No	No

(Continued)

10.3389/feduc.2023.1217167

TABLE 1 (Continued)

References		Definition of AL	Context							
			Country	Research aims	Theoretical/ conceptual framework	Research design	Study participants	Educational setting	Subject matter	Sample size
38.	Clark et al. (2022)	Popham's assessment literacy definition: "Assessment literacy consists of individuals' understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (2011, p. 267)	US	Explore teachers' diagnostic assessment literacy	Popham's assessment literacy definition	QL	No	No	No	17
39.	Ye (2022)	Willis et al.'s (2013) definition of AL	CHN	Investigate the characteristics of urban public teachers' assessment literacy in China	No	MX	No	SE	Morality and the Rule of Law	9
40.	Yan and Pastore (2022)	Pastore and Andrade's (2019) model of AL	HKG	Develop a self-report instrument to comprehensively assess teachers' formative assessment literacy; and examine the reliability and validity of that instrument.	Formative assessment and teacher assessment literacy	QT	No	PE, SE	NO	336 (PE) 145 (SE)
41.	Atjonen et al. (2022)	Xu and Brown's (2016) framework of AL	FI	Examine Finnish student teachers' assessment literacy	Xu and Brown's framework of AL.	QL	TE	PE	All	168
42.	Coombs et al. (2022)	[. . .] approaches to assessment is the ideal framework for conceptualizing how teachers think about and practice classroom assessment against the backdrop of educational reforms aimed at reshaping teachers' assessment practices	CHN	Investigate Chinese teachers' conceptions of classroom assessment and perceived skills	ACAI	QT	PD	PE, SE	No	746

Research design: QT = Quantitative; QL = Qualitative; MX = Mixed methods; Study participants: TE = Teacher Education; PD = Professional Development; Educational setting: K-12 = compulsory education; PE = Primary Education; SE = Secondary Education.

10.3389/feduc.2023.1217167

2022), three studies aimed to define a model of assessment literacy (Willis et al., 2013; Xu and Brown, 2016; Herppich et al., 2018; Pastore and Andrade, 2019), and two theoretical articles (Willis et al., 2013; Charteris and Dargusch, 2018). Among the review studies, two articles proposed a reconceptualization of the teacher assessment literacy construct (Xu and Brown, 2016; Looney et al., 2018).

Quantitative studies, instead, are generally purposed to examine teacher assessment literacy using standardized instruments such as questionnaires, inventories, or scales. The attempts to analyze how teachers develop assessment literacy focusing on differences between novices and experts can be traced back in time, especially in the US context where professional standards for teaching exerted a great influence on the debate on assessment literacy. The review of assessment literacy measures and the analysis of their psychometric evidence performed by Gotch and French (2014), for example, is aligned with this perspective. In their study Hailaya et al. (2014) used the Assessment Literacy Inventory (ALI) of Campbell et al. (2002). Other studies (Beziat and Coleman, 2015; Gotch and McLean, 2019), instead, used the Teacher Assessment Literacy Questionnaire (TALQ) of Plake et al. (1993). A more recent instrument linked to teacher standards (Classroom Assessment Standards, Joint Committee on Standards for Education Evaluation (JCSEE), 2015) is the Approaches to Classroom Assessment Inventory (ACAI) developed by DeLuca et al. (2016a) and used also in comparative studies (Coombs et al., 2020, 2022; DeLuca et al., 2021). A different instrument instead is offered in the study of Yan and Pastore (2022) who developed a scale focused only on teacher formative assessment literacy.

Most of the studies were conducted in English-speaking countries (e.g., the US, Canada, Australia, and New Zealand), while only three studies were located in the European area (Herppich et al., 2018; Christoforidou and Kyriakides, 2021; Atjonen et al., 2022).

Pre-service teacher students represent the target group more frequently investigated. In total, 11 studies are specifically focused on primary education, while 15 are on secondary education. Only five studies, instead, pondered in-service teachers and two of these studies consider both pre-service and professional development paths. Only two research studies consider, besides teacher candidates, teacher educators (Hill et al., 2014) and tutors (Akhtar et al., 2021).

The sample size, in quantitative and comparative studies, ranges from a minimum of 26 participants (Beziat and Coleman, 2015) to a maximum of 746 (Coombs et al., 2022). With the exclusion of the study of Atjonen et al. (2022) who collected data form 168 participants, generally, qualitative studies have a very reduced sample size (from 3 to 17 participants).

5.1. Teacher assessment literacy definitions

In the following, an overview of the expressions used in the selected articles to indicate teacher assessment literacy is presented. Then, a comparative and contrastive analysis of teacher assessment literacy definitions is provided. Finally, theoretical/conceptual frameworks and research methods used in the reported studies are analyzed.

There are four main expressions in the reviewed studies: teacher assessment literacy, teacher assessment competence, teacher assessment capability, and teacher assessment approaches. Only one study used the expression teacher assessment knowledge and skills (Edwards, 2017).

Teacher assessment literacy is the most frequent expression: it is used in 30 studies out of 42.

The expression teacher assessment capability, instead, is reported in five studies. A contextual element has to be highlighted in this case: with the exclusion of the comparative analysis of teacher professional standards in the article of DeLuca et al. (2016b), the studies which use the term assessment capability are generally located in Australia and New Zealand.

The expressions teacher assessment competence and teacher assessment approaches appeared in two (Herppich et al., 2018; DeLuca et al., 2019b) and four articles (DeLuca et al., 2018, 2019b, 2021; Coombs et al., 2022), respectively. All of these last articles report studies conducted using the Approaches to Classroom Assessment Inventory (ACAI).

A careful reading of teacher assessment literacy definitions shows how the identification of teacher assessment literacy with the basic principles of a sound assessment practice in the classroom, generally found in early definitions (Stiggins, 2002; Popham, 2011), is present also in more recent studies (Clark et al., 2022). Thus, while, for example, DeLuca and Bellara (2013) point out how assessment literate teachers integrate assessment practices, theories, and philosophies to support teaching and learning within a standards-based education framework, other authors (Gotch and French, 2014; Clark et al., 2022) recall, in their definitions of teacher assessment literacy, the importance of "understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions" (Popham, 2011, p. 267). This cluster of definitions presents a very essential conceptualization of assessment literacy influenced by teacher professional standards, as well as by the attempts to measure teacher assessment literacy levels. If, on the one hand, assessment literacy is defined only in terms of knowledge and skills to be implemented by teachers in their classrooms (Kaden and Patterson, 2014; Fulmer et al., 2015; Edwards, 2017), then, on the other hand, some authors focus their attention on assessment literacy measures (Gotch and French, 2014; Hailaya et al., 2014; Kaden and Patterson, 2014; Beziat and Coleman, 2015). Teacher assessment literacy definitions in this first cluster have a clear check-list nature which is influenced by the exert of professional standards as policy mechanisms for leveraging teacher development and educational quality. The concern behind these conceptualizations is to ensure a sound assessment practice in the classroom. The identified components (e.g., construct assessment tasks, interpret, report, and communicate assessment results) are, therefore, very essential to allow teachers to enact required assessment practice.

Two phenomena, however, impressed a substantial change in this way of defining teacher assessment literacy: the widespread interest in formative assessment and the recognition of the socio-cultural nature of assessment practice. The change reported in Willis et al.'s, study where assessment literacy is defined as a "dynamic context-dependent social practice that involves teachers articulating and negotiating classroom and cultural knowledges

with one another and with learners, in the initiation, development and practice of assessment to achieve the learning goals of students" (Willis et al., 2013, p. 242). Over the years, it is possible to detect the impact of this definition on the studies published after 2013. Within the articles of the present review, for example, Wills et al.'s definition of assessment literacy not only is reported in the theoretical/conceptual framework by the majority of the articles but six studies used this definition (Poskitt, 2014; Coombs et al., 2018; DeLuca et al., 2019a; Schneider et al., 2020; Clark et al., 2022; Ye, 2022). However, also the change impressed by the assessment for learning perspective in teacher assessment literacy conceptualization is evident. In their attempt to develop an instrument reflective of contemporary assessment practices and contexts (i.e., the JCSEE Classroom Assessment Standards of 2015), for example, DeLuca et al. (2016a) consider assessment literacy as "the skills and knowledge teachers require to measure and support student learning through assessment" (p. 248). Other authors, such as Looney et al. (2018), Adie et al. (2020), and Oo et al. (2022), instead, have stressed, in their definitions of assessment literacy, the use of assessment to support student learning. In this vein, it is interesting to note that three studies focused on specific assessment practices. While Edwards (2017) developed an analytic rubric for summative assessment literacy and Rogers et al. (2022) investigate the key aspect of formative assessment literacy, Yan and Pastore (2022) designed and validated a self-reported scale to assess teacher formative assessment literacy.

Since the publication of Willis et al.'s, study, more articulated and sophisticated definitions with different components have been created.

The first attempt, in this perspective, is represented by the study of Xu and Brown (2016) who provide a reconceptualization of teacher assessment literacy. Merging the research fields of education assessment and teacher education, these authors conceive assessment literacy as a combination of "cognitive traits, affective and belief systems, and socio-cultural and institutional influences" (p. 155) With an emphasis on assessment practice, the Teacher Assessment Literacy in Practice (TALIP) framework, which has a hierarchical structure and a cyclical nature, is functional to teacher learning progression in the educational assessment domain. The TALIP model has been found in the other three articles of the present review. These studies have in common the same aim: explore teacher assessment literacy development. More specifically, Gotch and McLean (2019) used the TALIP framework to examine the outcomes of a state education agency-sponsored teacher development initiative in the US. Doyle et al. (2021), instead, resorted to this model to reconsider the teacher's identity as an assessor; Atjonen et al. (2022), finally, used it to analyze Finnish student teachers' assessment literacy progression.

Among the new assessment literacy conceptualization attempts, there is the study provided by Looney et al. (2018). These authors affirm the need to investigate teacher identity as an assessor, as well as the role of conceptions, beliefs, experiences, and feelings to better understand assessment practices. As pointed out by Xu and Brown (2016), the focus on teachers' perspectives on values leads to questioning teacher assessment practice through the lenses of teacher identity and agency.

Aligned with the last two attempts to define teacher assessment literacy, the model of Pastore and Andrade (2019) emphasizes, instead, the socio-contextual, cultural, relational, and emotional dimensions of assessment practice. More specifically, assuming a holistic and adaptative competence perspective, this model has not a hierarchical structure. Its three-dimensional architecture (conceptual, practical, and socio-emotional assessment dimensions) is connected with local contextual factors, including teachers' professional wisdom and practice, and school and classroom contexts.

These last conceptualizations of assessment literacy have in common the attempts to balance very different aspects and components. Personal and professional identity, agency, and therefore aspects such as conceptions, values, beliefs, and ethical and moral responsibilities are all recognized as relevant for teachers who are called to develop new assessment repertories and practices in the classroom as a consequence of policy and social changes.

The more recent definitions of teacher assessment literacy are similar, especially in their components, to the definition of assessment competence reported in the study by DeLuca et al. (2020). Recognizing the socio-cultural nature of assessment, this competence corresponds to "knowledge, beliefs, and practices as situated within teaching and learning contexts and as influenced by multiple factors, including policy requirements, teacher professional development, learning environment and teacher-student negotiations" (p. 27). A slightly different perspective, instead, is offered by Herppich et al. (2018) who developed a very detailed analytical conceptualization of teacher assessment competence. Within teacher professional competencies, the assessment one is defined as a "measurable cognitive disposition that is acquired by dealing with assessment demands in relevant educational situations and that enables teachers to master these demands quantifiably in a range of similar situations in a relatively stable and relatively consistent way" (p. 185).

Teacher assessment capability definitions (Gunn and Gilmore, 2014; Hill et al., 2014) tend to emphasize the assessment from a learning perspective, as well as the role of student involvement and motivation. The notion of assessment capability, recalling the work of Absolum et al. (2009), includes the "ability to develop assessment that transform learning goals into assessment activities that accurately reflect student understand and achievement" (p. 9). Among these definitions, Cowie et al. (2014) also consider the teacher's ability to interpret assessment results and use these data to adjust and adapt instruction to student learning needs. In a very similar way, DeLuca et al. (2019c), recalling Bernstein's theory, point out how an assessment capable teacher makes a professional judgment based on learning and assessment theories and experiences. All these elements are functional to support teachers in designing, interpreting, and using assessment evidence in the service of student learning.

The ultimate expression found in this literature review is teacher assessment approaches. The first definition of teacher assessment approaches provided by DeLuca et al., in 2018 tends to largely overlap with the assessment capability definitions, as well as with the assessment literacy definitions provided by Xu and Brown (2016) and Looney et al. (2018). First of all, assessment approaches are referred only to the classroom context. Second, the

approaches to assessment are influenced by different factors such as "experiences with assessment (as students and teachers), their values and beliefs on what constitutes valid and useful evidence of student learning, their knowledge of assessment theory, and the prevalence of systemic assessment policies" (DeLuca et al., 2018, p. 367). In a more recent definition provided by Coombs et al. (2022), instead, an interesting element is added: teacher assessment approaches, (should) represent a glimmer of how teachers conceive and practice assessment against the backdrop of educational reforms which generally tend to reframe assessment practices.

5.2. Teacher assessment literacy components

The first and more conspicuous cluster of definitions of assessment literacy ($N = 16$) felt the effects of the professional standards perspective. Following Stiggins (1991) and Popham (1991), and a practical-oriented view, these definitions are focused on very essential aspects:

- Assessment design, construction, administration, and scoring;
- Interpretation and use of assessment results in support of instructional decision-making, as well as of student learning;
- Reporting and communicating assessment results.

The required set of knowledge and skills allows teachers to ensure a sound classroom assessment practice. It is interesting to note how the majority of these conceptualizations (Gotch and French, 2014; Kaden and Patterson, 2014; Deneen and Brown, 2016; DeLuca et al., 2016a,b; Edwards, 2017; Akhtar et al., 2021; Clark et al., 2022; Oo et al., 2022; Rogers et al., 2022) regard pre-service and novice teachers, and the efforts made in terms of initial teacher education to support teachers reaching a "professional requirement" (DeLuca et al., 2016b).

A different panorama emerges, instead, when the focus is on the second cluster which groups the farther conceptualizations of teacher assessment literacy.

Conversely, in previous conceptualizations, the definition of Willis et al. (2013) seems more nuanced. Although the list of essential knowledge and skills is not reported, the importance of initiation, development, and practice of assessment is pointed out by the authors. Stressing the socio-cultural and contextual nature of assessment, this definition highlights the importance of teachers' negotiation and mediation in the assessment process.

Moving from a clear recognition of the socio-cultural perspective addressed by Willis et al. (2013), in their definition of assessment literacy, Xu and Brown propose a hierarchical structure that balances the inclusion of "hard" and "soft" assessment components:

- A wide teacher knowledge base;
- Institutional and socio-cultural contexts;
- Teacher literature in practice;
- Teaching and learning;
- Teacher identity (re)construction;

This model has without doubt the merit of providing a detailed picture of assessment literacy components that act at different levels in the construction of teacher professional identity; a perspective that has been stressed by Looney et al.'s (2018) model of teacher assessor identity. Moreover, in this case, the foundational components, which, traditionally, defined assessment literacy (in terms of knowledge and skills), represent the base of an effective assessment practice. However, these aspects seem not sufficient to describe a competent practice. The emphasis on the assessor's identity catalyzes what was just pointed out by Willis et al.: assessment is a social context-dependent practice and it is expected that an assessment literate teacher knows how to differentiate and calibrate this practice in response to different situations and circumstances. The awareness of teaching as an evolving profession(alism) is the key to understanding these last conceptualizations of assessment literacy.

The model provided by Pastore and Andrade (2019), among the last definitions of assessment literacy, is aligned with this perspective. However, in this case, assessment literacy is conceived not in a hierarchical, bottom-up view, but as a nested interplay of components (i.e., knowledge, skills, and dispositions), grouped in three dimensions. The inner adaptative structure of this model allows teachers to be responsive to different educational contexts and to learn or refine/review knowledge, skills, and disposition that can emerge over time in response to institutional reforms or instructional innovations.

It is clear how the models of this second cluster assume that assessment literacy has an inner core that includes the fundamental knowledge and skills necessary to ensure a sound assessment. However, although necessary, these components are not sufficient to competently enact assessment. Other components are also relevant, such as teachers' dispositions, beliefs, conceptions, and values. Furthermore, these last definitions of assessment literacy tend to identify as fundamental components the same aspects reported in the assessment competence definitions (Herppich et al., 2018; DeLuca et al., 2020).

Assuming that an assessment-competent teacher should master different assessment situations, Herppich et al. (2018) design their model matching assessment process and product perspectives. More specifically, the authors here emphasize the role of the assessment process. This process can be performed by teachers systematically or not, and can be differentiated considering aspects such as teacher assessment aims, teacher assessment activities planning, and teacher assessment-based decision-making. Although a strong similarity appears between this model and Mandinach and Gummer's (2016) model of data literacy, the authors, echoing the work of Xu and Brown (2016) recognize the role of dispositions which are differentiated in cognitive dispositions (e.g., content knowledge, pedagogical content knowledge, teaching and learning knowledge, etc.) and other dispositions (e.g., beliefs, subjective theories, and self-concepts).

An analogous picture emerges with the definitions of teacher assessment capabilities. Within this cluster, Charteris and Dargusch (2018) point out the importance of teacher awareness and skills in assessment practice; Hill et al. (2014) and DeLuca et al. (2019c), instead, provide a definition which in part inherits the

components of traditional assessment literacy definitions (e.g., assessment design, interpretation, and use of assessment evidence), and reproposes, following the Bernstein's theory, the importance of teacher professional judgment, the role of teacher agency, as well as the student motivation and engagement.

The last cluster includes the assessment approaches definitions. If, on the one hand, knowledge and skills continue to be considered relevant to ensure an effective assessment practice in the classroom, then, on the other hand, as reported also in other expressions or definitions (Xu and Brown, 2016; Looney et al., 2018; Pastore and Andrade, 2019), factors such as values, beliefs, conceptions, and previous assessment experiences are recognized as shaping factors of assessment practice and teacher assessor identity.

5.3. Teacher assessment literacy theoretical/conceptual frameworks

The analysis of theoretical/conceptual approaches used to frame teacher assessment literacy reveals a composite and somewhat scattered scenario (Table 2). Although the socio-cultural perspective is constantly reported in the selected studies to stress the recognition of the socio-cultural nature of assessment, the Bernstein's socio-cultural theory is used to justify only four studies. Among these, only one is empirical research; the others, instead, include the theoretical attempt of Willis et al. to conceptualize assessment literacy, the systematic review provided by Xu and Brown (2016) as a step function to design their model of assessment literacy, and the comparative study of DeLuca et al. (2019c) on vertical and horizontal dimensions of assessment education systems. Interpreting assessment through semiotic categories, Willis et al. (2013) posed assessment literacy within the socio-cultural learning perspective and pointed out the complex nature of assessment. Depicted as an ethical practice, assessment is influenced by social, cultural, and dynamic variables. Accordingly, learning in the assessment domain is a complex process. In this vein, vertical structures which are fixed in official, schooled, formalized, and hierarchical learning are paralleled by the recognition of the importance of horizontal structures which, instead, indicate the context-dependent, tacit, experiential learning. The effects of this perspective on teacher assessment education have been remarkable because the traditional idea of preparing teachers to ensure sound assessments is replaced with the recognition of the situated nature of assessment and the representation of this practice as a critical process of inquiry performed by teachers. Thus, Willis et al. expanded what is meant by assessment literacy and laid the foundations of a new, complex, conceptualization of assessment literacy. Overcoming assessment knowledge and skills, the development of assessment literate teachers is linked to the professional wisdom of teachers, as well as to their values, conceptions, and their ethical and moral responsibilities. The impact of the socio-cultural approach becomes particularly evident in the studies which have investigated the role of teacher conceptions, values, beliefs (Gunn and Gilmore, 2014; Deneen and Brown, 2016), teacher agency (Clark, 2015), and teacher identity (Looney et al., 2018; Doyle et al., 2021). More specifically, the widespread interest in teacher assessment conceptions (Brown, 2004; Deneen and Brown, 2016) can be justified because conceptions represent, since ever in educational research, a crucial and powerful access to the modalities (how) and purposes (why) of teachers' practices. The attempts to identify teacher assessment conceptions have been relevant in understanding the dynamics involved in the implementation of educational policies.

TABLE 2 Overview of theoretical/conceptual approaches in the selected studies.

Theoretical/conceptual approach	Studies
Bernstein's socio-cultural theory	1, 18, 25, 30
Xu and Brown's (2016) TALIP	25, 33, 34, 41
Pastore and Andrade's (2019) 3-dimensional model	40
Guskey's model of professional learning evaluation (2002)	9
Carless' theoretical framework (2005)	9
Kozma's conceptual model (2003)	13
Teacher agency	11
Teacher identity	19
Practice theory	18
Initial teacher education framework or pre-service assessment education	2, 3, 4, 5, 14, 15, 17, 21, 23, 36, 37
Teacher assessment inventories	9 10, 35, 42
Teacher professional standards	13, 21, 31
Teacher' conceptions/approaches	14, 20, 29
Teacher competence	23, 28

A discrete number of studies move within the initial teacher education framework. Other studies, instead, rely on teacher professional standards as conceptual framework. However, all these studies have had a somewhat light impact on the conceptualizations or definitions of teacher assessment literacy. Indeed, the focus on initial teacher education programs and teacher professional standards led to exploring the extent to which teacher education in the assessment domain is aligned with policy expectations/requirements. Therefore, these studies although relevant in examining drivers and challenges in preparing "assessment-ready" teachers tend to reduplicate a practical view of assessment literacy. Furthermore, these studies, generally, are very close to studies framed within the perspective of teacher assessment inventories/instruments.

The teacher competence approach is recalled only in two studies: the first one, rooted in the analytical perspective (Herppich et al., 2018), assumes that individual elements of competence may be developed and improved by external

intervention (Rotthoff et al., 2021). The second one (Pastore and Andrade, 2019), instead of assuming a holistic perspective on assessment literacy, considers it as linked with other competencies and makes reference to complex real-life situations. The teacher competence approach, both in the analytic and holistic versions, is very close to the socio-cultural perspective. Not surprisingly, the same aspects (e.g., personal and contextual variables, and the role of negotiation and mediation) are considered relevant in the development of teacher assessment literacy.

Two interesting perspectives are offered by the literature review of Fulmer et al. (2015) who used Kozma's model as a conceptual framework to address the different contextual factors (at micro, meso, and macro level) that can influence teacher assessment literacy and classroom assessment practice, and by the study of Charteris and Dargusch (2018) who deploy the use of practice theory to investigate how, and to what extent, schooling settings affect teacher assessment capability development. More specifically, Kozma's framework, in the study of Fulmer et al., is used to investigate the relationships among teachers' views, knowledge, and practices to identify how to support teachers in developing assessment practices. The emphasis on contextual factors in this model tends to overlap with the socio-cultural perspective and its emphasis on teachers' values, conceptions, and knowledge. Kozma's notion of different levels of influence on practice evokes, in fact, the vertical and horizontal structures identified by Willis et al. as operating in the assessment domain. Recalling the importance of teacher identity as an assessor and the role of contexts in assessment capability development, Charteris and Dargush reflect on the schooling practice architectures in terms of cultural-discursive, material-economic, and socio-political arrangements. With a focus on initial teacher education programs, the authors refer to the concept of assessment capability as a concept directly influenced by teacher agency and identity. The development of the assessment capability is, therefore, deeply rooted in real contexts and encompasses social processes of mediation, interaction, and cultural negotiation. The stress on these aspects, however, is not new compared to the perspective of Willis et al.

Some studies ground on the new conceptualizations of assessment literacy. The study of Gotch and McLean, for example, merges the TALIP framework (Xu and Brown, 2016) with the perspective of a comprehensive assessment system developed by Herman's (2016) and investigates the role of summative and interim assessment practice. In their study, Yan and Pastore (2022), instead, on the backdrop of formative assessment literature, used the Pastore and Andrade (2019) three-dimensional model of assessment literacy to design and validate a scale on teacher formative assessment literacy.

Some other studies, instead, proposed remarkable intersections between different conceptual frameworks. For example, Christoforidou and Kyriakides (2021) adopt a measurement framework in the educational effectiveness research to quantitatively detect the teacher assessment literacy characteristics as defined in the TALIP model of Xu and Brown (2016). Furthermore, comparing and contrasting the dynamic and the competence-based approaches, the authors try to highlight which aspects should be considered for an effective professional development path.

TABLE 3 Overview of data collection methods.

Data collection methods	Studies
Document analysis	2, 13, 27, 31, 32, 41, 42
Focus groups	4, 38,
Interviews	4, 5, 6, 11, 14, 34
Observations	6, 8, 17, 39
Secondary data analysis	7, 12, 15, 16, 19, 36
Surveys (inventories, questionnaires, scales)	3, 4, 8, 9, 10, 20, 21, 23, 24, 25, 28, 29, 30, 32, 33, 34, 35, 39, 40, 42
Other (instruments to instigate student reflection)	4, 8, 6, 11, 26

5.4. Teacher assessment literacy research methods

Selected articles were coded according to the category of research methods used (Table 3). The analysis confirms how considerable research efforts have been made to explain, examine, and investigate assessment literacy, rather than theorize or conceptualize this teacher professional domain. Only eight studies try to define the concept of assessment literacy or to identify which components should or not be considered. These articles, except for Poskitt's (2014) study which used a mixed-method research approach, are all qualitative in nature.

The use of qualitative methods is approximately the same as quantitative methods, and these research methods are more commonly used than mixed methods. More specifically, 10 qualitative studies try to shed light on different aspects, namely: the alignment of teacher education policy, teacher professional standards, and teachers' expectations (DeLuca and Bellara, 2013); and the effects of pre-service courses or professional standards on teachers' assessment literacy (DeLuca et al., 2016b, 2019c; Adie et al., 2020; Atjonen et al., 2022; Oo et al., 2022), teachers' conceptions (Cowie et al., 2014; Clark et al., 2022; Rogers et al., 2022), or teachers' assessment decisions (Clark, 2015). Therefore, 14 studies out of 15 quantitative studies report systematic attempts to investigate assessment literacy components; nine of these studies analyze how teachers develop their assessment approaches focusing on personal or contextual components. In the first case, variables such as teachers' mindsets about learning (DeLuca et al., 2019b), the career stage (Coombs et al., 2018), or the personality of pre-service teachers (Schneider et al., 2020) can be recalled. In other studies, the focus is on contextual factors such as the influence of educational policy and educational cultures/systems (DeLuca et al., 2018, 2020, 2021). Only 2 studies examine the impact of teacher education on assessment literacy (Gotch and McLean, 2019; Christoforidou and Kyriakides, 2021); while two studies explore the efficacy of education systems on professional programs designed to prepare assessment literate teachers (Hailaya et al., 2014; Beziat and Coleman, 2015).

It is important to note that studies with a quantitative or mixed-method approach tend to use standardized or structured instruments for data collection (e.g., questionnaires, inventories, and scales). The latent assumption is that research performed

with this kind of data collection method can be more credible or scientific than observational, qualitative research has to be pointed out. However, a likely explanation could be also related to the traditional modality that has been used to investigate teacher assessment literacy (e.g., inventories and scales to measure and compare teachers' assessment literacy levels). While Gotch and French (2014) offer a review of past available assessment literacy measures, over the 10 years considered in the present review, only three new instruments have been developed. The ACAI instrument (DeLuca et al., 2016a), widely used also in cross-countries studies, offers a very refined standardized measure of assessment literacy. However, this assessment literacy concept is derived from policy documents and maintains a strong connection with professional standards in the classroom. The scale developed for teacher formative assessment literacy (Yan and Pastore, 2022), instead, has a theory-driven nature (formative assessment and teacher assessment literacy). Edwards (2017), finally, combining deductive observations from the literature and inductive observations based on findings from a case study, developed an analytical rubric of summative assessment literacy. This instrument is purposed to support preservice teachers with summative assessment practice.

Document analysis and secondary data analysis are reported in qualitative studies (including systematic review studies). Generally, these kinds of data collection methods are used also when the research focus is on policy documents (i.e., professional teacher standards). The other instruments used in qualitative studies are interviews, observations, and focus groups, respectively.

The analysis of research and data collection methods shows more clearly the prevalence of descriptive studies on teacher assessment literacy.

6. Discussion

The interest in preparing teachers to effectively practice assessment can be traced back to the years when the first definition attempts were mostly (even not exclusively) focused on practical issues/aspects (knowledge and skills necessary to test and evaluate student performance in the class). Not surprisingly, for Roeder (1972), assessment literacy corresponded to the proper use of evaluation techniques by teachers. Originally limited to measurement and testing practice (Gullickson and Hopkins, 1987), assessment literacy has been provocatively recalled by Stiggins (1991) and Popham (1991) as a crucial element to restrain the US educational crisis, and for a long period, this concept meant to be familiar with and knowledgeable about assessment.

The pressure exerted, first, by the global assessment policy environment and the associated discourses of accountability, and then, by the rapid diffusion of the assessment for learning agenda led to deem assessment literacy as a priority for teacher practice and teacher education. Therefore, what a teacher is expected to know and practice in terms of assessment begins to be more and more complex and demanding. The term assessment literacy, first limited to the assessment fundamentals, has progressively overlapped with the competency concept as a set of knowledge, skills, and dispositions germane to teacher assessment practice.

The widespread diffusion of the socio-cultural perspective (Willis et al., 2013), along with the practice turn (Schatzki et al., 2001; Kemmis et al., 2016) and the recognition of the context relevance in the development of teacher professionalism have impressed a substantial acceleration in this re-semantic process. As the contrastive analysis showed, assessment literacy and assessment competence are now used as synonymous. Furthermore, the emphasis added by the socio-cultural approach on aspects, such as the personal identity of teachers as assessors, and the role of beliefs, values, and conceptions led to more sophisticated working definitions of teacher assessment literacy. The teacher assessment approaches expression which incorporates the essence of assessment literacy and competence definitions stressed this perspective indicating the plural modalities through which teachers can interpret and enact assessment.

While teacher assessment literacy, competence, and approaches have a similar meaning and tend, especially considering the more recent definitions, to include the same foundational components, a discourse a part has to be made for the expression of teacher assessment capability. While some authors (Gunn and Gilmore, 2014; Hill et al., 2014) define assessment capability as a set of skills and understandings a teacher needs to support student learning, it is not clear if, and to what extent, the capability approach of Sen (1989) influenced this conceptualization.

The change in the assessment literacy conceptualizations and the identification of its foundational components offers an opportunity to ascertain what counts for teachers to be competent in the assessment domain. Conversely, to past definitions, the focus on teachers' conceptions demonstrates how, among the different theoretical and conceptual approaches identified in this review, the socio-cultural perspective really impressed a considerable shift in the assessment domain.

Drawing upon the socio-cultural perspective, assessment literacy is understood as socially distributed, context-dependent, and embedded in cultural artifacts, objects, and people (e.g., professional standards, school organization, students, and colleagues). Over the past years, research studies have, in a redundant way, pointed out the assessment illiteracy of teachers (both pre-and in-service) and called for improvement of teacher preparation to practice assessment. The new conceptualizations of assessment literacy, far from a check-list approach sometimes flattened against professional standards, instead, reinforce the idea of a dynamic connection between the vertical and the horizontal domains of assessment literacy and question how to investigate these aspects and their impact on the development of teacher assessment competence (Figure 2).

The detection of critical features on education/training paths, in this vein, should include institutional and social expectations (i.e., in professional standards), or teachers' professional learning needs in the assessment domain. Therefore, the great challenge for research in this field is to provide a better understanding of how teachers incorporate new ideas into their practice, how they transfer their learning into classrooms, and how pedagogies and mediating artifacts effectively drive teachers to become assessment literate.

7. Limitations

The findings of this review must be interpreted in light of the study's limitations. The first limitation pertains to the

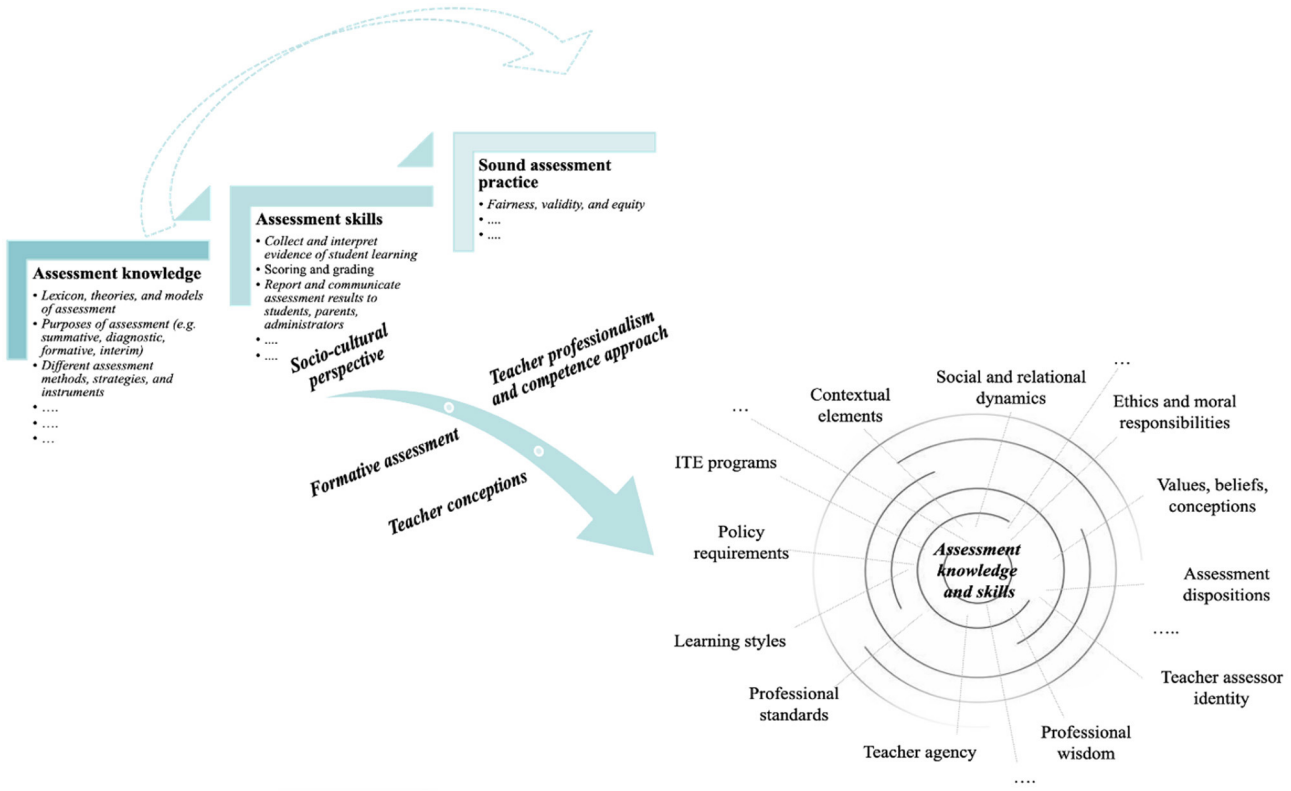


FIGURE 2
Teacher assessment literacy foundational components.

study's inclusion. Certainly, the choice to not include books, book chapters, dissertations, or other published works apart from peer-reviewed articles influenced the present findings. The search process constrained to education research journals led to omitting quality works such as, for example, the reviews provided by Barnes et al. (2015) and Bonner (2016). These well-designed studies, although captured in handbooks, overviewed the research on teachers' beliefs and conceptions about assessment and represent authoritative works in informing policy, pedagogy, and practice in teacher assessment literacy. As a threat to the viability of the present systematic review, this limitation has to be pointed out. Furthermore, only English language peer-reviewed articles were included because it was viewed as having scientific quality and rigor. Future research work should consider this other kinds of studies and other sources across languages to reduce publication bias and offer a more comprehensive understanding of teacher assessment literacy conceptualizations. Although the inclusion and exclusion criteria have been useful in the selecting literature phase, the lack of quality overview represents a further limitation. The exclusion of studies focused on the relationships between teacher assessment literacy and teacher subject domain (including English as a second language) has also to be mentioned. If, on the one hand, this limitation has been helpful in the identification of studies within the broader context of teacher professionalism and teacher education, then, on the other hand, understanding how subject-matter instruction and assessment education are related to each other and how they affect the development of teacher assessment literacy could represent a future research stream.

8. Conclusion

This review demonstrates that the theoretical shift impressed, since 2013, by the socio-cultural perspective has deeply affected how teacher assessment literacy is conceived. The recognition of personal, social, contextual, and cultural features shows how the assessment competence has a highly complex nature. Furthermore, this study highlights *a de facto* overlap between the assessment literacy and assessment competence definitions. While the growing attention deserved on teachers' conceptions seems to explain the use of the assessment approaches expression, the assessment capability appears more focused on a pedagogical version of formative assessment. Compared to other expressions, teacher assessment capability seems to not capture the complexity of current perspectives on assessment, although it recognizes the importance of cultural, social, political, and material factors in the development of teachers' identity as assessors.

The topic diversity in articles selected in this review remains stable over time, although it is possible to identify a clear trend toward specializations. The identification of what means to be an assessment literate teacher is the common core theme that is linked to different research attempts oriented to explore how to measure

this competence, how to promote it in pre- and in-service education paths, and how to manage different variables which could affect teacher learning in the assessment domain.

The current recognition of the importance of teacher assessment approaches has relevant implications in terms of educational research in both pre- and in-service contexts. Understanding how teachers deal with micro (classroom level), meso (school level), and macro (national school system) assessment situations is fundamental to ensure the design and implementation of more responsive assessment education paths. A caution, however, has to be addressed here. A strong interest in assessment conceptions, values, beliefs, risks to generate tunnel visions. An over-attention of teachers' values, beliefs, and conceptions should not replace the need for investigation on assessment education (e.g., teacher education contents, pedagogies, and professional learning drivers).

With a meaningful and powerful cross-fertilization in mind, a call for inter-sectional research between the two broad areas of teacher education and educational assessment is, in conclusion, launched. In this way, it would be possible to reduce the gap between vertical and horizontal components of teacher learning in the assessment domain bridging what teachers learn in public and formal contexts and what they need to practice assessment in daily school life. This review presents a map of the research field on teacher assessment literacy and shows how some themes become central while other themes first dominating the debate (e.g., the attempts to measure assessment literacy) become more peripherical. Overcoming descriptive questions regarding how assessment literacy can be promoted in practice, further research is necessary to determine the extent to which the components enumerated in current definitions adequately reflect a common idea of an assessment literate teacher. The debate flux, furthermore, calls also for different research attempts (especially in terms of research design). Few large-scale or cross-cultural comparative studies have been found in this review. Additional future studies would benefit from longitudinal studies to track the transfer of teacher learning about assessment to actual classroom practice. Finally, the need for a connection between different research clusters (e.g., pre- and in-service teacher education; personal and contextual factors, assessment pedagogies and curriculum; teacher assessment literacy measurement instruments and assessment literacy professional drivers; etc.) is also advocated to reinforce the development of an integrated and established teacher assessment literacy field of research.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares 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.

References

Absolum, M., Flockton, L., Hale, J., Hipkins, R., and Reid, I. (2009). *Directions for Assessment in New Zealand: Developing Students' Assessment Capabilities.* Wellington: Ministry of Education.

*Adie, L., Stobart, G., and Cumming, J. (2020). The construction of the teacher as expert assessor. *Asia Pac. J. Teach. Educ.* 48, 436–453. doi: 10.1080/1359866X.2019.1633623

*Akhtar, Z., Hussain, S., and Ahmad, N. (2021). Assessment literacy of prospective teachers in distance mode of education: a case study of Allama Iqbal Open University, Islamabad. *J. Educ. Educ. Dev.* 8, 218–234. doi: 10.22555/joeed.v8i1.114

Alexander, P. A. (2020). Methodological guidance paper: the art and science of quality systematic reviews. *Rev. Educ. Res.* 90, 6–23. doi: 10.3102/0034654319854352

Allal, L. (2020). Assessment and the co-regulation of learning in the classroom. *Assess. Educ. Princip. Policy Pract.* 27, 332–349. doi: 10.1080/0969594X.2019.1609411

Alonzo, D., Labad, V., Bejano, J., and Guerra, F. (2021). The Policy-driven dimensions of teacher beliefs about assessment. *Aust. J. Teach. Educ.* 46, 36–51. doi: 10.14221/ajte.2021v46n3.3

*Atjonen, P., Pöntinen, S., Kontkanen, S., and Ruotsalainen, P. (2022). In enhancing preservice teachers' assessment literacy: focus on knowledge base, conceptions of assessment, and teacher learning. *Front. Educ.* 7, 891391. doi: 10.3389/feduc.2022.891391

Ball, S. (2015). Education, governance and the tyranny of numbers. *J. Educ. Policy* 30, 299–301. doi: 10.1080/02680939.2015.1013271

Barnes, N., Fives, H., and Dacey, C. M. (2015). "Teachers' beliefs about assessment," in *International Handbook of Research on Teacher Beliefs*, eds H. Fives, and M. Gregoire Gill (New York, NY: Routledge), 284–300.

*Barnes, N., Gareis, C., DeLuca, C., Coombs, A., and Uchiyama, K. (2020). Exploring the roles of coursework and field experience in teacher candidates' assessment literacy: a focus on approaches to assessment. *Assess. Matters* 14, 5–41. doi: 10.18296/am.0045

*Beziat, T. L. R., and Coleman, B. K. (2015). Classroom assessment literacy: evaluating pre-service teachers. *Researcher* 27, 25–30.

Biesta, G., and Tedder, M. (2006). *How is Agency Possible? Towards An Ecological Understanding of Agency-As-Achievement, Working Paper 5.* Exeter: The Learning Lives Project.

Black, P., and Wiliam, D. (2018). Classroom assessment and pedagogy. *Assess. Educ. Princip. Policy Pract.* 25, 551–575. doi: 10.1080/0969594X.2018.1441807

Blïomeke, S., and Kaiser, G. (2017). "Understanding the development of teachers' professional competencies as personally, situationally and socially determined," in *The SAGE Handbook of Research on Teacher Education*, eds D. J. Clandinin, and J. Husu (London: SAGE Publications), 783–802.

Bonner, S. M. (2016). "Teachers' perceptions about assessment: competing narratives," in *Handbook of Human and Social Conditions in Assessment*, eds G. T. L. Brown, and L. R. Harris (New York, NY: Routledge), 21–39.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educ. Measure. Issue. Pract.* 30, 3–12.

Brooks, C. (2021). The quality conundrum in initial teacher education. *Teach. Teach.* 27, 131–146. doi: 10.1080/13540602.2021.1933414

Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assess. Educ.* 11, 301–318. doi: 10.1080/0969594042000304609

Brown, G. T. L. (2022). The past, present and future of educational assessment: a transdisciplinary perspective. *Front. Educ.* 7,1060633. doi: 10.3389/feduc.2022.1060633

Campbell, C., Murphy, J. A., and Holt, J. K. (2002). "Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers," in *Annual Meeting of the Mid-Western, Educational Research Association* (Columbus, OH).

Carless, D. (2005). Prospects for the implementation of assessment for learning. *Assess. Educa. Principl. Policies Pract.* 12, 39–45. doi: 10.1080/0969594042000333904

*Charteris, J., and Dargusch, J. (2018). The tensions of preparing pre-service teachers to be assessment capable and profession-ready. *Asia Pac. J. Teac. Educ.* 46, 354–368. doi: 10.1080/1359866X.2018.1469114

*Christoforidou, M., and Kyriakides, L. (2021). Developing teacher assessment skills: the impact of the dynamic approach to teacher professional development. *Stud. Educ. Eval.* 70, 101051. doi: 10.1016/j.stueduc.2021.101051

*Clark, A. K., Nash, B., and Karvonen, M. (2022). Teacher assessment literacy: implications for diagnostic assessment systems. *Appl. Meas. Educ.* 35, 17–32. doi: 10.1080/08957347.2022.2034823

*Clark, J. S. (2015). My assessment didn't seem real": the influence of field experiences on preservice teachers' agency and assessment literacy. *J. Soc. Stud. Educ. Res.* 6, 91–111. doi: 10.17499/jsser.91829

Cochran-Smith, M. (2023). What's the "problem of teacher education in the 2020s? *J. Teach. Educ.* 74,1–3. doi: 10.1177/00224871231160373

Cochran-Smith, M., Keefe, E. S., and Jewett Smith, R. (2021). A study in contrasts: multiple-case perspectives on teacher preparation at new graduate schools of education. *New Educ.* 17, 96–118. doi: 10.1080/1547688X.2020.1822485

*Coombs, A., DeLuca, C., LaPointe-McEwan, D., and Chalas, A. (2018). Changing approaches to classroom assessment: AN empirical study across teacher career stages. *Teach. Teach. Educ.* 71, 131–144. doi: 10.1016/j.tate.2017.12.010

*Coombs, A., DeLuca, C., and MacGregor, S. (2020). A person-centered analysis of teacher candidates' approaches to assessment. *Teach. Teach. Educ.* 87, 102952. doi: 10.1016/j.tate.2019.102952

*Coombs, A., Rickey, N., DeLuca, C., and Liu, S. (2022). Chinese teachers' approaches to classroom assessment. *Educ. Res. Policy Pract.* 21, 1–18. doi: 10.1007/s10671-020-09289-z

*Cowie, B., Cooper, B., and Ussher, B. (2014). Developing an identity as a teacher-assessor: three student teacher case studies. *Assess. Matters* 7, 64–89. doi: 10.18296/am.0128

Dall'Alba, G. (2018). "Reframing expertise and its development: a lifeworld perspective," in *The Cambridge Handbook of Expertise and Expert Performance, 2nd Edn*, eds K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams (Cambridge: Cambridge University Press), 33–39.

Darling-Hammond, L. (2016). Research on teaching and teacher education and its influences on policy and practice. *Educ. Res.* 45, 83–91. doi: 10.3102/0013189X16639597

Day, C. (2017). "Competency-based education and teacher professional development," in *Competence-Based Vocational and Professional Education*, ed M. Mulder (New York, NY: Springer), 165–182.

*Dekkers, R., Carey, L., and Langhorne, P. (2022). *Making Literature Reviews Work: A Multidisciplinary Guide to Systematic Approaches*. Cham: Springer. doi: 10.1007/978-3-030-90025-0_1

DeLuca, C. (2012). Preparing teachers for the age of accountability: toward a framework for assessment education. *Action Teac. Educ.* 34, 576–591. doi: 10.1080/01626620.2012.730347

*DeLuca, C., and Bellara, A. (2013). The current state of assessment education: aligning policy, standards, and teacher education curriculum. *J. Teach. Educ.* 64, 356–372. doi: 10.1177/0022487113488144

*DeLuca, C., Coombs, A., and LaPointe-McEwan, D. (2019b). Assessment mindset: exploring the relationship between teacher mindset and approaches to classroom assessment. *Stud. Educ. Eval.* 61, 159–169. doi: 10.1016/j.stueduc.2019.03.012

*DeLuca, C., Coombs, A., MacGregor, S., and Rasooli, A. (2019a). Toward a differential and situated view of assessment literacy: studying teachers' responses to classroom assessment scenarios. *Front. Educ.* 4, 94. doi: 10.3389/feduc.2019.00094

*DeLuca, C., LaPointe-McEwan, D., and Luhanga, U. (2016a). Approaches to classroom assessment inventory: a new instrument to support teacher assessment literacy. *Educ. Assess.* 21, 248–266. doi: 10.1080/10627197.2016.1236677

*DeLuca, C., LaPointe-McEwan, D., and Luhanga, U. (2016b). Teacher assessment literacy: a review of international standards and measures. *Educ. Asses. Eval. Account.* 28, 51–272. doi: 10.1007/s11092-015-9233-6

*DeLuca, C., Rickey, N., and Coombs, A. (2021). Exploring assessment across cultures: teachers' approaches to assessment in the U.S., China, and Canada. *Cogent. Educ.* 8, 1921903. doi: 10.1080/2331186X.2021.1921903

*DeLuca, C., Schneider, C., Coombs, A., Pozas, M., and Rasooli, A. (2020). A cross-cultural comparison of German and Canadian student teachers' assessment competence. *Assess. Educ. Princip. Policy Pract.* 27, 26–45. doi: 10.1080/0969594X.2019.1703171

*DeLuca, C., Valiquette, A., Coombs, A., LaPointe-McEwan, D., and Luhanga, U. (2018). Teachers' approaches to classroom assessment: a large-scale survey. *Assess. Educ. Principl. Policy Pract.* 25, 355–375. doi: 10.1080/0969594X.2016.1244514

*DeLuca, C., Willis, J., Cowie,., B., Harrison, C., Coombs, A., et al. (2019c). Policies, programs, and practices: exploring the complex dynamics of assessment education in teacher education across four countries. *Front. Educ.* 4, 132. doi: 10.3389/feduc.2019.00132

*Deneen, C.C., and Brown, G. T. L. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent. Educ.* 31, 1225380. doi: 10.1080/2331186X.2016.1225380

Desimone, L. (2009). Improving impact studies of teachers' professional development: toward better conceptualizations and measures. *Educ. Res.* 38, 181–199. doi: 10.3102/0013189X08331140

Dixon, H., and Haigh, M. (2009). Changing mathematics teachers' conceptions of assessment and feedback. *Teach. Develop.* 13, 173–186. doi: 10.1080/13664530903044002

*Doyle, A., Conroy Johnson, M., Donlon, E., McDonald, E., and Sexton, P. (2021). The role of the teacher as assessor: developing student teacher's assessment identity. *Aust. J. Teach. Educ.* 46, 52–68. doi: 10.14221/ajte.2021v46n12.4

*Edwards, F. (2017). A rubric to track the development of secondary pre-service and novice teachers' summative assessment literacy. *Assess. Educ. Princip. Policy Pract.* 24, 205–227. doi: 10.1080/0969594X.2016.1245651

Feiman-Nemser, S. (2008). "Teacher learning. How do teacher learn to teach?," in *Handbook of Research on Teacher Education*, eds M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre, and K. E. Demers (New York, NY: Routledge).

Flórez Petour, M. T. (2015). Systems, ideologies and history: a three-dimensional absence in the study of assessment reform processes. *Assess. Educ. Princip. Policy Pract.* 22, 3–26. doi: 10.1080/0969594X.2014.943153

*Fulmer, G. W., Lee, I. C. H., and Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers' assessment practices: an integrative review of research. *Assess. Educ. Princip. Policy Pract.* 22, 475–494. doi: 10.1080/0969594X.2015.1017445

*Gotch, C. M., and French, B. F. (2014). A systematic review of assessment literacy measures. *Educ. Meas. Issues Pract.* 33, 14–18. doi: 10.1111/emip.12030

*Gotch, C. M., and McLean, C. (2019). Teacher outcomes from a statewide initiative to build assessment literacy. *Stud. Educ. Eval.* 62, 30–36. doi: 10.1016/j.stueduc.2019.04.003

Gullickson, A. R., and Hopkins, K. D. (1987). Perspectives on educational measurement instruction for pre-service teachers. *Educ. Meas. Issues Pract.* 6, 12–16. doi: 10.1111/j.1745-3992.1987.tb00501.x

*Gunn, A. C., and Gilmore, A. (2014). Early childhood initial teacher education students' learning about assessment. *Assess. Matters* 7, 24–38. doi: 10.18296/am.0124

Guskey, T. R. (2002). Does it make a difference? Evaluating professional development. *Educ. Leadership* 5, 45–51.

*Hailaya, W., Alagumalai, S., and Ben, F. (2014). Examining the utility of assessment literacy inventory and its portability to education systems in the Asia Pacific region. *Aust. J. Educ.* 58, 297–317. doi: 10.1177/0004944114542984

Harrison, C. (2005). Teachers developing assessment for learning: Mapping teacher change. *Teach. Develop.* 9, 255–264.

Herman, J. (2016). *Comprehensive Standards Based-Assessment Systems Supporting Learning*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

*Herppich, S., Praetorius, A.-K., Forster, N., Glogger-Frey, I., Karst, K., Leutner, D., et al. (2018). Teachers' assessment competence: integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teach. Teach. Educ.* 76, 181–193. doi: 10.1016/j.tate.2017.12.001

*Hill, M. F., Ell, F., Grudnoff, L., and Limbrick, L. (2014). Practise what you preach: Initial teacher education students learning about assessment. *Assess. Matters* 7, 90–112. doi: 10.18296/am.0126

Hilton, G., Flores, M. A., and Niklasson, L. (2013). Teacher quality, professionalism and professional development: findings from a European project. *Teach. Dev.* 17, 431–447. doi: 10.1080/13664530.2013.800743

Joint Committee on Standards for Education Evaluation (JCSEE). (2015). *Classroom Assessment Standards: Practices for PK-12 Teachers.*

*Kaden, U., and Patterson, P. P. (2014). Changing assessment practices of teaching candidates and variables that facilitate that change. *Act. Teach. Educ.* 36, 406–420. doi: 10.1080/01626620.2014.977700

Kemmis, S., Edwards-Groves, C., Lloyd, A., Grootenboer, P., Hardy, I., and Wilkinson, J. (2017). "Learning as being 'Stirred into practices," in *Practice Theory Perspectives on Pedagogy and Education*, eds P. Grootenboer, C. Edwards-Groves, and S. Choy (Singapore: Springer), 45–65.

Kemmis, S., Wilkinson, J., and Edwards-Groves, C. (2016). "Roads not travelled, roads ahead: How the theory of practice architectures is travelling," in *Exploring Education and Professional Practice: Through the Lens of Practice Architectures*, eds K. Mahon, and S. Francisco (Singapore: Springer), 239–256.

Kemmis, S., Wilkinson, J., Edwards-Groves, C., Hardy, I., Grootenboer, P., and Bristol, L. (2014). *Changing Practices, Changing Education*. Singapore: Springer Science & Business Media.

Kozma, R. B. (2003). "ICT and educational change: A global phenomenon," in *Technology, Innovation, and Educational Change: A Global Perspective*, ed R. B. Kozma (Eugene, OR: International Society for Technology in Education), 1–18.

*Looney, A., Cumming, J., van Der Kleij, F., and Harris, K. (2018). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assess. Educ. Princip. Policy Pract.* 25, 442–467. doi: 10.1080/0969594X.2016.1268090

Mandinach, E. B., and Gummer, E. S. (2016). *Data Literacy for Teachers: Making it Count in Teacher Preparation and Practice*. New York, NY: Teachers College Press

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improv. Schools* 12, 101–113.

Mertler, C. A., and Campbell, C. (2005). "Measuring teachers' knowledge & application of classroom assessment concepts: Development of the assessment literacy inventory," in *Paper Presented at the Annual Meeting of the American Educational Research Association* (Montreal, QC).

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 21. doi: 10.1371/journal.pmed.1000097

Mouraz, A., Leite, C., and Fernandes, P. (2019). Between external influence and autonomy of schools: effects of external evaluation of schools. *Paidéia* 29, e2922. doi: 10.1590/1982-4327e2922

National Taskforce on Assessment Education for Teachers. (2016). *Assessment Literacy Defined*. Available online at: http://assessmentliteracy.org/national-task-force-assessmenteducation-teachers/

Neuman, S. B., and Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *Am. Educ. Res. J.* 46, 532–566. doi: 10.3102/0002831208328088

OECD (2022). *Trends Shaping Education 2022*. Paris: The Organization for Economic Cooperation and Development.

O'Neill, J., and Adams, P. (2014). The future of teacher professionalism and professionality in teaching. *New Zeal. J. Teachers Work* 11, 1–2.

*Oo, C. Z., Alonzo, D., and Asih, R. (2022). Acquisition of teacher assessment literacy by pre-service teachers: a review of practices and program designs. *Issues Educ. Res.* 32, 352–373. Available online at: hkp://www.iier.org.au/iier32/oo.pdf

*Pastore, S., and Andrade, H. L. (2019). Teacher assessment literacy: a three-dimensional model. *Teach. Teach. Educ.* 84, 128–138. doi: 10.1016/j.tate.2019.05.003

Petticrew, M., and Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*. Oxford: Blackwell.

Plake, B. S., Impara, J. C., and Fager, J. J. (1993). Assessment competencies of teachers: a national survey. *Educ. Meas. Issues Pract.* 12, 39. doi: 10.1111/j.1745-3992.1993.tb00548.x

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educ. Meas. Issues Pract.* 10, 12–15. doi: 10.1111/j.1745-3992.1991.tb00211.x

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educ. Leadersh.* 62, 82–83.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Pract.* 48, 4–11. doi: 10.1080/00405840802577536

Popham, W. J. (2011). Assessment literacy overlooked: a teacher educator's confession. *Teach. Educ.* 46, 265–273. doi: 10.1080/08878730.2011.605048

Popham, W. J. (2013). *Classroom Assessment: What Teachers Need to Know, 7th Edn.* Boston, MA: Pearson.

Popham, W. J. (2018). *Assessment Literacy for Educators in a Hurry*. Alexandria: ASCD.

*Poskitt, J. (2014). Transforming professional learning and practice in assessment for learning. *Curric. J.* 25, 542–566. doi: 10.1080/09585176.2014.981557

Roberts, C., Khanna, P., Lane, A. S., Reimann, P., and Schuwirth, L. (2021). Exploring complexities in the reform of assessment practice: a critical realist perspective. *Adv. Health Sci. Educ.* 26, 1641–1657. doi: 10.1007/s10459-021-10065-8

Roeder, H. H. (1972). Are today's teachers prepared to use tests? *Peaboby J. Educ.* 49, 239–240. doi: 10.1080/01619567209537858

*Rogers, A. P., Mitescu Reagan, E., and Ward, C. (2022). Preservice teacher performance assessment and novice teacher assessment literacy. *Teach. Educ.* 33, 175–193. doi: 10.1080/10476210.2020.1840544

Rotthoff, T., Kadmon, M., and Harendza, S. (2021). It does not have to be either or! Assessing competence in medicine should be a continuum between an analytic and a holistic approach. *Adv. Health Sci. Educ.* 26, 1659–1673. doi: 10.1007/s10459-021-10043-0

Schatzki, T., Knorr-Cetina, K., and von Savigny, E. (eds.) (2001). *The Practice Turn in Contemporary Theory*. London: Routledge.

Scheerens, J. (2016). *Educational Effectiveness and Ineffectiveness: A Critical Review of the Knowledge Base*. Dordrecht: Springer.

*Schneider, C., DeLuca, C., Pozas, M., and Coombs, A. (2020). Linking personality to teachers' literacy in classroom assessment: a cross-cultural study. *Educ. Res. Eval.* 26, 53–74. doi: 10.1080/13803611.2021.1902354

Sen, A. (1989). Development as capability expansion. *J. Dev. Plann.* 19, 41–58. doi: 10.1007/978-1-349-21136-4_3

Shepard, L. A. (2017). "Forward," in *SAGE Handbook of Research on Classroom Assessment*, ed J. H. McMillan (Thousand Oaks, CA: SAGE Publications), xix–xxii.

Stiggins, R. (1991). Assessment literacy. *Phi Delta Kappan.* 72, 534–539.

Stiggins, R. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan.* 77, 238–245.

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan.* 86, 22–27. doi: 10.1177/003172170408600106

Stiggins, R. (2017). *The Perfect Assessment System*. Alexandra: VA, ASCD.

Stiggins, R. J. (2002). Assessment crisis: the absence of Assessment for Learning. *Phi Delta Kappan* 83, 758–765. doi: 10.1177/003172170208301010

Tierney, R. D. (2006). Changing practices: Influences on classroom assessment. *Assess. Educ.* 13, 239–264.

Torres, A. C., and Weiner, J. M. (2018). The new professionalism? Charter teachers' experiences and qualities of the teaching profession. *Educ. Policy Anal. Arch.* 26. doi: 10.14507/epaa.26.3049

Webb, N. L. (2002). "Assessment literacy in a standards-based urban education setting," in *Paper Presented at the Annual Meeting of the American Educational Research Association* (New Orleans, LA).

*Willis, J., Adie, L., and Klenowski, V. (2013). Conceptualising teachers' assessment literacies an era of curriculum and assessment reform. *Aust. Educ. Res.* 40, 241–256. doi: 10.1007/s13384-013-0089-9

*Xu, Y., and Brown, G. T. L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teach. Teach. Educ.* 58, 149–162. doi: 10.1016/j.tate.2016.05.010

Yan, Z. (2021). Assessment-as-learning in classrooms: the challenges and professional development. *J. Educ. Teach.* 47, 293–295. doi: 10.1080/02607476.2021.1885972

*Yan, Z., and Pastore, S. (2022). Are teachers literate in formative assessment? The development and validation of the Teacher Formative Assessment Literacy Scale. *Stud. Educ. Eval.* 74, 101183. doi: 10.1016/j.stueduc.2022.101183

*Ye, W. (2022). Urban public-school teachers' assessment literacy in China's Zhengzhou city. *Education* 3–13. doi: 10.1080/03004279.2021.2025128

* References marked with an asterisk indicate studies included in the systematic review.

frontiers | Frontiers in Education

TYPE Perspective
PUBLISHED 17 July 2023
DOI 10.3389/feduc.2023.1211580

Check for updates

OPEN ACCESS

EDITED BY
Xinya Liang,
University of Arkansas, United States

REVIEWED BY
Alexander Robitzsch,
IPN - Leibniz Institute for Science and Mathematics Education, Germany
Yi-Fang Wu,
Cambium Assessment, Inc.,
United States

*CORRESPONDENCE
Gavin T. L. Brown
✉ gt.brown@auckland.ac.nz

RECEIVED 24 April 2023
ACCEPTED 28 June 2023
PUBLISHED 17 July 2023

CITATION
Brown GTL, Kannan P, Sinharay S, Zapata-Rivera D and Zenisky AL (2023) Challenges and opportunities in score reporting: a panel of personal perspectives. *Front. Educ.* 8:1211580. doi: 10.3389/feduc.2023.1211580

COPYRIGHT
© 2023 Brown, Kannan, Sinharay, Zapata-Rivera and Zenisky. 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.

Challenges and opportunities in score reporting: a panel of personal perspectives

Gavin T. L. Brown[1]*, Priya Kannan[2], Sandip Sinharay[3], Diego Zapata-Rivera[3] and April L. Zenisky[4]

[1]Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand, [2]WestEd, San Francisco, CA, United States, [3]Educational Testing Service, Princeton, NJ, United States, [4]College of Education, University of Massachusetts, Amherst, MA, United States

The field of score reporting continues to evolve because of new challenges, opportunities, and needs of society (e.g., COVID, remote teaching and learning). In this paper, the new challenges and opportunities in score reporting are discussed from the personal perspective of four experts who have previously conducted research in designing score reports in education. Comments are organized around four key questions concerning challenges raised by the Covid pandemic, how research will change, what current research is being conducted, and new directions in the field of score reporting.

KEYWORDS

post-COVID technology solutions, score reporting systems, testing, scientific communication, stakeholder comprehension, data-driven decision-making, assessment uses/purposes, data visualization

Introduction

It has been more than 4 years since the publication of *Score Reporting Research and Applications* (Zapata-Rivera et al., 2018). This book, part of the National Council for Measurement in Education (NCME) book series, includes work in areas such as validity in score reporting (Tannenbaum, 2018), cognitive affordances of graphical representations (Hegarty, 2018), evaluation of subscores (Sinharay et al., 2018), communicating measurement error information to teachers and parents (Zapata-Rivera et al., 2018), score reporting issues for licensure, certification, and admissions programs (O'Donnell and Sireci, 2018), communicating growth (Zenisky et al., 2018), score reports for large-scale testing programs (Slater et al., 2018), and evaluating the use of interactive reports and dashboards in formative contexts (Brown et al., 2018; Corrin, 2018; Feng et al., 2018).

In 2023, Diego Zapata-Rivera organized a panel of experts in NCME to discuss new challenges and opportunities in score reporting that respond to new trends in assessment due to changes in society and education. For example, we can see an increase in the use of digital learning and assessment systems which has resulted in the field of score reporting moving toward general reporting systems that provide teachers and learners with relevant insights based on an abundance of process and response data.

Four authors from the NCME book were contributors to the conference panel and are co-authors of this manuscript. They are: Gavin T. L. Brown (GTLB; the University of Auckland), Priya Kannan (PK; WestEd), April Zenisky (AZ; University of Massachusetts, Amherst), and Sandip Sinharay (SS; Educational Testing Service). Unfortunately, Linda Corrin (Deakin University), who had planned to participate in the panel discussion, was not able to participate. This manuscript captures what each panelist said in response to four questions:

Q1. How have the challenges of the last few years (e.g., COVID, remote teaching and learning) impacted the field of score reporting?
Q2. How will the nature of research in the area of score reporting change due to the availability of data from digital learning and assessment systems?
Q3. What aspects of your work can inform current work on designing and evaluating interactive reports and dashboards?
Q4. What new challenges and opportunities you expect will be present in the field of score reporting?

The order in which panelists answered questions was changed after each question give each of the panelists the opportunity to provide the first response to a question. The responses provided by the panelists are presented below followed by some final concluding remarks.

> Q1. How have the challenges of the last few years (e.g., COVID, remote teaching and learning) impacted the field of score reporting?

PK: The COVID-19 pandemic resulted in unprecedented disruptions to the ways in which kids experienced school (Parks et al., 2021; Krause et al., 2022). The drastic shift to virtual learning created learning setbacks for almost all children (Wyse et al., 2020), but particularly for children who were already underserved (Bailey et al., 2021; Goudeau et al., 2021), leading to disadvantages, both educationally and economically, that could last a lifetime (Dorn et al., 2020). The ways in which children across the various demographic and socio-economic subgroups experienced remote learning was dramatically different. With that, the learning loss varied significantly by access to remote learning, the quality of remote instruction, home support, and the degree of engagement (Dorn et al., 2020). Researchers hypothesized that the achievement gaps would start further widening (Bailey et al., 2021), and this became abundantly clear in the United States with the release of the 2022 NAEP report card.[1] There were greater score declines in reading and mathematics, particularly for Black and Hispanic students when compared to their White peers (Sparks, 2022). All of this has led to a reckoning of sorts in the score reporting community, particularly in how we look at student performance and achievement, specifically the dramatic and disparate learning setbacks experienced by students from under-served communities, and how we could shift the focus from achievement gap to opportunity gap by appropriately highlighting underlying systemic issues through our reporting.

Scholars have been discussing the impact of the expansive "opportunity gap" that exists across racial and associated class lines for several decades now (e.g., McDonnell, 1995; Porter, 1995). For example, in Ladson-Billings, 2006 AERA presidential address, Gloria Ladson-Billings highlighted how the focus on the "gap" is misplaced and called out the importance of paying attention to the "educational debt" that has accumulated over years of systemic disparities that have disproportionately impacted Black and Brown communities. However, the awareness and movement to really shift the narrative of how student test results are communicated, from a focus on 'achievement gap' to a focus on the 'opportunity gaps' for students across diverse communities (more precisely, gaps in their 'opportunities to learn') has really happened post COVID.

> To quote Marion (2020),
>
> "Opportunity-to-learn (OTL) is a more than 50-year-old concept that has evolved from a focus on whether students have had sufficient access to instruction or content linked to particular concepts, to a more robust conception regarding the conditions and resources provided to schools to enable students to succeed." (p. 2)

The key point being made is that OTL encompasses a much broader reflection on the set of resources that influence teaching and learning in the school setting. This includes factors such as school and classroom conditions, school climate, access to qualified teachers, opportunities for teacher professional development (PD) available across various districts, time scheduled for instruction across different districts, schools, and communities, student opportunities for out-of-school learning, student access to high-quality books in and out of school, student access to technological tools, and other resources.

The headlines emphasizing the performance/achievement gaps among student subgroups were persistent prior to the pandemic, but have been particularly so with the release of the most recent NAEP Report card (e.g., Mahnken, 2022; Modan, 2022). However, taking an OTL approach to score reporting and presenting high-quality OTL data alongside achievement results can help avoid the stereotyping of lower-performing groups, by pointing to some of the systemic and resource factors that influence performance. So, there is clearly a need to shift away from reporting average scores across disaggregated subgroups, which has the unintended consequence of reinforcing implicit racial, cultural, and economic stereotypes and deficit notions about groups of students, and move toward a focus on more systemic issues. We could begin to do this by using data from a variety of contextual OTL factors as the primary disaggregating variables and using effective visualization to present the within group differences among students from various demographic subgroups with different opportunities to learn (e.g., how do students from various demographic subgroups with similar access to instructional resources compare?). With such a view, we could push the needle in changing the narrative from a story that leads to deficit notions to a story that points to more systemic issues that need to be addressed. The hope is that this would then be the first step in moving toward a conversation around educational equity.

In the USA, NAEP score reports have already started considering these issues by incorporating their survey questionnaire data that includes several such OTL factors to see how these contextual variables can help explain the within group differences among student groups, and also to help policymakers identify the systemic opportunity gaps and address them appropriately. However, this shift away from focusing on achievement gaps and starting to unravel the systemic gaps in OTL should percolate beyond just NAEP reporting to score reporting on state summative assessments and beyond.

AZ: In reflecting back on the challenges of the past few years, one thing that has become abundantly clearer is the need to be realistic about test type and test purpose. It has long been the case that good data (and the communication of good data) was critically important in educational testing (Goodman and Hambleton, 2004). But, the reporting for summative tests in the context of K-12 education has

1 https://nces.ed.gov/nationsreportcard/

historically served more of a confirmatory/passive/high-level informative purpose, because it generally takes a long time to return results and thus when the reports come, they are far removed from instructional utility (e.g., the kids are moved up a grade, have different teachers, changed schools; Hattie and Brown, 2008; Brown et al., 2018).

So, while typical summative tests may play a role moving forward, the need for data - good data, in a timely manner - is what matters more than ever. End-of-year tests may serve a policy purpose, but what is needed now are tests with direct classroom relevance. In terms of reporting, that means the priority is probably not grand, highly stable scores and performance classifications, but data for instruction that is closely linked to what a teacher can do in the next day, the next week, or the next month. So, for those of us who work in reporting, the challenges of the past few years have reinforced that we need to reorient ourselves a little.

- First, we should not try to make summative *Individual Student Reports* (ISRs) into something they cannot be. Where those kinds of tests are administered and used going forward, the reports can and should be oriented to fulfill a descriptive or informational purpose relative to the intended users, without trying to extract deeply diagnostic information where is does not exist.
- And second, in terms of reporting, we should be paying a different kind of attention to different tests (interim or through-year tests, or formative assessments) in terms of how we can craft reporting resources for the user groups articulated for those tests that target and accomplish different goals, such as instructional planning at a sufficiently usable grain size (O'Donnell and Zenisky, 2020). The results of these assessments need to be immediately and obviously connected to a lesson or activities or next steps. With reporting, the greater the distance to action, the less likely *any* action is going to occur.

All of this is not to say that we should discard summative test reporting, but in the years to come I would like to see more acknowledgment of this reality, that some tests provide information that is backward-looking, and others are built to provide information that is forward-looking. It would be helpful to stop thinking that all tests can be all things, in terms of reporting, start recognizing the different purposes of assessments and their data, and play to the strengths of each type of test as a data source and different intended users and use cases.

GTLB: The covid pandemic required education online, including assessment of student learning. Ensuring the security of online testing (Dawson, 2021) is a *sine qua non* of valid reporting. The current state of online proctoring creates doubts as to the validity of invigilation (Alessio et al., 2017; Wuthisatian, 2020). Despite efforts to ensure academic integrity, dishonesty is widespread (Murdock et al., 2016). In the context of distance examination systems, there was more cheating than previously (Reisenwitz, 2020). Further, changes made to administration and scoring of tests and examinations during the pandemic to accommodate the lockdown did not always go to plan as seen in the UK's school exam grading controversy in which an algorithm created very different grades than estimated by schoolteachers[2]. Consequently, much research is needed on how to design and validate test reports that correctly identify and communicate the impact of construct irrelevant factors on the estimation of ability or proficiency.

Governments constantly review and revise curriculum to reflect evolving perspectives of what children need to be taught, and by implication what tests need to measure. Changes in curriculum require changes to test reporting interfaces. Underneath a change of report labels lies the question of whether Curriculum Label A really means the same thing as Label Q and if the items do measure Q when designed for A. In Brown et al. (2018), there is a report for achievement objectives in reading comprehension (Figure 8.3) showing results for finding information, knowledge, understanding, and connections derived from the New Zealand Ministry of Education (1994) curriculum framework. The New Zealand Ministry of Education (2007) updated the curriculum and positioned reading comprehension with a new set of categories (i.e., process and strategies, purposes and audiences, content and ideas, language features, structure). Expert content analysis (e.g., Does item 33 of A map onto Q?) is required to determine if the definition and operationalization of A is the same as Q. My suspicion is that this matters more to psychometricians than curriculum developers. Furthermore, at least in New Zealand, the New Zealand Ministry of Education has proceeded with a curriculum review during the government mandated lockdown. Consequently, the pressure to change badges on test reports without validating the mapping of old items and categories to new will impose significant pressures on reporting and equating of reports over time.

Covid has reminded us that reporting for school or policy administrators is very different to reporting for classroom teachers. Teachers need to know 'now' who needs to be taught what next so that they can adjust lesson plans, student groupings, or activities. Putting test reports in the hands of administrators is unlikely to lead to real classroom change. Indeed, a survey study in New Zealand showed that as administrators took over the use of a test system for school evaluation purposes, most teachers saw the tests as irrelevant for classroom use (Brown and Harris, 2009). Designing tests that prioritize teacher needs, while not ignoring those of administrators, is the ambition of educational testing (i.e., helping teachers teach better; Popham, 2000).

SS: There have been various types of impact. Test publishers now must include more caveats in score reports. For example, they may need to include caveats about state averages being based on smaller percentage of students since some years of pandemic led to more limited participation in testing. There are now more "holes" in reports such as score histories and student growth score reports because of missing test score data in spring 2020 and, in many cases, spring 2021. A big problem now is the determination and reporting of the loss in learning due to COVID and the determination of ways to estimate growth even with the loss. Fortunately, there is already quite a bit of research on these issues including Gajderowicz et al. (2022), Maldonado and De Witte (2020), and Toker (2022).

Given that many tests, since the start of the pandemic, now provide remote testing options, at least two questions related to score reporting are: (1) Whether and how should one consider the testing mode while score reporting? (2) Should there be separate reporting scales for test-center examinees versus remote examinees?

Test publishers may also have to take account of the mode when choosing a norm group. For example, consider a score report that shows the scaled score along with an average performance range

2 https://en.wikipedia.org/wiki/2020_United_Kingdom_school_exam_grading_controversy

(APR). If a test-taker took a test remotely, how should the APRs be computed? Should we compute the APRs based on only remote test-takers or should we combine both remote and test center test-taker, or report two sets of APRs? For many tests, remote test-takers perform better than test-center test-takers. So, if we compute APRs using only remote test-takers, then a particular remote test-taker will appear worse (relative to the APRs) than if we computed the APRs using both remote and test-center test-takers.

> Q2. How will the nature of research in the area of score reporting change due to the availability of data from digital learning and assessment systems?

GTLB: Work with schools makes it clear that there is no one data management system being used to handle student data including achievement or performance data. That means designing test reports that contextualize performance with useful information from background data is extremely messy and difficult. However, it may be possible to identify meaningful information, such as opportunity to learn data (e.g., was the student present when *xyz* was taught? And if not, what reason was there for the absence?), that sheds light on performance and supports appropriate responses by the teacher. Unfortunately, the cost of developing robust systems is such that there is resistance to open access between proprietorial systems that will delay the reporting of information that might help teachers understand why some students are not doing as well as expected.

A major challenge in score reporting is the churn in teacher recruitment, retirement, resignation, or transfer. This impacts the overall level of teacher assessment literacy (Xu and Brown, 2016). Although a test report system can be well designed to communicate effectively with teachers, that does not guarantee that post-deployment of the system, all teachers will be competent to understand the reports. There is a constant need to deploy support at the moment when teachers need to make decisions based on test reports. It is possible to forget what had been taught in initial teacher education (if anything was) and infrequent use of the reports creates challenges. Hence, test reporting systems need to include a variety of instructional and support resources that help teachers interpret reports correctly. Assuming that the test manual is enough simply is not warranted. The expression *RTFM* exists for a reason; users rarely consult the manual (Blackler et al., 2016). Multiple communication channels are needed to ensure teachers interpret reports as they ought to (Brown, 2019) and this must be maintained throughout the life of the test reporting system, a matter of potential economic impact as well.

SS: The nature of research will change (or maybe expand is a more appropriate word than change) in several ways. For example, it is possible to obtain additional data (e.g., timing data, key stroke data, eye tracking data, etc.) from digital assessment systems that are not available when tests are given on paper. Thus, more research and operational analyses (e.g., analysis on motivation, new test security analysis, new types of speededness analysis, etc.) can be done now that could not be done previously. Many tests now routinely conduct timing analysis and flag examinees if something appears suspicious, for example test taker(s) completing a 120-item test in 5 min (if they scored high, then they may have cheated, while if they scored low, they may have lacked motivation). Exciting research is being performed by people like Hongwen Guo and Kadriye Ercikan at ETS (Guo and Ercikan, 2021) and by Steven Wise at NWEA (Wise, 2017, 2021).

If a digital learning and assessment system can accurately measure learning progression, I suppose we have to explore ways to report the progression in an easily comprehensible manner. Some exciting work on this area has been done both at ETS by people like Aurora Graf and Peter van Rijn and outside ETS by people like Derek Briggs, but there is scope for a lot of further work. Collecting validity evidence for the utility of score reports is more difficult since we cannot just have focus groups reacting to static reports but have to have potential users navigate online, interactive reporting systems that offer different buttons/menus/choices for users to select in order to see how reports work.

PK: Online and digital learning environments, instructional technologies, and game-based learning environments have seen a rise in recent years (e.g., Heffernan and Heffernan, 2014; Feng et al., 2018; Sinatra et al., 2020; Rahimi and Shute, 2021). Students in several districts across the United States now have district-provided laptop and tablet devices which gives them access to various types of digital and online learning platforms. Their interactions in such online learning and testing environments result in much underlying background and log data such as: number of times a student accesses various features within the learning or assessment environment, where and when the student clicks, how the student navigates within the digital environment, the amount of time a student spends on the assigned task or question, the number of attempts a student makes to answer a question correctly, the number of hints and scaffolds that the student uses, to name just a few examples. Such data could provide a richer context and additional insights on a student's current state of understanding and could provide some opportunities to support more effective and personalized learning experiences for each student.

There is a great opportunity to provide feedback that is instructionally useful with the large amounts of data that can now be available in these digital contexts. There is already some interesting work in this area, and several Learning Analytic Dashboards (LADs) are being designed (Molenaar and Knoop-van Campen, 2018; Michaeli et al., 2020; Sahin and Ifenthaler, 2021) with the intent of providing real-time feedback on instructionally informative data such as students' time on task, progress toward goals, their overall level of conceptual understanding, and their strengths and needs relative to ongoing formative goals. This information can be scaffolded and presented to teachers in a real-time actionable dashboard (examples in Kannan et al., 2019; Kannan and Zapata-Rivera, 2022) with scaffolds and visualizations that can help teachers in tailoring their instruction to fill specific gaps in students' conceptual understandings.

However, with the overwhelming amount of data available, teachers are often left "data rich and information poor" (DRIP). The concept of DRIP was first extended to education about 10 or so years ago (Charman, 2009) when educators were beginning to get bombarded with large volumes of data from large-scale assessment and reporting systems. Such large amounts of data result in an unwanted increase in teachers' cognitive load when they are already strapped for time. It could also lead to several "curiosity-driven explorations" (Khosravi et al., 2021, pp. 3; Wise and Jung, 2019) of irrelevant questions, which again poses unwanted and unrequited demands on their limited time.

Therefore, it is very important that the data provided to educators is not overwhelming, and that the score reports and dashboards be designed in such a way that the information is scaffolded in an interpretable and usable way to suit the needs of the users (Kannan

and Zapata-Rivera, 2022). There are a number of ways in which big data can be scaffolded. One example is to use a question-based interaction format (VanWinkle et al., 2011), where questions that reflect the needs of the intended user group are unraveled through stakeholder-specific needs assessments. The dashboards are then designed such that users can pose specific questions based on their needs and pre-canned visualizations and actionable data chunks that support instructional decision-making are populated to support their immediate use (Zapata-Rivera, 2021; Kannan et al., 2022). LADs are increasing across the educational landscape (Papamitsiou and Economides, 2014; Sahin and Ifenthaler, 2021), and research in this area should continue to focus on how these dashboards can provide timely, interpretable, usable, actionable, and useful information to educators so that they do not pose a demand on their already limited time.

AZ: This is an exciting time to be in testing. It really is. Digital learning and digital assessment systems mean data, and data means, theoretically, more and different things to say about test-taker performance (DiCerbo, 2020). But we are not quite *there* yet, especially with respect to reporting. In terms of reports, we are still in the potential stage. In part, that is because of our training and nature as professionals: many of us in reporting grew up as psychometricians who by nature are very careful about what inferences are right and proper (in terms of reliability and validity) and which are not (poetically described by Sireci (2021) as "psychometric paralysis"). The kinds of data and volumes of data we can gather are still not well understood in terms of our established validity paradigms, and this has implications for communication of those data elements.

But, with uncertainty, I think there is opportunity. We can do research in terms of big data. What does it all mean? We can do research on how to communicate these new kinds and quantities of data - how do we display different data so that differentiated user groups can be supported in various informational and actionable goals (Hegarty, 2011; Zenisky, 2015)? And not just how do we display information but also, how might we structure the data to engage users relative to those goals? That's where we get into stakeholder research on the use of reports and dashboards and connect those research activities to what users actually do with the data (e.g., Wainer et al., 1999; Rick et al., 2016; Corrin, 2018).

To that end, we still have much to learn about the kinds of questions stakeholders might have and actions they might take in light of different kinds of data that is emerging. How do we package reports and results so that they can do what they need to do, in terms of anticipating those needs? That is one direction where reporting research could be heading. Some user questions will be informational in nature, some will be actionable in nature, and thus the task in front of us is to learn enough to build the systems that let people get what they need to get, to do what they need to do.

> Q3. What aspects of your work can inform current work on designing and evaluating interactive reports and dashboards?

SS: Interactive reports and dashboards aim to produce a wealth of information about test-taker's engagement, experiences, and performance on tasks (Kannan and Zapata-Rivera, 2022). Users of interactive reports and dashboards most likely would want to see the results on various aspects of learning and various types of interaction between a test-taker and the system, on the difference in performance, of the same group over multiple time periods, over multiple groups for the same time period, over multiple groups over several time periods, and often at the subscale level. My research on score reporting has focused mainly on subscores (Sinharay et al., 2011) and more generally on evaluating psychometric quality of the reported information (Sinharay and Johnson, 2019; Sinharay, 2021). So, I suppose my work may be relevant in the determination of:

- the information that is appropriate to include in interactive reports and dashboards,
- the appropriate interpretations of the information, and
- the appropriate uses of the information.

And I anticipate that my research may be helpful in answering questions like:

- Is it justified to report all the information that is intended to be reported or is demanded by clients?
- Are the various scores reliable enough to be reported or interpreted for their intended purposes?
- For situations when various scores for the individual examinees are not reliable enough, are the differences between average school-level scores reliable enough to be reported?

PK: In my current role at WestEd, I work mostly with state departments of education and regional educational laboratories in providing consultation, technical assistance, and other psychometric support particularly in areas related to score reporting. Assessment programs that we have recently been supporting for several states are in the early learning space (basically preschool through kindergarten assessments that assess children in broad domains such as social–emotional skills, independence, and motor coordination, and foundational knowledge and skills that prepare them for kindergarten). These assessments are often administered to fulfill the reporting requirements for the Office of Special Education Programs (OSEP), which requires that children with disabilities who enter a publicly funded preschool education program are assessed upon entry and exit to the program to demonstrate growth in the assessed domains. These assessments can also be used to monitor progress over time for all children enrolled in publicly funded preschool programs, but the OSEP reporting use-case is the most common as it is a federal reporting requirement in those states.

There are several challenges in assessing and reporting for this age group and this population. Educators in preschool contexts are often less familiar with large-scale assessments (Pretti-Frontczak et al., 2002; Ertle et al., 2016; Schachter et al., 2019) and reports from these large-scale assessments. Administering standardized assessments (which are often observation-based) in this context comes with its own set of challenges (Finello, 2011). Parents are often unfamiliar with the need for and context of assessment in this space. We have encountered several challenges in developing appropriate score reports and dashboards that are interpretable and useful for educators and parents. I can provide a few examples here to illustrate.

- Scale scores often do not mean anything to these stakeholders, particularly parents of preschoolers, and they are often baffled by

such numbers on reports. So, we have been working with our state department clients to identify ways in which we could report results from multiple interim assessments without indicating a scale score, but in such a way that parents can still see progress or movement made by their children.

- Teachers and parents want benchmarks and normative comparisons to compare their child's performance. As you can imagine, at this age group, stakeholders, particularly parents, are used to percentiles. Parents often take their child to the pediatrician's office, and are told that their child is 25th percentile in height, 50th percentile in weight, and so on. But, though these early learning assessments are based on underlying learning progressions, these assessments have not been normed and do not have normative samples to show such comparisons. Therefore, we have been working with our clients to identify other criterion-based benchmarks (e.g., average level for 3- and 5-year-olds, or average level that indicates kindergarten readiness) to provide benchmark comparisons for stakeholders.
- During a needs assessment study for a couple of the client states, we found that teachers and parents often want feedback at a nuanced skill-level, but reporting at this individual skill (item) level is not feasible when these measurements are often based on a single time-point and a single observation. We have been working with our clients to provide sufficiently detailed feedback that may be useful in an interim assessment context, while at the same time not providing item-level details that can lead to over-interpretation/misinterpretation of the results.
- Teachers/speech language pathologists and others who use these data often find the reporting dashboards confusing and overwhelming to navigate and to use in their practice. Our evaluations with teachers have revealed additional needs for professional development (PD) to understand and use the reporting dashboards. We have been working with our partners and clients to identify and create appropriate PD videos that can be accessed by teachers at any time.

Keeping the stakeholder (user) at the center of the design and evaluation process (Kannan, 2023), we are also working with teachers and parents in these states by implementing the iterative multi-step approach (Zapata-Rivera et al., 2012; Hambleton and Zenisky, 2013) and repeatedly evaluating iterations of the Individual Student Reports and dashboards with the intended stakeholders to evaluate the extent to which they are able to accurately interpret and appropriately use the results (Kannan, 2023). Parents are a particularly heterogeneous stakeholder group (Kannan et al., 2018, 2021), and it is critically important to ensure that the reports designed for parents provide them with interpretable information that they can appropriately use. Therefore, the extent to which our evaluations indicate gaps in interpretation and/or use by teachers and parents, we have been making additional revisions to scaffold and elucidate the information being provided to specifically cater to stakeholder needs, and cycle back in for additional rounds of evaluations.

AZ: At present, I'm working on several different projects in the adult education space that involve both formative and summative assessments, and in each project, I am closely engaged with teachers. I keep hearing the word "actionable" being used to describe assessments and assessment results, and that to me is a difficult-to-pin-down term if we consider that reporting traditionally is a top-down activity. To that end, again, there is a real need to view stakeholder groups where they are in terms of assessment literacy and also feedback literacy and recognize the presence of variability even within groups. Users *must* be centered in this process. In the idea of actionable there is a potential for interactivity in reporting, and that to me suggests building tools that do not necessarily answer specific questions but rather flexibly respond to the kinds of questions stakeholders might pose (Zenisky and Hambleton, 2013; Zenisky, 2016).

In terms of designing and evaluating interactive reports and dashboards, it's perhaps better from an efficiency perspective to have the programmers jump right to the code and build, but that likely means the specifications for the build are based on something other than users. I was in a meeting recently where a very nice project manager wanted a very specific list of changes to a report to hand to a developer to check a box on a deliverable, and they were perhaps less than thrilled when I declined to provide such a list. In that case, I myself am nowhere near the target user of the assessment. I can critique reports, and I can advise on what the literature speaks to as good practice generally, but the stakeholders are the ones we need to listen to (Zenisky and Hambleton, 2015; Corrin, 2018). That's where my current work in reporting is now. Defining "actionable" and using the words of stakeholders to guide reporting, rather than "you'll take what I give you and like it."

GTLB: My own research (Brown, 2004, 2008) into how teachers conceive of the purposes and nature of assessments, including tests, shows that their pre-existing conceptions matter to how they use tests. When the priority is on using assessment to improve learning, then informal formats are prioritized, especially if they reveal insights into students' deeper or higher-order learning (Brown, 2009). However, survey research with students showed that students who believe teachers use formal testing to improve teaching performed better (Brown et al., 2009). Further, students who accept that testing will legitimately evaluate their accomplishments tend to do better (Brown, 2011). Not surprising, how teachers understand assessment matters to their behaviors. Thus, how assessment is designed in any jurisdiction matters to how teachers and students will understand and respond to assessment (Brown and Harris, 2009). Environments are not equal, so there are few universalities in how teachers conceive of assessment (Brown et al., 2019), meaning that test reports cannot be universal either.

The pre-existing belief structures of teachers and administrators have been derived from their extended experience of assessments in formal and informal environments. This means that test reports must be designed in light of those factors (Brown and Hattie, 2012). Test reports must go beyond total score and rank order to provide usable instructional insights. Notwithstanding concerns about the validity and reliability of sub-score reports (Sinharay et al., 2018), sub-score reports from well-designed tests provide a more robust basis for decision making than teachers' own intuitions. This means that teachers need guidance from test reports to reduce the temptation to believe students' skill is less than what a well-designed test shows they can do. This was demonstrated in New Zealand where teachers judged almost all children to be worse than what their test performance showed (Meissel et al., 2017). For formative purposes, it is more likely that sub-scores will be educational.

Moreover, teachers have theories that explain why some students learn and others do not (e.g., *students do not learn because of poverty*).

Data-driven decision-making professional development strategies (Lai and Schildkamp, 2016) seem to help teachers re-examine their *a priori* beliefs about poor performance. Test reports or data visualizations that help teachers see that their favored explanation is invalid (e.g., scores by students with or without indicators of poverty) are needed so that teachers and leaders can grapple with the responsibilities and authority they have to ensure learning occurs. Without reports that provide information that helps teachers and leaders do their jobs and which simultaneously correct wrong thinking about learning, tests will not do much to improve equitable outcomes.

> Q4. What new challenges and opportunities you expect will be present in the field of score reporting?

AZ: Reporting, when we think about it, goes to the heart of the issue of why we do what we do – it facilitates the 'why' of testing. Right now, however, public appetite for summative tests that do not say much at a granular level is waning quickly. Quite honestly, sometimes it does feel like there are nuances to testing and test scores that only testing professionals care deeply about. So, for those of us who work in reporting, we may at times find ourselves in a netherworld of sorts, where we understand all too well the technicalities of data and psychometrics, while at the same time having one foot firmly in the world of users and real-world use of tests. That is *a* challenge, to figure out how to talk about test results considering the reality of the data and all the uncertainty that comes with it. But this is also an opportunity.

Relatedly, another challenge has been, and will be in the coming years, how to make tests relevant. When we say that a test cannot be used for this, or is not valid for that, despite taking hours of time to administer, and occupying a huge amount of mental space for educators, those tests are being relegated to noise. I believe we do need to push harder on being realistic about what specific tests can and cannot do, and advance reporting methods appropriate to different tests.

Lastly, I think a new challenge will be how to navigate the push for artificial intelligence as a way to solve everything and ensure that people are still involved in the process of reporting. There are certainly ways that technology can be involved in our work and that can help us with some of the tricky bits, but there still remains no substitute for the engagement with stakeholders and iterations that spring from that that are informed by real users and use cases (Kannan et al., 2018; Slater et al., 2018).

GTLB: As a research scientist, I am excited by the many intriguing possibilities generated by new technologies to examine what students are doing while they answer online tests. Technologies such as eye-tracking, process logs, galvanic skin responses, event related potentials, among others have the potential to reveal what the mind is really doing. Vast amounts of data can be generated by these technologies some of which will no doubt correlate with tested performance. However, there is a strong possibility that the associations of eye movements, skin responses, mouse usage, or brain electrical activity are not related with any meaningful principles that could inform instruction. A similar debate has been held around *f*MRI studies (Vul et al., 2009).

It was only when Greiff et al. (2015) developed a conceptual framework of how a problem should be solved under the scientific method that they could make sense of how students responded to an online test of reasoning. A recent study of process data on a computer-based test made sense only when time spent on an item and actions within the item were interpreted as evidence of persistent effort; analysis showed that effortful persistence contributed to overall better performance even when a specific item was answered incorrectly (Lundgren and Eklöf, 2020). These studies, among others, show that the data by themselves do not make sense of themselves (Pearl and Mackenzie, 2018). Zumbo et al. (2023) have identified the key issue with the promise of sensor data; it lacks a coherent psychological theory to explain how the movement of eyes, mice, electricity, or blood relates to instruction and learning. As a discipline, we cannot explain why these physiological or behavioral data mean anything for understanding teaching and learning. Thus, much needs to be done to not just display sensor data but communicate how that information can be usable by teachers and administrators. Does it matter? For now, we do not know.

SS: As I mentioned earlier, one challenge, given the availability of an ocean of data, is to determine exactly how much information from data on timing, eye-tracking, learning, growth, and so on is justified to be reported from the viewpoints of accuracy, reliability, or validity. This determination will require both psychometric analysis as well as focus groups, discussions with parents and teachers, and so on.

One challenge is that we have to improve data literacy as consumers of test scores get overwhelmed with more and more data/test score information from summative and interim assessments. An opportunity is that we could report a substantial amount of new information that would provide a context for interpreting a score. For K-12 assessments, for example, we could flag a score where the student appeared unmotivated (i.e., lots of omits, fast guessing, low engagement characteristics, etc.). A teacher (or parent) should know that a score under these conditions may not represent the student's best work. For writing samples, one may report flags indicating that a student used poor strategies in production (e.g., little revision, no signs of outlining). I admit, though, that these options could lead to challenges/lawsuits by the test takers against the testing company, similar to what testing companies fear tests-takers might do when their scores are put on hold or canceled due to possible unfair practices, such as cheating.

I mentioned research on reporting loss of learning due to the pandemic. An opportunity is to go one step further and conduct research on a more general topic—how to report and what to report when a major disruption or unexpected problem occurs, where an unexpected problem could be a disease outbreak, or could be a natural disaster, a huge computer problem, or something different.

PK: I agree with all of the points made by my fellow panelists here. From my perspective, I think that all of my previous responses are somewhat related to the challenges and opportunities that we are faced with, in the field of score reporting at this time.

First, within the current political climate and the overall push-back on assessments, score reports can be used as a vehicle for pushing the needle toward a conversation around educational equity by changing the narrative from 'achievement gaps' to 'opportunity gaps' and beginning to address the systemic issues in opportunities and access that persist across various racial and class lines. Score reports can and should be used as an effective tool in unraveling these systemic issues by not only using OTL variables for disaggregating data, but also by conducting intersectional analyses to identify bias and using effective visualizations to clearly report these results to various stakeholder groups so that it results

in much needed action. There are likely to be measurement challenges associated with these types of analyses, such as the over-interpretation of correlational results in associating OTL with educational outcomes. Still, this shift in focus could be the necessary first step in helping policymakers in identifying the opportunity gaps and enabling them to start addressing systemic issues through appropriate legislations.

Second, as I pointed out earlier, the availability of large amounts of data, such as time spent on task, number of attempts made, number of hints or scaffolds used, can have its own associated benefits and challenges. While such data can be promising in providing instructionally useful feedback to teachers and personalized learning experiences for students, it is important to ensure that such data is appropriately scaffolded for the users so that teachers and students are able to understand the results presented and take appropriate actions from these results. Using techniques such as the 'question-based interaction approach' could help scaffold this data for end-users making it easy to take appropriate actions.

Finally, the context-specificity and stakeholder-specificity of score reports and dashboards (Kannan, 2023) is further exemplified by the early learning contexts we currently work with at WestEd. This highlights the critical importance of using an audience-specific approach and implementing an iterative multistep framework (Zapata-Rivera et al., 2012; Hambleton and Zenisky, 2013) in the design and evaluation of dashboards to create reports that are interpretable and useful to the end users.

Conclusion

We have discussed several challenges and opportunities in the field of score reporting that result from changes in society and education. Some of these challenges and opportunities have to do with alignment between assessment type and appropriate use. This includes the types of decisions the assessment supports and the assessment information that it produces. Our experts elaborated on the challenges that COVID has imposed regarding the nature of research and practice on score reporting. These include: the need to shift away from reporting scores to providing insights that have clear relevance for instruction (e.g., what teachers need to know now to support the next instructional activities), co-designing reports with the intended audience considering equity and context factors, recognizing the different purposes of assessments and their intended use, and the implications for reporting assessment results due to test mode (e.g., remote vs. in-person testing).

Regarding the amount of data available using digital learning assessment and learning systems, the experts mentioned the opportunities that additional multimodal data will afford (e.g., assessing motivation and detecting cheating) and warned us about the need for developing appropriate methods that help us analyze rich process data to support decision making. Also, the authors elaborated on the importance of designing reporting systems (or dashboards) that take into account the needs of the audience and offer support for appropriate interpretation (Kannan and Zapata-Rivera, 2022).

The experts' current research can inform new developments in the field of score reporting in various ways including identifying information to include in interactive reports and dashboards, co-designing and evaluating new reporting systems with the audience to ensure the reports provide relevant insights, exploring the psychometric properties of assessment results, developing materials to support teacher interpretation of assessment results.

Finally, the experts consider that challenges mentioned can be opportunities for the creation of assessments that provide relevant insights for different audiences. For example, new psychometric methods that can deal with disruptive situations will be developed. It is expected that conversations around educational equity will impact the way reports are designed and evaluated.

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

GB: conceptualization, writing—original draft, writing—review and editing, and funding acquisition. PK, SS, and AZ: writing—original draft. DZ-R: methodology, supervision, writing—original draft, and writing—review and editing. All authors contributed to the article and approved the submitted version.

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.

References

Alessio, H. M., Malay, N. J., Maurer, K., Bailer, A. J., and Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning J.* 21. doi: 10.24059/olj.v21i1.885

Bailey, D. H., Duncan, G. J., Murnane, R. J., and Au Yeung, N. (2021). Achievement gaps in the wake of COVID-19. *Educ. Res.* 50, 266–275. doi: 10.3102/0013189X211011237

Blackler, A. L., Gomez, R., Popovic, V., and Thompson, M. H. (2016). Life is too short to RTFM: how users relate to documentation and excess features in consumer products. *Interact. Comput.* 28, 27–46. doi: 10.1093/iwc/iwu023

Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assess. Educ. Princip. Policy Pract.* 11, 301–318. doi: 10.1080/0969594042000304609

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. Hauppauge: Nova Science Publishers.

Brown, G. T. L. (2009). "Teachers' self-reported assessment practices and conceptions: using structural equation modelling to examine measurement and structural models" in *Structural equation modelling in educational research: concepts and applications*. eds. T. Teo and M. S. Khine (Rotterdam: SensePublishers), 243–266.

Brown, G. T. L. (2011). Self-regulation of assessment beliefs and attitudes: a review of the Students' conceptions of assessment inventory. *Educ. Psychol.* 31, 731–748. doi: 10.1080/01443410.2011.599836

Brown, G. T. L. (2019). Technologies and infrastructure: costs and obstacles in developing large-scale computer–based testing. *Educ. Inq.* 10, 4–20. doi: 10.1080/20004508.2018.1529528

Brown, G. T. L., Gebril, A., and Michaelides, M. P. (2019). Teachers' conceptions of assessment: a global phenomenon or a global localism. *Front. Educ.* 4:e0016. doi: 10.3389/feduc.2019.00016

Brown, G. T. L., and Harris, L. R. (2009). Unintended consequences of using tests to improve learning: how improvement-oriented resources heighten conceptions of assessment as school accountability. *J. MultiDisc. Eval.* 6, 68–91. doi: 10.56645/jmde.v6i12.236

Brown, G. T. L., and Hattie, J. (2012). "The benefits of regular standardized assessment in childhood education: guiding improved instruction and learning" in *Contemporary educational debates in childhood education and development*. eds. S. Suggate and E. Reese (London: Routledge), 287–292.

Brown, G. T. L., O'Leary, T. M., and Hattie, J. A. C. (2018). "Effective reporting for formative assessment: the asTTle case example" in *Score reporting: research and applications*. ed. D. Zapata-Rivera (London: Routledge)

Brown, G. T. L., Peterson, E. R., and Irving, S. E. (2009). "Beliefs that make a difference: adaptive and maladaptive self-regulation in students' conceptions of assessment" in *Student perspectives on assessment: what students can tell us about assessment for learning*. eds. D. M. McInerney, G. T. L. Brown and G. A. D. Liem (Charlotte: Information Age Publishing)

Charman, P. (2009). Data rich, information poor: creative and innovative approaches to results analysis to support teaching and learning. Paper presented at the 35th Annual Conference of the International Association for Educational Assessment, Brisbane, Australia.

Corrin, L. (2018). Evaluating students' interpretation of feedback in interactive dashboards. In D. Zapata-Rivera (Ed.), *Score reporting: research and applications* (pp. 145–159). London: Routledge.

Dawson, P. (2021). *Defending assessment security in a digital world: preventing e-cheating and supporting academic integrity in higher education*. London: Routledge).

DiCerbo, K. (2020). Assessment for learning with diverse learners in a digital world. *Educ. Meas. Issues Pract.* 39, 90–93. doi: 10.1111/emip.12374

Dorn, E., Hancock, B., Sarakatsannis, J., and Virulea, E. (2020). COVID-19 and student learning in the United States: the hurt could last a lifetime. McKinsey & Company. Available at: https://www.childrensinstitute.net/sites/default/files/documents/COVID-19-and-student-learning-in-the-United-States_FINAL.pdf (Accessed April 19, 2023).

Ertle, B., Rosenfeld, D., Presser, A. L., and Goldstein, M. (2016). Preparing preschool teachers to use and benefit from formative assessment: the birthday party assessment professional development system. *Math. Educ.* 48, 977–989. doi: 10.1007/s11858-016-0785-9

Feng, M., Krumm, A., and Grover, S. (2018). "Applying learning analytics to support instruction" in *Score reporting research and applications*. ed. D. Zapata-Rivera (New York, NY: Routledge), 145–159.

Finello, K. M. (2011). Collaboration in the assessment and diagnosis of preschoolers: challenges and opportunities. *Psychol. Sch.* 48, 442–453. doi: 10.1002/pits.20566

Gajderowicz, T. J., Jakubowski, M. J., Patrinos, H. A., and Wrona, S. M. (2022). Capturing the educational and economic impacts of school closures in Poland. 10253Policy Research Working Paper Series, The World Bank.

Goodman, D., and Hambleton, R. K. (2004). Student test score reports and interpretive guides: review of current practices and suggestions for future research. *Appl. Meas. Educ.* 17, 145–220. doi: 10.1207/s15324818ame1702_3

Goudeau, S., Sanrey, C., Stanczak, A., Manstead, A., and Darnon, C. (2021). Why lockdown and distance learning during the COVID-19 pandemic are likely to increase the social class achievement gap. *Nat. Hum. Behav.* 5, 1273–1281. doi: 10.1038/s41562-021-01212-7

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

Guo, H., and Ercikan, K. (2021). Comparing test-taking behaviors of English language learners (ELLs) to non-ELL students: use of response time in measurement comparability research (research report no. RR-21-25). ETS. doi: 10.1002/ets2.12340,

Hambleton, R., and Zenisky, A. (2013). *Reporting test scores in more meaningful ways: a research-based approach to score report design. APA handbook of testing and assessment in psychology*. Washington, DC: APA.

Hattie, J. A., and Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: development principles from New Zealand. *J. Educ. Technol. Syst.* 36, 189–201. doi: 10.2190/ET.36.2.g

Heffernan, N., and Heffernan, C. (2014). The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* 24, 470–497. doi: 10.1007/s40593-014-0024-x

Hegarty, M. (2011). The cognitive science of visual-spatial displays; implications for design. *Top. Cogn. Sci.* 3, 446–474. doi: 10.1111/j.1756-8765.2011.01150.x

Hegarty, M. (2018). "Advances in cognitive science and information visualization" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Kannan, P. (2023). "Score reporting: design and evaluation methods informed by research" in *International encyclopedia of education*. eds. R. J. Tierney, F. Rizvi and K. Ercikan. *4th* ed (Amsterdam: Elsevier)

Kannan, P., Beigman-Klebanov, B., Shao, S., and Long, R. (2019). Evaluating teachers' needs for on-going feedback from a technology-based book reading intervention. Paper presented at the 2019 Annual Meeting of the National Council for Measurement in Education, Toronto, ON.

Kannan, P., Deane, P., and Phelps, G. (2022). Validating writing traits for classroom assessment purposes. Paper presented at the 2022 annual meeting of the American Educational Research Association.

Kannan, P., and Zapata-Rivera, D. (2022). Facilitating the use of data from multiple sources for formative learning in the context of digital assessments: informing the design and development of learning analytic dashboards. *Front. Educ.* 7:913594. doi: 10.3389/feduc.2022.913594

Kannan, P., Zapata-Rivera, D., and Bryant, A. D. (2021). Evaluating parent comprehension of measurement error information presented in score reports. *Practical Assessment, Evaluation, & Research.* 26. doi: 10.7275/rgwg-t355

Kannan, P., Zapata-Rivera, D., and Leibowitz, E. A. (2018). Interpretation of score reports by diverse subgroups of parents. *Educ. Assess.* 23, 173–194. doi: 10.1080/10627197.2018.1477584

Khosravi, H., Shabaninejad, S., Bakaria, A., Sadiq, S., Indulska, M., and Gasevic, D. (2021). Intelligent learning analytics dashboards: automated Drill-down recommendations to support teacher data exploration. *J. Learn. Anal.* 8, 133–154. doi: 10.18608/jla.2021.7279

Krause, K. H., Verlenden, J. V., Szucs, L. E., Swedo, E. A., Merlo, C. L., Niolon, P. H., et al. (2022). Disruptions to school and home life among high school students during the COVID-19 pandemic - adolescent behaviors and experiences survey, United States, January-June 2021. *Morbid. Mortal. Week. Rep. 1* 71, 28–34. doi: 10.15585/mmwr.su7103a5

Ladson-Billings, G. (2006). From the achievement gap to the education debt: understanding achievement in U. S. Schools. 2006 presidential address. *Educ. Res.* 35, 3–12. doi: 10.3102/0013189X035007003

Lai, M. K., and Schildkamp, K. (2016). "In-service teacher professional learning: use of assessment in data-based decision-making" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (London: Routledge)

Lundgren, E., and Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educ. Res. Eval.* 26, 275–301. doi: 10.1080/13803611.2021.1963940

Mahnken, K. (2022). 'Nation's report card': two decades of growth wiped out by two years of pandemic. The 74 million. Available at: https://www.the74million.org/article/nations-report-card-two-decades-of-growth-wiped-out-by-two-years-of-pandemic/ (Accessed April 19, 2023).

Maldonado, J., and De Witte, K. (2020). The effect of school closures on standardised student test outcomes. Discussion paper series, no. DPS20.17, KU Leuven Department of economics, Leuven Available at: https://feb.kuleuven.be/research/economics/ces/documents/DPS/2020/dps2017.pdf (accessed on 13 October 2020).

Marion, S. (2020). Using opportunity-to-learn data to support educational equity. Center for Assessment. Available at: https://www.nciea.org/wp-content/uploads/2021/11/CFA-Marion.OTL_.Indicators.pdf (Accessed April 19, 2023).

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educ. Eval. Policy Anal.* 17, 305–322. doi: 10.3102/01623737017003305

Meissel, K., Meyer, F., Yao, E. S., and Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educ.* 65, 48–60. doi: 10.1016/j.tate.2017.02.021

Michaeli, S., Kroparo, D., and Hershkovitz, A. (2020). Teachers' use of education dashboards and professional growth. *Int. Rev. Res. Open Distrib. Learn.* 21, 61–78. doi: 10.19173/irrodl.v21i4.4663

Modan, N. (2022). These 6 charts highlight COVID-19's impact on NAEP scores. K-12 dive. Available at: https://www.k12dive.com/news/these-6-charts-highlight-covid-19s-impact-on-naep-scores/634852/ (Accessed April 19, 2023).

Molenaar, I., and Knoop-van Campen, C. A. N. (2018). How teachers make dashboard information actionable. *IEEE Trans. Learn. Technol.* 12, 347–355. doi: 10.1109/TLT.2018.2851585

Murdock, T. B., Stephens, J. M., and Grotewiel, M. M. (2016). "Student dishonesty in the face of assessment: who, why, and what we can do about it" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (London: Routledge).

New Zealand Ministry of Education (1994). *English in the New Zealand curriculum.* Wellington, NZ: Learning Media.

New Zealand Ministry of Education (2007). *The New Zealand curriculum for English-medium teaching and learning in years 1–13.* Wellington, NZ: Learning Media.

O'Donnell, F., and Sireci, S. G. (2018). "Score reporting issues for licensure, certification, and admissions programs" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

O'Donnell, F., and Zenisky, A. L. (2020). Digital module 21: results reporting for large-scale assessments. *Educ. Meas. Issues Pract.* 39, 137–138. doi: 10.1111/emip.12408

Papamitsiou, Z., and Economides, A. (2014). Learning analytics and educational data Mining in Practice: a systematic literature review of empirical evidence. *Educ. Technol. Soc.* 17, 49–64.

Parks, S. E., Zviedrite, N., Budzyn, S. E., Panaggio, M. J., Raible, E., Papazian, M., et al. (2021). COVID-19-related school closures and learning modality changes - United States, august 1-September 17, 2021. *Morbid. Mortal. Week. Rep.* 70, 1374–1376. doi: 10.15585/mmwr.mm7039e2

Pearl, J., and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. New York: Hachette Book Group.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (*6th*). Boston: Allyn & Bacon.

Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. *Educ. Res.* 24, 21–27. doi: 10.2307/1176117

Pretti-Frontczak, K., Kowalski, K., and Brown, R. D. (2002). Preschool teachers' use of assessments and curricula: a statewide examination. *Except. Child.* 69, 109–123. doi: 10.1177/001440290206900108

Rahimi, S., and Shute, V. (2021). "Learning analytics dashboards in educational games" in *Visualizations and dashboards for learning analytics. Advances in analytics for learning and teaching*. eds. M. Sahin and D. Ifenthaler (Cham: Springer).

Reisenwitz, T. H. (2020). Examining the necessity of proctoring online exams. *J. High. Educ. Theory Pract.* 20. doi: 10.33423/jhetp.v20i1.2782

Rick, F., Slater, S., Kannan, P., Sireci, S. G., Dickey, J., and Zenisky, A. L. (2016). Parents' perspectives on summative test score reports. Center for Educational Assessment Research Report no. 937. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Sahin, M., and Ifenthaler, D. (2021). "Visualization and dashboards: challenges and future directions" in *Visualizations and dashboards for learning analytics. Advances in analytics for learning and teaching*. eds. M. Sahin and D. Ifenthaler (Cham: Springer)

Schachter, R. E., Strang, T. M., and Piasta, S. B. (2019). Teachers' experiences with a state-mandated kindergarten readiness assessment. *Early Years* 39, 1–17. doi: 10.1080/09575146.2017.1297777

Sinatra, A., Graesser, A. C., Hu, X., Goldberg, B., and Hampton, A. J. (2020). Design recommendations for intelligent tutoring systems: Volume 8-data visualization. US Army Combat Capabilities Development Command–Soldier Center.

Sinharay, S. (2021). Score reporting for examinees with incomplete data on large-scale educational assessments. *Educ. Meas. Issues Pract.* 40, 79–91. doi: 10.1111/emip.12396

Sinharay, S., and Johnson, M. S. (2019). Measures of agreement: reliability, classification accuracy, and classification consistency. In DavierM. von and Y. S. Lee (Eds.), *Handbook of diagnostic classification models*. Singapore: Springer Nature.

Sinharay, S., Puhan, G., and Haberman, S. J. (2011). An NCME instructional module on subscores. *Educ. Meas. Issues Pract.* 30, 29–40. doi: 10.1111/j.1745-3992.2011.00208.x

Sinharay, S., Puhan, G., Haberman, S. J., and Hambleton, R. K. (2018). "Subscores: when to communicate them, what are their alternatives, and some recommendations" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Sireci, S. G. (2021). NCME presidential address 2020: valuing educational measurement. *Educ. Meas. Issues Pract.* 40, 7–16. doi: 10.1111/emip.12415

Slater, S., Livingston, S. A., and Silver, M. (2018). "Score reports for large-scale testing programs: managing the design process" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Sparks, S. D. (2022). Two decades of Progress, nearly gone: National Math, Reading Scores Hit Historic Lows. Education Week. Available at: https://www.edweek.org/leadership/two-decades-of-progress-nearly-gone-national-math-reading-scores-hit-historic-lows/2022/10 (Accessed April 19, 2023).

Tannenbaum, R. J. (2018). "Validity aspects of score reporting" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Toker, T. (2022). Detecting possible learning losses due to COVID-19 pandemic: an application of curriculum-based assessment. *Int. J. Contemp. Educ. Res.* 9, 78–86. doi: 10.33200/ijcer.985992

VanWinkle, W., Vezzu, M., and Zapata-Rivera, D. (2011). *Question-based reports for policymakers (research memorandum no. RM-11–16)*. Princeton, NJ: Educational Testing Service.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290. doi: 10.1111/j.1745-6924.2009.01125.x

Wainer, H., Hambleton, R. K., and Meara, K. (1999). Alternative displays for communicating NAEP results: a redesign and validity study. *J. Educ. Meas.* 36, 301–335. doi: 10.1111/j.1745-3984.1999.tb00559.x

Wise, S. L. (2017). Rapid-guessing behavior: its identification, interpretations, and implications. *Educ. Meas. Issues Pract.* 36, 52–61. doi: 10.1111/emip.12165

Wise, S. L. (2021). Six insights regarding test-taking disengagement. *Educ. Res. Eval.* 26, 328–338. doi: 10.1080/13803611.2021.1963942

Wise, A. F., and Jung, Y. (2019). Teaching with analytics: towards a situated model of instructional decision-making. *J. Learn. Anal.* 6, 53–69. doi: 10.18608/jla.2019.62.4

Wuthisatian, R. (2020). Student exam performance in different proctored environments: evidence from an online economics course. *Int. Rev. Econ. Educ.* 35:100196. doi: 10.1016/j.iree.2020.100196

Wyse, A. E., Stickney, E. M., Butz, D., Beckler, A., and Close, C. N. (2020). The potential impact of COVID-19 on student learning and how schools can respond. *Educ. Meas. Issues Pract.* 39, 60–64. doi: 10.1111/emip.12357

Xu, Y., and Brown, G. T. L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teach. Teach. Educ.* 58, 149–162. doi: 10.1016/j.tate.2016.05.010

Zapata-Rivera, D. (2021). Open student modeling research and its connections to educational assessment. *Int. J. Artif. Intell. Educ.* 31, 380–396. doi: 10.1007/s40593-020-00206-2

Zapata-Rivera, D., Kannan, P., and Zwick, R. (2018). "Communicating measurement error information to teachers and parents" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Zapata-Rivera, D., VanWinkle, W., and Zwick, R (2012). Applying score design principles in the design of score reports for CBAL™ teachers. Research memorandum no. RM-12-20. Princeton, NJ: Educational Testing Service.

Zenisky, A. L. (2015). "Visual displays for reporting test data: making sense of test performance" in *Use of visual displays in research and testing: Coding, interpreting, and reporting data*. eds. M. McCrudden, G. Schraw and C. Buckendahl (Charlotte: Information Age Publishing).

Zenisky, A. L. (2016). Choose your own (data) adventure: perils and pitfalls, and lots of promise. *The NERA Researcher* 54, 11–14.

Zenisky, A. L., and Hambleton, R. K. (2013). "From "Here's the story" to "You're in charge": developing and maintaining large-scale online test and score reporting resources" in *Improving large-scale assessment in education*. eds. M. Simon, M. Rousseau and K. Ercikan (New York, NY: Routledge).

Zenisky, A. L., and Hambleton, R. K. (2015). "Test score reporting: best practices and issues" in *Handbook of test development*. eds. S. Lane, M. Raymond and T. Haladyna. *2nd* ed (New York, NY: Routledge).

Zenisky, A. L., Keller, L. A., and Park, Y. (2018). "Reporting student growth: challenges and opportunities" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Zumbo, B. D., Maddox, B., and Care, N. M. (2023). Process and product in computer-based assessments. *Eur. J. Psychol. Assess.* doi: 10.1027/1015-5759/a000748

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 21 July 2023
DOI 10.3389/feduc.2023.1182508

Check for updates

OPEN ACCESS

EDITED BY
Anastasiya A. Lipnevich,
The City University of New York, United States

REVIEWED BY
Terrence Calistro,
The City University of New York,
United States
Drew Gitomer,
Rutgers, The State University of New Jersey,
United States
Arminda Wey,
Rutgers University, United States
contributed to the review of DG

*CORRESPONDENCE
Brent Bridgeman
✉ bridge8@msn.com

RECEIVED 08 March 2023
ACCEPTED 11 July 2023
PUBLISHED 21 July 2023

CITATION
Bridgeman B, Olivera-Aguilar M and Holtzman S (2023) The GRE as a predictor of persistence to a PhD.
Front. Educ. 8:1182508.
doi: 10.3389/feduc.2023.1182508

COPYRIGHT
© 2023 Bridgeman, Olivera-Aguilar and Holtzman. 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.

The GRE as a predictor of persistence to a PhD

Brent Bridgeman*, Margarita Olivera-Aguilar and Steven Holtzman

Educational Testing Service, Princeton, NJ, United States

Because dropout from PhD programs is common and is a problem with serious consequences for both students and institutions, identifying indicators of likely dropout would be very valuable. Scores on admissions tests might be useful, but existing data on their utility is contradictory and typically based on highly restricted small samples from just a single institution or a small set of institutions. Programs dropping the GRE as an admissions requirement noted this lack of convincing evidence that the test was useful in predicting the criterion of primary interest–persistence in graduate school PhD programs versus early dropout. HLM and quartile analyses were used to provide that evidence with a sample of 1,672 graduate programs containing 157,924 students. GRE Verbal and Analytical Writing scores, but not Quantitative scores, are shown to predict persistence versus dropout in a variety of majors with especially strong results in business, engineering, and the physical sciences (e.g., in the physical sciences only 40% of the students with low GRE Analytical Writing scores in their programs persist while 78% of the students with high scores do so).

KEYWORDS

dropout, persistence, graduate school, PhD, HLM

Introduction

About 50% of students who begin doctoral programs drop out before receiving the degree (Cassuto, 2013). Being able to identify students who are at greatest risk of dropping out would have clear benefits for both students and graduate institutions. Both students and institutions make a significant financial investment in a graduate education and early drop out wastes these investments. Institutions that identified students with a higher risk of dropping out could commit more resources to helping these students succeed in graduate school. Scores on graduate admissions tests such as the GRE would seem to be one way of identifying these at-risk students, but the existing evidence on the value of such tests in identifying students likely to persist in PhD programs is equivocal and typically based on small samples with severe range restriction.

Showing a relationship between scores on a cognitive test such as the GRE and dropout from graduate school is challenging for several reasons. First, most dropout is for reasons other than a lack of cognitive ability. A study of students who dropped out of graduate programs by the National Center for Educational Statistics (Nevill and Chen, 2007) indicated that the top reasons for leaving were: change in family status, conflict with job or military, dissatisfied with program, needed to work, personal problems, other financial reasons, taking time off, and other career interests. A cognitive test could not be expected to predict a change in family status or job conflict. Second, most graduate programs have already selected students based on strong indicators of cognitive skill, so it is impossible to predict how students without those skills would have done had they been admitted. Third, generalizing from studies done in a single institution or a small handful of institutions is difficult and can lead to contradictory and confusing results.

Despite these challenges, a number of studies have attempted to evaluate the relationship of GRE scores to persistence in graduate programs. One study noted that GRE scores were higher for men who dropped out of PhD STEM programs than for those who remained enrolled (Petersen et al., 2017). But because this study used data from four flagship state universities that had selected students based on strong GRE quantitative (GRE-Q) scores virtually no one in the sample had low scores. The men who dropped out had average scores of 742 on the old 200–800 GRE scale and the students who persisted had average scores of 723. Thus, the results cannot tell us anything about the likely success of students with low or mediocre scores, or the potential value of considering GRE scores as part of a holistic review process. A study of biomedical graduate programs at the University of North Carolina (Hall et al., 2017) similarly reached a conclusion of no relationship of GRE scores to persistence but had the same problem of attempting to reach conclusions in a sample from highly selective programs that include few if any students with low scores. The title of another recent study (Miller et al., 2019), "Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion," seemed to suggest no relationship between GRE scores and persistence in a doctoral program. Nevertheless, the text concluded that significant associations exist. Using a multivariate logistic model with the 3,692 physics students in their sample, the study abstract noted, "Significant associations with completion were found for undergraduate GPA in all models and for GRE Quantitative in two of four studies models." Specifically, in the model that included all students the significant odds ratio for GRE-Q was 1.013 with a standard error of 0.004. The "fail to predict doctoral completion" in the original article title is actually contradicted in the article text that notes, "the traditional admissions metrics of undergraduate grade point average (GPA) and the Graduate Record Examination (GRE) Quantitative, Verbal, and Physics Subject Tests do not predict completion as effectively [as] admissions committees presume." "Fail to predict" and not predicting as effectively as committees presume are not the same thing. Furthermore, a number of flaws in the original analysis that tended to underestimate the effects for GRE-Q were identified (Weisman, 2019). Some of these flaws were addressed in a response by the authors (Miller et al., 2020) but a major concern remained–the severe range restriction on GRE-Q scores. As the authors acknowledged in their response to the comment, "Undergraduate physics majors' GRE-Q scores are nearly all within just a few standard errors (SEs) of a perfect score. This strong range restriction necessarily limits the strength of any correlation between GRE-Q and any other variable."

A study of 344 applicants to a top-five PhD program in economics (Grove and Stephen, 2007) indicated a significant positive relationship of both GRE verbal (GRE-V) and quantitative (GRE-Q) scores to program completion. The Probit for GRE-V was 0.064 and for GRE-Q the Probit was 0.152 with standard errors of 0.031 and 0.064 for V and Q, respectively. A study using data from a flagship public university and an Ivy League university (Bridgeman and Cline, 2022) indicated practically and statistically significant predictions of persistence for students in PhD programs in chemistry ($n = 315$) from GRE-Q, but not from GRE-V. Among students in the top quartile of GRE-Q scores in the chemistry programs 14% dropped out compared to 30% dropout in the bottom quartile of GRE scores, yielding a 2(drop/stay) x2(top quartile/bottom quartile) Chi-square with Yates correction of 5.28 with a $p < 0.03$. A study of 203 applicants (over a 7-year period) to a math PhD program at a Tier I public university in California noted significant GRE-V correlations with program completion (Ma et al., 2018); the coefficient from the logistic regression was 0.051 ($p = 0.02$). For GRE-Q, the authors reported a coefficient of 0.065 ($p = 0.07$) and noted that the coefficient "is only marginally significant, possibly due to the small variation in the GREQ scores among math PhD students (all scores tend to be near the maximum possible score of 166, with about a half within 2 points, and three quarters within 7 points of the maximum)."

In summary, some studies find a significant relationship of GRE scores to PhD program completion while others do not, and there are substantial limitations in all of these studies. Attempting to generalize from the existing studies is difficult because they may focus on only a single program at one university or data from just a few highly selective universities. The current study examines early dropout versus persistence to a PhD degree with a very large sample of programs and students by using comprehensive data from the National Student Clearinghouse.

Method

Sample

From ETS files, we identified GRE test takers with scores from 2012 to 2016 so that sufficient time would have elapsed for them to enroll in graduate programs and make some progress toward a degree. Names were sent to the National Student Clearinghouse that then provided data on where and when these students were enrolled in higher education institutions. These data did not indicate whether students' postsecondary enrollment was specific to undergraduate or graduate studies, and therefore we implemented a data selection procedure that would allow us to select students who were likely enrolled in a graduate program. Specifically, using data from the biographical questionnaire that students complete when they register to take the GRE, only students who indicated a doctoral degree goal and who were enrolled in a higher education institution within 2 years of taking the GRE were retained in the sample.

Procedures

We identified students who were still enrolled 4 years after taking the GRE (persisters) or were no longer enrolled (early dropouts). Note that some of the dropouts may have obtained master's degrees in less than 4 years after taking the GRE, but because they had stated a doctoral degree goal they could still appropriately be labeled as dropouts from their stated degree goal. Students were classified according to their intended graduate majors on the biographical questionnaire. It is possible that some students may have switched from their original stated intentions, but because the intention was recorded just as the students were about to apply to graduate school we assume that most students stayed with their stated intentions. We defined these intentions within institutions as programs, and only programs with at least 20 students were retained in the sample because analyses had to be conducted at the program-within-institution level before taking the sample size weighted average over programs. As indicated in Table 1, the sample consisted of 1,672 graduate programs

TABLE 1 Number of programs, students, and program means and SDs.

Program area	# of programs	# of students	Mean	SD
GRE verbal				
Business	22	578	156	3.6
Education	198	10,372	154	2.9
Engineering	137	14,683	155	3.1
Humanities	190	12,835	159	2.6
Life sciences	521	64,574	153	3.2
Physical sciences	198	21,120	156	2.7
Social sciences	406	33,762	155	3.2
GRE quantitative				
Business	22	578	159	5.1
Education	198	10,370	149	3.2
Engineering	137	14,684	162	2.6
Humanities	190	12,812	150	2.8
Life sciences	521	64,565	152	3.2
Physical sciences	198	21,120	159	3.0
Social sciences	406	33,759	150	3.7
GRE analytic writing				
Business	22	582	4.1	0.4
Education	199	10,410	4.0	0.3
Engineering	141	14,989	3.8	0.3
Humanities	190	12,873	4.3	0.3
Life sciences	522	64,709	3.9	0.2
Physical sciences	198	21,294	3.9	0.2
Social sciences	407	33,888	4.1	0.3

containing 157,924 students. Although students are expected to have scores on the three GRE sections, Verbal (GRE-V), Quantitative (GRE-Q), and Analytical Writing (GRE-AW), occasionally a student would leave one section blank resulting in no score for that section and explaining the very slight variations in sample sizes across sections in Table 1. The Mean column is based on the weighted average of the means within each program/institution. The highest GRE-V mean, as might be expected, was in the humanities, but there was relatively little variation across the program means from a low of 153 to a high of 159. On the other hand, there was considerable variation in the GRE-Q mean scores from a low of 149 to a high of 162 with the highest means in business, engineering, and the physical sciences. We also looked at the median scores in each program/institution and noted that on-average medians were within one point of the means.

Analyses

Two methods were used to analyze the results. First, we conducted a series of hierarchical linear models (HLM) predicting the 0–1 criterion of persisting or dropping out from a binary indicator of whether the test takers were international or domestic (United States citizens or resident aliens), GRE-V, GRE-Q, and GRE-AW scores. Then, to make these results easier to visualize, within each program/institution we divided the students by GRE score quartile separately for V, Q, and AW, and identified the percent of students in each quartile who were still enrolled 4 years after taking the GRE (persisters) and took the within-institution sample size weighted average over all institutions. Note that both the HLM and quartile analyses account for students embedded in different programs/institutions; the GRE quartiles reflected different score levels in different programs/institutions and persistence v dropout was unique to each program/institution.

Hierarchical linear models

To account for the clustering of test takers within institutions, we conducted a series of generalized estimating equation logistic regressions in each of the study majors: business, education, engineering, humanities, life sciences, physical sciences, and social sciences. For each major, first an unconditional random effects analysis of variance (ANOVA) with PhD persistence as the outcome variable was conducted to compute the ICC for dichotomous outcomes (Snijders and Bosker, 1999; as cited in O'Connell et al., 2008) and determine the proportion of variance due to institutional variability. ICC values above 0.10 indicate the need to use multi-level modeling to account for the clustering of the data (Lee, 2000). In a second set of models, we added a binary variable for international versus domestic student status and each of the three GRE scores. Based on recommendations by Enders and Tofighi (2007), the continuous predictors were group mean centered. To be consistent with the sample selection for the quartile analysis, we restricted the analysis to programs within institutions with at least 20 individuals.

Quartile analysis

Within each program/institution, and separately for each of the three GRE scores, we identified students the bottom score quartile through the top score quartile. We computed the mean GRE score in each quartile in each program/institution and took the sample size weighted average of these means. These means and their standard deviations are in Table 2. Note that in the full sample of GRE test-takers the standard deviations for V, Q, and AW are 8.6, 9.6, and 0.9 (GRE, 2022), so the differences in means between the first and fourth quartiles are substantial. We also used the same procedure with undergraduate point average (UGPA) and with combined scores, e.g., students who were in top quartile on three measures (UGPA, GRE-V, and GRE-AW) or in the bottom quartile in all three measures.

Results

Hierarchical linear models

Although the unconditional models revealed that the proportion of between institution variance was over the suggested value of 0.10

10.3389/feduc.2023.1182508

TABLE 2 Program means and SDs by GRE quartile.

	Q1		Q2		Q3		Q4	
Program	Mean	SD	Mean	SD	Mean	SD	Mean	SD
GRE-V								
Business	147	3.7	154	3.7	159	4.2	165	3.0
Education	146	3.0	152	3.2	157	3.3	163	2.8
Engineering	148	3.4	154	3.1	159	2.8	164	2.2
Humanities	152	3.2	158	2.9	162	2.7	167	1.9
Life sciences	147	2.9	152	3.2	156	3.3	161	3.3
Physical sciences	149	3.1	155	2.9	160	2.7	165	2.1
Social sciences	147	3.4	154	3.6	158	3.5	163	3.1
GRE-Q								
Business	148	6.0	157	6.2	163	4.9	168	3.2
Education	141	3.0	147	3.3	151	3.5	158	3.6
Engineering	156	3.0	162	2.6	165	2.4	169	1.4
Humanities	143	2.7	149	2.9	153	3.0	160	3.0
Life sciences	146	3.0	151	3.2	155	3.3	160	3.4
Physical sciences	154	3.4	160	3.3	164	3.1	168	2.0
Social sciences	143	3.3	150	3.9	154	4.3	160	4.6
GRE-AW								
Business	3.1	0.4	3.8	0.4	4.4	0.4	5.1	0.4
Education	3.1	0.3	3.9	0.3	4.3	0.3	5.0	0.3
Engineering	3.0	0.3	3.6	0.3	4.2	0.3	4.9	0.3
Humanities	3.4	0.3	4.2	0.3	4.7	0.3	5.4	0.3
Life sciences	3.1	0.3	3.8	0.3	4.2	0.3	4.8	0.3
Physical sciences	3.1	0.2	3.8	0.3	4.3	0.3	5.0	0.3
Social sciences	3.2	0.3	3.9	0.3	4.4	0.3	5.1	0.3

(Lee, 2000) only for business (ICC = 0.12), we nevertheless ran HLM models for all of the programs. The results indicate an improvement in better model fit on the models including the predictors over the unconditional model (Tables 3–9). Overall, the results from the model with predictors indicate that GRE-V was a significant positive predictor and GRE-Q was a significant negative predictor of PhD persistence in every major. While the effect of GRE-AW was positive in every major (except in life sciences where GRE-AW had a significant negative relationship), its relationship was statistically significant only in Education, Engineering and Physical Sciences. It should be noted that while the regression coefficients for GRE-V and GRE-Q represents the single point increase on a scale that goes from 130 to 170, the GRE-AW scale is from 0 to 6, hence, appearing to have a larger effect in some majors.

Quartile analysis

The quartile results for persistence in graduate school are in Figures 1–3. For both GRE-V and GRE-W the percent of persisters in the fourth GRE quartile was always larger, often much larger, than the percent of persisters in the first quartile. For example, in engineering only 25% of the students in the first GRE-AW quartile persisted while 73% of the fourth quartile students were still enrolled 4 years after taking the GRE; in the physical sciences the percent persisting was 40% in the first quartile and 78% in the fourth. Surprisingly (but consistent with the HLM analysis), the results were reversed for GRE-Q with lower scoring students more likely to persist. For example, in the physical sciences 75% of the first quartile students persisted while only 50% of the fourth quartile students did so. Programs show a monotonic increase from the first to forth quartiles with the exception of GRE-AW in the life sciences with a decline from the first to third quartiles, but then a positive jump in the fourth quartile. This odd fit could explain the negative Logit weight and odds ratio less than 1.0 for GRE-AW in the HLM analyses for the life sciences while the weight is positive in all other majors.

Another way to look at these data is to determine the percent of programs in each program area in which there are more persisters in the top GRE score quartile than in the bottom quartile. These percents are in Table 10. Results mirrored the conclusions from Figures 1–3 with especially strong results for programs in business, engineering, and physical sciences for both GRE-V and GRE-AW scores, but results in the opposite direction for GRE-Q.

We wondered if the pattern would be the same in more and less selective institutions, so we divided programs in thirds based on mean GRE-Q scores at the institution (i.e., highly selective, average, and less selective institutions). But the pattern was essentially the same for the more and less selective institutions.

TABLE 3 HLM results for Business ($N = 571$).

	Unconditional model	Model with predictors	
	Logit (SE)	Logit (SE)	Odds ratio
Fixed effects			
Intercept	0.296 (0.0168)*	1.097*** (0.175)	
International		−2.421*** (0.279)	0.089
GRE-Q		−0.055** (0.019)	0.946
GRE-V		0.077*** (0.023)	1.08
GRE-AW		0.045 (0.179)	1.046
Covariance estimates			
Level − 2 intercept (τ_{00})	0.444 (0.195)	0.253 (0.166)	
ICC*	0.119	0.071	
Model fit			
-2 LL	753.70	551.09***	
AIC	757.70	563.09	
BIC	759.89	569.64	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The p values attached to the likelihood are related to a likelihood ratio test.

TABLE 4 HLM results for Education ($N = 10,246$).

	Unconditional model	Model with predictors	
	Logit (SE)	Logit (SE)	Odds ratio
Fixed effects			
Intercept	1.914 (0.046)	2.152*** (0.046)	
International		−3.265*** (0.134)	0.038
GRE-Q		−0.013* (0.006)	0.987
GRE-V		0.022*** (0.006)	1.022
GRE-AW		0.114* (0.053)	1.121
Covariance estimates			
Level − 2 intercept (τ_{00})	0.190 (0.037)	0.145 (0.035)	
ICC*	0.055	0.042	
Model fit			
-2 LL	8247.64	7224.44***	
AIC	8251.64	7236.44	
BIC	8258.22	7256.17	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The p values attached to the likelihood are related to a likelihood ratio test.

TABLE 5 HLM results for Engineering ($N = 14,612$).

	Unconditional model	Model with predictors	
	Logit (SE)	Logit (SE)	Odds ratio
Fixed effects			
Intercept	0.173 (0.045)	1.238*** (0.042)	
International		−3.134*** (0.060)	0.044
GRE-Q		−0.043*** (0.005)	0.957
GRE-V		0.032*** (0.005)	1.032
GRE-AW		0.188*** (0.041)	1.206
Covariance estimates			
Level − 2 intercept (τ_{00})	0.207 (0.035)	0.086 (0.022)	
ICC*	0.059	0.026	
Model fit			
-2 LL	19838.03	12761.69***	
AIC	19842.03	12773.69	
BIC	19847.87	12791.21	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The p values attached to the likelihood are related to a likelihood ratio test.

TABLE 6 HLM results for humanities ($N = 12,754$).

	Unconditional model	Model with predictors	
	Logit (SE)	Logit (SE)	Odds ratio
Fixed effects			
Intercept	1.267 (0.034)	1.480*** (0.037)	
International		−2.793*** (0.096)	0.061
GRE-Q		−0.018*** (0.004)	0.982
GRE-V		0.037*** (0.005)	1.037
GRE-AW		0.018 (0.037)	1.018
Covariance estimates			
Level − 2 intercept (τ_{00})	0.111 (0.021)	0.126 (0.024)	
ICC*	0.033	0.037	
Model fit			
-2 LL	13676.79	12268.01***	
AIC	13680.79	12280.01	
BIC	13687.29	12299.49	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The p values attached to the likelihood are related to a likelihood ratio test.

TABLE 7 HLM results for life sciences ($N = 64,254$).

	Unconditional Model	Model with predictors	
	Logit (SE)	Logit (SE)	Odds ratio
Fixed effects			
Intercept	1.251 (0.028)	1.403*** (0.031)	
International		−2.655*** (0.052)	0.070
GRE-Q		−0.034*** (0.002)	0.967
GRE-V		0.044*** (0.002)	1.045
GRE-AW		−0.075*** (0.018)	0.928
Covariance estimates			
Level − 2 intercept (τ_{00})	0.320 (0.026)	0.374 (0.030)	
ICC*	0.088	0.102	
Model fit			
-2 LL	67166.57	62329.09***	
AIC	67170.57	62341.09	
BIC	67179.09	62366.62	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The p values attached to the likelihood are related to a likelihood ratio test.

TABLE 8 HLM results for physical sciences ($N = 21,025$).

	Unconditional model	Model with predictors	
	Logit (SE)	Logit (SE)	Odds ratio
Fixed effects			
Intercept	0.738 (0.036)	1.472*** (0.032)	
International		−3.199*** (0.056)	0.041
GRE-Q		−0.021*** (0.004)	0.979
GRE-V		0.028*** (0.004)	1.029
GRE-AW		0.089** (0.032)	1.093
Covariance estimates			
Level − 2 intercept (τ_{00})	0.172 (0.025)	0.085 (0.018)	
ICC*	0.050	0.018	
Model fit			
-2 LL	26711.11	18666.42***	
AIC	26715.11	18678.42	
BIC	26721.68	18698.15	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The p values attached to the likelihood are related to a likelihood ratio test.

TABLE 9 HLM results for Social sciences ($N = 33,477$).

	Unconditional model	Model with predictors	
	Logit (SE)	Logit (SE)	Odds ratio
Fixed effects			
Intercept	1.204 (0.028)	1.396*** (0.029)	
International		−2.525*** (0.059)	0.080
GRE-Q		−0.012*** (0.003)	0.988
GRE-V		0.051*** (0.003)	1.052
GRE-AW		0.026 (0.023)	1.027
Covariance estimates			
Level − 2 intercept (τ_{00})	0.219 (0.022)	0.220 (0.024)	
ICC*	0.062	0.063	
Model fit			
-2 LL	36404.09	32975.44***	
AIC	36408.09	32987.44	
BIC	36416.1	33011.44	

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. The p values attached to the likelihood are related to a likelihood ratio test.

Next we evaluated persistence for top and bottom quartiles of UGPA and for students who were in the top or bottom quartiles on multiple measures, that is, for example, a student would need to be in the top quartile in their program/institution on GRE-V and GRE-AW and GRE-Q or in the bottom quartile on all three scores. These results are in Table 11. Note that because results for GRE-Q are in the opposite direction there is more of an increase in persistence for the combination of just GRE-V and GRE-AW than for the combination that includes all three GRE scores. Also note that UGPA has a very limited impact as an indicator of persistence. This may be due to the very restricted range and generally high scores for UGPA with mean UGPAs of 3.6 or higher on the 0–4 scale for all majors.

Discussion and conclusions

In both the HLM and quartile analyses, and with a much larger sample than in any prior research, greater persistence rates were associated with higher GRE-V and GRE-AW scores, but with lower GRE-Q scores. We do not have enough information to fully understand this result for GRE-Q, but one speculation is that students with very strong quantitative skills may drop out with just a master's degree (despite their originally stated intention to get a PhD) when they realize that they already have the potential to earn a high salary without putting in the additional years needed to earn a PhD. Although exact data on the number of students choosing this path are not available, anecdotal reports are common. For example, a CNN report

TABLE 10 Percent of Programs in which the percent of persisters is higher in the fourth quartile than in the first quartile.

GRE verbal	
Program area	% of programs with more students persisting in top quartile
Business	86
Education	55
Engineering	97
Humanities	64
Life sciences	69
Physical sciences	88
Social sciences	76
GRE quantitative	
Business	14
Education	32
Engineering	15
Humanities	38
Life sciences	32
Physical sciences	22
Social sciences	45
GRE analytical writing	
Business	77
Education	60
Engineering	97
Humanities	65
Life sciences	53
Physical sciences	90
Social sciences	67

focused on the number of PhD dropouts getting high-paying jobs in Silicon Valley and noted:

> Dropouts are nothing new to the Valley. Quite the opposite: The tech turk–characteristically, someone too brilliant, too arrogant, too obsessed for the classroom–is key to the Valley's creation myth and the stories it tells about itself. Sergey Brin, Larry Page and Jerry Yang dropped out of graduate programs (Ozy, 2014).

Consistent with this speculation are the large declines in persistence in business, engineering, and physical sciences for students where there are many employment opportunities for students with strong quantitative skills, and the relatively flat profiles for education, humanities, and the social sciences.

Additional research is also needed to better understand the apparently low persistence rates for international students. Our data did not indicate whether students left graduate school after receiving a master's degree. Although we considered any such students to be non-persisters given a stated degree goal of a doctorate, international students may have been more likely to intentionally or unintentionally indicate a doctoral degree goal when they were actually seeking a master's degree.

These results suggest that GRE scores could have a place as part of a holistic review of potential PhD candidates. We fully support the conclusions in a 2014 *Nature* article (Miller and Stassun, 2014) that GRE scores should not have a disproportionate weight in admissions decisions; many factors should be considered in making holistic admissions decisions, and relatively low GRE scores should not be used to reject an otherwise clearly qualified candidate. But the relationship of GRE scores to persistence demonstrated in this research suggests that while GRE scores should not have disproportionate weight they also should not have zero weight whether used as part of an admissions decision or in identifying admitted students who may need extra support to avoid early dropout.

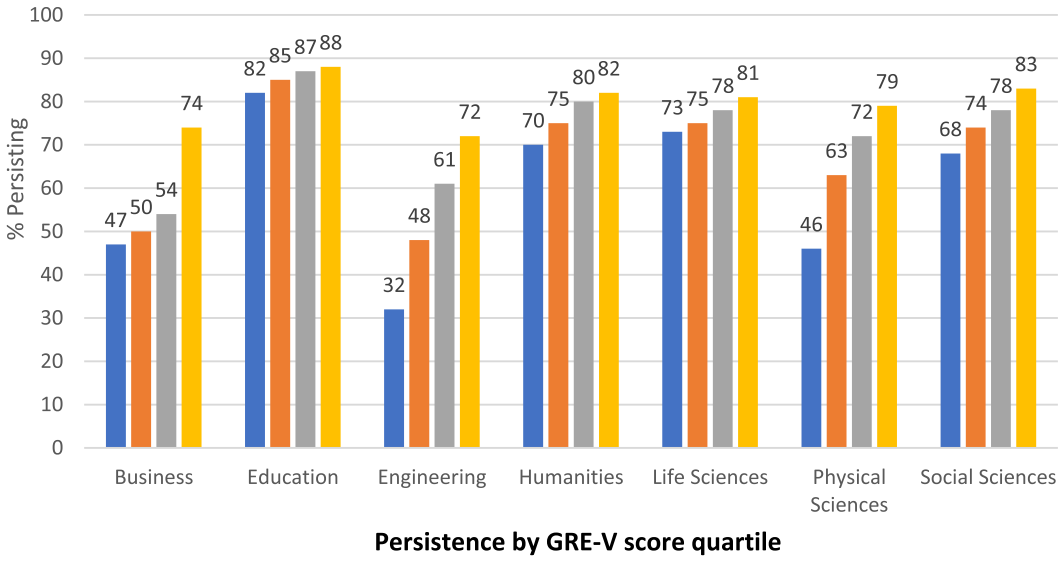


FIGURE 1
Persistence by GRE-V score quartile.

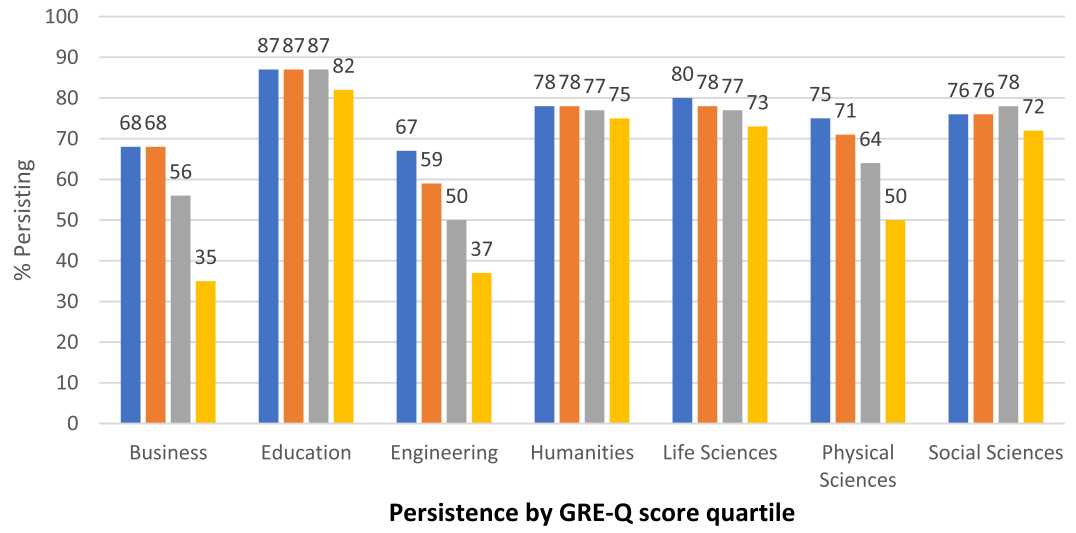


FIGURE 2
Persistence by GRE-Q score quartile.

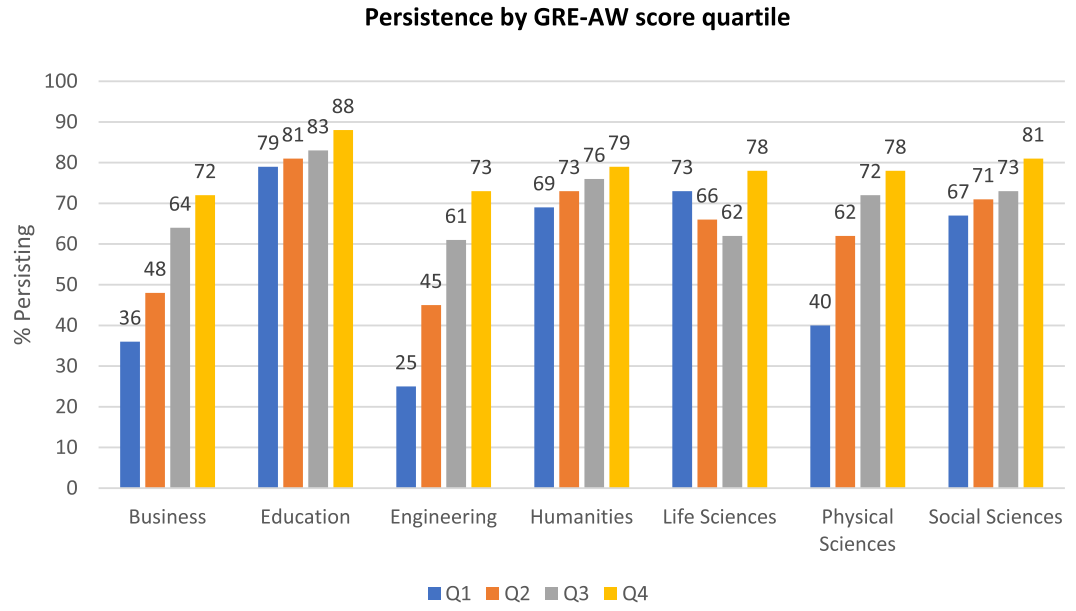


FIGURE 3
Persistence by GRE-AW score quartile.

TABLE 11 Percent persisting in first and fourth quartiles for combined predictors with cell sizes in parentheses.

Study major	Q1 in VWQ	Q4 in VWQ	Q1 in VW	Q4 in VW	Q1 in UGPA	Q4 in UGPA	Q1 in VW and UGPA	Q4 in VW and UGPA
Business	56% (34)	86% (21)	34% (74)	88% (60)	61% (132)	61% (117)	33% (21)	94% (16)
Education	85% (691)	88% (572)	78% (1180)	89% (1056)	87% (2240)	86% (3024)	82% (399)	90% (457)
Engineering	45% (568)	73% (519)	21% (1966)	76% (1746)	56% (3084)	53% (3906)	27% (452)	77% (514)
Humanities	72% (829)	81% (681)	65% (1397)	81% (1239)	75% (2829)	79% (742)	66% (511)	79% (101)
Life sciences	77% (3873)	82% (3546)	70% (6813)	82% (6523)	80% (15189)	75% (19462)	73% (2176)	82% (2581)
Physical sciences	59% (817)	81% (860)	31% (2778)	81% (2255)	68% (4757)	66% (5262)	41% (690)	82% (720)
Social sciences	71% (1769)	83% (1441)	61% (2917)	83% (2706)	73% (5496)	76% (5207)	66% (1108)	82% (793)

V is GRE-V, W is GRE-AW, Q is GRE-Q, UGPA is self-reported undergraduate grade point average.

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 ETS IRB. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

BB, MO-A, and SH contributed to the conceptualization, analyses, and manuscript writing for this study. All authors contributed to the article and approved the submitted version.

Funding

The study received funding from ETS.

Conflict of interest

ETS researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in this report do not necessarily reflect ETS positions or policies.

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.

References

Bridgeman, B., and Cline, F. (2022). Can the GRE predict valued outcomes? Dropout and writing skill. *PLoS One* 17:e0268738. doi: 10.1371/journal.pone.0268738

Cassuto, L. (2013). *Ph.D. attrition: how much is too much?* The Chronicle of Higher Education. Available at: https://www.chronicle.com/article/ph-d-attrition-how-much-is-too-much/.

Enders, C. K., and Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol. Methods* 12, 121–138. doi: 10.1037/1082-989X.12.2.121

GRE. (2022). *Interpreting your GRE scores: 2022-23*. Available at: https://www.ets.org/pdfs/gre/interpreting-gre-scores.pdf.

Grove, W. A., and Stephen, W. (2007). The search for economics talent: doctoral completion and research productivity. *Am. Econ. Rev.* 97, 506–511. doi: 10.1257/aer.97.2.506

Hall, J. D., O'Connell, A. B., and Cook, J. G. (2017). Predictors of student productivity in biomedical graduate school applications. *PLoS One* 12:e0169121. doi: 10.1371/journal.pone.0169121

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: the case of school effects. *Educ. Psychol.* 35, 125–141. doi: 10.1207/S15326985EP3502_6

Ma, T., Wood, K. E., Xu, D., Guidotti, P., Pantano, A., and Komarova, N. L. (2018). Admission predictors for success in a mathematics graduate program. *ArXiv* 2018. doi: 10.48550/arXiv.1803.00595

Miller, C., and Stassun, K. (2014). A test that fails. *Nature* 510, 303–304. doi: 10.1038/nj7504-303a(2014)

Miller, C. W., Zwickl, B. M., Posselt, J. R., Silvestrini, R. T., and Hodapp, T. (2019). Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion. *Sci. Adv.* 5:eaat7550. doi: 10.1126/sciadv.aat7550

Miller, C. W., Zwickl, B. M., Posselt, J. R., Silvestrini, R. T., and Hodapp, T. (2020). Response to comment on typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion. *Sci. Adv.* 6:4647. doi: 10.1126/sciadv.aba4647

Nevill, S. C., and Chen, X. (2007). *The path through graduate school: a longitudinal examination 10 years after bachelor's degree. (NCES No. 2007-162)*. Washington, DC: National Center for Education Statistics, US Department of Education.

O'Connell, A. A., Goldstein, J., Rogers, H. J., and Peng, C. Y. J. (2008). "Multilevel logistic models for dichotomous and ordinal data" in *Multilevel modeling of educational data*. eds. A. A. O'Connell and D. B. McCoach (Charlotte, NC: Information Age Publishing, Inc), 199–242.

Ozy, P. B. (2014). *PhD dropout?* Lucrative Silicon Valley career. Available at: https://money.cnn.com/2014/09/09/news/economy/ozy-dropout-career/index.html.

Petersen, S. L., Erenrich, E. S., Levine, D. L., Vigoreaux, J., and Gile, K. (2017). Multi-institutional study of GRE scores as predictors of STEM PhD degree completion: GRE gets a low mark. *PLoS One* 13:e0206570. doi: 10.1371/journal.pone.0206570

Snijders, T. A. B., and Bosker, R. J. (1999). Multilevel analysis An introduction to basic and advanced multilevel modeling. Thousand Oaks: Sage Publications.

Weisman, M. B. (2019).*Do GRE scores help predict getting a physics Ph.D.: a comment on a paper by Miller et al.* Available at: https://arxiv.org/ftp/arxiv/papers/1902/1902.09442.pdf.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 29 August 2023
DOI 10.3389/feduc.2023.1229129

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Diana Pereira,
University of Minho, Portugal
Michalis P. Michaelides,
University of Cyprus, Cyprus

*CORRESPONDENCE
Juan Carlos Pérez-Morán
✉ jucarpint@gmail.com

RECEIVED 25 May 2023
ACCEPTED 11 August 2023
PUBLISHED 29 August 2023

CITATION
Henriquez PS, Pérez-Morán JC, del Cid García CJ and Zamora JE (2023) Factor structure and invariance of the scale to measure teaching performance in the area of social sciences.
Front. Educ. 8:1229129.
doi: 10.3389/feduc.2023.1229129

COPYRIGHT
© 2023 Henríquez, Pérez-Morán, del Cid García and Zamora. 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.

Factor structure and invariance of the scale to measure teaching performance in the area of social sciences

Patricio Sebastián Henríquez[1], Juan Carlos Pérez-Morán[2]*, Carlos Javier del Cid García[1] and Jesús Enrique Zamora[1]

[1]Autonomous University of Baja California, Ensenada, Mexico, [2]Promoter Network of Diagnostic Evaluation Methods and Educational Innovation, Ensenada, Mexico

The use of scales to evaluate teaching from the students' perspective is a method frequently used in educational systems around the world. The objective of this study is to analyze the factorial structure of the Teaching Performance Evaluation Scale (EEDDocente, by acronyms in Spanish) designed with the purpose of providing information that favors decision-making based on evidence for the improvement of teaching in the area of Social Sciences, as well as measuring the invariance by *School stage* and *Educational Program*. The sample consisted of 1,849 students of the Bachelor's Degrees in Law, Psychology, Accounting, Administration, Education Sciences, Communication Sciences, Computer Science, and Sociology of the School of Social and Administrative Sciences (FCAyS) of the Autonomous University of Baja California, Mexico. Based on a three-factor model that meets the fit and quality criteria, a Multi-group Confirmatory Factor Analysis (MGCFA) was performed to measure the invariance of the EEDDocente by *School stage* and *Educational program*. It is concluded that the three-factor model can be used to measure, from the students' perspective, the performance of teachers in the *Area of Social Sciences*. Likewise, it is concluded that the invariance of the simultaneous measurement is achieved, providing evidence to perform mean difference analysis between the different *Educational programs*.

KEYWORDS

students' evaluation of teaching, higher education students, validity, confirmatory factor analysis, invariance

1. Introduction

The evaluation of teaching in Higher Education Institutions (HEIs) is one of the most relevant components linked to the assumption of improving educational quality (Calatayud, 2021; Torquemada, 2022; Bleiberg, 2023; Silva, 2023). The measurement of the effectiveness of teaching practice occupies a central place in HEIs strategies, which allows the generation of information on the teaching and learning process that serves as an input to trace routes for improving the quality, relevance, effectiveness and accountability of education systems around the world (Chen and Hoshower, 2003; Liebowitz, 2021; Seivane and Brenlla, 2021; Camacho, 2022; Zhao et al., 2022). In addition, it is a transcendental input for the improvement and feedback of teacher performance in its multiple dimensions, thus attending to the formative function of this process (Marsh, 2007; Luna and Torquemada, 2008; Liebowitz, 2022; Silva et al., 2022).

Around the world, accountability and the growing demand to ensure the improvement of learning of future professionals graduating from universities has placed the evaluation of teaching performance as the dominant axis of educational policies (Vaillant, 2016; Liebowitz, 2022). However, there is a growing interest in the methodological aspects, techniques and instruments for collecting information (questionnaires, attitude scales, interviews, focus groups, classroom observation), and the subjects (students, teachers, managers, external experts) best suited to obtain reliable, valid, sufficient and relevant data on the evaluation of teaching (Marsh, 1984; Cruz Ávila, 2007; Romero and Martínez, 2017; Zamora, 2021; Bleiberg, 2023).

In particular, there is growing concern about the use of Students' Evaluations of Teaching (SET) to make high-impact or consequential decisions in processes such as promotion, tenure, and awarding of stimuli and incentives (Boring et al., 2016; Hornstein, 2017; Wang and Guan, 2017; Benton, 2018; Ching, 2018; Mitchell and Martin, 2018; Bazán-Ramírez et al., 2021). Likewise, the purposes of the evaluation of teacher performance have been mainly oriented to condition the hiring or dismissal processes of academic staff, deciding who gets an economic incentive or job promotion based on the result of the teacher evaluation (Stroebe, 2016; Gómez and Valdés, 2019). However, more and more decision makers, education systems and HEIs see an opportunity and advantage in using relatively brief SETs, mid- or end-of-course, to provide formative feedback on teacher performance and competencies (Marsh, 2007; Silva et al., 2022).

The study of university teacher performance began to become widespread internationally in the 1980s, as part of the accountability processes derived from changes in government policies for financing higher education (Cisneros-Cohernour and Stake, 2010; Zamora, 2021; García-Olalla et al., 2022). The evaluation of teaching performance has its genesis in the first student learning assessment systems in the United States during the 1920s; by the second half of this decade, learning assessment served as a tool to evaluate teaching (Alcaraz-Salarirche, 2015; Zhao et al., 2022). For its part, the SET was an innovation in HEIs in the United States, a consequence of the consumer orientation of the capitalist system: students, as users of the educational service, are the ones who should evaluate it (García, 2000). During the 1980s and early 1990s, numerous studies were carried out on the subject, for example, those of Cohen (1981, 1983), Feldman (1988, 1989a,b, 1990, 1992, 1993), and Marsh (1984, 1986, 1987, 1993, 2007).

Teacher performance evaluation must maintain high and solid technical quality standards to fulfill its main purpose, which is linked to improving teaching practices and student learning. However, in the mid-1990s, studies began to emerge that questioned the reliability and usefulness of quantitative instruments to evaluate teaching (Theall et al., 2001; García, 2014; Boring et al., 2016; Benton, 2018; Ching, 2018). Among the most recurrent criticisms by researchers are those related to the idea that the SET presents problems of logic and structure in terms of components and characteristics to test the effectiveness of teaching, and random and biased responses, as well as a subjective judgment on the part of students about teaching (Stroebe, 2016, 2020; Wang and Guan, 2017; Ching, 2018; Zhou and Qin, 2018; Gu et al., 2021).

Despite criticisms to the contrary, it is undeniable that the use of scales and questionnaires has been the most widely used mechanism to evaluate university teachers (Wang and Guan, 2017; Zamora, 2021), and that questionnaires as evaluation instruments are viable tools to measure the effectiveness of teaching in HEIs (García, 2014; Mohammadi, 2021). However, it is necessary that evaluation questionnaires maintain validity, reliability, relevance, and pertinence for uses and consequences in the educational context (Messick, 1995; Kane, 2006; International Test Commission, 2013; Spooren et al., 2013; American Educational Research Association, 2018; Reyes et al., 2020; Lera et al., 2021).

The evaluation of the quality of teaching practice through experience, certifications, academic degrees, among other factors, shows little correlation with the effectiveness of teaching performance (Williams and Hebert, 2020). Thus, the evaluation of teaching, based on the perception of students, currently has a preponderant role in the processes of improving the quality of teaching in universities (Aleamoni, 1999; Salazar, 2008; Mohammadi, 2021; Zamora, 2021). The SET allows HEIs to have a reference for the improvement of teaching practice, as long as the performance measures maintain a high level of objectivity, methodological rigor and relationship with the implementation dimensions of the academic objectives of educational systems (House, 1998; Navarro and Ramírez, 2018; Seivane and Brenlla, 2021). At this point, it is important to mention that most of the criteria or dimensions of the SET are defined by committees of specialists, which are based on models of indicators of teaching quality and effectiveness, but with a strong influence of philosophical and pedagogical principles, and of the policy and regulations of the functions of the academic staff of each HEIs.

Among the first syntheses of criteria to design SET are those proposed by Feldman (1976) and Hildebrand et al. (1971). By analyzing students' points of view, Feldman (1976) proposed three categories for effective teaching: (a) *Presentation*, which includes teachers' enthusiasm for teaching, their knowledge of the subject, and clarity of presentation and organization of the course; (b) *Facilitation*, which refers to the availability of teachers for consultation, respect for students and their ability to encourage students to achieve learning in class; and (c) *Regulation*, which includes teachers' ability to set clear objectives, classroom management skills, appropriateness of course materials and activities, and fairness in student assessment and feedback. For their part, Hildebrand et al. (1971) and Hildebrand (1973) propose five factors to measure teaching effectiveness: (a) *Analysis and synthesis skills*, which refers to the teacher's mastery of class content; (b) *Clarity and organization*, which refer to the teacher's ability to present course topics; (c) *Interaction with the group*, which refers to the teacher's ability to interact with students and maintain the active participation of the group; (d) *Interaction with each student*, which refers to the teacher's ability to establish trust and respect with each individual student; and (e) *Dynamism and enthusiasm*, which refers to the teacher's enthusiasm and pleasure in teaching the subject. More recently, authors such as Marsh (1987), Marsh and Dunkin (1997), Richardson (2005), and Schellhase (2010) have proposed models of up to nine to 10 criteria (*assignments and readings, breadth of coverage, examinations and grading, group interaction, individual rapport, instructor enthusiasm, learning and academic value, organization and clarity, workload and difficulty,* and *summative evaluation*) to assess the quality of instruction.

In Ibero-America, several authors have focused on the design and validity of the measurement of teacher performance through scales considering various criteria models of the effectiveness and quality of teaching; in particular, on obtaining the psychometric properties of the measurement instruments, and on the evidence of their internal

consistency and reliability. In this sense, the studies by García-Gómez-Heras et al. (2017), in Spain; Estrada et al. (2019), in Nicaragua; Bazán-Ramírez et al. (2021), in Peru; and Márquez and Madueño (2016), and Bazán-Ramírez and Velarde-Corrales (2021), in Mexico; they are noteworthy. For their part, García-Gómez-Heras et al. (2017) focused on revealing which professors' behaviors were most appreciated by first-year students of studies taught at the Faculty of Health Sciences of the Rey Juan Carlos University of Madrid (Degrees in Medicine, Nursing, Physiotherapy, Dentistry Psychology and Occupational Therapy). The authors applied the questionnaire developed by Tuncel (2009) on the behaviors of university teachers that influence the academic performance of students. This questionnaire is made up of 48 items organized into six factors: (a) *Emotional aptitude of university teachers*, (b) *Teacher-student interaction*, (c) Achievement of educational objectives, (d) *Theory-practice relationship*, (e) *Organization and planning*, and (f) *Feedback*.

Likewise, Estrada et al. (2019), Gómez and Valdés (2019) conducted a study to establish the psychometric properties of the *Opinion Questionnaire for the Evaluation of Teaching Performance* (OQETP) composed of 38 items, focused on evaluating teaching performance from the students' perception, at the National University of Trujillo, Nicaragua. The OQETP items are presented on a Likert-type scale with five response categories and are organized into eight questionnaire dimensions: (a) *Formal Responsibility*, (b) *Methodology*, (c) *Communication*, (d) *Materials*, (e) *Attitude*, (f) *Evaluation*, (g) *Motivation*, and (h) *Satisfaction*.

In Peru, Bazán-Ramírez et al. (2021) analyzed the factorial structure of the *Teaching Performance Scale for Psychology Teachers* (EDDPsic) and measured the invariance between groups (according to *gender*, *age* and *academic stage*). This instrument was designed based on the model of five didactic performance criteria (Carpio et al., 1998; Silva et al., 2014). In total, the EDDpsic is made up of 18 items ($K = 18$) organized into subscales: (a) *Competence Exploration* ($k = 3$), (b) *Criteria Explanation* ($k = 5$), (c) Illustration ($k = 3$), (d) Feedback ($k = 4$), and (e) Evaluation ($k = 3$). Their study involved 316 Psychology students, from basic cycles (fourth and sixth semesters) and disciplinary-professional *cycles* (eighth and tenth semesters), from two public universities in Peru. They also performed a Multigroup Confirmatory Factor Analysis (MGCFA) with the five-factor model that showed the best fit indices. Based on their results, they determined the invariance of the scale measure across the three study variables (*age, sex* and *academic stage*), for which the participants were divided into independent groups. The results revealed adequate fitness tests for the *Configural* model in each of the three variables (χ^2 $p > 0.05$, CFI < 0.01, RMSEA $\leq$0.06), so it was considered that the structure of the model was the same for each group. Similarly, evidence of factorial invariance was obtained for the *Weak* (M1), *Strong* (M2) and *Strict* (M3) models, in the variables of *age* (M1: CFI = −0.004; M2: CFI = -0.001; M3: CFI = −0.001), and *gender* (M1: CFI = −0.001; M2: CFI = −0.001; M3: CFI = −0.001). In the case of the *academic stage* variable, evidence of invariance was obtained for the *Weak* and *Strong* models (M1: CFI = −0.004; M2: CFI = −0.000) but not for the *Strict* model (M3: CFI = −0.018).

In Mexico, Márquez and Madueño (2016) analyzed the psychometric properties of an instrument made up of 16 items ($K = 16$) applied to students at a university in Sonora, to recover their opinion on the basic competencies of teachers in the teaching of undergraduate courses. From the 30,224 questionnaires answered, the construct validity of the instrument was determined using the principal components method with Varimax rotation, extracting two factors: (a) *Pedagogical mediation* ($k = 11$), and (b) *Teaching attitudes* ($k = 5$). For their part, Bazán-Ramírez and Velarde-Corrales (2021) evaluated the performance of teachers and students in their didactic interactions through the self-report of 124 psychology students in Mexico. The authors obtained the construct validity, convergent and divergent, of five *didactic performance criteria*, both of the teacher and the student, by means of EFA and CFA. The validation confirmed the theoretical structuring of five factors that correspond to the five dimensions: (a) *Exploration of competencies,* (b) *Explicitness of criteria,* (c) *Illustration,* (d) *Feedback*, and (e) *Evaluation*, derived from the models of didactic performance (Carpio et al., 1998; Silva et al., 2014). The authors also performed descriptive analyses of the students' responses to the didactic performance criteria, according to their *academic stage*, *age* and *sex*.

In summary, it can be said that the models of criteria and instruments to evaluate teaching in the HEIs present a wide diversity of components and characteristics. Likewise, these instruments generally present acceptable psychometric properties of reliability and validity. However, most of them are made up of a large number of criteria and items, which results in instruments that can help in a diagnosis with greater depth and granulation, but make it difficult to apply in cases where students have to answer repeatedly an instrument for each of their teachers at the end of each school year and throughout their university studies. Although, it is important to highlight that most measurement models based on more than five criteria do not satisfactorily meet all the necessary technical quality criteria. In this regard, several authors mention that one of the problems of SET is that multidimensional models that try to cover a large number of criteria based on theory present internal structure problems (Stroebe, 2016, 2020; Ching, 2018). This is explained to some extent when universities include in their teacher evaluations criteria that refer to affective components such as *student satisfaction with the class*, *interest in the subject content*, *teacher's capacity for empathy*, among others. So far it can be concluded that the instruments for measuring the effectiveness and quality of teaching that seek to include a wide variety of criteria present problems of logic and internal structure, as well as difficulties for their application in evaluation strategies where it is required that students respond repeatedly at a specific time in the school year. Another important problem to mention is that most of the SETs evaluate different criteria, making it impossible to make comparative studies that help to evaluate the policies to improve the quality of teaching among different educational programs, schools, and universities.

This paper analyzes the psychometric properties and evidence of construct validity of internal structure and invariance of the Teaching Performance Evaluation Scale (EEDDocente, by acronyms in Spanish) applied at the middle of each school stage to assess the performance of each one of the teachers in the different educational programs of the School of Administrative and Social Sciences (FCAyS, for its acronym in Spanish) of the Autonomous University of Baja California (UABC, for its acronym in Spanish). The EEDDocente is applied biannually with the purpose of identifying strengths and weaknesses of teaching performance from the students' perspective and thus provide feedback on teaching and design teacher training and updating courses (Cashin and Downey, 1992; Liebowitz, 2021; Zamora, 2021; Silva et al., 2022).

Despite the variety of instruments for the evaluation of teaching practice, the relevance of the EEDDocente lies in its purpose, design and objective that seek to maintain coherence between the instrument

and the use of results (Stroebe, 2016; Estrada et al., 2019; Gómez and Valdés, 2019; Aravena-Gaete and Gairín, 2021). The EEDDocente was designed to provide information to identify teachers' needs for updating and continuous training, and to influence the improvement of performance and teaching practices at the classroom level. Among its specific characteristics, the EEDDocente focuses on student-centered teaching and, based on this, the information provided by the scale seeks to generate processes of reflection on teaching practice and a change in the conception and vision of how they develop university teaching (Tomás-Folchy and Durán-Bellonch, 2017).

However, there is no evidence related to the internal structure and invariance of this instrument. This paper aims to address this problem and contribute to the existing literature by analyzing the internal structure of the three-factor model of a reduced version ($K = 15$) of the EEDDocente that is based on categories of solid theoretical models: (a) *Classroom organization*, (b) *Teaching quality*, and (c) *Learning assessment/feedback* (Hildebrand et al., 1971; Feldman, 1976; König et al., 2017; Nasser-Abu, 2017; Chan, 2018; Bazán-Ramírez et al., 2021; Henríquez et al., 2023). Likewise, with invariance analysis it is possible to reduce student bias when evaluating teaching among the different *educational programs* in which they are enrolled.

2. Method

2.1. Participants

We analyzed the responses of a focused sample of 1849 students out of a total of 4,226 enrolled in the FCAyS of the UABC who participated in the internal strategy of teaching performance evaluation 2022–1 (conducted in the first semester of the year). For the selection of the sample of participants, the FCAyS Teaching Evaluation Coordination randomly selects, during the second semester of each *school stage*, two groups from each semester of the eight current educational programs (Law, Psychology, Accounting, Business Administration, Educational Sciences, Communication Sciences, Computer Science, and Sociology), one from the morning shift and one from the afternoon shift. In addition, it randomly chooses four groups of the *Common core of the Areas of knowledge of Legal Sciences, Accounting and Administrative Sciences* and *Social Sciences*, two from the morning shift and two from the afternoon shift. Table 1 shows the distribution of the study sample by *Educational program, Common core* and *Area of knowledge*. Note that the number of participants by subject area shows a wide difference. In particular, between the *Area of Legal Sciences*, with 366 participating students, and the other two *Areas of knowledge*, where almost twice as many students participated in each of them. Likewise, there is a considerable difference between the sample of participating students per *School stage* [*Basic stage* (1st and 2nd semester) $N = 632$, *Disciplinary stage* (3rd, 4th, 5th and 6th semester) $N = 816$, *Terminal stage* (7th and 8th semester) $N = 392$].

2.2. Measurement

Scale for the Evaluation of Teaching Performance (EEDDocente) was designed by the coordinators of teacher evaluation of the FCAyS (Henríquez et al., 2017, 2018; Henríquez and Arámburo, 2021) with the purpose of providing at the middle of each school stage relevant information, based on the opinion of the students, on the performance of each teacher who teaches classes in the current educational programs, favoring continuous training and decision-making to improve teaching. In total, a student can answer the EEDDocente up to seven times, depending on the number of teachers who teach the different classes in the semester in progress. The EEDDocente is a typical performance test made up of 25 ordered response items ($K = 25$) with four categories: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Agree*, 4 = *Strongly agree*. During the design of the EEDDocente, a committee made up of teachers and graduates of the Social Sciences area of the University where the scale is applied was formed, who participated together with specialists in the writing of the items. With this, it was sought to ensure that the scale items were designed from a student-centered teaching approach. The items are organized into three subscales in which the dimensions underlie: (a) *Course organization*, refers to the teacher's ability to explain in a clear and organized manner the contents of the subject matter and the objectives and activities in class, as well as to use didactic strategies in an adequate manner to awaken the students' interest in the learning objectives; (b) *Quality of teaching*, refers to the teacher's ability to relate the contents of the subject matter with those of other classes,

TABLE 1 Distribution of the sample of participating students by *Educational program, Common core* and *Area of knowledge* of the FCAyS.

Area of knowledge	Educational program	Population (%)	Sample (%)
Accounting and Administrative Sciences $N = 1,404$ (33.2%)	Bachelor's Degree in Business Administration	496 (11.7%)	293 (15.8%)
	Bachelor's Degree in Accounting	390 (9.2%)	295 (16%)
	Bachelor's Degree in computer science	92 (2.2%)	57 (3.1%)
	Common trunk of the area of accounting and administrative sciences	426 (10.1%)	162 (8.8%)
Legal Sciences $N = 1,174$ (27.8%)	Bachelor's Degree in Law	1,174 (27.8%)	366 (19.8%)
Social Sciences $N = 1,648$ (39%)	Bachelor's Degree in communication Sciences	191 (4.5%)	117 (6.3%)
	Bachelor's Degree in Education Sciences	255 (6%)	111 (6%)
	Bachelor's Degree in Psychology	696 (16.5%)	287 (15.5%)
	Bachelor's Degree in Sociology	34 (0.8%)	26 (1.4%)
	Common trunk of the area of social sciences	472 (11.2%)	135 (7.3%)
FCAyS	Total students	4,226 (100%)	1849 (100%)

TABLE 2 Items of the Scale for the Evaluation of Teaching Performance (EEDDocente).

Factores	Items ID	Items of the EEDDocente The teacher...
F1. Course organization	Q4.1	explains the contents of the subject clearly and presents them with an appropriate sequence. [explica los contenidos de la materia con claridad y los presenta con una secuencia adecuada.]
	Q4.2	explain at the beginning of each class the objective and activities of the day. [explica al inicio de cada clase el objetivo y actividades del día.]
	Q4.10	has theoretical command of the subject. [tiene dominio teórico de la asignatura.]
	Q4.11	use digital tools appropriately. [utiliza las herramientas digitales de manera adecuada.]
	Q4.12	maintains timely communication with students. [mantiene comunicación oportuna con los estudiantes.]
	Q4.13	demonstrates communication skills and ease of speech. [demuestra habilidades de comunicación y facilidad de palabra.]
	Q4.14	respect the plan and the framework of the course. [respeta el plan y el encuadre del curso.]
	Q4.15	resolves student doubts appropriately. [resuelve dudas de los estudiantes de manera pertinente.]
	Q4.16	reflects commitment and enthusiasm for their work. [refleja compromiso y entusiasmo por sus labores.]
F2. Quality of Teaching	Q4.3	promotes the connection of the contents of the subject with situations, experiences or problems of daily life (for example, situations to be faced in the work context). [promueve la conexión de los contenidos de la materia con situaciones, experiencias o problemas de la vida cotidiana (por ejemplo, situaciones a enfrentar en el contexto laboral).]
	Q4.4	relates the contents of the subject with the contents of other subjects. [relaciona los contenidos de la materia con los contenidos de otras materias.]
	Q4.5	encourages student participation in class development, for example, through questions, presentations, discussion of ideas, opinions, etc. [fomenta la participación de los estudiantes en el desarrollo de la clase, por ejemplo, a través de preguntas, exposiciones, debate de ideas, opiniones, etcétera.]
	Q4.6	establishes rules and norms of socialization with students. [establece reglas y normas de socialización con los estudiantes.]
	Q4.7	fosters an atmosphere of coexistence based on trust and respect among all. [fomenta un ambiente de convivencia basado en la confianza y respeto entre todos.]
	Q4.8	make adaptations at the request or in favor of the learning needs of the group. [realiza adecuaciones a petición o a favor de las necesidades de aprendizaje del grupo.]
	Q4.9	awakens the group 's interest in the contents and purposes of the subject. [despierta el interés del grupo por los contenidos y propósitos de la asignatura.]
F3. Evaluation and Feedback of Learning	Q10.1	use evaluation strategies that I like. [utiliza estrategias de evaluación que me agradan.]
	Q10.2	use evaluation methods with which I learn better. [utiliza métodos de evaluaciones con los que aprendo de mejor forma.]
	Q10.3	promote forms of learning support parallel to partial exams: for example, advice, clarifications, doubts, post-evaluation feedback, among others. [promueve formas de apoyo al aprendizaje paralelas a los exámenes parciales: por ejemplo, asesorías, aclaraciones, dudas, retroalimentación post-evaluación, entre otras.]
	Q10.4	is interested in improving student learning, beyond the final grade obtained. [se interesa en el mejoramiento del aprendizaje de los estudiantes, más allá de la calificación final obtenida.]
	Q10.5	is concerned with establishing forms of evaluation related to real-life problems. [se preocupa por establecer formas de evaluación relacionadas con problemáticas de la vida real.]
	Q10.6	is concerned with differentiating between students who learn more and less easily, adapting their teaching strategies and forms of evaluation. [se preocupa por diferenciar entre los estudiantes que aprenden con mayor y menor facilidad, adaptando sus estrategias de enseñanza y las formas de evaluación.]
	Q10.9	shows openness for corrections and adaptations with respect to non-conforming grades or errors. [muestra apertura para correcciones y adecuaciones respecto con calificaciones inconformes o errores.]

Source: Self-made.

encourage group participation in class activities, establish norms of coexistence in the classroom and make adjustments to favor the achievement of the group's learning objectives; and (c) *Evaluation and feedback of learning,* refers to the teacher's ability to apply strategies for evaluation and feedback of learning with a formative approach, differentiate between students who learn more and less easily, adapt their teaching strategies and forms of evaluation, establish forms of evaluation related to real-life problems, and show openness to corrections and adjustments regarding non-conforming grades or errors. As a foundation, the EEDDocente is based on multidimensional conceptual models consolidated and commonly reported in the literature related to the evaluation of teaching by students (Marsh, 1984, 1993, 2007; Feldman, 1988, 1993; Centra, 1993; Braskamp and Ory, 1994; Arreola, 2007; Fink, 2008; Bazán-Ramírez et al., 2021). Table 2 shows the items that make up the three subscales of the EEDDocente.

2.3. Procedure

The protocol and procedure for applying the instrument was approved by the FCAyS-UABC Management and supervised by the FCAyS Teacher Evaluation Coordination, in accordance with current institutional research ethical standards. It should be noted that the application of the EEDDocente is part of the internal strategy of evaluation of the teaching performance of the FCAyS that is applied

at the middle of each school cycle by the Coordination of Teaching Evaluation of said faculty. In particular, the application is carried out with the support of students who provide their professional social service and who are previously trained to apply the evaluation instrument in the classroom. The students to whom the instruments are applied are previously informed of the objectives and procedures of the evaluation strategy, and of the confidentiality, safeguarding, and use of their answers in order to promote the continuous training of teachers, research, and decision-making to improve the performance of teachers who teach classes at the FCAyS. On this occasion, the EEDDocente was administered during school hours in each of the classrooms of the 80 randomly selected groups that make up the sample. On average, the explanation of the purpose of the teacher evaluation, the instructions and the application of the EEDDocente lasted 25 min. In addition, at the end of the application an effort was made for the students to answer all the items on the scale.

2.4. Data analysis

The data analysis is organized in four stages: (1) purification of the database, descriptive statistics, elimination of atypical cases; (2) verification of the preliminary assumptions of normality, reliability and linearity; (3) explained variance, measure of sample adequacy and analysis of the internal structure through the application of the Confirmatory Factor Analysis (CFA); and (4) measurement of invariance using Multi-Group Confirmatory Factor Analysis (MGCFA). Following the recommendations of Hu and Bentler (1999) and Hirschfeld and Von-Brachel (2014), statistical analyses were performed with the support of the *dplyr* (Wickham et al., 2019), *psych* (Revelle and Revelle, 2015), *lavaan* (Rosseel, 2012) and *semTools* (Jorgensen et al., 2022) from the open source software RStudio version 1.4 (R Core Team, 2022).

In the first stage, the database was cleaned, eliminating missing and atypical cases based on the Tukey Fences test. As a result of said procedure, 1,679 of the 1,849 original cases remained, of which 549 are from the *Basic stage*, 748 from the *Disciplinary stage*, and 374 from the *Terminal stage*. Consecutively, the mean, standard deviation, standard error and *item-total* correlation (*rpbis*) of each one of the items, and the *general index* and by subscales of the EEDDocente were estimated.

In the second stage, the assumption of normality was tested by applying the *Multivariate normality* test for asymmetry and kurtosis by Mardia (1970) with an acceptance criterion ≥ 0.05. The Kolmogorov–Smirnov test with Lilliefors correction was performed consecutively. With the kurtosis coefficient it is possible to identify the tendency of the participants to respond in a biased way toward one of the response categories (Vance et al., 1983), while with the symmetry coefficient the degree of concentration of responses to a central area of the distribution. In the Kolmogorov–Smirnov test, if the value of p is less than α (0.05, default value), the null hypothesis is rejected (the distribution is normal) (Dallal and Wilkinson, 1986). As a result of said procedure, items Q4.3, Q4.7, Q4.10, Q4.12, Q4.13, Q4.14, Q4.15, and Q4.16 were eliminated, which presented values well outside the boundaries of the kurtosis and skewness coefficients between −1 and +1 recommended by Hair et al. (2019). Likewise, items Q10.7 and Q10.8, which did not meet the cutoff criterion of $rpbis \geq 0.2$, were eliminated (Brown, 2012).

For its part, global and subscale reliability was verified with the estimation of the standardized *Rho Alpha* coefficient (ρ) and the McDonald *Omega* coefficient (ω) together with Cronbach's *Alpha* (α) (Cronbach, 1951, 1988; McNeish, 2018; Raykov and Marcoulides, 2019). The quality criteria for the reliability coefficients were $\rho \geq 0.70$, $\omega \geq 0.80$, and $\alpha \geq 0.70$ (Cronbach, 1951, 1988; Katz, 2006; Zhang and Yuan, 2016; Nájera-Catalán, 2019). Once the preliminary analysis and the quality criteria were taken into account, we proceeded to analyze the model of three factors [*Course Organization* (F1), *Quality of Teaching* (F2), and *Evaluation and Feedback of Learning* (F3)] that underlie in the internal structure of the EEDDocente through the CFA application. For this, the Weighted Least Squares (WLS) and Robust Weighted Least Squares (WLSMV) estimation methods were applied (Jöreskog and Sörbom, 1979; Brown, 2015; Kline, 2015; Gazeloglu and Greenacre, 2020). On the other hand, in the evaluation of the adjustment indexes, the recommendations of Hu and Bentler (1999) and Hair et al. (2019). In particular, the adjustment indices and criteria were the Comparative Fit Index (CFI) ≥ 0.95, the Tucker-Lewis Index (TLI) ≥ 0.95, the Normalized Mean Square Residual (SRMR) ≤ 0.08 and the Mean Square Error of Approximation (RMSEA) ≤ 0.06 (Browne and Cudeck, 1993; Schreiber et al., 2006). For the subsequent analysis, only items with factor loadings ≥ 0.43 were considered.

Finally, an MGCFA was carried out to measure the invariance by *School stage* and *Educational program* based on the adjusted model of three factors of the EEDDocente. To verify the assumption of invariance depending on the *School stage*, three groups were considered: (a) students of the *Basic stage* (1st and 2nd semester), (b) students of the *Disciplinary stage* (3rd, 4th, 5th and 6th semester), and students of the *Terminal stage* (7th and 8th semester). On the other hand, to verify the assumption of invariance based on the *Knowledge Area*, three groups were considered: (a) students enrolled in the *Accounting and Administrative Sciences Educational* programs, (b) students enrolled in the *Legal Sciences Educational* programs, and (c) students enrolled in the *Educational* programs of *Social Sciences*. The *Configural*, *Weak*, *Strong* and *Strict* invariance models were contrasted (Dimitrov, 2010; Milfont and Fischer, 2010). For this, the recommendations of Byrne et al. (1989) and Vandenberg and Lance (2000) focus the analysis on the increasingly restrictive comparison of the model parameters. To consider the factorial invariance between models adequate, it was established as a criterion that the Chi-square difference ($\Delta\chi^2$) was not significant ($p > 0.05$). However, since the $\Delta\chi^2$ is affected by the sample size, the recommendations of Vandenberg and Lance (2000), Cheung and Rensvold (2002) and Dimitrov (2010) were followed, establishing RMSEA parameters close to the cutoff criterion of 0.08, a difference in RMSEA parameters between models less than 0.015 (ΔRMSEA ≤ 0.015), and a difference in CFI and TLI parameters between models less than 0.010 (ΔCFI and ΔTLI < 0.010) (Chen, 2007; Dimitrov, 2010; Putnick and Bornstein, 2016).

3. Results

3.1. Descriptive results and preliminary analyses

The average of the general index of the EEDDocente was 86.61, with a standard deviation of 11.05. Likewise, the average score of

TABLE 3 Descriptive statistics ($n = 1,679$, $K = 15$).

Subscale	Item (k)	M	SD	Skewness	Kurtosis	*rpbis*
1. Course organization	Q4.1	3.56	0.69	−1.34	0.64	0.69
	Q4.2	3.31	0.87	−0.98	−0.08	0.63
	Q4.9	3.44	0.78	−1.23	0.62	0.75
	Q4.11	3.62	0.69	−1.80	2.57	0.46
2. Quality of Teaching	Q4.4	3.07	0.95	−0.58	−0.83	0.53
	Q4.5	3.57	0.72	−1.61	1.79	0.55
	Q4.6	3.40	0.84	−1.23	0.55	0.58
	Q4.8	3.46	0.80	−1.34	0.90	0.69
3. Evaluation and Feedback of Learning	Q10.1	3.43	0.68	−1.09	1.04	0.75
	Q10.2	3.29	0.74	−0.83	0.27	0.76
	Q10.3	3.38	0.73	−1.03	0.67	0.75
	Q10.4	3.49	0.68	−1.18	0.88	0.73
	Q10.5	3.49	0.66	−1.09	0.77	0.69
	Q10.6	3.26	0.82	−0.89	0.09	0.70
	Q10.9	3.39	0.73	−1.07	0.80	0.60
Average		3.41	0.76	−1.15	0.8	0.66
General index of the EEDDocente		86.61	11.05	−1.10	1.02	

TABLE 4 Overall and subscale internal consistency values of EEDDocente.

Subscale	α	ρ	ω
1	0.77	0.77	0.78
2	0.75	0.75	0.77
3	0.90	0.90	0.90
Overall	0.92	0.92	0.93

the scale items was 3.41 (4 = *Strongly agree*), with item Q4.4 being the one with the lowest average score (3.07) and Q4.11 the one with the highest average score. (3.62). For its part, the average *item-total* correlation of the scale was 0.64, meeting the cut-off criterion ($rpbis \geq 0.2$). Likewise, the items presented, on average, moderate correlations among themselves (0.42) with correlation coefficients that oscillated between 0.26 and 0.74. Table 3 shows the descriptive results of the items and the general index of the EEDDocente.

Regarding the assumption of normality, significant results ($p < 0.001$) were obtained with the multivariate normality test of asymmetry and kurtosis by Mardia (1980), rejecting the assumption of multivariate normality in the study sample. Likewise, the results of the Kolmogorov–Smirnov test with Lilliefors correction yielded values that reject the normal distribution of the *general index* ($D = 0.12$, $p < 0.001$). The global reliability coefficients of the EEDocente ($\alpha = 0.92$, $\rho = 0.92$ and $\omega = 0.93$) meet the quality criteria established *a priori*. Likewise, the three subscales meet the quality criteria $\alpha \geq 0.70$, $\rho \geq 0.70$. However, regarding the McDonald *Omega* coefficient (ω), subscales 1 and 2 [*Course Organization* (F1) and *Quality of Teaching* (F2)] present values below the $\omega \geq 0.80$ criterion. Table 4 shows the values obtained from the general reliability coefficient and by subscale of the EEDocente.

3.2. Confirmatory factor analysis

The fit indices estimated using the WLS ($\chi^2 = 251.21$; $df = 87$, $p = 0.000$; CFI = 0.868; TLI = 0.841; GFI = 0.936; NNFI = 0.814; RMSEA = 0.034; SRMR = 0.057) and WLSMV ($\chi^2 = 52.80$, $df = 87$, $p = 0.999$, CFI = 1.0, TLI = 1.0, GFI = 0.999, RMSEA = 0.000, SRMR = 0.21) were adequate for the three-factor model of the EEDDocente. In turn, the factors presented on average moderate correlations among themselves (0.64) with correlation coefficients ranging between 0.58 and 0.76. In addition, the standardized factor loadings of the three-factor model showed significant and adequate values (see Figure 1).

3.3. Factorial invariance

Table 5 shows the results of the adjustment of the factorial invariance parameters of the three-factor model of the EEDDocente based on the *School stage* and by *Knowledge Area*. It is shown that the three-factor model of teacher performance from the perception of the students was adequate for the groups according to the *School stage* (*Basic Stage, Disciplinary, Stage and Terminal Stage*) and by *Knowledge Area* (*Accounting and Administrative Sciences, Legal Sciences*, and

10.3389/feduc.2023.1229129

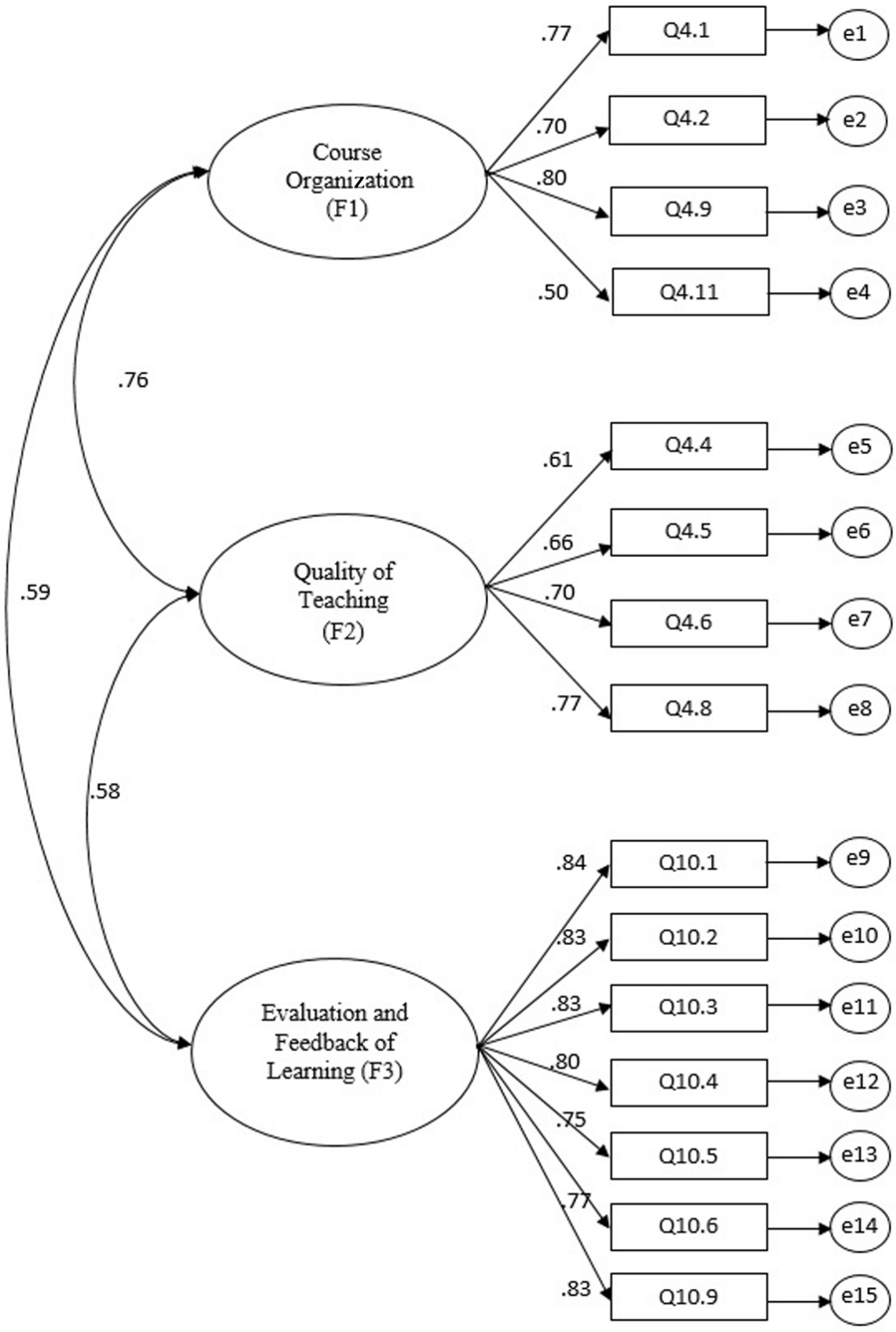


FIGURE 1
Three-factor first-order CFA model of EEDDocente.

Social Sciences). The *Configural* invariance model presented a good fit for all study groups. In particular, the differences between the *Weak*, *Strong* and *Strict* models, both for the groups based on *School stage* and *Knowledge Area*, meet the cut-off criteria (ΔCFI <0.010, ΔRMSEA ≤0.015) (Chen, 2007; Dimitrov, 2010; Putnick and Bornstein, 2016). With the differences obtained between the *Weak* (ΔRMSEA = −0.002 and ΔCFI = −0.001), *Strong* (ΔRMSEA = −0.001 and ΔCFI = −0.002) and Strict (ΔRMSEA = 0.000 and ΔCFI = −0.005) models for the groups depending on the *School stage* and the *Weak* (ΔRMSEA = −0.002) models and ΔCFI = −0.001, *Strong* (ΔRMSEA = −0.002 and ΔCFI = −0.002) and *Stric* (ΔRMSEA = 0.002 and ΔCFI = −0.008) for the groups depending on the educational programs by *Knowledge Area*, factorial invariance is verified.

4. Discussion

The development and validation of the EEDocente represents an important contribution to the study and measurement of teacher performance from the perspective of students (Shevlin et al., 2000;

TABLE 5 Fit indices to evaluate the factorial invariance by *school stage* and *area of knowledge* of the three-factor model of the EEDDocente.

Model	χ^2	*df*	CFI	ΔCFI	TLI	ΔTLI	RMSEA	ΔRMSEA	SRMR	BIC	ΔBIC
School stage											
Configural	812.09	261	0.954	---	0.944	---	0.062	---	0.033	49935.1	---
Week	852.58	285	0.953	−0.001	0.948	0.003	0.060	−0.002	0.043	49806.5	−137.15
Strong	902.89	309	0.950	−0.002	0.949	0.002	0.059	−0.001	0.044	49677.9	−127.34
Strict	989.25	339	0.946	−0.005	0.949	0.000	0.059	0.000	0.047	49543.7	−135.68
Area of knowledge											
Configural	816.14	261	0.950	---	0.940	---	0.065	---	0.034	50125.3	---
Week	856.90	285	0.949	−0.002	0.943	0.003	0.063	−0.002	0.043	49981.6	−135.09
Strong	907.61	309	0.946	−0.002	0.945	0.002	0.062	−0.001	0.045	49858.7	−125.13
Strict	1029.83	333	0.938	−0.008	0.942	−0.003	0.063	0.002	0.052	49779.2	−97.59

Whittington, 2001; Campbell et al., 2005; Richardson, 2005). In particular, this study provides evidence of reliability, internal structure, and factorial invariance that allow for further comparative studies and thus evidence-based decision-making. Contrary to high-stakes assessments, the use of this type of assessment for the purpose of performance improvement and continuous teacher training is a rare practice, but vital for the improvement of classroom education in all education systems around the world. By way of discussion, the most relevant findings of the study are presented and contrasted with the results of other similar studies.

In particular, the reduced version ($K = 15$) of the EEDDocente complies with the psychometric quality criteria of reliability and internal structure. The global reliability coefficients of the EEDDocente meet the cut-off criteria ($\alpha = 0.92$, $\rho = 0.92$ and $\omega = 0.93$), and the reliability coefficients per subscale are very close to what was expected. Likewise, with the results of the CFA obtained, the three-factor structure proposed by the Coordination of teacher evaluation of the FCAyS is corroborated (Henríquez et al., 2017, 2018; Henríquez and Arámburo, 2021). The multidimensional model of three factors with 15 items presents adequate factor loadings (between 0.50 and 0.84) and an acceptable. With this, the structure of the EEDDocente, which addresses some of the most relevant teaching competencies throughout the educational levels, endorses and consolidates its underlying theoretical model. This is consistent with other studies of similar instruments that present similar theoretical dimensions in their structure (Marsh, 1984, 1993, 2007; Feldman, 1988; Centra, 1993; Feldman, 1993; Braskamp and Ory, 1994; Fink, 2008; Silva et al., 2014; Irigoyen et al., 2016; Bazán-Ramírez and Velarde-Corrales, 2021; Bazán-Ramírez et al., 2021). It is important to mention that the items eliminated do not affect the interpretation of the construct, maintaining the three basic dimensions of the EEDDocente defined at the beginning by the design committee. Likewise, with a smaller scale, the time and possible problems related to the average number of times a regular student of the FCAyS must answer the EEDDocente per semester are reduced.

Added to this, the study provides new findings on factorial invariance depending on the School stage and *Educational programs* in the *Knowledge Areas* of *Accounting and Administrative Sciences*, *Legal Sciences*, and *Social Sciences* in samples of university students. The *Configural* invariance model presented a good fit for all study groups, and the differences in the parameters of the *Weak* and *Strong* models are adequate. This guarantees that the EEDDocente can be considered on the same scale for the different groups under study and confirms that the three-factor model measures in the same way in all of them (Vandenberg and Lance, 2000; Wang and Wang, 2012). In addition, the Bayesian Information Criterion (BIC) presented a sequential reduction, which can be interpreted as a sign of equivalence between the samples (Cheung and Rensvold, 2002). In this regard, Chen (2007) mentions that the RMSEA and SRMR tend to reject invariant models when the sample size is not equal between the groups, so it is advisable to use the CFI as the main criterion to establish invariance.

It must be recognized that the main limitation of the study has to do with the fact that, although the student samples are large, they are not equitable between the study groups. In particular, it is important to remember that there is a difference greater than 100 individuals between the groups of the *School stage* of the *Basic stage* ($N = 632$) and the *Disciplinary stage* ($N = 816$), and that this difference increases when compared with the number of students participating in the *Terminal stage* ($N = 392$). The same happens with the number of participants in the educational programs by *Area of Knowledge*, where 366 students from the *Knowledge Area* of *Legal Sciences* participated, and in *Accounting and Administrative Sciences* and *Social Sciences* almost twice as many participated ($N = 807$ and $N = 676$ respectively).

5. Conclusion

By way of conclusion, it can be said that the findings derived from the reliability analysis and the CFA provide evidence that supports an adequate adjustment of the three-factor structural model [*Class Organization* (F1), *Teaching quality* (F2), and *Assessment and Feedback on learning* (F3)] of the reduced version ($K = 15$) of the EEDDocente to evaluate teaching performance throughout the *School stage* (*Basic stage*, *Disciplinary stage*, and *Terminal stage*) and the *Areas of knowledge* (*Accounting and Administrative Sciences, Legal Sciences*, and *Social Sciences*). In addition, factorial invariance analysis based on the *School stage* and the *Educational programs* by *Areas of Knowledge* in samples of university students show an adequate adjustment of the *Configural* model, and the differences in the parameters of the *Weak*, *Strong*, and *Strict* models. These results indicate that none of the study groups presents a systematic tendency to answer the items higher or lower than the rest of the groups (Vandenberg and Lance, 2000; Meredith and Teresi, 2006; Wang and Wang, 2012), providing

evidence to carry out comparative studies. With all this, it is guaranteed that the EEDDocente complies with the standards of reliability, internal structure validity and invariance, and its use as a brief and easy-to-administer instrument is supported, presenting an important contribution to the study and measurement of teaching from the students' perspective. It is recommended for future research ensure the equivalence of the samples of the study groups to favor the analysis of the metric invariance and factorial invariance of the EEDDocente and to carry out comparative and predictive studies. It is also important to consider the application of the EEDDocente in other schools and universities in order to have a tool for brief application with the purpose of providing relevant information at the end of each school stage, based on the opinion of the students, on teaching performance, favoring continuous training and decision making to improve the effectiveness and quality of teaching.

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 protocol and procedure of the evaluation strategy on which the present study was based were approved by the FCAyS-UABC administration and supervised by the FCAyS-UABC Teaching Evaluation Coordination, in accordance with the institutional norms in force regarding research ethics. For the application of the EEDDocente, a group of students was trained as applicators of the instrument and the voluntary participation of the students of each educational program of the FCAyS-UABC was requested, who were previously informed about the objectives and procedures of the study and about the confidentiality, safeguarding and use of the information obtained based on their answers.

Author contributions

PH and JP-M contributed to the idea of research, its conceptualization, implementation, and methodology, were in charge of writing the manuscript, and also contributed to the analysis and interpretation of data, revising the English version, and writing in Frontiers format. CC provided support in the search for additional bibliographic information, was responsible for drafting the introduction to the manuscript, and reviewed the writing style of the article. JZ contributed to the design and implementation of the EEDDocente, and also provided elements for the delimitation of the participants. All authors contributed to the article and approved the submitted version.

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.

References

Alcaraz-Salarirche, N. (2015). *Aproximación Histórica a la Evaluación Educativa: De la Generación de la Medición a la Generación Ecléctica [Historical Approach to Educational Evaluation: From the Measurement Generation to the Eclectic Generation]. Revista Iberoamericana de Evaluación Educativa. 8:1.* Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=5134142 (Accessed January 21, 2023).

Aleamoni, L. (1999). Student rating myths versus research facts from 1924 to 1998. *J. Pers. Eval. Educ.* 13, 153–166. doi: 10.1023/a:1008168421283

American Educational Research Association (2018). *American Psychological Association y National Council on Measurement Education [AERA, APA, NCME]*. Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Aravena-Gaete, M. E., and Gairín, J. (2021). Evaluación del desempeño docente: Una mirada desde las agencias certificadoras. [Evaluation of teacher performance: a look from the certification agencies]. *Prof. Rev. Currículum Form. Prof.* 25, 297–317. doi: 10.30827/profesorado.v25i1.8302

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems (3rd)*. Bolton, MA: Anker.

Bazán-Ramírez, A., and Velarde-Corrales, N. (2021). Auto-reporte del estudiantado en criterios de desempeño didáctico en clases de Psicología [students self-report within didactic performances criteria in psychology classes]. *J. Behav. Health Soc.* 13, 22–35. doi: 10.22201/fesi.20070780e.2021.13.1.78071

Bazán-Ramírez, A., Pérez-Morán, J. C., and Bernal-Baldenebro, B. (2021). Criteria for teaching performance in psychology: invariance according to age, sex, and academic Stage of Peruvian students. *Front. Psychol.* 12:764081. doi: 10.3389/fpsyg.2021.764081

Benton, S. L. (2018). Best practices in the evaluation of teaching. *Best Pract. Eval. Teach.* 69, 1–18.

Bleiberg, J. (2023). "Revisiting teacher evaluation a decade after reforms" in *International encyclopedia of education*. eds. R. Tierney, F. Rizvi and K. Ercikan. *4th* ed (Amsterdam, Netherlands: Elsevier)

Boring, A., Ottoboni, K., and Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Sci. Open Res.* 1, 1–11. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Boštjančič, E., Komidar, L., and Johnson, R. B. (2018). Factorial validity and measurement invariance of the slovene version of the cultural intelligence scale. *Front. Psychol.* 9:1499. doi: 10.3389/fpsyg.2018.01499/full

Braskamp, L. A., and Ory, J. C. (1994). *Assessing faculty work: Enhancing individual and institutional performance*. San Francisco, CA: Jossey-Bass.

Brown, J. D. (2012). "Classical test theory" in *The Routledge handbook of language testing measurement*. eds. G. Fulcher and F. Davidson (Abingdon: Routledge)

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford Publications.

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit, testing structural equation models" in *Testing estructural models*. eds. K. A. Bollen and J. S. Long (Thousand Oaks, California: Sage)

Byrne, B. M., Shavelson, R. J., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Calatayud, M. (2021). Evaluación docente y mejora profesional. Descubrir el Encanto de su complicidad. [teacher evaluation and professional improvement. Discover the charm of their complicity]. *Rev. Iberoam. Eval. Educ.* 14, 87–100. doi: 10.15366/riee2021.14.1.005

Camacho, I. (2022). El desempeño docente y su implicación en la enseñanza. Teacher performance and their involvement in teaching. *Form. Estrat.* 6, 105–120.

Campbell, H. E., Steiner, S., and Gerdes, K. (2005). Student evaluations of teaching: how you teach and who you are. *J. Public Aff. Educ.* 11, 211–231. doi: 10.1080/15236803.2005.12001395

Carpio, C., Pacheco, V., Canales, C., and Flores, C. (1998). Comportamiento inteligente y juegos de lenguaje en la enseñanza de la psicología [Intelligent behavior and language games in the teaching of psychology]. *Acta Comport* 6, 47–60.

Cashin, W. E., and Downey, R. G. (1992). Using global student rating items for summative evaluation. *J. Educ. Psychol.* 84, 563–572. doi: 10.1037/0022-0663.84.4.563

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.

Chan, W. M. (2018). Teaching in HIGHER education: students' perceptions of effective teaching and good teachers. *Soc. Sci. Educ. Res. Rev* 5, 40–58.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Chen, Y., and Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: an assessment of student perception and motivation. *Assess. Eval. High. Educ.* 28, 71–88. doi: 10.1080/02602930301683

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Model.* 9, 233–255. doi: 10.1097/NNR.0b013e3182544750

Ching, G. (2018). A literature review on the student evaluation of teaching: an examination of the search, experience, and credence qualities of SET. *High. Educ. Eval. Dev.* 12, 63–84. doi: 10.1108/HEED-04-2018-0009

Cisneros-Cohernour, E., and Stake, R. (2010). La evaluación de la docencia en educación superior: de evaluaciones basadas en opiniones de estudiantes a modelos por competencias [the evaluation of teaching superior education: from evaluations based on opinions of students in models of competences]. *Rev. Iberoam. Eval. Educ.* 3, 218–231.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. *Rev. Educ. Res.* 51, 281–309. doi: 10.2307/1170209

Cohen, P. A. (1983). Comment on a selective review of the validity of student ratings of teaching. *J. High. Educ.* 54, 448–458. doi: 10.2307/1981907

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Cronbach, L. J. (1988). Internal consistency of tests: analyses old and new. *Psychometrika* 53, 63–70. doi: 10.1007/BF02294194

Cruz Ávila, M. (2007). *Una propuesta para la evaluación del profesorado universitario*. A proposal for the evaluation of university teaching staff.Trabajo de investigación para la obtención del grado de doctor. Universitat Autonoma de Barcelona. Available at: https://www.tdx.cat/bitstream/handle/10803/5285/mca1de1.pdf (Accessed January 16, 2023).

Dallal, G., and Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefors's test statistic for normality. *Am. Stat.* 40, 294–296. doi: 10.1080/00031305.1986.10475419

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Meas. Evaluat. Counsel. Dev.* 43, 121–149. doi: 10.1177/0748175610373459

Estrada, L. A., Yglesias, L. A., Miranda, A. E., Díaz, J. K., and Díaz, S. M. (2019). Propiedades Psicométricas de un Cuestionario sobre Evaluación del Desempeño Docente Universitario desde la Percepción del Estudiante [psychometric properties of a questionnaire on evaluation of university teaching performance from the perception of the student]. *Rev. Investig. Estadíst.* 2, 92–102.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Res. High. Educ.* 5, 243–288. doi: 10.1007/BF00991967

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: matched or mismatched priorities? *Res. High. Educ.* 28, 291–329. doi: 10.1007/BF01006402

Feldman, K. A. (1989a). Instruccional effectiveness of college teachers as judged by teacher themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Res. High. Educ.* 30, 137–194.

Feldman, K. A. (1989b). The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisectional validity studies. *Res. High. Educ.* 30, 583–645.

Feldman, K. A. (1990). An afterword for "the association between student ratings of specific instructional dimensions and student achievement": refining and extending the synthesis of data from multisection validity students. *Res. High. Educ.* 31, 315–317.

Feldman, K. A. (1992). College students' views of male and female college teachers: part I- evidence from the social laboratory experiments. *Res. High. Educ.* 33, 317–375.

Feldman, K. A. (1993). College students' views of male and female teachers: part II-evidence from students' evaluations of their classroom teachers. *Res. High. Educ.* 34, 151–211.

Fink, D. (2008). "Evaluating teaching: a new approach to an old problem" in *To improve the academy: Resources for faculty, instructional, and organizational development*. eds. S. Chadwick-Blossey and D. R. Robertson (Hoboken, NJ: Jossey-Bass)

García, J. M. (2000). "Las dimensiones de la efectividad docente, validez y confiabilidad de los cuestionarios de evaluación de la docencia: síntesis de investigación internacional [the dimensions of teaching effectiveness, validity and reliability of teacher evaluation questionnaires: international research synthesis]" in *Evaluación de la docencia*. eds. M. Rueda and Y. F. Díaz-Barriga (Mexico: Paidós), 41–62.

García, J. M. (2014). *¿Para qué sirve la evaluación de la docencia? Un estudio exploratorio de las creencias de los estudiantes [how useful is the evaluation of teaching? An exploratory study on students' beliefs]*. Education policy analysis archives, 22. pp. 1–20. Available at: https://www.redalyc.org/pdf/2750/275031898010.pdf (Accessed January 14, 2023).

García-Gómez-Heras, S., Gil-Madrona, P., Hernandez-Barrera, V., Lopez-de-Andres, A., and Carrasco-Garrido, P. (2017). Importance of university teacher behaviour in the faculty of health science. *Aust. Med. J.* 10, 800–810. doi: 10.21767/AMJ.2017.3128

García-Olalla, A., Villa-Sánchez, A., Aláez, M., and Romero-Yesa, S. (2022). Aplicación y resultados de un sistema para evaluar la calidad de la docencia universitaria en una década de experimentación [Implementation and results of a system to evaluate the quality of University teaching in a decade of experimentation]. *Rev. Investig. Educ.* 40, 51–68. doi: 10.6018/rie.401221

Gazeloglu, C., and Greenacre, Z. A. (2020). Comparison of weighted least squares and robust estimation in structural equation modeling of ordinal categorical data with larger sample sizes. *Cumhuriyet Sci. J.* 41, 193–211. doi: 10.17776/csj.648054

Gómez, L. F., and Valdés, M. G. (2019). The evaluation of teacher performance in higher education. *Propósitos Represent.* 7, 479–515. http://dx.doi.org/10.20511/pyr2019.v7n2.255.

Gu, R., Wang, H. N., and Lou, L. S. (2021). Optimization and application of data analysis strategy for college students' evaluation of teaching. *J. Zhejiang Univ. Tech.* 20, 201–207.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2019). *Multivariate Data Analysis*. Hampshire: Cengage Learning EMEA.

Henríquez, P., and Arámburo, V. (2021). Evaluación del desempeño docente por áreas de conocimiento: El Caso de la Facultad de Ciencias Administrativas y Sociales de la Universidad Autónoma de Baja California, México. [evaluation of teaching performance by areas of knowledge: the case of the Faculty of Administrative and Social Sciences of the Autonomous University of Baja California, Mexico]. *Act. Investig. Educ.* 21, 1–20. doi: 10.15517/aie.v21i3.46294

Henríquez, P., Arámburo, V., and Boroel, B. (2018). *Análisis de las estrategias de enseñanza según áreas de conocimiento en el nivel educativo superior: Percepciones de estudiantes de la Facultad de Ciencias Administrativas y Sociales (FCAYS), Universidad Autónoma de Baja California (UABC), México. [Analysis of teaching strategies according to areas of knowledge at the higher education level: Perceptions of students from the School of Administrative and Social Sciences (FCAYS), Autonomous University of Baja California (UABC), Mexico]*. Compendio Investigativo de Academic Journals Celaya 2018, 14. pp. 2320–2325.

Henríquez, P., Arámburo, V., and Dávila, E. (2017). *Percepción de los estudiantes universitarios acerca de las estrategias pedagógicas y de evaluación del aprendizaje utilizadas por sus profesores: el Caso de la FCAYS de la UABC. [perception of university students about the pedagogical and learning assessment strategies used by their professors: the case of the FCAYS at UABC]*. In Memorias electrónicas del XIV Congreso Nacional de Investigación Educativa, COMIE. San Luis Potosí: México. 1–14. Available at: https://www.comie.org.mx/congreso/memoriaelectronica/v14/doc/1956.pdf (Accessed January 24, 2023).

Henríquez, P., Pérez-Morán, J. C., Cid, C. D., and Zamora, J. E. (2023). *Reporte técnico de las propiedades psicométricas de estructura factorial e invarianza de la Escala de Evaluación del Desempeño Docente (EEDDocente) de la FCAyS*. Technical report on the psychometric properties of factorial structure and invariance of the Teaching Performance Evaluation Scale (EEDDocente) of the FCAyS. Documento interno [internal document].

Hildebrand, M. (1973). The character and skills of the effective professor. *The Journal of Higher Education*. 44, 41–50. doi: 10.2307/1980624

Hildebrand, M., Wilson, R.C., and Dienst, E.R. (1971). *Evaluating university teaching*. Center for Research and Development in Higher Education. Berkeley, CA.

Hirschfeld, G., and Von-Brachel, R. (2014). Multiple-group confirmatory factor analysis in R–a tutorial in measurement invariance with continuous and ordinal indicators. *Pract. Assess. Res. Eval.* 19, 1–12. doi: 10.7275/qazy-2946

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Educ.* 4, 1–8. doi: 10.1080/2331186x.2017.1304016

House, E. R. (1998). Acuerdos institucionales para la evaluación. [Institutional arrangements for evaluation]. *Perspectivas* 28, 123–131.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model A Multidiscip. J.* 6, 1–55. doi: 10.1080/10705519909540118

International Test Commission. (2013). *ITC Guidelines on Test Use. ITC-G-TU-20131008*. Available at: https://www.intestcom.org/files/guideline_test_use.pdf. (Accessed January 15, 2023).

Irigoyen, J. J., Jiménez, M., and Acuña, K. (2016). Discurso didáctico e interacciones sustitutivas en la enseñanza de las ciencias [Didactic discourse and substitute interactions in teaching Sciences]. *Enseñ. Investig. Psicol* 21, 68–77.

Jöreskog, K. G., and Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: ABT Books.

Jorgensen, T. D., Mansolf, M., and Enders, C. K. (2022). *Pooled Score Tests for Multiply Imputed Data Using semTools*. Available at: https://rdrr.io/cran/semTools/man/lavTestScore.mi.html (Accessed January 23, 2023).

Kane, M. (2006). "Validation" in *Educational measurement*. ed. R. L. Brennan (New Jersey: National Council on Measurement in Education and American Council on Education)

Katz, M. H. (2006). *Multivariable analysis* (*2nd Edn.*). Cambridge: Cambridge University Press.

Kline, R. B. (2015). *Principles and practice of structural equation modeling. 4th Edn.* New York: Guilford publications.

König, J., Ligtvoet, R., Klemenz, S., and Rothland, M. (2017). Effects of opportunities to learn in teacher preparation on future teachers' general pedagogical knowledge: analyzing program characteristics and outcomes. *Stud. Educ. Eval.* 53, 122–133. doi: 10.1016/j.stueduc.2017.03.001

Lera, M., León, J., and Ruiz, P. (2021). Adaptation of the teacher efficacy scale to measure effective teachers' educational practices through students' ratings: a multilevel approach. *Psicothema* 33, 509–517. doi: 10.7334/psicothema2020.262

Liebowitz, D. (2021). Teacher evaluation for accountability and growth: should policy treat them as complements or substitutes? *Labour Econ.* 71:102024. doi: 10.1016/j.labeco.2021.102024

Liebowitz, D. (2022). Teacher evaluation for growth and accountability: under what conditions does it improve student outcomes? *Harv. Educ. Rev.* 92, 533–565. doi: 10.17763/1943-5045-92.4.533

Luna, E., and Torquemada, A. (2008). Los cuestionarios de evaluación de la docencia por los alumnos: balance y perspectivas de su agenda. [questionnaires for the evaluation of teaching by students: balance and perspectives of their agenda]. *Rev. Electrón. Investig. Educ. Espec.* 10, 1–11.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519–530. doi: 10.1093/biomet/57.3.519

Mardia, K. V. (1980). *"Tests for univariate and multivariate normality" in handbook of statistics*. Amsterdam, Netherlands: Elsevier.

Márquez, L., and Madueño, M. L. (2016). Propiedades psicométricas de un instrumento Para apoyar el proceso de evaluación del docente universitario. [psychometric properties of an instrument to support the evaluation process of the university professor]. *Rev. Electrón. Investig. Educ.* 18, 53–61.

Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *J. Educ. Psychol.* 76, 707–754. doi: 10.1037/0022-0663.76.5.707

Marsh, H. W. (1986). Applicability paradigm: students' evaluations of teaching effectiveness in different countries. *J. Educ. Psychol.* 78, 465–473. doi: 10.1037/0022-0663.78.6.465

Marsh, H. W. (1987). Students' evaluations of university teaching: research findings, methodological issues, and directions for future research. *Int. J. Educ. Res.* 11, 253–388. doi: 10.1016/0883-0355(87)90001-2

Marsh, H. W. (1993). Multidimensional students' evaluations of teaching effectiveness. *J. High. Educ.* 64, 1–18. doi: 10.1080/00221546.1993.11778406

Marsh, H. W. (2007). "Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness" in *The scholarship of teaching and learning in higher education: An evidence-based perspective*. eds. R. P. Perry and J. C. Smart (Berlin, Germany: Springer Science and Business Media)

Marsh, H. W., and Dunkin, M. J. (1997). "Students' evaluations of university teaching: a multidimensional perspective" in *Effective teaching in higher education: Research and practice*. eds. R. P. Perry and J. C. Smart (New York: Agathon), 241–320.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychol. Methods* 23, 412–433. doi: 10.1037/met0000144

Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Med. Care* 44, S69–S77. doi: 10.1097/01.mlr.0000245438.73837.89

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Milfont, T. L., and Fischer, R. (2010). Testing measurement invariance across groups: applications in cross-cultural research. *Int. J. Psychol. Res.* 3, 111–130. doi: 10.21500/20112084.857

Mitchell, K., and Martin, J. (2018). Gender Bias in Student Evaluations. PS. *Polit. Sci. Polit.* 51, 648–652. doi: 10.1017/S104909651800001X

Mohammadi, M. (2021). Dimensions of teacher performance evaluation by students in higher education. *Shanlax Int. J. Educ.* 9, 18–25. doi: 10.34293/education.v9i2.3673

Nájera-Catalán, H. E. (2019). Reliability, population classification and weighting in multidimensional poverty measurement: a Monte Carlo study. *Soc. Indic. Res.* 142, 887–910. doi: 10.1007/s11205-018-1950-z

Nasser-Abu, F. (2017). Teaching in higher education: good teaching through students' lens. *Stud. Educ. Eval.* 54, 4–12. doi: 10.1016/j.stueduc.2016.10.006

Navarro, C., and Ramírez, M. (2018). Mapeo sistemático de la literatura sobre evaluación docente (2013-2017). [systematic mapping of the literature on teacher evaluation (2013-2017)]. *Educ. Pesqui.* 44, 1–23. doi: 10.1590/S1678-4634201844185677

Putnick, D. L., and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev. Rev.* 41, 71–90. doi: 10.1016/j.dr.2016.06.004

Raykov, T., and Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educ. Psychol. Meas.* 79, 200–210. doi: 10.1177/0013164417725127

Revelle, W., and Revelle, M. W. (2015). Package 'psych'. *Comprehens. R Arch. Netw.* 337:338.

Reyes, E., Luna, E., and Caso, J. (2020). Evidencias de validez del Cuestionario de Evaluación de la Competencia Docente Universitaria. [evidence of validity of the university teaching competence assessment questionnaire]. *Perfiles Educ.* 42, 106–122. doi: 10.22201/iisue.24486167e.2020.169.58931

Richardson, J. T. (2005). Instruments for obtaining student feedback: a review of the literature. *Assess. Eval. High. Educ.* 30, 387–415. doi: 10.1080/02602930500099193

Romero, T., and Martínez, A. (2017). Construcción de instrumentos de evaluación del desempeño docente universitario desde una perspectiva cualitativa. [Construction of evaluation instruments of university teaching performance from a qualitative perspective]. *Rev. Univ. Caribe* 18, 34–43. doi: 10.5377/ruc.v18i1.4800

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36.

Salazar, J. (2008). Diagnóstico Preliminar sobre Evaluación de la Docencia Universitaria. Una Aproximación a la Realidad en las Universidades Públicas y/o Estatales de Chile. [Preliminary Diagnosis on Evaluation of University Teaching. An Approach to Reality in Public and/or State Universities in Chile]. *Rev. Iberoam. Eval. Educ.* 1:e3.

Schellhase, K. C. (2010). The relationship between student evaluation of instruction scores and faculty formal educational coursework. *Athl. Train. Educ. J.* 5, 156–164. doi: 10.4085/1947-380X-5.4.156

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338. doi: 10.3200/JOER.99.6.323-338

Seivane, M. S., and Brenlla, M. E. (2021). Evaluación de la calidad docente universitaria desde la perspectiva de los estudiantes. [evaluation of university teaching quality from the students' perspective]. *Rev. Iberoam. Evalu. Educ.* 14, 35–46. doi: 10.15366/riee2021.14.1.002

Shevlin, M., Banyard, P., Davies, M., and Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assess. Eval. High. Educ.* 25, 397–405. doi: 10.1080/713611436

Silva, L. A. (2023). Congruencia entre la práctica docente y la evaluación de la docencia por parte del estudiantado en educación superior: estudio de casos en la Universidad Veracruzana. [congruence between teaching practice and the evaluation of teaching by students in higher education: case study at the Universidad Veracruzana]. *Rev. Educ.* 47, 114–128. doi: 10.15517/revedu.v47i1.51978

Silva, H., Morales, G., Pacheco, V., Camacho, A., Garduño, H., and Carpio, C. (2014). Didáctica como conducta: una propuesta para la descripción de las habilidades de enseñanza [Didactic as behavior: a proposal for the description of teaching skills]. *Rev. Mexic. Anál. Conduc.* 40, 32–46. doi: 10.5514/rmac.v40.i3.63679

Silva, J., Solís, P., Huaman, J., and Quispe, F. (2022). La evaluación formativa en el desempeño docente universitario: Revisión sistemática de literatura. [Formative evaluation in university teaching performance: Systematic literature review]. *Tecnohum. Rev. Cien.* 2, 1–14. doi: 10.53673/th.v2i4.177

Spooren, P., Brockx, B., and Mortelman, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Rev. Educ. Res.* 83, 598–642. doi: 10.3102/0034654313496870

Stroebe, W. (2016). *Student evaluations of teaching: no measure for the TEF*. Times Higher Education. Available at: https://www.timeshighereducation.com/comment/student-evaluations-teaching-no-measure-tef (Accessed July 3, 2023).

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: a theoretical and empirical analysis. *Basic Appl. Soc. Psychol.* 42, 276–294. doi: 10.1080/01973533.2020.1756817

R Core Team. (2022). *Writing R extensions. R Foundation for Statistical Computing*, pp. 1–208. Available at: https://cran.r-project.org/doc/manuals/R-exts.html (Accessed July 4, 2023).

Theall, M., Abrami, C., and Lisa, A. (2001). *The student ratings debate: Are they valid? How can we best use them*. San Francisco, California: Jossey Bass Press.

Tomás-Folchy, M., and Durán-Bellonch, M. (2017). Comprendiendo los factores que afectan la transferencia de la formación permanente del profesorado. Propuestas de mejora [understanding the factors affecting transfer of university teachers' permanent training. Proposals for improvement]. *Rev. Electrón. Interuniv. Form. Prof.* 20, 145–157. doi: 10.6018/reifop/20.1.240591

Torquemada, A. (2022). *Evaluación, desarrollo, innovación y futuro de la docencia universitaria. [evaluation, development, innovation and future of university teaching]*. De la Red Iberoamericana de Investigadores en Evaluación de la Docencia. Perfiles educativos, 44. pp. 200–204.

Tuncel, S. D. (2009). Determining effective teacher behavior contributing to students' academic success. *Int. J. Phys. Educ.* 1, 15–18.

Vaillant, D. (2016). El fortalecimiento del desarrollo profesional docente: una mirada desde Latinoamérica. [the strengthening of teacher professional development: a look from Latin America]. *J. Supranational Policies Educ.* 5:5-21. doi: 10.15366/jospoe2016.5

Vance, R. I., Winne, P. S., and Wright, E. S. (1983). A longitudinal examination of rater and ratee effects in performance ratings. *Pers. Psychol.* 36, 609–620. doi: 10.1111/j.1744-6570.1983.tb02238.x

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Wang, D. F., and Guan, L. (2017). Higher education quality evaluation from the perspective of students: theoretical construction and reflection. *J. Nat. Inst. Educ. Admin.* 5:75. doi: 10.3969/j.issn.1672-4038.2017.05.005

Wang, J., and Wang, X. (2012). *Structural equation modeling applications using Mplus*. Chichester: John Wiley & Sons Ltd..

Whittington, L. (2001). Detecting good teaching. *J. Public Aff. Educ.* 7, 5–8. doi: 10.1080/15236803.2001.12023490

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., et al. (2019). Welcome to the Tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Williams, K., and Hebert, D. (2020). Evaluation systems: a literature review on issues and impact. Louisiana educational research association. *Research Issues in Contemporary Education* 5, 42–50.

Zamora, E. (2021). La evaluación del desempeño docente mediante cuestionarios en la universidad: Su legitimidad según la literatura y los requerimientos Para que sea efectiva [the evaluation of teaching performance through questionnaires at the university: its legitimacy according to the literature and the requirements for it to be effective]. *Rev. Actual. Investig. Educ.* 21, 1–23. doi: 10.15517/aie.v21i3.46221

Zhang, Z., and Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: methods and software. *Educ. Psychol. Meas.* 76, 387–411. doi: 10.1177/0013164415594658

Zhou, J. L., and Qin, Y. (2018). The basic types of college students' teaching evaluation behavior deviation and its relationship with students' background characteristics. *Fudan Educ. Forum.* 2018:6. doi: 10.3389/fpsyg.2022.1004487

Zhao, L., Xu, P., Chen, Y., and Yan, S. (2022). A literature review of the research on students' evaluation of teaching in higher education. *Front. Psychol.* 13:1004487. doi: 10.3389/fpsyg.2022.1004487

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 14 September 2023
DOI 10.3389/feduc.2023.1197324

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Reza Kafipour,
Shiraz University of Medical Sciences, Iran
Maria Cutumisu,
McGill University, Canada

*CORRESPONDENCE
Margot Chauliac
✉ margot.chauliac@odisee.be

RECEIVED 30 March 2023
ACCEPTED 14 August 2023
PUBLISHED 14 September 2023

CITATION
Chauliac M, Willems J, Gijbels D and Donche V (2023) The prevalence of careless response behaviour and its consequences on data quality in self-report questionnaires on student learning.
Front. Educ. 8:1197324.
doi: 10.3389/feduc.2023.1197324

COPYRIGHT
© 2023 Chauliac, Willems, Gijbels and Donche. 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.

The prevalence of careless response behaviour and its consequences on data quality in self-report questionnaires on student learning

Margot Chauliac*, Jonas Willems, David Gijbels and Vincent Donche

Department of Training and Education Sciences, University of Antwerp, Antwerp, Belgium

Introduction: Self-report questionnaires are widely used in high schools and universities to gain insights into students' learning strategies and enhance the quality of their education. However, it is important to acknowledge the possibility of respondents being inattentive when completing these questionnaires. While reliability analyses are typically performed at the group level, when providing individual feedback, it is crucial that each respondent's results are reliable. This study aimed to evaluate the prevalence of careless response behaviour in a questionnaire concerning student learning.

Methods: Data analysis encompassed a substantial sample of 12,000+ students in their final two years of secondary education, averaging around 17 years of age. Employing five complementary detection techniques, the study identified instances of careless responding present in the questionnaire data.

Results: Our results underscore a notable prevalence of careless response behaviour among the surveyed students. Application of the five detection techniques revealed a substantial number of instances indicating inattentive responding. Furthermore, the questionnaire's measurement scales were evaluated for reliability. The study noted the presence of carelessness but observed minimal impact on group-level results.

Discussion: The outcomes of this study hold important implications for using self-report questionnaires in education. The prevalence of careless responding emphasizes the need for scrutinizing individual responses. Despite careless responses, their influence on overall group-level data integrity seems restricted. Nonetheless, the study underscores the importance of cautiously interpreting individual-level outcomes, particularly when using these results for individual feedback.

KEYWORDS

careless responding, data quality, self-report questionnaires, detection, processing strategies

1. Introduction

One of the most important aims of educational research is to understand and enhance the quality of learning (Lonka et al., 2004; Kendeou and Trevors, 2012; Gijbels et al., 2014; Dinsmore, 2017). Therefore, gaining insights into how students engage in the learning process is crucial. To date, empirical research has predominantly relied on self-report questionnaires for this (Catrysse et al., 2016; Fryer and Dinsmore, 2020). These questionnaires offer a practical means to survey large groups of respondents in a short period of time and, importantly, provide unique information about many critical aspects of students' learning processes (Fryer and Dinsmore, 2020).

Despite the advantages and widespread use of questionnaires, there have been negative critiques from scholars regarding their reliability (Veenman et al., 2006). Some criticisms highlight respondents' hesitance to accurately report on their processing strategies (Schellings, 2011; Veenman, 2011), leading to a response behaviour that can be labelled as careless (Huang et al., 2012; Meade and Craig, 2012). Careless respondents may disregard survey instructions, misinterpret content or not take the survey seriously, resulting in inaccurate and unusable data (Marasi et al., 2019).

The presence of careless respondents can have a significant impact on data quality, increasing the likelihood of observing relationships between variables that are not actually correlated (Huang et al., 2015). Furthermore, including careless respondents in study results can create disruptions at the individual level. In educational settings such as high schools and universities, where self-report questionnaires on student learning are often used to provide students with feedback, careless respondents may receive inappropriate advice. On an individual level, respondents who are aware of their careless completion of the questionnaire may not attach significance to the feedback provided. However, carelessness can lead to other issues. The outcomes of a questionnaire are sometimes used by educators to evaluate and adjust their teaching methods. Therefore, it is crucial to be able to identify careless respondents so that practitioners are aware of which results should be handled with extra caution. In what follows, we further detail the phenomenon of careless response behaviour, its implications for data quality, and psychometric analysis techniques that can be employed to detect careless response behaviour.

2. Theoretical framework

2.1. Careless respondents

The phenomenon of respondents failing to read and pay attention to the content of a survey is not a new occurrence, and it has been addressed by previous studies using various terms (Huang et al., 2012; Goldammer et al., 2020). The term *random respondents* has the longest history and refers to respondents whose responses demonstrate a pattern of inattention or negligence (Thompson, 1975; Johnson, 2005). However, even seemingly random respondents may exhibit a non-random pattern in their response behaviour (Neuringer, 1986; Meade and Craig, 2012). As a result, more recent literature tends to refer to these respondents as *inattentive* (Bowling et al., 2016; McGonagle et al., 2016) or *careless* (Meade and Craig, 2012). In what follows we will consistently use the latter term.

Careless responding can manifest in various ways, including socially desirable answering and acquiescence bias. Due to social desirability bias, respondents show the tendency to answer questions in such a way that their answers look favourably towards others (Krumpal, 2013). Acquiescence bias refers to a response behaviour where respondents have a tendency to select positive responses to the questions posed (Krosnick, 1999), also referred to as *yeah saying* (Lechner and Rammstedt, 2015). While these behaviours are often categorised as forms of careless responding, this is by definition not the case. Respondents must, after all, invest effort in processing questionnaire items to respond to them in a biased manner. There are two common types of typical careless response behaviour: straightlining and random responding (DeSimone et al., 2018). Both types involve respondents not giving sufficient attention to questions, but they have distinct characteristics. Straightlining occurs when respondents consistently provide identical or nearly identical answers to a set of questions, irrespective of the positive or negative wording of the items (Herzog and Bachman, 1981; Kim et al., 2019). It is important to note that the intensity of straightlining can vary, with some respondents consistently choosing the same response option and others alternating between similar answers (e.g., *often* and *very often*) (Dunn et al., 2018). Random responding (DeSimone et al., 2018; Arias et al., 2020) is more difficult to detect than straightlining. Random respondents choose their answers arbitrarily and intentionally use all available response categories to appear credible (DeSimone et al., 2018).

Various indicators have been developed and employed to identify careless respondents and extract their responses from datasets (Curran, 2016). These indicators can be proactive or reactive. Proactive indicators include specific items or scales that are incorporated in a questionnaire before administration. Reactive indicators refer to a wide range of *post hoc* statistical analyses aimed at identifying careless respondents. Providing an extensive overview of all possible detection indicators and their corresponding cut-off values is beyond the scope of this study. We refer to Curran (2016) for an extensive review of various methods that can be used to identify careless respondents. However, we will highlight five indicators that are often employed in other fields to detect carelessness in self-report questionnaires (Huang et al., 2012; Meade and Craig, 2012).

2.1.1. Control items

Control items, also known as directed response or instructed response items, serve as proactive indicators for detecting careless response behaviour. These items instruct respondents to provide a specific response, such as 'mark *Strongly disagree* to this item' (Lavrakas, 2008; Dunn et al., 2018). As these questions have only one correct answer, it becomes possible to identify careless response behaviour (Huang et al., 2012; Meade and Craig, 2012; Huang et al., 2015). To mask their presence, control items are often scattered throughout a survey (Marasi et al., 2019). These items are considered a sensitive approach for detecting carelessness because it is unlikely that the instructed response will be given without reading the question (Niessen et al., 2016).

When using control items, researchers must establish on a criterion for identifying and eliminating respondents based on apparent careless response behaviour. One first approach is to use a cut-off score based on the number of control items that respondents answer incorrectly. Those who answer more control items incorrectly

are more likely to have responded carelessly (Dunn et al., 2018). Respondents whose scores on incorrectly answering control items exceed a predetermined threshold are then classified as careless (Maniaci and Rogge, 2014; Bowling et al., 2016). This approach is lenient in detecting carelessness. Alternatively, a stricter approach is to require correct answers to all control items (Hauser and Schwarz, 2016). A single incorrect answer to a control item would result in a respondent being classified as careless.

2.1.2. Response time

A reactive indicator commonly used to detect carelessness is response time. The underlying assumption is that there is a minimum amount of time required to read a question and choose a response option that aligns with one's opinions and beliefs (Huang et al., 2012). When response time falls below this minimum threshold, it is unlikely that the respondent read and answered the item seriously, compromising the reliability of the response.

However, the use of response time as an indicator of carelessness presents several methodological challenges. A review by Matjašic et al. (2018) examined 28 studies that employ response time as an indicator of carelessness and found a lack of consensus on how to measure response time. Should it be analysed on a per-item, per-page, or per-questionnaire basis? There is no agreement on what exactly constitutes answering too quickly. Multiple factors, such as perceived questionnaire difficulty, reading speed, decision-making speed, fatigue, and distraction, can introduce significant variations in response times within a sample (Dunn et al., 2018). Response times are likely to differ across different surveys, making it challenging to establish concrete guidelines for using response time as an indicator of careless response behaviour (Huang et al., 2012).

Additionally, it is important to note that response time analyses for detecting carelessness are typically carried out as one-tail analyses; flagging only respondents who answer too quickly. Huang et al. (2012) and Meade and Craig (2012) investigated short response times as a potential indicator of carelessness. In their study, Huang et al. (2012) set a cut-off score based on an educated guess, considering it highly improbable for a respondent to answer a question in less than two seconds. Meade and Craig (2012) took a different approach by examining outliers in the distribution of response times.

2.1.3. Long-string analysis

Long-string analysis is a reactive indicator that examines the invariability of respondents' response patterns. It detects extreme straightlining by observing how frequently a respondent consecutively chooses the same response option (Herzog and Bachman, 1981; Niessen et al., 2016; DeSimone et al., 2018; Kim et al., 2019). The underlying rationale is that careful respondents are expected to choose different response options for different items. Thus, a response pattern that shows minimal or no variability can suggest careless responding. Long-string analysis is particularly suited for questionnaires that include different scales measuring distinct constructs, especially when they include reverse-coded items and when the items from different scales are randomly dispersed throughout the survey (Dunn et al., 2018). However, establishing a clear-cut cut-off value for long-string analysis is challenging (Johnson, 2005), because there is no specific point at which a string of identical responses can be considered excessively long.

2.1.4. Even-odd index

The even-odd index involves dividing the items of a questionnaire's scale into two subscales based on even and odd numbers. Each subscale is then scored separately, and then the correlations between the subscales are calculated to assess the respondent's consistency (Meade and Craig, 2012). For reliable results, it is necessary to reverse possibly negatively worded items before analysis. The even-odd index requires scales with sufficient items to form the two subscales (Curran, 2016). Typically, scales with at least four items measuring the same construct are needed to calculate a correlation. This analysis assumes that responses to items within the same scale should be similar. Therefore, a high correlation suggests more careful completion of the questionnaire, while a low correlation suggests more inconsistent response behaviour. Jackson (1977) argued that even-odd index scores lower than 0.30 may indicate careless response behaviour. However, other scholars consider response behaviour to be careless if the even-odd index deviates from the mean by more than two standard deviations (Iaconelli and Wolters, 2020).

2.1.5. Mahalanobis distance

Mahalanobis distance is a multivariate outlier statistic (Mahalanobis, 1936) that measures the distance between observations and the centre of the data, accounting for the correlational structure between the items. The distance is smallest when the vector of a person's responses is similar to the vector of the sample means. Conversely, a larger Mahalanobis distance may indicate carelessness (Ward and Meade, 2018). The underlying assumption is that response patterns deviating significantly from the sample norm could be indicative of careless response behaviour (Meade and Craig, 2012; DeSimone et al., 2015). This technique has shown some promise as an indicator of careless response behaviour (Ehlers et al., 2009; Meade and Craig, 2012; Maniaci and Rogge, 2014).

2.2. Carelessness and data quality

Multiple studies have highlighted the substantial impact of including careless respondents in a dataset, leading to significant alterations in variable correlations, statistical power, and effect sizes (Woods, 2006; Rammstedt et al., 2013; Wood et al., 2017; DeSimone and Harms, 2018). The prevalence of careless responding can vary widely depending on the indicators used to detect carelessness. Previous research has reported estimates ranging from as low as 1% (Gough and Bradley, 1996) to as high as 46% of respondents (Oppenheimer et al., 2009). Despite these large variations, there is some consensus among scholars that it is reasonable to expect approximately 10 to 15% of respondents to exhibit careless behaviour (Meade and Craig, 2012; Curran, 2016). Even a low percentage of careless responses can have a significant impact on data quality (Oppenheimer et al., 2009; Huang et al., 2015). However, to the best of our knowledge, there is a scarcity of studies in the field of student learning that specifically examine the relationship between carelessness and data quality (Iaconelli and Wolters, 2020).

3. Present research

High schools and universities commonly use online self-assessment instruments to gain insights into learners' needs and assess

the quality of student learning in order to provide feedback (Ruohoniemi et al., 2017; Vermunt and Donche, 2017; Demulder et al., 2019). Questionnaires serve as institutional tools to monitor learning quality and as instruments to provide individual feedback to students. While completion of these questionnaires may be voluntary in some cases, they can also be mandatory, particularly when included in coaching trajectories or used as online self-assessment instruments for students transitioning into higher education. Irrespective of the specific goal, it is of utmost importance that the collected data accurately represent the respondents' genuine thoughts and beliefs (Tourangeau et al., 2000; Schwarz, 2007). This validity becomes even more critical when the questionnaire results are used to provide individual feedback or guidance for future learning endeavours.

The number of respondents that are classified as careless depends not only on the method used for detection but also on the specific cut-off value applied. Different studies employ different cut-off values for carelessness indicators. Researchers may choose a particular cut-off value based on intuition or through statistical analysis. For instance, Huang et al. (2012) used an intuitive cut-off value of two seconds per item for response time, while Iaconelli and Wolters (2020) derived their cut-off value from the average completion time of their survey. Cut-off values used in measures like long-string analysis and even-odd index also vary across studies (Jackson, 1977; Meade and Craig, 2012; Iaconelli and Wolters, 2020). Additionally, some studies employ multiple cut-off values, allowing researchers to adopt both lenient and strict approaches (McGonagle et al., 2016; Silber et al. 2019). This cautious approach provides flexibility in investigating carelessness.

In the field of student learning, research on the prevalence of careless response behaviour is scarce. Therefore, we aim to address this gap by conducting a study using a large dataset of students in their final years of secondary education. These students completed a self-report questionnaire on their cognitive processing strategies. Our study has three main research goals. Firstly, we seek to estimate the proportion of careless respondents within the dataset. To accomplish this, we will employ five carelessness indicators identified in previous research (Meade and Craig, 2012; Maniaci and Rogge, 2014; DeRight and Jorgensen, 2015). We will use different cut-off values for each indicator, adopting both a strict and more lenient approach to identify respondents whose responses should be treated with more caution. Secondly, we will examine the results of the different indicators and assess their coherence with one another. By comparing the outcomes of each indicator, we aim to gain insights into the consistency of the identified careless respondents across multiple measures. Lastly, we will investigate how the inclusion or exclusion of careless respondents from the dataset impacts the reliability of the data. By analysing the relationship between careless response behaviour and reliability estimates, we aim to assess the influence of carelessness on the overall quality and trustworthiness of the collected data.

4. Methodology

To study carelessness in online self-report questionnaires, we used data derived from the Columbus project, a substantial research initiative undertaken by the Flemish Department of Education and Training. The data Columbus refers to both the name of the exploration tool developed to enhance the career decision-making processes of students nearing the completion of secondary education and the overall project itself (Demulder et al., 2021). Data were collected by means of an online exploration instrument. Students can complete this questionnaire during their final 2 years in secondary education. The instrument aims to aid students in their study choice process and to facilitate their transition into higher education by measuring their key competencies while also providing them with feedback on their learning (Demulder et al., 2021).

4.1. Participants

The total sample for this study comprised of 18,386 respondents from four different educational streams: general (60.29%), technical (34.04%), vocational secondary education (3.69%) and arts (1.98%). However, our analysis focused specifically on the students in the general and technical education streams as they constituted the majority of the sample population. Thus, our analysis was conducted on a subset of the total sample, specifically on 13,628 students who completed the section on cognitive processing strategies.

To ensure data quality, the total response time for each respondent was automatically recorded by the survey software. This measure encompassed the time from when the first question was displayed to when the last question was answered, including periods when the survey was open, but respondents were not actively engaged. An outlier analysis was conducted to identify extreme response times that could potentially distort the mean. Consequently, respondents whose response times fell outside the range identified in the box-and-whisker plot were excluded from the analysis. This resulted in the exclusion of 1,050 respondents, leaving a final sample of 12,578 respondents for further analysis. The age range of the respondents varied from 15 to 21 years, with a mean age of 17.01 years. Of the respondents, 56.5% identified as female.

4.2. Instrument

The questionnaire used in this study consisted of five scales, which were selected from two already validated questionnaires on student learning. Three scales were taken from the short version of the *Inventory of Learning Patterns of Students* (ILS-SV) (Donche and Van Petegem, 2008; Vermunt and Donche, 2017), while the other two scales stemmed from the *Learning and Study Strategies Inventory* (LASSI) (Weinstein et al., 1988). In total, the questionnaire included 26 items, with 24 items tapping 5 different cognitive processing strategies and two additional control items aimed at detecting the prevalence of carelessness (Table 1). Both control items followed a uniform format: "*Choose the first response option*." One control item was placed after the questions from the ILS scales, and the other was inserted after the questions from the LASSI scales. For all five scales, respondents provided their answers using a five-point Likert scale. The response options for the ILS-scales ranged from *I rarely or never do this* to *I almost always do this* while the response options for the LASSI-scales ranged from *Not applicable to me at all* to *Very applicable to me*. In order to check the expected dimensional structure of the selected scales, two CFA analyses were carried out on the total sample (N = 12.578), showing adequate model fit for, respectively, the three ILS-SV scales (12 items; CFI = 0.945, RMSEA = 0.060, SRMR = 0.043) and two LASSI-scales (12 items; CFI = 0.945, RMSEA = 0.067,

TABLE 1 Scales, item examples, number of items and reliability (internal consistency).

Scale	Sample question	# of items	Cronbach's alpha
ILS			
Relating and structuring	I relate facts to the bigger picture of a lesson or of the subject matter.	4	0.79
Concrete processing	I use what I learn in a lesson in my activities outside school.	4	0.71
Memorising	I learn definitions as literally as possible by heart.	4	0.67
LASSI			
Information processing	I try to relate what I learn to my own experiences.	6	0.82
Selecting main ideas	During the lesson, I am able to pick up the most important information from the lesson.	6	0.83

SRMR = 0.071). To assess the internal consistency of each scale, Cronbach's alpha coefficient was calculated (Table 1).

4.3. Analyses

To investigate the prevalence of carelessness, we employed a combination of proactive and reactive indicators. Careless response behaviour was identified using five different, yet complementary, indicators: control items, response time, long-string analysis, even-odd index and Mahalanobis distance. For each of these indicators, both strict and lenient cut-off values were established. In the case of the control items, the lenient approach classified respondents as careless if they answered both items incorrectly. The strict approach required only one incorrect response to classify a respondent as careless. For the reactive indicators, the strict approach classified respondents as careless if their results on the respective indicator deviated from the mean by one standard deviation. In the lenient approach, the cut-off values were set at two standard deviations from the mean. By employing these different indicators and cut-off values, we aimed to adopt a cautious approach to identify and capture various degrees of careless response behaviour among the respondents. This comprehensive approach allowed us to assess the prevalence of carelessness in a robust manner.

To examine the relationship between the various indicators of carelessness, we adopted a two-pronged approach. Firstly, we assessed the number of indicators that categorised respondents as careless. Secondly, we compared the results of the proactive and reactive indicators to each other. To analyse the relationship between the different indicators, we employed Silber et al. (2019) procedure, which involved dividing the respondents into three distinct groups. The first group comprised respondents who correctly answered both control items and were categorised as *high attentive*. The second group, labelled as the *medium attentive group*, consisted of respondents who answered one of the two control items correctly. The third group, classified as the *low attentive group*, failed to answer either of the control items correctly. Subsequently, a one-way ANOVA analysis was conducted to investigate how differently these three groups scored on the four reactive detection techniques. To further explore the variations among the groups, Games-Howell *post hoc* comparisons were performed. Effect sizes were calculated using omega squared, providing an estimation of the magnitude of the observed differences.

To address our third research goal, which focused on how the inclusion or exclusion of careless respondents affected the reliability of the measured scales, we conducted a reliability analysis. Cronbach's alpha was chosen as the measure of internal consistency for the scales. To assess the impact of including or excluding careless respondents on the internal consistency levels of the scales, we performed Feldt's tests. These tests allowed us to determine whether there were significant differences in the internal consistency levels among different subsamples, including the careful respondents, the careless respondents, and the full sample (Diedenhofen and Musch, 2016).

5. Results

5.1. The prevalence of carelessness

Figure 1 illustrates how many respondents were categorised as careless by the lenient and strict approach for each of the five different carelessness indicators: control items, response time, long-string analysis, the even-odd index and Mahalanobis distance.

Table 2 details the descriptive statistics on the presence of careless respondents according to the five indicators.

In relation to the lenient cut-off values, the percentage of respondents identified as careless varied from 0.62 to 8.68%, depending on the specific indicator used. Among the indicators, the even-odd index categorised the highest number of respondents as careless, while the response time indicator yielded the fewest careless respondents. On the other hand, when employing strict cut-off values, the range of careless respondents extended from 0.90% (for control items) to 23.25% (for response time). It follows logically that employing stricter criteria for identifying careless responses allows for the identification of a larger number of respondents. However, the counterintuitive finding emerges in relation to the control items indicator. Notably, more respondents were found to answer two control items incorrectly rather than answering incorrectly only once. As a result, the lenient approach yields a higher number of identified careless respondents in this specific case.

5.2. The relationship between different indicators

The relationship between the different indicators was examined in two phases. In the first phase, we assessed the level of agreement among the indicators and investigated whether they identified the same respondents, despite their complementary nature. In the second

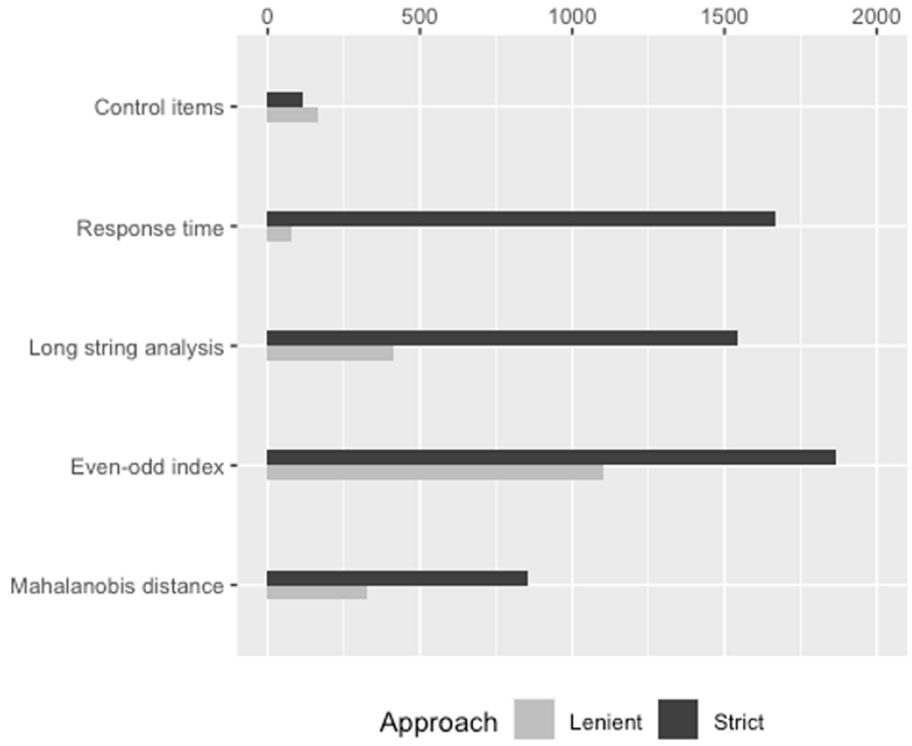


FIGURE 1
Number of careless respondents according to strict and lenient approaches to five separate indicators.

phase, an ANOVA analysis was conducted to explore the relationship between the proactive and reactive indicators.

5.2.1. The reactive indicators

Given the use of five distinct carelessness indicators, we conducted an additional analysis to determine the number of respondents identified as careless by one, two, three, four, or all five indicators. This investigation aimed to enhance our understanding of the complementarity of these various techniques. The findings are presented in Table 3.

Under both the lenient and strict cut-off values, the majority of respondents identified as careless were flagged by only one indicator. When applying lenient cut-off values, approximately 12.27% of respondents were categorised as careless based on a single indicator, while with strict cut-off values, this percentage increased to 30.51%. Relatively fewer respondents were classified as careless by two or more indicators. When lenient cut-off values were used, around 1.50% of respondents were flagged by two indicators, whereas with strict cut-off values, this percentage rose to 6.54%. The number of respondents labelled as careless by three or more indicators was negligible, with the percentage not exceeding 1.19% in both lenient and strict approaches.

Specifically, 189 respondents were identified as careless by two lenient indicators, while 822 respondents were classified as careless by two strict indicators. A summary of these indicator combinations is presented in Table 4.

The use of lenient or strict cut-off values leads to significant discrepancies in the number of respondents categorised as careless. For instance, only four respondents were identified as careless by both the lenient cut-off value for response time and the lenient cut-off value for long-string analysis. However, when employing strict cut-off values, the number of respondents labelled as careless by both response time and long-string analysis increased to 250. Similar substantial differences can be observed for other combinations of indicators, as shown in Table 4. When lenient cut-off values were used, the combination of control items and Mahalanobis distance classified the highest number of respondents as careless, totalling 91 individuals. On the other hand, when employing strict criteria, the combination of response time and long-string analysis identified the largest number of respondents as careless, with a total of 250 individuals.

5.2.2. The relationship between proactive indicators and reactive indicators

To investigate the relationship between proactive and reactive carelessness indicators, the respondents were initially divided into three attentiveness groups: high, medium, and low. This categorisation was based on the calculation of the number of correct and incorrect responses to control items. Subsequently, the relationship between these attentiveness groups and each reactive indicator (response time, long-string analysis, even-odd index, and Mahalanobis distance) was examined using ANOVA. Effect sizes (such as omega squared) and multiple pairwise comparisons (Games-Howell Post-hoc Tests) were calculated to further analyse the differences in scores across the various reactive indicators. The results of these analyses, showcasing how the attentiveness groups performed on the different reactive indicators, are summarised in Table 5.

Attentiveness had a significant effect on response time [F (2, 12,575) = 18.89, p = <0.001], indicating that high attentive respondents took longer to complete the survey compared to low and medium attentive groups. The average response time for high attentive respondents was 237.50 s, while it was 214.95 s for low attentive

10.3389/feduc.2023.1197324

TABLE 2 Descriptive statistics of the careful and careless groups using lenient versus strict cut-off values.

	Indicators for carelessness				
	Control items	Response time	Long-string analysis	Even-odd index	Mahalanobis distance
Full sample					
N	12,578	12,578	12,578	12,578	12,578
(%)	(100)	(100)	(100)	(100)	(100)
M (SD)	*NA*	236.88	4.62	0.49	26.00
		(77.16)	(1.86)	(0.57)	(19.99)
Lenient approach					
cut-off	2 control items wrong	82.56	8.34	−0.65	65.98
Careful					
N	12,412	12,500	12,168	11,475	12,255
(%)	(98.68)	(99.38)	(96.74)	(91.32)	(97.43)
M (SD)	*NA*	237.99	4.40	0.63	23.64
		(76.11)	(1.39)	(0.36)	(10.66)
Careless					
N	166	78	410	1,103	323
(%)	(1.32)	(0.62)	(3.26)	(8.68)	(2.57)
M (SD)	*NA*	60.03	11.05	−0.95	115.34
		(15.21)	(2.43)	(0.10)	(55.45)
Strict approach					
cut-off	1 control item wrong	159.72	6.48	−0.08	45.99
Careful					
N	12,465	10,912	11,032	10,713	11,725
(%)	(99.10)	(76.75)	(87.71)	(85.17)	(93.22)
M (SD)	*NA*	252.36	4.10	0.70	22.29
		(70.52)	(1.08)	(0.26)	(8.68)
Careless					
N	113	1,666	1,546	1865	853
(%)	(0.90)	(23.25)	(12.29)	(14.83)	(6.78)
M (SD)	*NA*	135.53	8.39	−0.70	76.90
		(23.23)	(2.11)	(0.33)	(45.64)

TABLE 3 Overview of the number of respondents considered careless by different indicators.

	Careless by				
	1 indicator	2 indicators	3 indicators	4 indicators	5 indicators
Lenient approach					
N (%)	1,543 (12.27)	189 (1.50)	41 (0.33)	9 (0.07)	0 (0)
Strict approach					
N (%)	3,837 (30.51)	822 (6.54)	150 (1.19)	23 (0.18)	4 (0.03)

respondents. *Post hoc* comparisons indicated that the mean score for the medium attentive ($M = 201.73$, SD = 97.49) and low attentive ($M = 214.95$, SD = 105.70) groups did not differ significantly from each other.

The three groups differed significantly in their scores on the long-string careless response indicator [F (2, 12,575) = 28.93, $p = <0.001$]. High attentive respondents tended to choose the same response option consecutively an average of 4.60, while the medium and low attentive groups had values of 5.91 and 4.84, respectively. The mean score of the high attentive group ($M = 4.60$, SD = 1.83) did not significantly differ from the mean score of the low attentive group ($M = 4.84$, SD = 2.60).

TABLE 4 Respondents categorised as careless by two indicators for both lenient and strict approaches[a].

	Control items	Response time	Long-string analysis	Even-odd index	Mahalanobis distance
Control items		3 (2)	2 (2)	5 (1)	91 (33)
Response time			4 (250)	7 (164)	6 (83)
Long-string analysis				33 (108)	15 (36)
Even-odd index					23 (143)
Mahalanobis distance					

[a]The values for the strict approach between brackets.

TABLE 5 Results of the ANOVA analysis.

	N (%)	*M* (SD)	*df*	Mean Square	*F*	*p*	ω^2
Response time							
High attentive	12,299 (97.78)	237.50(76.38)[a]	2	112147.16	18.89	<0.001	0.003
Medium attentive	113 (0.90)	201.73 (97.49)[b]					
Low attentive	166 (1.32)	214.95 (105.70)[b]					
Long-string analysis							
High attentive	12,299 (97.78)	4.60 (1.83)[a]	2	99.77	28.93	<0.001	0.005
Medium attentive	113 (0.90)	5.91 (3.15)[b]					
Low attentive	166 (1.32)	4.84 (2.60)[a]					
Even-odd index							
High attentive	12,299 (97.78)	0.49 (0.56)[a]	2	5.55	17.40	<0.001	0.003
Medium attentive	113 (0.90)	0.27 (0.65)[b]					
Low attentive	166 (1.32)	0.31 (0.67)[b]					
Mahalanobis distance			2	1106259.36	4941.48	<0.001	0.440
High attentive	12,299 (97.78)	24.00 (11.77)[a]					
Medium attentive	113 (0.90)	120.60 (77.84)[b]					
Low attentive	166 (1.32)	109.31 (51.27)[b]					

Different superscripts identify significant mean differences; significance level: $p < 0.05$.

Even-odd index was also sensitive to differences in attentiveness [F (2, 12,575) = 17.40, p = <0.001]. The responses of high attentive respondents exhibited a higher correlation (0.49) between the even and odd items of a scale compared to the low attentive group (0.31). There was no significant difference in the mean score between the medium attentive group (M = 0.27, SD = 0.65) and the low attentive group (M = 0.31, SD = 0.67).

The attentiveness groups differed significantly on the Mahalanobis distance indicator [F (2, 12,575) = 4941.48, p = <0.001]. The Mahalanobis distance index of the high attentive group (M = 24.00, SD = 11.77) was lower than that of the low attentive group (M = 109.31, SD = 51.27). *Post hoc* comparisons revealed significant differences between the different groups, except for the difference between the medium and the low attentive group.

5.3. The relationship between carelessness indicators and reliability estimates

In this stage of the analysis, we examined the relationship between careless response behaviour and data quality. Specifically, we explored the impact of including or excluding careless respondents on the reliability of the measured scales. Reliability estimates for the scales were calculated for each subgroup, allowing for a comparison of the reliability estimates, particularly in terms of internal consistency (Cronbach's Alpha) with those of the complete dataset. The results of these reliability estimates are presented in Table 6.

The analysis of the relationship between careless response behaviour and data quality revealed different trends among the five detection indicators. Depending on the specific carelessness indicator used, the reliability scores either increased or decreased compared to the reliability scores of the complete sample. This pattern was observed for both the lenient and strict approaches.

When using the control items as indicators of carelessness, the reliability of the careless group was lower than that of the careful group, while the reliability of the careful group remained comparable to the reliability of the full sample. In the case of the lenient approach for response time, the reliability score of the careless group was significantly lower than that of the full sample. However, with the strict approach, the reliability of the careless

10.3389/feduc.2023.1197324

TABLE 6 Comparison of the scales' internal consistencies for both lenient and strict cut-off approaches.

	Full sample	Lenient approach		Strict approach	
		careful	careless	careful	careless
Control items	*Cronbach's Alpha*				
Relating and structuring	0.79[a]	0.79[a]	0.73[b]	0.79[a]	0.77[b]
Concrete processing	0.71[a]	0.71[a]	0.67[a]	0.71[a]	0.60[b]
Memorising	0.67	0.67	0.60	0.67	0.63
Information processing	0.82	0.82	0.80	0.82	0.78
Selecting main ideas	0.83[a]	0.83[a]	0.78[b]	0.83[a]	0.70[b]
Response time	*Cronbach's Alpha*				
Relating and structuring	0.79[a]	0.79[a]	0.73[c]	0.78[a]	0.83[b]
Concrete processing	0.71	0.71	0.74	0.71	0.74
Memorising	0.67	0.67	0.63	0.66	0.72
Information processing	0.82[a]	0.82[a]	0.69[b]	0.81[a]	0.85[a]
Selecting main ideas	0.83[a]	0.83[a]	0.49[b]	0.82[a]	0.85[a]
Long-string analysis	*Cronbach's Alpha*				
Relating and structuring	0.79[a]	0.78[a]	0.94[c]	0.77[a]	0.88[b]
Concrete processing	0.71[a]	0.70[a]	0.92[b]	0.69[a]	0.85[c]
Memorising	0.67[a]	0.67[a]	0.84[b]	0.66[a]	0.74[c]
Information processing	0.82[a]	0.81	0.99[b]	0.80[c]	0.95[d]
Selecting main ideas	0.83[a]	0.83[a]	0.81[a]	0.81[a]	0.90[b]
Even-odd index	*Cronbach's Alpha*				
Relating and structuring	0.79[a]	0.81[b]	0.37[b]	0.82[a]	0.45[c]
Concrete processing	0.71[a]	0.73[b]	0.22[b]	0.75[c]	0.32[d]
Memorising	0.67[a]	0.70[b]	0.01[c]	0.71[d]	0.21[e]
Information processing	0.82[a]	0.83[b]	0.68[c]	0.83[b]	0.70[d]
Selecting main ideas	0.83[a]	0.83[a]	0.68[c]	0.84[b]	0.69[c]
Mahalanobis distance	*Cronbach's Alpha*				
Relating and structuring	0.79[a]	0.79[a]	0.69[b]	0.80[a]	0.68[b]
Concrete processing	0.71[a]	0.71[a]	0.62[b]	0.72[a]	0.62[b]
Memorising	0.67[a]	0.67[a]	0.62[b]	0.68[a]	0.58[b]
Information processing	0.82[a]	0.82[a]	0.72[b]	0.82[a]	0.69[b]
Selecting main ideas	0.83[a]	0.83[a]	0.71[b]	0.83[a]	0.74[b]

Different superscripts identify significant mean differences; significance level: $p < 0.05$.

group increased for certain scales, while the reliability of the careful group remained comparable to the full sample. The analysis of long-string analysis revealed higher reliability scores for the careless group compared to the full sample and the overall group of careless respondents. Additionally, the lenient approach yielded higher reliability scores than the strict approach. However, for both cut-off values, the reliability score of the careful group decreased. In contrast, the even-odd index exhibited an opposite pattern. The careless group identified using the lenient approach had the lowest reliability score, while the reliability score of the careful group increased considerably. The Mahalanobis distance indicator demonstrated similar patterns to the even-odd index, but the even-odd index appeared to be more strongly associated with lower reliabilities in the group of careless responders.

Furthermore, when considering a significant number of respondents identified as careless by at least one indicator using the lenient or strict approach, the relationship with reliability estimates was examined within these subgroups as opposed to the full sample. Both careful responder subgroups showed significantly higher levels of internal consistency on all scales compared to the careless responder subgroups. Grouping responders using the lenient approach as careless resulted in the lowest levels of reliabilities compared to the strict approach. These findings highlight how the inclusion of careless respondents in samples can considerably affect the reliability estimates of scales (see Table 7).

TABLE 7 Comparison of reliability estimates of the scales for the full sample ($N = 12.578$) and for the subsamples based on lenient and strict cut-off approaches.

	Full sample	Lenient approach 2SD		Strict approach 1SD	
		Careful	Careless	Careful	Careless
Careless by at least 1 indicator		10,796 (85.83)	1782 (14.17)	7,742 (62.55)	4,836 (37.45)
Scales	*Cronbach's Alpha*				
Relating and structuring	0.79[a]	0.80[b]	0.68[c]	0.80[b]	0.77[d]
Concrete processing	0.71[a]	0.73[b]	0.59[c]	0.73[b]	0.68[d]
Memorising	0.67[a]	0.69[b]	0.46[c]	0.70[b]	0.61[d]
Information processing	0.82[a]	0.82[a]	0.81[c]	0.81[b]	0.83[d]
Selecting main ideas	0.83[a]	0.84[b]	0.73[c]	0.83	0.82[d]

Different superscripts identify significant mean differences; significance level: $p < 0.05$.

6. Discussion and conclusion

The assumption is often made that completing questionnaires is a simple and uncomplicated task. It is commonly believed that respondents merely read the questions, provide their answers, and proceed to the next question until the questionnaire is finished. Nonetheless, this perception can deceive us into thinking that this process is effortlessly manageable for every respondent. In reality, the completion of a questionnaire is a complex endeavour that engages various cognitive processes. Unfortunately, not all respondents invest the necessary effort in these processes, whether intentionally or unintentionally (Tourangeau, 1984; Karabenick et al., 2007). Respondents who complete questionnaires carelessly are a concern for researchers as even a low percentage of carelessness can significantly alter the data (Oppenheimer et al., 2009; Huang et al., 2015) in ways that change the correlations between variables, statistical power, and effect sizes they observe in their studies (Woods, 2006; Rammstedt et al., 2013; Wood et al., 2017; DeSimone and Harms, 2018). Consequently, the study aims to achieve two primary objectives: first, to investigate the prevalence of careless response behaviour, and second, to explore its potential consequences on data quality. To illustrate these aspects, a large dataset is employed for analysis in this study.

The first objective of the study was to detect the presence of careless response behaviour by examining pro-active and reactive indicators. The identification of careless respondents varied depending on the specific indicator used and whether a lenient or strict approach was taken The outcomes obtained from different indicators were influenced by the particular type of careless response behaviour exhibited by the respondents (DeSimone and Harms, 2018). By employing multiple complementary detection techniques, various types of careless response behaviour were identified in the dataset, consistent with previous research findings (Curran, 2016). Our study expanded on these results by examining the differences between employing a lenient or strict cut-off approach when identifying careless responders. When lenient criteria were used, between 0.62 and 8.68% of respondents were classified as careless. Among the indicators, response time flagged the fewest respondents, while the even-odd index flagged the largest number. However, when strict cut-off values were applied, significantly higher percentages of respondents were categorised as careless. The control items identified 0.90% of respondents, whereas response time flagged 23.25% of respondents. Comparing the percentages of careless respondents reported in other studies reveals similarly diverse findings: Gough and Bradley (1996) identified 1% of their respondents as careless while Oppenheimer et al. (2009) identified 46% of their respondents as careless.

To gain deeper insights into the effectiveness of these indicators in detecting careless response behaviour, we also examined their individual contributions to the selection process. Results indicate that, with both lenient and strict cut-off values, the majority of respondents were classified as careless based on just one indicator. A smaller number of respondents were labelled as careless by two indicators and even fewer were identified as such by three indicators. Very few respondents were labelled as careless by four or five indicators. This suggests that the indicators used are highly complementary and allow researchers to discern different types of careless respondents. Overall, 14.17% of respondents were considered careless by at least one of the five indicators when employing the lenient approach, while the strict criteria considered 37.45% of respondents as careless. These percentages are lower than the 49% reported by Oppenheimer et al. (2009). Additionally, the lenient approach's result falls within the range of 10 to 15%, which is considered a reasonable percentage of careless respondents according to scholarly consensus (Meade and Craig, 2012; Curran, 2016). Despite the substantial difference between the two approaches and the fact that the results of the lenient approach fall within the expected range, the questionnaire responses alone do not provide sufficient information to determine if any of the strict cut-off values were excessively stringent. The study further demonstrated that identification of careless respondents is not solely dependent on the selection of pro-active and reactive indicators, but is also influenced by decisions made regarding the choice of cut-off values.

The second objective of the study was to investigate the relationship between careless response behaviour and data quality, specifically focusing on the internal consistency of the scales. The analyses revealed significant differences in the alpha coefficients of the scales when examined within different subsamples. Among respondents identified as careless by the long-string analysis, their responses still demonstrated internal consistency. In contrast, respondents flagged by the even-odd index showed inconsistent responses. This finding is expected since the long-string analysis

detects identical answers, which typically yield consistent results, whereas the even-odd index identifies random answering patterns. The impact on reliability scores appears to depend on the specific carelessness indicator used, as well as the number of respondents labelled as careless. In a larger sample like ours, these differences may go unnoticed, emphasising the importance of examining careless behaviour at the individual level as well. By considering careless responding on an individual basis, researchers can gain a more comprehensive understanding of how it affects data quality and the internal consistency of scales.

The findings of the study also revealed that the inclusion or exclusion of careless respondents in different sample groups had an impact on the reliability estimates of the data samples. Specifically, we examined whether different groups of respondents, classified based on their attentiveness as determined by their responses to the control items, differed in terms of reactive indicators of carelessness. The analyses demonstrated that response time, long-string analysis, even-odd index, and Mahalanobis distance were closely associated with the level of attentiveness. The group classified as highly attentive exhibited longer response times compared to both the medium and low attentive groups, who completed the questionnaire more quickly. Moreover, the high attentive group provided shorter strings of consecutive identical answers compared to the medium and low attentive groups. The high attentive group also scored higher on the even-odd index and Mahalanobis distance compared to the other two groups. Regarding the relationship between attentiveness and scale reliability, the internal consistency of the scales was higher for the high attentive group compared to the medium and low attentive groups. Irrespective of the specific scale being measured, the reliability was consistently lowest for the medium or low attentive group. These findings suggest that the inclusion of careless respondents can significantly impact the reliability estimates in different sample groups and highlights the importance of considering attentiveness levels and the presence of careless responding when assessing the reliability of scales and interpreting the data obtained from different respondent groups.

Our study successfully demonstrated the prevalence of careless responding in a large database using student surveys, leading to several important observations for research practice. Firstly, employing different and complementary detection methods, both proactive and reactive, proves to be a fruitful approach for identifying careless respondents with diverse answering patterns beyond simple straightlining behaviour (Huang et al., 2012; Curran, 2016). Secondly, the results further underscore the significance of considering careless responders when examining data quality concerns in student surveys, as their presence can also impact the accuracy of subsequent analyses and results (also see Ward and Mead, 2023). Thirdly, incorporating multiple indicators of carelessness in the analyses is valuable, but it is equally important to inspect the cut-off values, as indicated by the presented results. These discussed indicators allow researchers to detect careless response behaviour at an initial level. However, more in-depth techniques are needed to gain deeper insights into this matter. Future research should shed light on the actual process of completing questionnaires to understand why respondents exhibit specific answering behaviours. Conducting post-questionnaire interviews could be an initial step towards uncovering the reasons behind certain response patterns. Additionally, employing eye-tracking studies could provide an unobtrusive means for investigating the completion process of the questionnaire (Chauliac et al., 2020, 2022). This would enable a better understanding of individual differences in cognitive processing during questionnaire completion, and the findings could then be incorporated into reliability analyses.

Despite the limitations of the current research, it highlights that the detection techniques examined can serve as an initial step in identifying respondents whose answering patterns careless response behaviour, which is crucial for researchers using self-report questionnaires. It is important to note that the presence of careless respondents in a dataset is not inherently negative. When researchers are aware of the existence of careless response behaviour, they can perform analyses that consider its influence without immediately excluding careless respondents from the dataset. For researchers aiming to detect carelessness and enhance data quality in survey research, it is like walking a tightrope, where one must tread with exceptional care. On one hand, researchers should avoid being overly strict in identifying carelessness to prevent the exclusion of reliable respondents from the dataset. On the other hand, being too lenient may result in including respondents who genuinely exhibited careless behaviour. While using these indicators to detect carelessness can be a useful starting point for identifying potential suspect cases, it should not be the sole endpoint or final determination. Further research and methodologies are necessary to delve deeper into understanding and addressing careless response behaviour in survey research. Researchers should continuously refine and improve detection techniques, considering additional factors such as respondent interviews and eye-tracking studies, to enhance understanding of the underlying reasons for careless responses and incorporate this knowledge into reliability analyses.

Data availability statement

The datasets presented in this article are not readily available given the confidential nature of the data. Requests to access the datasets should be directed to margot.chauliac@odisee.be.

Ethics statement

In accordance with local legislation and institutional requirements, this study involving human participants did not necessitate ethical review and approval. The datasets used in this study are derived from the Columbus project, commissioned by the Flemish Ministry of Education and Training. All participants provided written informed consent. During the registration process, students agreed to the terms and conditions developed in collaboration with the Data Protection Officer of the Ministry of Education and Training, aligning with the General Data Protection Regulation (GDPR). The Ministry strictly adheres to GDPR principles when handling personal data. The sharing of personal data follows a designated protocol, outlined in the Columbus

protocol (January 27, 2021), which is accessible at https://data-onderwijs.vlaanderen.be/documenten/bestand.ashx?id=13051.

Author contributions

MC, JW, DG, and VD contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by BOF (UA/FFB170229) and the Flanders Research Foundation (FWO/K801922N).

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.

References

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., and Arias, B. (2020). A little garbage in, lots of garbage out: assessing the impact of careless responding in personality survey data. *Behav. Res. Methods* 52, 2489–2505. doi: 10.3758/s13428-020-01401-8

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., and Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *J. Pers. Soc. Psychol.* 111, 218–229. doi: 10.1037/pspp0000085

Catrysse, L., Gijbels, D., Donche, V., De Maeyer, S., Van den Bossche, P., and Gommers, L. (2016). Mapping processing strategies in learning from expository text: An exploratory eye tracking study followed by a cued recall. *Frontline learning research* 4, 1–16. doi: 10.14786/flr.v4i1.192

Chauliac, M., Catrysse, L., Gijbels, D., and Donche, V. (2020). It is all in the "surv-eye": can eye tracking data shed light on the internal consistency in self-report questionnaires on cognitive processing strategies?. *Frontline Res. Educ.* 8, 26–39.

Chauliac, M., Catrysse, L., Gijbels, D., De Maeyer, S., and Donche, V. (2022). Self-report questionnaires scrutinised: Do eye movements reveal individual differences in cognitive processes while completing a questionnaire? *Int. J. Soc. Res. Methodol.* 26, 391–407.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *J. Exp. Soc. Psychol.* 66, 4–19. doi: 10.1016/j.jesp.2015.07.006

Demulder, L., Lacante, M., and Donche, V. (2021). "Large scale measurements to support students in their transition to higher education. The importance of including a non-cognitive perspective" in *Research on teaching and learning in higher education*. ed. E. Braun (Waxman), 11–20.

Demulder, L., Willems, J., Verschueren, K., Lacante, M., and Donche, V. (2019). Het studiekeuzeproces voor hoger onderwijs in kaart: validering van de Verkorte Vragenlijst Studiekeuzetaken (VVST). *Pedagogische Studiën* 96, 312–328.

DeRight, J., and Jorgensen, R. S. (2015). I just want my research credit: frequency of suboptimal effort in a non-clinical healthy undergraduate sample. *Clin. Neuropsychol.* 29, 101–117. doi: 10.1080/13854046.2014.989267

DeSimone, J. A., DeSimone, A. J., Harms, P. D., and Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Appl. Psychol.* 67, 309–338. doi: 10.1111/apps.12117

DeSimone, J. A., and Harms, P. D. (2018). Dirty data: the effects of screening respondents who provide low-quality data in survey research. *J. Bus. Psychol.* 33, 559–577. doi: 10.1007/s10869-017-9514-9

DeSimone, J. A., Harms, P. D., and DeSimone, A. J. (2015). Best practice recommendations for data screening. *J. Organ. Behav.* 36, 171–181. doi: 10.1002/job.1962

Diedenhofen, B., and Musch, J. (2016). Cocron: a web Interface and R package for the statistical comparison of Cronbach's alpha coefficients. *Int. J. Internet Sci.* 11, 51–60.

Dinsmore, D. L. (2017). Toward a dynamic, multidimensional research framework for strategic processing. *Educ. Psychol. Rev.* 29, 235–268. doi: 10.1007/s10648-017-9407-5

Donche, V., and Van Petegem, P. (2008). The validity and reliability of the short inventory of learning patterns. in *Style and cultural differences: how can organisations, regions and countries take advantage of style differences*. eds. E. Cools, H. van den Broeck and T. Redmond (Leuven University Press), 49–59.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., and Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: comparison to other indicators and relationships with individual differences. *J. Bus. Psychol.* 33, 105–121. doi: 10.1007/s10869-016-9479-0

Ehlers, C., Greene-Shortridge, T. M., Weekley, J., and Zajack, M. D. (2009). *The exploration of statistical methods in detecting random responding*. Atlanta, GA: Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology.

Fryer, L. K., and Dinsmore, D. L. (2020). The promise and pitfalls of self-report: development, research design and analysis issues, and multiple methods. *Frontline Learn. Res.* 8, 1–9. doi: 10.14786/flr.v8i3.623

Gijbels, D., Donche, V., Richardson, J. T. E., and Vermunt, J. D. H. M. (2014). *Learning patterns in higher education*. Dimensions and research perspectives: Routledge.

Goldammer, P., Annen, H., Stöckli, P. L., and Jonas, K. (2020). Careless responding in questionnaire measures: detection, impact, and remedies. *Leadersh. Q.* 31:101384. doi: 10.1016/j.leaqua.2020.101384

Gough, H. G., and Bradley, P. (1996). *Manual for the California psychological inventory. 3rd edn.* Palo Alto, CA: Consulting Psychologists Press.

Hauser, D. J., and Schwarz, N. (2016). Semantic prosody and judgment. *J. Exp. Psychol. Gen.* 145, 882–896. doi: 10.1037/xge0000178

Herzog, A. R., and Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opin. Q.* 45, 549–559. doi: 10.1086/268687

Huang, J., Curran, P., Keeney, J., Poposki, E., and DeShon, R. (2012). Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* 27, 99–114. doi: 10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., and Bowling, N. A. (2015). Insufficient effort responding: examining an insidious confound in survey data. *J. Appl. Psychol.* 100, 828–845. doi: 10.1037/a0038510

Iaconelli, R., and Wolters, C. A. (2020). Insufficient effort responding in surveys assessing self-regulated learning: nuisance or fatal flaw? *Frontline Learn. Res.* 8, 104–125. doi: 10.14786/flr.v8i3.521

Jackson, D. N. (1977). Reliability of the Jackson personality inventory. *Psychol. Rep.* 40, 613–614. doi: 10.2466/pr0.1977.40.2.613

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *J. Res. Pers.* 39, 103–129. doi: 10.1016/j.jrp.2004.09.009

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., et al. (2007). Cognitive processing of self-report items in educational research: do they think what we mean? *Educ. Psychol.* 42, 139–151. doi: 10.1080/00461520701416231

Kendeou, P., and Trevors, G. (2012). "Quality learning from texts we read: what does it take?" in *Enhancing the quality of learning: Dispositions, instruction, and learning processes*. eds. J. R. Kirby and M. J. Lawson (Cambridge University Press), 251–275.

Kim, Y., Dykema, J., Stevenson, J., Black, P., and Moberg, D. P. (2019). Straightlining: overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Soc. Sci. Comput. Rev.* 37, 214–233. doi: 10.1177/0894439317752406

Krosnick, J. A. (1999). Survey research. *Annu. Rev. Psychol.* 50, 537–567. doi: 10.1146/annurev.psych.50.1.537

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Qual. Quant.* 47, 2025–2047. doi: 10.1007/s11135-011-9640-9

Lavrakas, P. J. (Ed.) (2008). *Encyclopedia of survey research methods. Vols. 1-10*. Sage Publications. doi: 10.4135/9781412963947

Lechner, C. M., and Rammstedt, B. (2015). Cognitive ability, acquiescence, and the structure of personality in a sample of older adults. *Psychol. Assess.* 27, 1301–1311. doi: 10.1037/pas0000151

Lonka, K., Olkinuora, E., and Mäkinen, J. (2004). Aspects and prospects of measuring studying and learning in higher education. *Educ. Psychol. Rev.* 16, 301–323. doi: 10.1007/s10648-004-0002-1

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Natl. Acad. Sci. India* 2, 49–55.

Maniaci, M. R., and Rogge, R. D. (2014). Caring about carelessness: participant inattention and its effects on research. *J. Res. Pers.* 48, 61–83. doi: 10.1016/j.jrp.2013.09.008

Marasi, S., Wall, A., and Brewer, K. (2019). Participant carelessness: is it a substantial problem with survey data? *Electron. J. Bus. Res. Methods* 17, 1–16. Available at: https://academic-publishing.org/index.php/ejbrm/article/view/1385/1348

Matjašic, M., Vehovar, V., and Manfreda, K. L. (2018). Web survey paradata on response time outliers: a systematic literature review. *Adv. Methodol. Stat.* 15, 23–41. doi: 10.51936/yoqn3590

McGonagle, A. K., Huang, J. L., and Walsh, B. M. (2016). Insufficient effort survey responding: an under-appreciated problem in work and organisational health psychology research. *Appl. Psychol.* 65, 287–321. doi: 10.1111/apps.12058

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455. doi: 10.1037/a0028085

Neuringer, A. (1986). Can people behave "randomly?": the role of feedback. *J. Exp. Psychol. Gen.* 115, 62–75. doi: 10.1037/0096-3445.115.1.62

Niessen, A. S. M., Meijer, R. R., and Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: which method to use? *J. Res. Pers.* 63, 1–11. doi: 10.1016/j.jrp.2016.04.010

Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* 45, 867–872. doi: 10.1016/j.jesp.2009.03.009

Rammstedt, B., Kemper, C. J., and Borg, I. (2013). Correcting big five personality measurements for acquiescence: an 18-country cross-cultural study. *Eur. J. Personal.* 27, 71–81. doi: 10.1002/per.1894

Ruohoniemi, M., Forni, M., Mikkonen, J., and Parpala, A. (2017). Enhancing quality with a research-based student feedback instrument: a comparison of veterinary students' learning experiences in two culturally different European universities. *Qual. High. Educ.* 23, 249–263. doi: 10.1080/13538322.2017.1407401

Schellings, G. (2011). Applying learning strategy questionnaires: problems and possibilities. *Metacogn. Learn.* 6, 91–109. doi: 10.1007/s11409-011-9069-5

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Appl. Cogn. Psychol.* 21, 277–287. doi: 10.1002/acp.1340

Silber, H., Danner, D., and Rammstedt, B. (2019). The impact of respondent attentiveness on reliability and validity. *Int. J. Soc. Res. Methodol.* 22, 153–164. doi: 10.1080/13645579.2018.1507378

Thompson, A. H. (1975). Random responding and the questionnaire measurement of psychoticism. *Soc. Behav. Personal. Int. J.* 3, 111–115. doi: 10.2224/sbp.1975.3.2.111

Tourangeau, R. (1984). "Cognitive sciences and survey methods" in *Cognitive aspects of survey methodology: building a bridge between disciplines*. eds. T. B. Jabine, M. L. Straf, J. M. Tanur and R. Tourangeau (Washington, DC: National Academy Press), 73–100.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The psychology of survey response* Cambridge University Press.

Veenman, M. V. J. (2011). Alternative assessment of strategy use with self-report instruments: a discussion. *Metacogn. Learn.* 6, 205–211. doi: 10.1007/s11409-011-9080-x

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., and Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacogn. Learn.* 1, 3–14. doi: 10.1007/s11409-006-6893-0

Vermunt, J. D. H. M., and Donche, V. (2017). A learning patterns perspective on student learning in higher education: State of the art and moving forward. *Educ. Psychol. Rev.* 29, 269–299. doi: 10.1007/s10648-017-9414-6

Ward, M. K., and Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Appl. Psychol.* 67, 231–263. doi: 10.1111/apps.12118

Ward, M. K., and Meade, A. W. (2023). Dealing with careless responding in survey data: prevention, identification, and recommended best practices. *Annu. Rev. Psychol.* 74, 577–596. doi: 10.1146/annurev-psych-040422-045007

Weinstein, C. E., Zimmermann, S. A., and Palmer, D. R. (1988). "Assessing learning strategies: the design and development of the LASSI" in *Learning and study strategies: Issues in assessment, instruction, and evaluation*. eds. C. E. Weinstein, E. T. Goetz and P. A. Alexander (San Diego, California: Academic Press), 25–40.

Wood, D., Harms, P. D., Lowman, G. H., and DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Soc. Psychol. Personal. Sci.* 8, 454–464. doi: 10.1177/1948550617703168

Woods, C. M. (2006). Careless responding to reverse-worded items: implications for confirmatory factor analysis. *J. Psychopathol. Behav. Assess.* 28, 186–191. doi: 10.1007/s10862-005-9004-7

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 22 September 2023
DOI 10.3389/feduc.2023.1241998

Check for updates

OPEN ACCESS

EDITED BY
Jeffrey K. Smith,
University of Otago, New Zealand

REVIEWED BY
Erkan Er,
Middle East Technical University, Türkiye
Liyan Liu,
Northeast Normal University, China

*CORRESPONDENCE
Gavin T. L. Brown
✉ gavin.brown@umu.se

RECEIVED 18 June 2023
ACCEPTED 05 September 2023
PUBLISHED 22 September 2023

CITATION
Brown GTL, Andersson C, Winberg M and Palm T (2023) Predicting formative feedback practices: improving learning and minimising a tendency to ignore feedback.
Front. Educ. 8:1241998.
doi: 10.3389/feduc.2023.1241998

COPYRIGHT
© 2023 Brown, Andersson, Winberg and Palm. 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.

Predicting formative feedback practices: improving learning and minimising a tendency to ignore feedback

Gavin T. L. Brown[1,2]*, Catarina Andersson[3,4], Mikael Winberg[3,5] and Torulf Palm[3,4]

[1]Department of Applied Educational Science, Umeå University, Umeå, Sweden, [2]Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand, [3]Department of Mathematics and Science Education, Umeå University, Umeå, Sweden, [4]Umeå Mathematics Education Research Centre (UMERC), Umeå University, Umeå, Sweden, [5]Umeå Science Education Research (UmSER), Umeå University, Umeå, Sweden

Introduction: Teacher conceptions of feedback ideally predict their feedback practices, but little robust evidence identifies which beliefs matter to practices. It is logical to presume that teacher conceptions of feedback would align with the policy framework of an educational jurisdiction. The *Teacher Conceptions of Feedback* (TCoF) inventory was developed in New Zealand which has a relatively low-stakes, formative policy framework.

Methods: This study surveyed 451 Swedish teachers working in Years 1-9, a policy context that prioritises teachers using multiple data sources to help students learn. The study used a translated version of the TCoF inventory, but isolated six items related to formative feedback practices from various factors.

Results: A six-factor TCoF was recovered (Praise, Improvement, Ignore, Required, PASA, and Timely), giving partial replication to the previous study. A well-fitting structural equation model showed formative practices were predicted by just two conceptions of feedback (i.e., feedback improves learning and students may ignore feedback).

Discussion: This study demonstrates that the TCoF inventory can be used to identify plausible relations from feedback beliefs to formative feedback practices.

KEYWORDS

feedback, formative practices, classroom teachers, Sweden, beliefs, perceptions, conceptions

Introduction

Teacher conceptions of or beliefs about feedback are likely to matter to how feedback is implemented and whether it contributes to greater learning or better teaching (Barnes et al., 2015). However, The Theory of Planned Behaviour (Ajzen, 1991) indicates that cognitive and affective attitudes towards intentions within a behavioural phenomenon are constrained by social norms and perceptions of behavioural control. Thus, as teacher feedback practices are likely to be shaped by the shared conceptions of feedback within a specific educational system, it is important to examine these relationships within a specific context. The present paper contributes to the field by being the first survey study to explicitly link teacher conceptions of feedback to their self-reported feedback practices. Moreover, it reports a structural equation model of these relationships within the education system of Sweden, which takes a low-stakes formative approach to assessment, rather than implementing a high-stakes testing regime.

Incidentally, the study also provides insights into the generalisability of a teacher self-report inventory about feedback conceptions developed in New Zealand and deployed in Swedish. Thus, a stronger emphasis on understanding feedback as a mechanism for improving educational outcomes might be expected in the Swedish educational system than in educational systems where assessments are more high-stake. In addition, it is reasonable to expect that a conception that emphasizes feedback as a vehicle for improvement will be reflected in teachers' feedback practices. Thus, the goal of this study was to measure teacher self-reported formative feedback practices and identify possible predictor beliefs from their conceptions of the nature and purpose of feedback. Our hypotheses were:

1. Swedish teachers will strongly exhibit improvement-oriented beliefs about feedback.
2. We expected improvement-oriented beliefs about feedback to influence the self-reported formative feedback practices.

Literature review

Based on a conventional definition, feedback is consequential to performance (Hattie and Timperley, 2007). An important function of feedback then is to provide information to learners and teachers about what each party needs to do in the classroom to meet curricular objectives of schooling. Performance data can be interactions in a classroom (e.g., question and answer) but also includes more formal diagnostic testing (Brown and Hattie, 2012) or analysis of errors made in classroom or home practice (Bejar, 1984). From these kinds of performance data teachers can formatively make appropriate adjustments to their classroom instruction and to student learning activities (Lai and Schildkamp, 2016). This includes giving learners feedback as to the task, the learning process, and the metacognitive self-awareness students have about the instructional objectives (Hattie and Timperley, 2007). Naturally, this approach to assessment requires resources (i.e., assessments that diagnose needs and time to plan responses), policies that prioritise using assessment formatively rather than solely for administrative or summative purposes, and teacher commitment to generating and providing feedback formatively for improvement.

Further, considering the importance of the teacher's active role in using performance data in this way, it is logical to imagine that teacher beliefs about feedback matter to the efficacy of these processes. Ajzen's (1991) Theory of Planned Behaviour (TPB) identifies the importance of teachers' beliefs and attitudes toward the phenomena of assessment and feedback as essential to understanding their intentions and actions. TPB points to the importance of attitudes, social norms, and perceptions of behavioural control as predictors of intentions, behaviours, and outcomes. This aligns well with Fives and Buehl's (2012) model in which teacher beliefs act as filters, frames, and guides to cognitive resources that impact their actions. Of course, teacher beliefs about the proper role of feedback are not universal; they are context bound by the policy and practice framework in which they are employed (Brown and Harris, 2009; Fulmer et al., 2015). As Bonner (2016) makes clear, teacher assessment practices are pressed by policy and regulatory pressures. Where external agencies permit teachers a great deal of control and autonomy in teaching (e.g., Sweden), it is highly likely that their beliefs will very much shape their practices. Hence, it is plausible that teacher conceptions of feedback will influence their self-reported feedback practices.

Research into teacher beliefs about feedback indicates that teachers have multiple conceptions about it with varying degrees of intensity (Brown et al., 2012). Because of the multiple purposes and uses of feedback, teachers have multiple and complex attitudes or conceptions in response to those uses within any jurisdiction. The relative strength of these varying conceptions appears to be ecologically rational in that teachers in general endorse the policies and purposes that apply to their level of employment (Rubie-Davies et al., 2012).

Relationship of feedback perceptions to self-reported practices

The research on how conceptions of feedback relate to behaviours is largely limited to self-reported practices rather than to actual observed practices. This raises the possibility that such data are invalid because of memory failure or ego-protective responses. To minimise that threat, valid data collection uses multiple items for each potential latent cause, designed to present theoretically important stimuli that are analysed mathematically to determine fit to the theory (Brown, 2023). Reliance on observation alone cannot expose what the most knowledgeable informant knows about what lies in the beliefs, thoughts, ideas, emotions, and attitudes of that individual; hence, self-report (Brown, 2023).

Thus, survey research still produces limited information about how beliefs about feedback relate to feedback practices. For example, a survey of 54 Tanzanian mathematics teachers (Kyaruzi et al., 2018), found that endorsement of approaches to feedback that focus on monitoring (e.g., asking students to indicate what went well and what went badly with their assignments) and scaffolding (e.g., adjusting instruction whenever I notice that students do not understand a topic) had strong prediction on high-quality feedback delivery practices (e.g., being supportive when giving students feedback and encouraging students to ask for feedback whenever they are uncertain). In contrast, a small survey of 61 Ethiopian teachers found that the relationship of teacher beliefs about feedback had a statistically not significant correlation with their practices (Dessie and Sewagegn, 2019).

In contrast, based on responses to the *Teachers Conceptions of Feedback* inventory (TCoF; Harris and Brown, 2008), a nationally representative survey of 518 New Zealand teachers found that endorsement of feedback about learning processes and involving students in feedback predicted greater use of non-teacher feedback methods (Brown et al., 2012). The same study reported that the use of praise in feedback predicted feedback actions that protected students from negative evaluative consequences. A survey of 390 Pakistani teachers, using the TCoF, found that endorsement of feedback as encouragement led to greater use of protective evaluation practices such as giving positive messages to students and not making critical comments (Aslam and Khan, 2021).

The Swedish context

According to the joint European Values Study and World Values Survey 2005–2022 (EVS/WVS, 2022), Sweden is a strongly secular-rational and individualistic country with a strong emphasis on equality and the individual's freedom and wellbeing. A similar description is

found in the six-dimensional Index of National Culture (INC) by Hofstede et al. (2010). This is reflected in teachers' relatively high degree of freedom to interpret and concretize the objectives of the national curriculum and to decide on appropriate teaching methods to help students achieve these goals (Helgøy and Homme, 2007). The Swedish egalitarian ideal is also reflected in teacher–student relations. Teachers do not receive, nor demand, respect solely based on their position/role in society. In the classroom, the average student is the norm and discrimination between students in terms of, for example, special classes or educational tracks for gifted or underperforming students, is rare. Rating of students in public, whether explicit or implicit, based on their school achievement is not in line with the Swedish culture.

The Swedish curriculum is goal-oriented, with national standards for student learning in years 3, 6, and 9 (nominally aged 9, 12, and 15). Grades are given in school years 6–9, but only the year 9 grades are high stakes because they matter for admission to upper secondary school. The grades are criterion-referenced, meaning if the standards are achieved, any number of students can receive that grade. Legally, the municipality is responsible for providing adequate resources for education and to conduct systematic evaluations [Utbildningsdepartementet (Ministry of Education), 2010]. In order to support schools and teachers to fulfil their obligations, the Swedish National Agency for Education (SNAEd) provides national screening materials, assessment support material, and standardized national tests. These assessment materials and tests serve various purposes: to inform decisions about support and adaptations of teaching, grading, and, at an aggregated level, to provide point estimates of student achievement at school or system level to support between-school equivalence in grading and for trend analysis [Skolverket (National Agency for Education), 2020].

The national standardized tests (NSTs) in years 3, 6, and 9 are mandatory. When grading students, the teachers are required to use all available information about students' knowledge and skills, with particular consideration of the results from the NSTs [Utbildningsdepartementet (Ministry of Education), 2010]. Thus, the SNAEd advises teachers to design and use different types of assessment situations for formative and summative purposes [Skolverket (National Agency for Education), 2022], and that teachers on all school levels choose, design and implement their own classroom assessment.

The research on Swedish teachers' conceptions of feedback is limited in both number and scope. However, by interviewing and surveying approximately 70 teachers and principals with different qualifications and experiences at seven schools in four of the largest cities in Sweden and Norway, Helgøy and Homme (2007) found that Swedish teachers to a higher extent than Norwegian teachers perceived NSTs as valuable tools in grading and in the improvement of their teaching. Moreover, unlike the Norwegian teachers, the Swedish teachers did not perceive NSTs as limiting their autonomy in the interpretation of national goals and how to organize the teaching to help students reach those goals.

Methods

This study used a self-administered, self-reported survey inventory administered with a forced-choice ordinal agreement response scale. A survey was used for several reasons: (a) human beliefs are not directly observable, meaning self-report is viable; (b) observations of teacher practices cannot be done anonymously nor easily surreptitiously, meaning accuracy and completeness cannot be guaranteed; and (c) a reliable measure of a teacher's feedback based on classroom observations would require many hours of lesson observation per teacher. Consequently, survey methodology was deemed to be appropriate and feasible methodology for the present study. Furthermore, a contribution of this study is to examine whether the inventory including both conceptions factors and a practice factor has validity (see the section Adaptation below). Analysis was done within the multiple indicators, multiple causes (MIMIC; Jöreskog and Goldberger, 1975) framework in which each survey item response is explained by a latent factor and a residual capturing the universe of unexplained variance and in which each latent construct is manifested by multiple indicators. The study uses confirmatory factor analysis and structural equation modeling to establish both the structure of responses and the relations of factors to each other. While the data are non-experimental, we consider that there is a causal path of influence from precedent conceptions of feedback to self-reported practices of feedback.

Participants

A total of 461 teachers working between years 1 and 9 in a northern city in Sweden responded to the survey. This is a 62% response rate from the municipality. Prior to analysis, data preparation involved identifying and removing from consideration participants who had more than 10% of responses missing per instrument. This sample was chosen because the research group and the municipality had previously decided to initiate a larger research project on improving assessment practices in these school years. This meant investigation of teacher conceptions was included as part of a multi-method, multi-study project. After deletion of 11 cases for high number of missing values, 450 teachers were retained (82% women, 17% men, 1% missing). Most participants had a teaching degree (78%), with 5% not having such a degree, and 17% not answering. Length of teaching experience was grouped by year ranges: 5% <2 years, 16% 2–5 years, 12% 6–10 years, and 66% >10 years. Distribution across the grade levels taught was almost equal (Years 1–3, $n = 149$; Years 4–6, $n = 156$, Years 7–9, $n = 141$).

Instruments

Teachers conceptions of feedback

The *Teachers Conceptions of Feedback* inventory (TCoF; Harris and Brown, 2008) probes nine different aspects of how teachers perceive or conceive of the nature and purpose of feedback. In response to Hattie and Timperley's (2007) description of feedback, factors were developed for the levels of feedback and the assumption that feedback exists to improve learning and the legitimate expectation that feedback will exist (Table 1). The Improvement factor focuses on students using the feedback they receive. Reporting and Compliance contain statements indicating the existence of feedback is expected by stakeholders (e.g., leaders and parents) and should inform parents about student progress. The Task factor focused on giving students information about aspects of their work that could be improved. The Process factor focused on allowing students to engage actively in responding to feedback. The Self-regulation factor included items about student autonomy and agency in evaluating their own work. The

TABLE 1 Sample items, number of items, and Teacher Conceptions of Feedback (TCoF) factor names.

TCoF Factor	No of items	Sample item
Encouragement	6	praise1. *The point of feedback is to make students feel good about themselves*
Ignore	4	ignore1. *Feedback is pointless because students ignore my comments and directions*
Improvement	4	imp1. *Students use the feedback I give them to improve their work*
Peer and self-feedback	3	pasa1. *Students are able to provide accurate and useful feedback to each other and themselves*
Process	5	process1. *I give students opportunities to respond to my feedback*
Reporting and compliance	3	acc1. *At my school, teachers are expected to give both spoken and written feedback to students*
Self-regulation	5	srl1. *I encourage students to correct/revise their own work without my prompting*
Task	3	task1.*My comments help students create the kind of work I expect from them*
Timeliness	5	time1. *I aim to deliver feedback to students within two days of receiving their work*

Encouragement factor included statements suggesting that providing students with praise would boost their self-esteem. Two additional factors drew on assessment *for* learning emphases of involving students in assessment and providing timely feedback. The Peer and Self-Feedback factor focused on students actively giving themselves and each other feedback. The Timeliness factor included items relating to the importance of prompt response to student work.

This inventory was developed in New Zealand with a statistically invariant measurement model for both primary and secondary teachers (Brown et al., 2012). Positive endorsement (i.e., mean score > 4.00 out of 6.00) was seen in both New Zealand primary and secondary groups for Improvement, Task, and Process factors, with substantial differences (i.e., Cohen's $d > 0.60$) in mean in favour of primary teachers for feedback is Required, Peer and Self-feedback, Process, and Timeliness. The survey related TCoF conceptions to self-reported practices of feedback that had been aggregated into four types. Consistent with the notion that beliefs predict behaviours, the Improvement factor had a positive loading on Teacher Formative feedback practices (i.e., giving detailed written comments, writing hints, tips, and reminders on work, discussing work with students, and giving spoken comments in class). The Encouragement conception of feedback predicted teachers' Protective-Evaluation feedback involving giving stickers, stamps, or smiley faces on student work and praising students for how hard they have worked. Emphasis on feedback Reporting and Compliance with expectations increased the prevalence of Parent Reporting practices (i.e., Parent-teacher conferences and reports to parents).

Adaptation

While the TCoF focused on teacher conceptions of the nature and purpose of feedback, a close reading of the inventory suggested that embedded within the TCoF, there were eight statements that described specific behaviours teachers might enact. Six of these practices were from the Hattie and Timperley (2007) process (Process1, 2, 5), self-regulation (SRL1, 3), and praise (Praise6) factors. The two other possibilities were from Timeliness (Time1, 4). To test the possibility that these items formed a Practices factor they were disaggregated from their original scale and aggregated into a new separate scale of Feedback Practices, which theoretically would be predicted by the remaining TCoF conceptions of feedback factors. Hence, a major contribution of this study is to examine whether this adaptation had validity. An advantage to this approach would be to minimise the number of items needed to elicit both beliefs about feedback and practices.

Survey design

The instruments were translated into Swedish by the authors, prioritising functional equivalence rather than literal equivalence. After translation, the functional equivalence of the items was validated by three external reviewers, who were fluent in both languages. Items were presented in jumbled order seen by the item number in Supplementary material. Participants responded using a positively packed, 6-point agreement scale. This type of scale has two negative options (Strongly Disagree and Moderately Disagree, scored 1 and 2 respectively) and four positive options (Slightly Agree, Moderately Agree, Mostly Agree and Strongly Agree, scored 3–6, respectively). This approach gives greater ability to discriminate the degree of positivity participants hold for positively valued statements and is appropriate when participants are likely to endorse statements (Lam and Klockars, 1982; Klockars and Yamagishi, 1988; Masino and Lam, 2014). Hence, in circumstances when participants are expected to respond positively to a stimulus (e.g., teachers responding to a policy expectation), giving them more choices in the positive part of the response continuum produces good results.

Analysis

After deleting participants with more than 10% missing responses, we imputed missing values with the expectation maximisation algorithm (Dempster et al., 1977). The imputation had a statistically not significant result ($\chi^2 = 1508.122$, $df = 1{,}453$, $p = 0.153$) for Little's (1988) Missing Completely at Random (MCAR) test showing the distribution of missing responses was random. Hence, all analyses were conducted without missing values. All but one variable met accepted standards for skew (<2.00) and kurtosis (<7.00), meaning variables were normal (Kim, 2013). Because item Irr4 had kurtosis =11.40, it was transformed using a Box-Cox transformation (Courtney and Chang, 2018) in the normalr ShinyApp.[1] This produced kurtosis = −0.40 and that version of the item was used in all analyses.

1 https://autopsych.shinyapps.io/normalr/

Coherent with the MIMIC framework, a two-step process (Anderson and Gerbing, 1988) of testing measurement models for each construct (i.e., conceptions of feedback and feedback practices) through confirmatory factor analysis (CFA) was implemented before testing a structural model that linked the beliefs to practices. Once a well-fitting model for each construct was found, a structural equation model (SEM) was tested in which beliefs about feedback were positioned as predictors of feedback practices on the assumption that beliefs are a predictor of behaviours (Ajzen, 1991).

We tested a correlated model of eight conceptions of feedback factors but this had poor fit. Consequently, we inspected modification indices to identify items that violated simple structure (Revelle and Rocklin, 1979) or independence of residual assumptions (Barker and Shaw, 2015), while aiming to retain the eight conceptions factors. Items with weak loadings on their intended factors (i.e., <0.30) were candidates for deletion. Items with strong modification indices (i.e., MI > 20) to other factors or whose residuals are strongly attracted to those of other items were also candidates for deletion (Bandalos and Finney, 2010). Although, most researchers expect three items per factor, it is possible in multi-factorial inventories to estimate factors that have only two items (Bollen, 1989).

Fit of both CFA and SEM models was established by inspection of multiple fit indices (Hu and Bentler, 1999). Because the chi-square measure of discrepancy between a model and its underlying data is sensitive to sample size and model complexity, we accepted not statistically significant values for the normed chi-square (i.e., χ^2/df) as support for a model (Wheaton et al., 1977). Further evidence for non-rejection of a model arises when the comparative fit index (CFI) is >0.90 and the root mean square error of approximation (RMSEA) is <0.08. However, both the CFI and RMSEA indices are sensitive to models with more than three factors, with the CFI entering reject space and the RMSEA entering not reject space under those conditions (Fan and Sivo, 2007). Thus, greater reliance is put on the gamma hat >0.90 and the standardized root mean residual (SRMR) <0.08 because these are more robust against sample size, model complexity, and model misspecification than the CFI or RMSEA indices. Scale reliability was estimated using the Coefficient H maximal reliability index, which is based on an optimally weighted composite using the standardised factor loadings (Hancock and Mueller, 2001). Factor mean scores were calculated by averaging the raw score for each item contributing to the factor, an appropriate method when simple structure (i.e., items belong to only one factor) is present (DiStefano et al., 2009).

CFA and SEM were conducted in the Jamovi Project (2022) platform using the lavaan package (Rosseel, 2012). Because six-point ordinal scales function similarly to continuous variables (Finney and DiStefano, 2006), maximum likelihood estimation was used. Model syntax is provided in Supplementary material.

Post-factor member checking

After analysis was completed, member checking (Tong et al., 2007) was conducted in two interview groups in the same city in which the survey had been administered. This was done to examine our interpretation of items that did not load as expected and the meaning of the unexpected paths. Ten teachers were recruited face-to-face on a volunteer, convenience basis to participate in two group interviews ($n = 4$ and 6, respectively) at separate occasions. In the interviews, the teachers were first given time to think individually about each issue and then shared their thinking in a joint discussion. The teachers had completed the survey, but their responses had been anonymous and so they were commenting on aggregate data results to which they had contributed. Groups were led by two of the authors. In these discussions, we explored how respondents had interpreted the items and how they understood the unexpected relations we had detected. The authors took field notes during the 1 h conversations and conversations were audio-recorded. The interviews were verbatim transcribed from which themes were identified. Aggregation of responses to the focus issues was carried out by authors 2 and 4 and manual coding of themes was carried out.

Results

Sweden feedback conceptions and practices model

An inter-correlated factor model had promising fit, but still below expectations. Because of the high correlation values merging of SRL, Task, Process, and Improvement items into a single factor of Improvement helped fit. Further modifications removed items that violated simple structure or were strongly correlated with other items, resulting in improved fit to the data. Paths that were not statistically significant were also removed. These modifications created six correlated conceptions of feedback and one practices of feedback factor.

The teacher conceptions of feedback factors were:

I. Feedback praises students (Praise),
II. Feedback improves student learning (Improvement),
III. Students ignore feedback (Ignore),
IV. Feedback is expected or required by school policy (Required),
V. Feedback is generated by involving peers and the self (PASA), and
VI. Feedback is prompt or timely (Timely).

The teacher Feedback Practices (Practices) factor consisted of three process items, two self-regulation of learning items, and one praise item. Together, these items create a set of formative practices that focus on giving students information and time to think about and improve their work, while taking responsibility for their own outcomes. Additionally, the feedback teachers provide includes commenting on the effort students put into their work, as well as noting how it can be improved.

Items, standardised loadings, and scale coefficient H statistics for all seven factors are given in Table 2.

The inter-correlation of conceptions of feedback factors is shown in Table 3. As expected, the Students Ignore factor had negative values to three other factors (i.e., Improvement, Required, and PASA) and non-significant values to Praise and Timely. In contrast, all other factors were moderately and positively inter-correlated with values ranging from $0.36 \leq r \leq 0.70$. This suggests commitment to feedback for improvement is simultaneously weakly related to feedback being required, using praise, feedback from peers and self, and being timely, while not being something that students ignore. This suggests teacher beliefs are generally adaptive and in line with feedback theory.

Factor means are shown in Table 4, with between factor effect sizes (Cohen's d; Cohen, 1992). Values $d > 0.60$ are considered large

TABLE 2 TCoF Factors and items: Sweden.

Factors & items	Order of presentation number & wording	Loading
Praise (H = 0.79)		
praise2	15. The goal in giving feedback is to protect and enhance the student's self-esteem	0.835
praise1	10. The point of feedback is to make students feel good about themselves	0.708
praise4	23. Feedback should be full of encouraging and positive comments	0.581
Improvement (H = 0.77)		
imp2	4. I can see progress in student work after I give feedback to students	0.752
imp1	1. Students use the feedback I give them to improve their work	0.642
imp4	26. Giving students feedback is important because it helps them learn	0.533
task1	6. My comments help students create the kind of work I expect from them	0.648
task2	16. My feedback is specific and tells students what to change in their work	0.432
Ignore (H = 0.65)		
irr1	9. Feedback is pointless because students ignore my comments and directions	0.755
irr2	12. Students rarely make changes in their work in response to my feedback	0.517
irr3	20. I seldom give written feedback because students throw it away	0.355
Required (H = 0.64)		
acc1	5. Parents can tell how well their child is learning from my feedback	0.790
acc3	32. Feedback practices at my school are monitored by school leaders	0.343
PASA (H = 0.55)		
pasa1	18. Students are able to provide accurate and useful feedback to each other and themselves	0.659
pasa2	24. Students can be critical of their own work and can find their own mistakes	0.507
pasa3	31. Peers are the best source of feedback	0.296
Timely (H = 0.60)		
time4	19. I give students feedback immediately after they finish	0.651
time3	11. Students should not have to wait for feedback	0.555
time5	35. Quality feedback happens interactively and immediately in the classroom while students are learning	0.504
Formative feedback practices (H = 0.74)		
srl3	17. My feedback reminds each student to self-assess his or her own work	0.743
process1	22. I give students opportunities to respond to my feedback	0.560
process5	36. I organise time in class for students to revise, evaluate, and give themselves feedback about their own individual work	0.537
process2	28. In feedback, I describe student work to stimulate discussion about how it could improve	0.494
srl1	2. I encourage students to correct/revise their own work without my prompting	0.422
praise6	34. My feedback includes comments on the effort students put into their work	0.410

in educational research (Hattie, 2009). In general, teachers gave strongest endorsement to the conception of feedback for improvement, with large effects compared to all other feedback factors and practices. Praise and Prompt feedback, with scores less than moderately agree, had large differences to both Expected and Students Ignore feedback factors. Both Students Involved in PASA and Expected, above slightly agree, had large differences only to Students Ignore feedback, which was close to mostly disagree. Note that low score indicated that on the average teachers rejected the notion that students ignore feedback. The Formative feedback practices, with a score just above moderately agree, was much larger than Students Involved in PASA, Expected, and Students Ignore feedback factors, had medium to small differences with Prompt and Praise factors, respectively, and was much smaller than the Feedback Improvement conception.

Structural paths were retained in the model only if they were statistically significant. The only statistically significant predictors of Practice were the Improvement (β = 1.24) and Ignored (β = 0.61) (Figure 1). This model had acceptable to good fit indices ($\chi^2 = 465.334$, $df = 258$, $\chi^2/df = 1.80$, $p = 0.18$; CFI = 0.90; gamma hat = 0.96; RMSEA = 0.042, 90%CI = 0.036–0.048; SRMR = 0.050) and so was not rejected. Note that the correlation between Improvement and Ignore was negative but their common positive loading on Practices is a non-transitive relationship, something possible given the relatively modest value ($r = -0.65$) (Kim and Mueller, 1978). This suggests that the implementation of the formative feedback practices depends both on seeing feedback as something students tend to ignore and thus may need to be overcome through these practices and something which will help them improve their learning. This conclusion was supported by the member checking process, in that all teachers expressed the view that feedback is an inherent

TABLE 3 TCoF factor inter-correlation matrix: Sweden.

TCoF Factor	I	II	III	IV	V	VI
I. Praise	–	0.39***	-0.11 *ns*	0.51***	0.36***	0.56***
II. Improvement		–	−0.65***	0.68***	0.70***	0.45***
III. Ignore			–	−0.33***	−0.33***	−0.12 *ns*
IV. Required				–	0.56***	0.55***
V. PASA					–	0.43***
VI. Timely						–

ns = not statistically significant; ***$p < 0.001$.

TABLE 4 TCoF factor means: Sweden.

			Mean score differences (Cohen's *d*)						
Feedback conceptions	*M*	*SD*	I	II	III	IV	V	VI	VII
I. Praise	3.91	1.11	–	**0.90**	0.14	0.48	**0.74**	**1.76**	−0.32
II. Improvement	4.69	0.63		–	**1.14**	**1.76**	**1.91**	**3.43**	**0.70**
III. Ignore	2.22	0.81			–	**1.01**	**1.57**	**1.70**	**−2.55**
IV. Required	3.13	1.00				–	0.34	**0.63**	**−1.23**
V. PASA	3.46	0.77					–	0.34	**−0.99**
VI. Timely	3.76	1.00						–	−0.51
Practices									
VII. Formative practices	4.21	0.75							–

Values in bold are effect sizes $d > 0.60$.

part of good teaching and something they are expected to provide. Moreover, in case students ignore feedback, the teachers saw this as an indication that they need to either improve the characteristics of their feedback itself or the learning situation in which the feedback is provided.

Discussion

This study surveyed teacher self-reported conceptions of feedback and related those conceptions to self-reported formative feedback practices. Confirmatory factor analysis of the Swedish primary and junior secondary school teachers partially recovered previously published results from New Zealand teachers (Brown et al., 2012). The factor structure of the TCoF model was to a large extent similar, albeit reduced to six factors and the removal of the practices items into a new separate scale. Moreover, in general, the teachers moderately to mostly agreed with feedback for improved teaching and learning, while rejecting the ideas that students ignored feedback. In addition, the Swedish teachers moderately agreed with the formative feedback practices, which is a construct not previously identified. It is important to remember that the conception that Students Ignore Feedback has a moderate negative correlation ($r = -0.65$) with the Improvement Feedback conception. This exposes a non-transitive relationship in how negatively correlated factors both have positive loadings on the same outcome ($\beta = 0.71$ and 1.39, respectively).

Unlike, the small-scale survey in Ethiopia (Dessie and Sewagegn, 2019), this study found that teacher beliefs about feedback did have statistically significant relationships to practices, a result reported elsewhere (Brown et al., 2012; Kyaruzi et al., 2018; Aslam and Khan, 2021). Like Kyaruzi et al. (2018), this study found that formative feedback practices were supported by improvement-oriented beliefs about the purpose of feedback. Previous studies (Brown et al., 2012; Aslam and Khan, 2021) found that feedback as Praise loaded onto Protective practices. This was not replicated, most likely because feedback practices were operationalised here as formative, improvement-oriented practices. A novel result was the supportive role of beliefs that students might ignore feedback had on formative feedback practices.

Teachers' conceptions of feedback are related to each other and to self-reported feedback practices. Most importantly, this study showed that formative feedback practices were increased only by two conceptions of feedback. Specifically, the belief that feedback should contribute to improved learning and the belief that students tend to ignore feedback explain substantial variation in practices ($R^2 = 0.94$). Together with the results from the member checking, this paper indicates that if teachers are concerned students might ignore formative feedback, they mitigate that concern by engaging in these formative feedback practices. Furthermore, if they want feedback to support improved learning they claim to use these feedback practices.

These results may be understood by the function of beliefs on teacher actions (Fives and Buehl, 2012). The theory of planned behaviour (TPB; Ajzen, 1991) identifies the importance of social norms, attitudes, and perceptions of behavioural control as essential predictors of intentions and actions. Teacher beliefs about the role of feedback are context bound by the policy and practice framework in which they are employed (Brown and Harris, 2009; Fulmer et al., 2015). Swedish teachers work within an education system that claims that all students can learn, and which provides a high degree of freedom for teachers to interpret and concretize the objectives of the national curriculum and selection of appropriate teaching methods and materials. The education system also supports teachers to use multiple data sources to monitor student learning and the stakes of the assessments are moderate or low for both teachers and

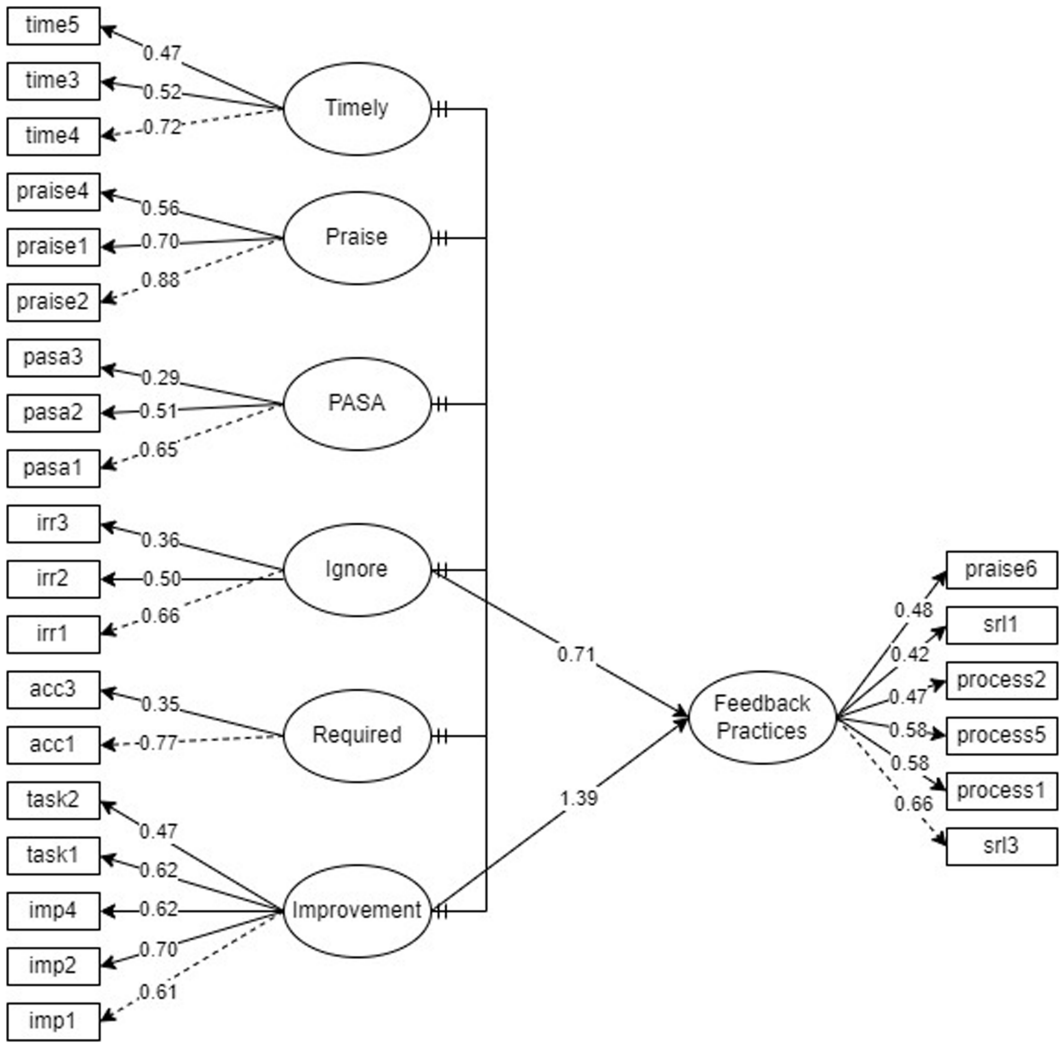


FIGURE 1
TCoF conceptions as predictors of self-reported feedback practices. Dashed line is seed value path; values are standardised loadings; residuals removed for simplicity; no correlated residuals; inter-correlation values in Table 2.

students. In such an educational system, feedback with the main purpose of learning and improvement may be seen as an integral part of the day-to-day teaching and assessment practice,

In contexts that de-emphasise consequences around achievement, there is opportunity to use error and failure productively for greater achievement and performance. However, in cultures and educational systems with high-stakes testing regimes and policies that use assessment mostly for demonstrating accountability and summative purposes, teacher perceptions of what feedback is and how it functions will be coloured by the first known effect of accountability (i.e., compliance with superiors; Lerner and Tetlock, 1999). Consequently, the teacher perceptions of feedback and their associations with formative feedback practices reported in this study may be similar only to those of other jurisdictions that have similar low-stakes frameworks for assessment (e.g., New Zealand).

The Swedish context avoids pronounced incentives for students to ignore feedback because there is little risk of looking bad if they fail to use it successfully. Public ranking of students, as is the case in more competitive environments, may inculcate a culture of ignoring information that induces shame. Nonetheless, student autonomy permits the possibility of choosing to ignore feedback that could be perceived as threatening to ego enhancement or well-being (Harris et al., 2018). Hence, it is reasonable for teachers to consider this possibility and act to minimise ego-protective reasons to disregard important information in feedback. Thus, it seems legitimate for teachers to signal that their feedback practices incorporate minimising the possibility that students would treat feedback maladaptively.

Implications for teacher education

Teachers generally agreed with the conception that the purpose of feedback is to enhance students' self-esteem and, hence, should be full of encouraging and positive comments (i.e., Praise). Although this goal is commendable, a focus on giving praise may hamper learning. Studies have shown that focussing on praise may come at the expense of identifying students' learning needs and suggestions on how to improve learning (Brown et al., 2012). Indeed, research has shown that praise commonly does not have a positive effect on students' achievement (Hattie and Timperley, 2007). These results are corroborated by our

model (Figure 1), where praise is not directly associated with the formative feedback practices that focus on the improvement of students work and self-regulation in learning. Hence, the goal of caring about the student and giving formative feedback may be in conflict. Since both goals are present in most curriculum statements, the solution is not to exclude either one, but instead to learn how to accomplish both. Thus, teacher education needs to address this tension and discuss how to circumvent it.

This could, for example, involve prospective teachers having to learn how to establish a classroom culture in which shortcomings (e.g., failure or not knowing) are seen as a natural part of learning and that attending to them is essential for learning. A prerequisite for the establishment of such a culture would be to find strategies to counteract the tendency of many students to link their school achievement with their self-esteem. Then, person-centered praise aiming at making students feel good about themselves could be replaced by positive comments aimed at linking positive outcomes to causes controllable by the student. Indeed, helping students to make adaptive attributions and to experience learning progress is associated with wellbeing (Winberg et al., 2014). Thus, teacher education must ensure that prospective teachers understand and enact caring for students by helping them to develop competence, rather than simply protect them from "bad" news.

Limitations of the study

We consider that the results presented here are likely to be typical of Swedish teachers rather than just the teachers participating in this study. While school administration is very localised, the policy and resource constraints exist equally for teachers elsewhere in the nation working in publicly funded primary and junior secondary schools. Nonetheless, a national survey would be needed to assure of generalisability claims made here. The relatively modest coefficient H values (i.e., all $H < 0.80$) suggest that the stability of these results is less than ideal. The stability of the factor measurement models needs to be tested in a further sample of Swedish teachers.

Of course, given the data are from a survey, we have made informed interpretations, corroborated by a small-scale member checking exercise, of what the factors mean and why the path values are what they are. Follow-up qualitative studies with teachers exploring their understanding of the results may provide further confidence in our explanations. Potentially, providing teachers with their own factor scores may provide further insights as to the meaning of teacher confidence in our findings.

Conclusion

This study contributes to our understanding of how teachers conceive of feedback. A significant contribution is the identification of a separate self-reported feedback practices scale within the TCoF inventory. The study shows clearly that teacher concerns to use feedback to improve learning and to minimise student tendencies to ignore feedback explain the formative feedback practices they implement.

Data availability statement

The datasets presented in this article are not readily available because, in accordance with Swedish law and GDPR, the datasets cannot be published until the data have been anonymized. Anonymization of the data can be done 10 years after the completion of the research project at the earliest. Requests to access the datasets should be directed to the corresponding author.

Ethics statement

The research took take place in accordance with the research ethics guidelines recommended by the Swedish Research Council (codex.vr.se). Informed consent was obtained from participating teachers. Before they were are asked for consent to participate in the study, they were informed about the project's intended implementation and purpose, that they could cancel their participation at any time, that data would be handled by the researchers only and that results would only be reported in such a way that individuals could not be identified. The data collected did not contain sensitive personal data and was processed in accordance with Umeå University's data management plan. This plan was developed to ensure that data is processed in accordance with the General Data Protection Regulation (GDPR). This kind of project with this kind of Swedish data does not require approval from the Swedish Ethics review board.

Author contributions

GB: conceptualisation, methodology, software, formal analysis, data curation, and writing – original draft. CA: validation, investigation, and writing – review and editing. MW: validation, investigation, writing – review and editing. TP: conceptualisation, validation, investigation, writing – review and editing, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by a grant from the Swedish Research Council (#2019-04349).

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1241998/full#supplementary-material

References

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice – a review and recommended 2-step approach. *Psychol. Bull.* 103, 411–423. doi: 10.1037/0033-2909.103.3.411

Aslam, R., and Khan, N. (2021). Secondary school teachers' knowledge and practices about constructive feedback: evidence from Karachi. *Pakistan. Cakrawala Pendidikan* 40, 532–543. doi: 10.21831/cp.v40i2.35190

Bandalos, D. L., and Finney, S. J. (2010). "Factor analysis: exploratory and confirmatory" in *The Reviewer's guide to quantitative methods in the social sciences*. eds. G. R. Hancock and R. O. Mueller (New York: Routledge), 93–114.

Barker, L. E., and Shaw, K. M. (2015). Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. *Am. J. Clin. Nutr.* 102, 533–539. doi: 10.3945/ajcn.115.113498

Barnes, N., Fives, H., and Dacey, C. M. (2015). "Teachers' beliefs about assessment" in *International handbook of research on teacher beliefs*. eds. H. Fives and M. Gregoire Gill (New York: Routledge), 284–300.

Bejar, I. I. (1984). Educational diagnostic assessment. *J. Educ. Meas.* 21, 175–189. doi: 10.1111/j.1745-3984.1984.tb00228.x

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons, Inc.

Bonner, S. M. (2016). "Teachers' perceptions about assessment: competing narratives" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge), 21–39.

Brown, G. T. L. (2023). Principles and assumptions of psychometric measurement. *Revista Digital de Investigación en Docencia Universitaria [Digital Journal of University Teaching Research]* 17:e1834. doi: 10.19083/ridu.2023.1834

Brown, G. T. L., and Harris, L. R. (2009). Unintended consequences of using tests to improve learning: how improvement-oriented resources heighten conceptions of assessment as school accountability. *J. Multidiscip. Eval.* 6, 68–91. doi: 10.56645/jmde.v6i12.236

Brown, G. T. L., Harris, L. R., and Harnett, J. (2012). Teacher beliefs about feedback within an assessment for learning environment: endorsement of improved learning over student well-being. *Teach. Teach. Educ.* 28, 968–978. doi: 10.1016/j.tate.2012.05.003

Brown, G. T., and Hattie, J. (2012). "The benefits of regular standardized assessment in childhood education: guiding improved instruction and learning" in *Contemporary educational debates in childhood education and development*. eds. S. Suggate and E. Reese (London: Routledge), 287–292.

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037//0033-2909.112.1.155

Courtney, M. G. R., and Chang, K. C. (2018). Dealing with non-normality: an introduction and step-by-step guide using R. *Teach. Stat.* 40, 51–59. doi: 10.1111/test.12154

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Ser. B* 39, 1–38. doi: 10.1111/j.2517-6161.1977.tb01600.x

Dessie, A. A., and Sewagegn, A. A. (2019). Moving beyond a sign of judgment: primary school teachers' perception and practice of feedback. *Int. J. Instr.* 12, 51–66. doi: 10.29333/iji.2019.1224a

DiStefano, C., Zhu, M., and Mîndrilă, D. (2009). Understanding and using factor scores: considerations for the applied researcher. *Pract. Assess. Res. Eval.* 14. doi: 10.7275/da8t-4g52

EVS/WVS. (2022). *European values study and world values survey: Joint EVS/WVS 2017–2022 dataset (joint EVS/WVS)*. GESIS Data Archive, Cologne. ZA7505. Dataset Version 4.0.0,. doi: 10.4232/1.14023

Fan, X., and Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivar. Behav. Res.* 42, 509–529. doi: 10.1080/00273170701382864

Finney, S. J., and DiStefano, C. (2006). "Non-normal and categorical data in structural equation modeling" in *Structural equation modeling: A second course*. eds. G. R. Hancock and R. D. Mueller (Greenwich, CT: Information Age Publishing), 269–314.

Fives, H., and Buehl, M. M. (2012). "Spring cleaning for the "messy" construct of teachers' beliefs: what are they? Which have been examined? What can they tell us?" in *APA Educational Psychology handbook: individual differences and cultural and contextual factors*. eds. K. R. Harris, S. Graham and T. Urdan, *Vol. 2* (Washington, DC: APA), 471–499.

Fulmer, G. W., Lee, I. C. H., and Tan, K. H. K. (2015). Multi-level model of contextual factors and teachers' assessment practices: an integrative review of research. *Assessment in Education: Principles, Policy & Practice* 22, 475–494. doi: 10.1080/0969594X.2015.1017445

Hancock, G. R., and Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, ToitS. Du and D. Sörbom (Eds.), *Structural equation Modeling: Present and future – a festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International Inc.

Harris, L. R., and Brown, G. T. (2008). *Teachers' conceptions of feedback inventory (TCoF)* [Measurement instrument]. Auckland, NZ: University of Auckland, Measuring Teachers' Assessment Practices (MTAP) Project. doi: 10.17608/k6.auckland.4800901.v1

Harris, L. R., Brown, G. T. L., and Dargusch, J. (2018). Not playing the game: student assessment resistance as a form of agency. *Aust. Educ. Res.* 45, 125–140. doi: 10.1007/s13384-018-0264-0

Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses in education*. London: Routledge.

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Helgøy, I., and Homme, A. (2007). Towards a new professionalism in school? A comparative study of teacher autonomy in Norway and Sweden. *European Educational Research Journal* 6, 232–249. doi: 10.2304/eerj.2007.6.3.232

Hofstede, G., Hofstede, G. J., and Minkov, M. (2010). *Cultures and organizations: Software of the mind* (*3. ed.*). New York: McGraw-Hill.

Hu, L.-T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Jamovi Project. (2022). jamovi. In (Version 2.3.21) [Computer Software]. Available at: https://www.jamovi.org

Jöreskog, K. G., and Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *J. Am. Stat. Assoc.* 70, 631–639. doi: 10.1080/01621459.1975.10482485

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics* 38, 52–54. doi: 10.5395/rde.2013.38.1.52

Kim, J.-O., and Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues* (*14*). Thousand Oaks, CA: Sage.

Klockars, A. J., and Yamagishi, M. (1988). The influence of labels and positions in rating scales. *J. Educ. Meas.* 25, 85–96. doi: 10.1111/j.1745-3984.1988.tb00294.x

Kyaruzi, F., Strijbos, J.-W., Ufer, S., and Brown, G. T. L. (2018). Teacher AfL perceptions and feedback practices in mathematics education among secondary schools in Tanzania. *Stud. Educ. Eval.* 59, 1–9. doi: 10.1016/j.stueduc.2018.01.004

Lai, M. K., and Schildkamp, K. (2016). "In-service teacher professional learning: use of assessment in data-based decision-making" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (New York: Routledge), 77–94.

Lam, T. C. M., and Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *J. Educ. Meas.* 19, 317–322. doi: 10.1111/j.1745-3984.1982.tb00137.x

Lerner, J. S., and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psycho. Bull.* 125, 255–275. doi: 10.1037/0033-2909.125.2.255ra

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* 83, 1198–1202. doi: 10.2307/2290157

Masino, C., and Lam, T. C. M. (2014). Choice of rating scale labels: implication for minimizing patient satisfaction response ceiling effect in telemedicine surveys. *TELEMEDICINE and e-HEALTH* 20, 1–6. doi: 10.1089/tmj.2013.0350

Revelle, W., and Rocklin, T. (1979). Very simple structure: an alternative procedure for estimating the optimal number of interpretable factors. *Multivar. Behav. Res.* 14, 403–414. doi: 10.1207/s15327906mbr1404_2

Rosseel, Y. (2012). Lavaan: an R package for structural equation Modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Rubie-Davies, C. M., Flint, A., and McDonald, L. G. (2012). Teacher beliefs, teacher characteristics, and school contextual factors: what are the relationships? *Br. J. Educ. Psychol.* 82, 270–288. doi: 10.1111/j.2044-8279.2011.02025.x

Skolverket (National Agency for Education). (2020). *Att planera, bedöma och ge återkoppling: stöd för undervisning [To plan, assess, and give feedback: Support for teaching]*. Stockholm: Skolverket [National Agency for Education].

Skolverket (National Agency for Education) (2022). "Betyg och prövning. Kommentarer till Skolverkets allmänna råd om betyg och prövning" in *Grades and examination. Comments to the National Agency for Education's general advice on grades and examination* (Stockholm: Skolverket [National Agency for Education])

Tong, A., Sainsbury, P., and Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int. J. Qual. Health Care* 19, 349–357. doi: 10.1093/intqhc/mzm042

Utbildningsdepartementet (Ministry of Education) [(2010): 800]. *Skollag [The Education Act]* Stockholm: Utbildningsdepartementet [Ministry of Education].

Wheaton, B., Muthén, B., Alwin, D. F., and Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociol. Methodol.* 8, 84–136. doi: 10.2307/270754

Winberg, T. M., Hellgren, J. M., and Palm, T. (2014). Stimulating positive emotional experiences in mathematics learning: influence of situational and personal factors. *Eur. J. Psychol. Educ.* 29, 673–691. doi: 10.1007/s10212-014-0220-y

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 26 September 2023
DOI 10.3389/feduc.2023.1239751

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Kerstin Kremer,
University of Giessen, Germany
Chia-Lin Tsai,
University of Northern Colorado, United States

*CORRESPONDENCE
Gavin Fulmer
✉ gavin-fulmer@uiowa.edu

RECEIVED 13 June 2023
ACCEPTED 08 September 2023
PUBLISHED 26 September 2023

CITATION
Ding C, Fulmer G, Hoffman L, Hand B and Suh JK (2023) The dimensionality of the epistemic orientation survey and longitudinal measurement invariance for the short form of EOS (EOS-SF).
Front. Educ. 8:1239751.
doi: 10.3389/feduc.2023.1239751

COPYRIGHT
© 2023 Ding, Fulmer, Hoffman, Hand and Suh. 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.

The dimensionality of the epistemic orientation survey and longitudinal measurement invariance for the short form of EOS (EOS-SF)

Chenchen Ding[1,2], Gavin Fulmer[2,3]*, Lesa Hoffman[4], Brian Hand[2] and Jee K. Suh[5]

[1]Department of Curriculum and Instruction, College of Education, Zhejiang Normal University, Jinhua City, Zhejiang, China, [2]Department of Teaching and Learning, College of Education, University of Iowa, Iowa City, IA, United States, [3]Northwest Evaluation Association (NWEA), Portland, OR, United States, [4]Department of Psychological and Quantitative Foundations, College of Education, University of Iowa, Iowa City, IA, United States, [5]Department of Curriculum and Instruction, College of Education, University of Alabama, Tuscaloosa, AL, United States

The dimensionality of the epistemic orientation survey (EOS) was examined across four occasions with item factor analysis (IFA). Because of an emphasis on the knowledge generation of epistemic orientation (EO), four factors were selected and built into a short form of EOS (EOS-SF) including knowledge generation, knowledge replication, epistemic nature of knowledge, and student ability. To track the stability of the factor structure for each factor of EOS-SF, longitudinal invariance models were conducted. Partial measurement invariance was obtained for each of the four factors of EOS-SF. This study provides an example of ongoing instrument development in the field of applied assessment research.

KEYWORDS

epistemic orientation, factor structure, longitudinal invariance, item factor analysis, knowledge generation

1. Introduction

Substantial work in the learning sciences and especially in science education emphasizes teachers eliciting students' prior knowledge and using this as a basis to support generative classrooms in which students' knowledge can grow and develop (Chin, 2007; National Research Council, 2012; NGSS Lead States, 2013; Chen and Techawitthayachinda, 2021). Shifts in classroom practice that adopt this approach represent a change in the role of the teacher from a fount of established knowledge to a resource for accessing and validating knowledge (Duschl and Bybee, 2014; Ash and Hand, 2022; Seung et al., 2023), which necessarily require support for teachers' changing practices, understandings, and beliefs (Hashweh, 1996; Fulmer, 2008; Maggioni and Parkinson, 2008; Desimone, 2009). One foundational element in this change is a teacher's epistemic orientations, which are beliefs about the epistemic nature of knowledge (Suh et al., 2022) that influence their planning and practice (Windschitl, 2002; Buehl and Fives, 2009). The epistemic nature of knowledge addresses the extent to which a teacher believes that knowledge is evolving, and that students' own thinking and abilities are malleable through learning experiences (Muis, 2007; Suh et al., 2022).

Recent work has shown that teachers' epistemic orientation plays a critical role in implementing reform-oriented classroom environments (Bae et al., 2022) by helping teachers move beyond the status quo to think more deeply about their instructional practice and their

classroom environment (Morandi et al., 2022; Lammert et al., 2022b) to embrace adaptiveness rather than routines (Suh et al., 2023). Attention to epistemic orientation has also explored how it can be developed over time through teacher professional development and ongoing support (Lammert et al., 2022a, 2023).

The substantial work in defining epistemic orientation and studying its role in classroom practice shows a successful and impactful line of study around the import of epistemic orientation. However, one area of concern is whether the measurement of epistemic orientation is itself stable across time and can be used to study teacher change over the medium to long terms. This is one example of the need for ongoing validation and interpretation of measurement instruments (Liu and Fulmer, 2008; Ding et al., 2023). To address this issue, the present study examines the factor structure of an existing epistemic orientation survey (EOS; Suh et al., 2022). The EOS includes four dimensions measuring teachers' orientation towards knowing, knowledge, teachers' instruction, and students' learning ability. Based on the focus of this study, we will select certain domains of epistemic orientation developed by Suh et al. (2022) and build a short form of the questionnaire of EOS.

2. Literature review

2.1. Defining epistemic orientation

Suh et al. (2022) defined epistemic orientation among teachers as beliefs about knowing and teaching. A focus on teachers' epistemic orientation towards constructivism is argued to help researchers explore how their classrooms allow students to engage in knowledge construction (Weiss et al., 2022). Epistemic orientation toward teaching for knowledge generation is defined as "a particular direction of thinking concerning how to deal with knowledge and knowledge generation processes when a teacher aims to create generative learning environments" (Suh et al., 2022, p. 1653) that is informed by beliefs while also drawing on preferences and tendencies that influence one's thinking and actions.

With the definition of epistemic orientation for teachers, Suh et al. (2022) developed an instrument to measure teachers' epistemic orientation, which has four domains: *epistemic alignment*, *classroom authority*, *epistemic nature of knowledge*, and *student ability*. Two domains, classroom authority and epistemic nature of knowledge, had similar beliefs of knowledge defined by Hofer and Pintrich (1997). Specifically, the epistemic nature of knowledge means that knowledge is continuously changing. Suh et al. (2022) defined the nature of knowledge as a set of beliefs about what knowledge is, which emphasizes that scientific knowledge is open to revision with new evidence. Moreover, the classroom authority domain was more aligned with the idea that knowledge could be challenged and created through critical thinking. Classroom authority is a set of beliefs about power relations between teachers and students, which emphasizes that students are active learners and have ownership of knowledge. Classroom authority also emphasizes that teachers should give students opportunities to develop ideas and construct their own knowledge.

In addition to the epistemic nature of knowledge and classroom authority, two key elements in epistemic orientation (Palma et al., 2018), Suh et al. (2022) also created another two domains for epistemic orientation: epistemic alignment and study ability. Suh et al. (2022) defined epistemic alignment as a set of beliefs about knowing, learning, and teaching, which emphasizes scientific investigation and knowledge justification by argumentation with evidence. Student ability is a set of beliefs about students' competence to learn, emphasizing that students can overcome learning challenges. Those two domains connect teaching and learning and go beyond the theories of knowledge.

2.2. Measuring epistemic orientation

Epistemic orientation is conjectured to comprise a continuum from replication to constructivism (Weiss et al., 2022). Teachers with an epistemic orientation towards constructivism tend to provide students with more opportunities to construct their knowledge and engage students in epistemic science practice (Bae et al., 2022). Also, such teachers will be aware that students are active agents in knowledge construction and have the power to shape knowledge production (Stroupe, 2014; Miller et al., 2018). Finally, they will facilitate a learning environment for students to construct new knowledge in science for meaningful learning. Teachers' epistemic orientation guides their decisions to create a learning environment for students and employ teaching approaches.

The original questionnaire by Suh et al. (2022), the Epistemic Orientation Survey (EOS), has four subscales to address the dimensions of epistemic orientation. The instrument consists of 44 Likert-type items in four subscales representing the four domains—epistemic alignment, authority relations, nature of knowledge, and student ability—with five response anchors: strongly disagree, disagree, unsure, agree, and strongly agree. The initial study provided evidence about the instrument's domain analysis and functionality.

Epistemic orientation is a trait that may change. Howard et al. (2000) developed constructivist approaches in a teacher training course. They found that three aspects of teachers' epistemology changed, and teachers tended to accept that students can: (1) examine complex knowledge and draw conclusions; (2) learn by discovering or doing rather than from textbooks and well-designed curricular material; and (3) develop their concepts through construction and clarifying misconceptions. Thus, they concluded that epistemology was a less stable trait than was previously supposed. More recently, Morandi et al. (2022) found that 2 years of teacher professional development showed changes in teachers' epistemic orientation toward knowledge generation, although not necessarily on all subscales with the greatest change in the area of epistemic alignment and student ability. They did not find significant differences after only 1 year of professional development. So, they argued that change in epistemic orientation may be slow and could be uneven. Taken together, these disparate findings emphasize the need for studies that can examine epistemic orientation over time while controlling for potential changes in the factor structure that could affect measures and interpretations. Our science education program conducted a two-year professional development and collected data with the survey developed by Suh et al. (2022), so it is possible to track the change in teachers' epistemic orientation.

2.3. Factor analysis and measurement invariance of instrument development

One way to study and manage potential changes in a factor structure is repeated analysis of the factor structure in different

samples or at different times, which is an ongoing process that provides further evidence for an instrument (Betts et al., 2010; Byrne and van de Vijver, 2010; Kaufman et al., 2016). Suh et al. (2022) developed the EOS and examined the factor structure of the instrument with the sample of pre-and in-service teachers at a single occasion. Examining the factor structure over time would provide a stronger basis for understanding the stability and validity of the construct. In this study, the factor structure analysis was repeatedly examined with in-service teachers on four occasions. This not only expands the investigation of the construct validity evidence of the EOS but also provides a broader perspective on the topic.

First, the present study examined whether the hypothesized structure of EOS was consistent across time. If the factor structure of epistemic orientation changes over time, then a new structure may be needed to track the changing nature of the construct. However, if the factor structure stays stable, this indicates that the instrument measures the same construct over time. Second, the study examined the invariance of the EOS over time, aiming to determine the extent to which the relationship between indicators and underlying factors for each subscale of EOS remains consistent across occasions (Millsap and Yun-Tein, 2004; Liu et al., 2017). For instance, if the indicator measurement properties were not invariant across time, then inferences drawn from scores on EOS may not relate to just factor-level change as intended. Failure to adequately comprehend temporal variation in the measurement model parameters could result in inaccurate interpretations of teachers' teaching orientation (Liu et al., 2017). Additionally, the variation in measurement also offers valuable insights for researchers about teachers' epistemic orientation. Moreover, if the instrument demonstrates a comparable factor structure across time *via* evidence for measurement invariance, it will be better suited to offering valuable insights for future studies. For example, that makes it more likely that an observed change could be due to intervention rather than shifts in the construct definition.

3. Research questions

The purpose of this paper is to investigate the factor structure of an instrument of epistemic orientation and investigate the measurement invariance of the instrument. There are three questions:

1. What is the dimensionality of the EOS at each occasion?
2. Based on the factor structure of the EOS, what would be an appropriate short form of the EOS (EOS-SF) for potential future use?
3. To what extent does the EOS-SF provide for measurement invariance across time for elementary science teachers? This means that each dimension of EOS-SF measures the same construct over time.

4. Method

This research investigated the EOS (Suh et al., 2022) with epistemic orientation as an overall latent variable with four hypothesized subfactors, using data from four waves of teacher data gathered during a longitudinal professional development project. All analyses were completed in Mplus Version 8.3 (Muthén and Muthén, 2021).

4.1. Data collection and participants

Respondents in the study were teacher participants in a two-year professional development (PD) workshop on the Science Writing Heuristic that focused on elementary teachers' orientation to generative learning in the teaching of science. In the workshops, teachers explored theories of learning and how to apply teaching approaches in science classrooms (Lammert et al., 2022a) by focusing on certain epistemic tools (Fulmer et al., 2021, 2023). All elementary teachers were from midwestern or southeastern U.S. districts, and they joined the program based on their interest in knowing more about learning theories and new teaching approaches. The Epistemic Orientation Survey (EOS) was distributed to teachers through email during the initial workshop and then approximately every six months afterward. Each survey response was gathered through the online survey platform Qualtrics.

The current study sample consisted of 123 elementary science teachers with years of teaching experience ranging from 1 to 32 years ($M = 14$, SD = 8) with complete data at baseline (Occasion 1). On occasion 2, there were 111 respondents with years of teaching experience ranging from 1 to 32 years ($M = 14$, SD = 9). On occasion 3, there were 123 respondents with years of teaching experience ranging from 1 to 32 years ($M = 14$, SD = 9). On occasion 4, there were 104 respondents with years of teaching experience ranging from 1 to 32 years ($M = 15$, SD = 9). Most teachers identified as women; three teachers identified as men participated in the workshop four times. The time interval between any two adjacent occasions is about 6 months. The sample size is adequate to estimate the IFA for each dimension per occasion (Singh and Masuku, 2014).

4.2. Measures

The Epistemic Orientation Survey (EOS; Suh et al., 2022) has been designed to measure teachers' orientation toward constructivist learning. The self-report measure consists of 44 items on a five-point Likert scale response format, and each scored on a scale of 0 to 4. Participants responded to the statement that best describes their understanding of learning in the workshops. Negatively-wording items were reverse-coded before conducting analyses, such that higher scores indicate higher epistemic orientation and lower scores indicate lower epistemic orientation. The EOS includes four subscales: Epistemic nature of knowledge (ENK), measured by eight items; Epistemic Alignment (EA), measured by 24 items; Classroom Authority (CA), measured by eight items; and Student Ability (SA), measured by four items.

4.3. Data analysis

Item factor analysis (IFA) was first conducted for the original EOS with 44 items. Then a short form of EOS (EOS-SF) was created based on this study's definition of epistemic orientation. Measurement

invariance was conducted for each factor of the EOS-SF after examining the stability of each factor.

4.3.1. Item factor analysis (IFA)

Item factor analysis (IFA) with a limited-information diagonally weighted least squares (WLSMV) estimator was used to examine each of the four hypothesized EOS scales separately (Liang and Yang, 2014). For each subscale, polychoric correlations were used to investigate the items' associations (Maydeu-Olivares, 2013). All measurement models in this study were scaled with a Z-score method, which sets factor means to 0 and factor variances to 1 and estimates all item loadings and thresholds (Ferrando and Lorenzo-Seva, 2018). We evaluated global model fit based on the comparative fit index (CFI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA). Good model fit is indicated by a CFI value ≥0.95, and SRMR and RMSEA values ≤0.06 (Brown, 2015). For nested models, the DIFFTEST function was used to identify the best model (Asparouhov et al., 2006; Muthén and Muthén, 2021).

Sources of local misfit were identified by examining the correlation matrix of residuals, which shows how far the model-predicted polychoric correlations were off from the polychoric correlations in the data and can be used to refine the factor structure (Shi et al., 2018; Bandalos, 2021). For the correlation matrix of residuals, the presence of relatively large positive residual correlations would indicate that these items were more related than was predicted by the latent factors. Relatively large negative residual correlations indicate that these items were less related than was predicted by the latent factors. Also, modification indices suggest which pair of items should be more related (Hill et al., 2007). Items' residual correlations in which the standardized expected parameter change index was greater than.3 were considered for inclusion, but decisions to add error covariance were made by combining local misfit information with a review of the item content. When three or more items have additional error covariance, this can be accounted for using a new latent factor (McNeish, 2017), that measures a different construct than the rest of items. Otherwise, given only pairs of two items, error covariances can be added to account for their additional relationship, as in the present study.

In IFA models, reliability is trait-specific and most often characterized by a quantity known as test information per factor. With the WLSMV estimator, test information was calculated for each factor. Reliability for the test information per factor can range from 0 to 1. Reliability for each factor was calculated by using the formula: reliability = information/ (information +1) (Milanzi et al., 2015).

4.3.2. Item selection for a short form of EOS (EOS-SF)

Based on the analysis of the hypothesized four-scale EOS factor structures, the study can identify items appropriate for a short form of EOS (EOS-SF). Creating an effective short form can make the survey focus more on the generative orientation of teaching while reducing the response burden of participants. Indicators that are suitable for inclusion in the short form must closely relate to the idea of knowledge generation and how teachers prepare the learning environment, and they must also have strong factor loadings contributing to the test information per factor. Factors that are not chosen include redundancy compared to existing factors or unrelated meaning to the defined epistemic orientation (Lorenzo-Seva and Ferrando, 2021).

4.3.3. Longitudinal measurement invariance

The longitudinal invariance of each subscale of EOS was examined within the IFA framework (Muthén and Asparouhov, 2002). Three steps were used for assessing measurement invariance: configural invariance, metric invariance, and scalar invariance (Byrne and van de Vijver, 2010). At each step of the analysis, the estimated model has additional restrictions imposed on the measurement model. Each less-restricted model is then compared with the comparable model with the more-restrictive model with the DIFFTEST function. Invariance is met if the fit of the more restrictive model was not significantly worse than the less restricted model (Widaman et al., 2010). Examples of Mplus input scripts for measurement invariance are available in supplementary material.

First, *configural invariance* allows all measurement model parameters (loadings, thresholds) to be estimated at each of the four occasions. This is a baseline model for further comparison with the following measurement invariance models. Factor means and factor variances for all occasions were constrained as 0 and 1, respectively, and all item residual variances across occasions were initially fixed to 1. Factor covariances and same-item residual covariances were estimated.

Second, *metric invariance*, also called weak factorial invariance (Hirschfeld and Von Brachel, 2014), evaluates the equality of factor loadings across time (Milfont and Fischer, 2010). To do so, all factor variances at the reference occasion were fixed to 1 and were freely estimated at other occasions. Factor means for all occasions were fixed to 0. The first metric model was a full metric model with equal factor ladings for the same item across time. Modification indices suggested whose loadings tend to differ across occasions, based on which one loading at a time was released. The model was then re-estimated and compared with the configural model. The procedure was repeated until the metric invariance model was not worse than the configural invariance model. At this step, at least a partial metric invariance model was achieved.

Third, *scalar invariance*, also called strong factorial invariance (McGrath, 2015), evaluates the equality of thresholds across time (Milfont and Fischer, 2010; McGrath, 2015). For the first occasion, the factor mean was fixed to 0, and the factor variance was fixed to 1. Factor means, and factor variances at other occasions were freely estimated. For measurement invariance at this stage, the thresholds of items that failed to have equal loadings in the metric invariance model were freely estimated in the scalar model. Other thresholds were then freed based on modification indices. Models for scalar invariance were compared with the last model in metric invariance. At each re-examination, only one item's threshold was freely estimated. The procedure was repeated until the scalar invariance model was not worse than the metric invariance. The final resulting model offers a set of factors that have stable structure over time and measurement invariance over time. The supported scalar invariance model indicates that the survey measures the same construct overtime and teachers' epistemic orientations across time are comparable (Bollen, 1989; Byrne and van de Vijver, 2010).

5. Results

5.1. Item factor analysis (IFA)

To answer research question 1, the factor structure of the EOS was examined with item factor analysis. Although each of the four

dimensions was designed to be unidimensional (Suh et al., 2022), multiple factors were found for two dimensions of EOS on each of the four occasions. As we outline below, two dimensions (epistemic alignment, EA; epistemic nature of knowledge, ENK) had different factor structures across the four occasions, and two dimensions (student ability, SA; classroom authority, CA) had consistent factor structures across the four occasions. Table 1 summarizes the results for the factor structures; the detailed results for EA, ENK, SA, and CA follow.

The EA dimension broke into three factors across four occasions: attitudes towards correct answers (AtCA), knowledge generation (KG), and knowledge replication (KR). The three-factor models had acceptable model fit (Table 2). The test reliability values for AtCA fluctuated over time. Specifically, the reliability values were more informative (0.70 ~ 0.90) on occasions 2 and 4 for people with traits from −3.0 to 2.0 SD. Meanwhile, reliabilities on occasions 1 and 3 were lower (0.65 ~ 0.80) than on occasions 2 and 4. The test reliabilities for KG were consistently informative (0.80 ~ 0.95) for people with lower traits from −4.4 to 1.6 SD on occasions 2, 3, and 4. The reliability of KG at occasion 1 was lower (0.70 ~ 0.80) for people with traits from −4.0 to 3.6 SD compared to the other three occasions. The test reliabilities for KR were informative (0.80 ~ 0.92) for people with lower traits from −4.0 to 1.2 SD on occasions 2, 3, and 4. The reliability of KR at occasion 1 was lower (0.75 ~ 0.85) for people with traits from −4.2 to 2.2 SD. This means that the hypothesized one-factor EA includes three unidimensional factors, and each factor measures one construct.

The ENK indicators had two-factor models on occasions 1 and 2 and had one-factor models on occasions 3 and 4. Models for each occasion had good model fit (Table 2). Two factors of ENK at occasions 1 and 2 had test reliability (0.70 ~ 0.92) for people with lower medium level ability from −4.0 to 1.2 SD. On occasions 3 and 4, the ENK had higher test reliabilities (0.75 ~ 0.90) for people with lower and medium level ability from −3.8 to 1.8 SD. This means that the factor structure of ENK changes over time.

The SA indicators had a one-factor structure across time. The CFIs were good, but the RMSEAs were unacceptable (Table 2). The test reliabilities for SA across time were stable and high (0.70 ~ 0.93) to measure students with lower traits from −3.2 to 0.3 SD but lacked information at higher trait levels. This means that the hypothesized one-factor SA has a stable one-factor structure over time.

TABLE 1 Summary of factor structures for four dimensions across four occasions.

	No. factors (No. Error covariances)			
Domain	Occasion 1	Occasion 2	Occasion 3	Occasion 4
EA (24 items)	3 (2)	3 (3)	3 (4)	3 (3)
ENK (8 items)	2 (2)	2	1 (1)	1 (1)
SA (4 items)	1	1	1	1
CA (7 items)	1	1 (2)	1 (2)	1

EA, Epistemic alignment; ENK, Epistemic nature of knowledge; SA, Student ability; CA, classroom authority.

For the CA dimension, item I23, "Teachers are responsible for managing classroom environments," was unclear about what kind of classroom environment teachers should create for students and had low polychoric correlations. This item was deleted for further analysis. The rest of the seven items had one-factor models across four occasions with a good model fit (Table 2). Reliabilities for CA across time were stable and high (0.75 ~ 0.90), measuring people with lower or medium traits from −4.0 to 1.8 SD. This means that the hypothesized one-factor CA has a stable one-factor structure over time.

These results indicate that the number of factors for the four proposed dimensions of the original EOS is not equally stable across measurement occasions. This is not only a matter of the number of items in the proposed dimensions but also of the structure. Therefore, attempting to create a short form of the EOS, as described in the next step, would require analyzing and building off the stable factor structure to select dimensions that are consistent with the theoretical definition of epistemic orientation.

5.2. Short form of EOS (EOS-SF)

To solve research question 2, a short form of EOS was created with the results from the factor structure of EOS. As shown, the hypothesized four-factor structure of EOS can be represented by six stable factors across time: knowledge generation (KG), knowledge replication (KR), attitudes towards correct answers (AtCA), classroom authority (CA), epistemic nature of knowledge (ENK), and student ability (SA). Not all of these empirical factors are fully consistent with the theoretical definition of epistemic orientation that this study focused on. In this study, we define epistemic orientation as a continuum from knowledge replication to construction. We chose the orientation towards knowing and knowledge and did not examine the orientation towards teaching and learning defined by Suh et al. (2022).

Teachers with a more informed orientation towards knowledge construction give students control over their learning because they know that scientific knowledge is constructed by scientists and develops with the efforts of the scientific community. Based on this definition of epistemic orientation, four factors were chosen to form a short form of EOS (EOS-SF): ENK, which relates to scientific knowledge's constructive and evolving nature; KG and KR, which relate to the two ends of the orientation continuum; and CA, which relates to student and teacher control of learning in the classroom. Two factors were not chosen: AtCA, which shows redundancy with KR, but was less reliable across measurement occasions, and SA, which measured beliefs about students' own competence but did not closely relate to the core idea of stance toward knowledge in the classroom.

Taking these four factors, a shortened version of the EOS was created that consisted of 35 items, and together named the short form of epistemic orientation survey (EOS-SF; Appendix A) to measure teachers' epistemic orientation. The factors are classroom authority (CA) with seven indicators, epistemic nature of knowledge (ENK) with eight indicators, knowledge generation (KG) with twelve indicators, and knowledge replication (KR) with eight indicators. The four factors together focus on teachers' beliefs in knowing and knowledge: more toward generative or replicative

TABLE 2 Model fit for each dimension of the EOS per occasion.

	Occasion 1		Occasion 2		Occasion 3		Occasion 4	
	CFI	RMSEA	CFI	RMSEA	CFI	RMSEA	CFI	RMSEA
EA	0.92	0.07	0.94	0.08	0.95	0.07	0.95	0.08
ENK	0.94	0.06	0.95	0.12	0.98	0.08	0.99	0.04
SA	0.98	0.25	0.95	0.25	1.00	0.00	0.98	0.19
CA	0.94	0.10	0.99	0.07	0.98	0.09	0.99	0.06

EA, Epistemic alignment; ENK, Epistemic nature of knowledge; SA, Student ability; CA, classroom authority.

belief. Then, these four factors were analyzed for measurement invariance over time.

5.3. Measurement invariance of each factor in EOS-SF

To solve research question 3, the measurement invariance of each factor of EOS-SF was investigated. The factor structure over time for the EOS-SF can be tested for measurement invariance to examine to what extent the measurement is effective, even as respondents change in their levels of the underlying attribute over time. Two sets of structural models were tested for the measurement invariance of the four factors. The first structural model was a consistent one-factor model over time for three of the factors (CA, KG, and KR). The second structural model was for ENK, which had a two-factor structure on occasions 1 and 2 but a one-factor structure on occasions 3 and 4 (Table 1).

The first type of structural model for longitudinal measurement invariance was a one-factor model where each factor's latent variables across time were correlated, such as for CA, KG, and KR. In addition, the responses for each item across four occasions were also dependent, so each item's residuals across four occasions were correlated. The reference occasion's factor variance and factor mean were fixed to 1 and 0, respectively. Detailed tables of model comparisons are available in supplementary material, and summarized below.

The subdomain classroom authority (CA) was measured by seven ordinal indicators (items 03, 08, 27, 13, 18, 31, 35). By conducting configural invariance, metric invariance, and scalar invariance testing, it was found that a partial scalar invariance model (model 3b in Table 3) is the best-fitting invariance model. For model 3b, factor loadings of five items (08, 13, 18, 31, 35) were kept invariant across time. Also, thresholds of four items (13, 18, 31, 35) were kept invariant across time. With four out of seven items having the same structure with the latent factor over time, seven items measured something in common. The invariant items served as anchors for the relations between items and latent factors at each occasion to the same scale, which made the subdomain comparable across occasions (Byrne and van de Vijver, 2010).

The subdomain knowledge generation (KG) was measured by 12 ordinal indicators (items 05, 10, 12, 17, 21, 25, 32, 36, 37, 39, 41, 43). By conducting configural invariance, metric invariance, and scalar invariance testing, it was found that the partial scalar model (model 3b in Table 4) is the best-fitting model for invariance. For model 3b, factor loadings and thresholds of two items (5, 43) were kept invariant across time.

Knowledge replication (KR) was measured by eight ordinal indicators (items 02, 07, 09, 22, 26, 29, 33, 40). By conducting configural invariance, metric invariance, and scalar invariance testing, it was found that the partial scalar model (model 3b in Table 5) was the final model. For model 3b, factor loadings of three items (22, 26, 40) kept invariant across time. Also, the thresholds of one item (26) were invariant across time.

The second type of structural model for measurement invariance is for ENK because this factor had different factor structures across occasions. On occasions 1 and 2, ENK had the same two-factor structure: one factor included four items (1, 6, 14, 19), and the other included four indicators (11, 16, 24, 28). However, the subdomain ENK had a one-factor structure on occasions 3 and 4. To conduct measurement invariance, the factor structure of ENK was kept the same across time. Therefore, eight items of ENK had a one-factor structure with four items (1, 6, 14, 19) correlated on each occasion to approximate the second factor as needed. In addition, the residuals for the same items across time were also correlated.

The subdomain epistemic nature of knowledge (ENK) was measured by eight ordinal indicators (items 01, 06, 14, 19, 11, 16, 24, 28). By conducting configural invariance, metric invariance, and scalar invariance testing, it was found that the partial scalar model (model 3a in Table 6) was the last invariance model. For model 3a, factor loadings and thresholds of 6 items (01, 14, 19, 11, 16, 28) were kept invariant across time.

In sum, this study first used item factor analysis (IFA) to examine the four dimensions of EOS developed by Suh et al. (2022). It was found that epistemic alignment (EA) had three factors across time: knowledge generation (KG), knowledge replication (KR), and attitudes toward correct answers (AtCA). The dimension of student ability (SA) and classroom authority (CA) had a one-factor structure over time. The dimension epistemic nature of knowledge (ENK) had a two-factor structure on occasions 1 and 2 and a one-factor structure on occasions 3 and 4. Based on the definition of epistemic orientation, four factors were chosen to shape a short form of EOS (EOS-SF), including KG, KR, ENK, and *CA*. Longitudinal invariance testing was conducted for each of the four factors of the EOS-SF.

6. Discussion

The present study examined the dimensionality of the Epistemic Orientation Survey (EOS) across four occasions. We found that the hypothesized four dimensions included six factors, which described different aspects of epistemic orientation. Based on our focus on generative learning as the target of teachers' epistemic orientation, four factors were selected to be included in a short form version,

TABLE 3 Model fit and model comparison for longitudinal invariance of *CA*.

		DIFFTEST							
Model name	Freed item constraint	H1 Model	χ^2 (df)	*p*	χ^2 (df)	*p*	CFI	RMSEA	SRMR
M1. Configural		N/A	N/A	N/A	384.369 (302)	0.001	0.964	0.043	0.090
M2a. Metric		M1	24.441 (18)	0.1411	397.977 (320)	0.002	0.966	0.053	0.097
M2b. Metric	E27T2	M1	16.040 (17)	0.521	385.155 (319)	0.007	0.971	0.050	0.095
M2c. Metric	E03T2	M1	11.849 (16)	0.754	379.425 (318)	0.010	0.973	0.049	0.094
M3a. Scalar		M2c	80.468 (49)	0.0031	442.760 (367)	0.004	0.967	0.049	0.096
M3b. Scalar	E08T1$2	M2c	56.026 (48)	0.199	430.808 (366)	0.011	0.972	0.047	0.096

CA, Classroom authority. M2c is the last metric invariance model. M3b is the last scalar invariance model.

TABLE 4 Model fit and model comparison for longitudinal invariance of KG.

		DIFFTEST							
Model name	Freed item constraint	H1 Model	χ^2 (df)	*p*	χ^2 (df)	*p*	CFI	RMSEA	SRMR
M1. Configural		N/A	N/A	N/A	1256.489 (1002)	< 0.0001	0.961	0.041	0.095
M2a. Metric		M1	85.745 (33)	< 0.0001	1353.100 (1035)	< 0.0001	0.952	0.046	0.112
M2b. Metric	E10T1	M1	71.761 (32)	< 0.0001	1324.231 (1034)	< 0.0001	0.956	0.044	0.110
M2c. Metric	E32T2	M1	64.268 (31)	< 0.0001	1310.173 (1033)	< 0.0001	0.958	0.043	0.108
M2d. Metric	E36T1	M1	56.512 (30)	0.002	1295.884 (1032)	< 0.0001	0.960	0.042	0.107
M2e. Metric	E12T4	M1	49.595 (29)	0.010	1283.196 (1031)	< 0.0001	0.962	0.041	0.106
M2f. Metric	E39T1	M1	45.847 (28)	0.018	1277.759 (1030)	< 0.0001	0.962	0.040	0.105
M2g. Metric	E25T3	M1	40.013 (27)	0.051	1267.801 (1029)	< 0.0001	0.964	0.040	0.103
M2h. Metric	E41T1	M1	34.265 (26)	0.129	1257.965 (1028)	< 0.0001	0.965	0.039	0.102
M2i. Metric	E37T1	M1	30.255 (25)	0.215	1252.318 (1027)	< 0.0001	0.966	0.039	0.101
M2j. Metric	E21T1	M1	25.203 (24)	0.395	1244.347 (1026)	< 0.0001	0.967	0.038	0.100
M2k. Metric	E25T4	M1	20.163 (23)	0.632	1235.848 (1025)	< 0.0001	0.968	0.037	0.098
M2l. Metric	E36T3	M1	15.641 (22)	0.833	1228.082 (1024)	< 0.0001	0.969	0.037	0.097
M2m. Metric	E17T4	M1	12.910 (21)	0.912	1224.072 (1023)	< 0.0001	0.970	0.036	0.097
M3a. Scalar		M2m	71.680 (53)	0.045	1281.550 (1076)	< 0.0001	0.969	0.036	0.098
M3b. Scalar	E12T2$3	M2m	62.802 (52)	0.145	1278.035 (1075)	< 0.0001	0.969	0.036	0.098

KG, Knowledge generation. M2m is the last metric invariance model. M3b is the last scalar invariance model.

EOS-SF. We found that there was evidence for longitudinal measurement invariance of each factor in EOS-SF, suggesting that the subscales of the short form survey could capture the change in epistemic orientation over time.

Researchers can choose certain subscales or the whole survey depending on their research purposes. The six factors of the original EOS could be regarded as subscales measuring different aspects of the epistemic orientation. Taking the present study as an example, the short form survey was created by choosing four of the six factors from the EOS originally developed by Suh et al. (2022) to match our focus on the learning orientation to generate knowledge and the evolving nature of science knowledge. Therefore, the attitude toward correct answers (AtCA) was not chosen because its content appeared redundant for the knowledge replication (KR) factor, and student ability (SA) was not chosen because it is more about a mindset toward students' intelligence. The short form proposed in this current study is consistent with the stance that all students can generate their own knowledge of science regardless of their perceived intelligence level.

One finding from the study of the factor structure of time is that the factor structure of evolving nature of knowledge (ENK) showed changes over time. On the first two occasions, ENK included two sub-factors: revisable knowledge and absolute knowledge. On the last two occasions, ENK had one factor: the evolving nature of knowledge, which indicates there was no distinction between revisable and absolute knowledge. The changes of factor structure of the ENK factor might indicate that teachers gradually adopt a more informed orientation toward knowledge generation (Suh et al., 2022) after 1 year of participation in the PD workshops—thereby developing a closer connection between the aspects of ENK over time.

A second finding is that the partial measurement invariance was obtained for each of the four factors of EOS-SF. We found that CA had four out of seven items that kept strong invariance. KG had two out of twelve items with strong invariance, which was barely

TABLE 5 Model fit and model comparison for longitudinal invariance of KR.

			DIFFTEST						
Model name	Freed item constraint	H1 Model	χ^2 (df)	*p*	χ^2 (df)	*p*	CFI	RMSEA	SRMR
M1. Configural		N/A	N/A	N/A	559.727 (410)	< 0.0001	0.958	0.050	0.086
M2a. Metric		M1	54.573 (21)	< 0.0001	611.976 (431)	< 0.0001	0.949	0.053	0.104
M2b. Metric	E07RT2	M1	41.651 (20)	0.003	586.687 (430)	< 0.0001	0.956	0.050	0.100
M2c. Metric	E02RT1	M1	36.516 (19)	0.009	578.473 (429)	< 0.0001	0.958	0.049	0.098
M2d. Metric	E29RT1	M1	30.858 (18)	0.030	570.084 (428)	< 0.0001	0.960	0.047	0.096
M2e. Metric	E33RT3	M1	25.460 (17)	0.085	561.542 (427)	< 0.0001	0.962	0.046	0.094
M2f. Metric	E09RT1	M1	19.805 (16)	0.229	553.500 (426)	< 0.0001	0.964	0.045	0.093
M2g. Metric	E02RT2	M1	14.705 (15)	0.473	544.460 (425)	< 0.0001	0.966	0.044	0.091
M3a. Scalar		M2g	74.009 (50)	0.015	603.050 (475)	< 0.0001	0.964	0.043	0.092
M3b. Scalar	E40RT3$2	M2g	67.377 (49)	0.042	599.720 (474)	< 0.0001	0.964	0.042	0.092
M3c. Scalar	E22RT1$3	M2g	61.162 (48)	0.096	596.612 (473)	< 0.0001	0.965	0.042	0.092

KR. Knowledge replication. M2g is the last metric invariance model. M3c is the last scalar invariance model.

TABLE 6 Model fit and model comparison for longitudinal invariance of ENK.

			DIFFTEST						
Model name	Freed item constraint	H1 Model	χ^2 (df)	*p*	χ^2 (df)	*p*	CFI	RMSEA	SRMR
M1. Configural		N/A	N/A	N/A	488.036 (386)	< 0.0001	0.965	0.042	0.087
M2a. Metric		M1	23.857 (21)	0.300	492.561 (407)	< 0.0001	0.970	0.038	0.094
M2b. Metric	E24RT3	M1	19.515 (20)	0.489	486.431 (406)	< 0.0001	0.972	0.037	0.093
M2c. Metric	E06T2	M1	14.104 (19)	0.778	477.338 (405)	< 0.0001	0.975	0.035	0.092
M3a. Scalar		M2c	50.200 (51)	0.505	527.243 (456)	< 0.0001	0.975	0.032	0.093

ENK, Epistemic nature of knowledge. M2c is the last metric invariance model. M3a is the last scalar invariance model.

invariant. KR had three out of eight items that had strong invariance, and ENK had six out of eight items that kept strong invariance. The partial invariance is an important empirical finding that supports an underlying assumption of many studies that aim to track change in epistemic orientation over time. The finding of partial measurement invariance, coupled with tools for estimating factor scores that account for measurement non-invariance, provide a strong case for using the questionnaire across repeated occasions to study growth. This is because that the invariant items put items on different occasions on the same scale, so we can compare the change of the latent trait over time. That supports other research that has aimed to study changes in epistemic orientation across time (e.g., Bae et al., 2022), and how this may connect to experiences of professional learning or to enacted instructional practice (e.g., Lammert et al., 2022b; Morandi et al., 2022). We suggest that subsequent work is needed to use invariant factor scores as part of their analyses of teacher growth of epistemic orientation as an essential next step.

Thirdly, the methods used in this study could be used as an example for the measurement of how constructs change over time. Factor structure and measurement invariance were both used to examine the stability of the factors. Usually, researchers choose one method to examine the dimensionality of an instrument (Lei et al., 2020). When we use the instrument multiple times, we can gain a comprehensive understanding of the dimensionality and stability of the instrument. In addition, the measurement invariance provides opportunities for researchers to compare the change of latent traits after several training of a specific teaching theory or instructional methods.

Fourthly, the EOS-SF measuring epistemic orientation may be applied to teachers in other grade levels, even though the EOS-SF was developed with data from elementary science teachers. With epistemic orientation, teachers tend to have a high possibility to create a generative learning environment for students to construct their understanding of science concepts (Bae et al., 2022; Weiss et al., 2022), which is also important for middle and high school students. Moreover, the EOS-SF would be a useful tool for PD developers and implementers to track teachers' levels of epistemic orientation at the beginning of a PD experience and across time, as well as use this information to tailor the learning opportunities to teachers' different levels of epistemic orientation. Teachers with more informed levels of epistemic orientation may need different types of support and experiences during PD than teachers with epistemic orientations less supportive of knowledge generation.

Lastly, this work is part of an emerging tradition in the field to engage in applied assessment research using measurement principles as an ongoing area of inquiry. This differs from work that would release a "finalized" instrument or questionnaire. Rather, the field

shows a keen interest in the persistent study of the application and interpretation of tools (Liu and Fulmer, 2008; Ding et al., 2023; Fulmer et al., 2023), whether for the improvement of that specific tool or to inspire new patterns of assessment development and use (Ruiz-Primo and Shavelson, 1996; Mislevy, 2018; Harris et al., 2022). In the present case, this enables us to gain a richer interpretation of the instruments and understanding of epistemic orientation. It also provides a stronger basis for future work that can study changes of epistemic orientation over time. Thus, the field's effort in the reexamination of instruments can provide additional evidence for the quality and effectiveness of any existing assessment.

7. Limitation

In this study, error covariances in CFA and IFA were added to the factor structure. Usually, the common variance among items is explained by the factor loadings. Theoretically, the residuals of items are uncorrelated (Barker and Shaw, 2015) because the latent trait is the only reason why they relate to each other. However, some researchers suggest adding error covariances when two items have something else in common and cannot be explained by the latent trait (Cattell and Tsujioka, 1964). The residual covariances among items partition the measurement noise covariance for a better fit of the latent trait (Deng et al., 2019). Researchers can choose to add a new factor for three or more overly-related items with reasonable choices (McNeish, 2017). Take this study as an example, we added error covariances for three factors in epistemic alignment construct (See supplementary material). Researchers could decide the factor structure based on the residual covariance and statements of items.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.openicpsr.org/openicpsr/project/184283.

Ethics statement

The studies involving humans were approved by Institutional Review Boards (IRBs). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

CD conducted data analysis and drafted the paper. GF assisted in revising the writing. LH provided help with the data analysis when CD had questions about data analysis. GF, BH, and JS, who are co-PIs for the funded grant, designed the project, recruited teachers, conducted PD workshops, and collected data used in this research. All authors contributed to the article and approved the submitted version.

Funding

The project was funded by the National Science Foundation (Grant Agreement Number: DRL-1812576). This grant is about developing the adaptive expertise of elementary science teachers.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1239751/full#supplementary-material

References

Ash, A. M., and Hand, B. (2022). Knowledge evaluation and disciplinary access: mutually supportive for learning. *Learn. Res. Pract.* 8, 116–132. doi: 10.1080/23735082.2022.2092645

Asparouhov, T., Muthén, B., and Muthén, B. O. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. *Matrix* 1, 1–6. Available at: http://www.statmodel.com/download/webnotes/webnote10.pdf

Bae, Y., Hand, B. M., and Fulmer, G. W. (2022). A generative professional development program for the development of science teacher epistemic orientations and teaching practices. *Instr. Sci.* 50, 143–167. doi: 10.1007/s11251-021-09569-y

Bandalos, D. L. (2021). Item meaning and order as causes of correlated residuals in confirmatory factor analysis. *Struct. Equ. Model. Multidiscip. J.* 28, 903–913. doi: 10.1080/10705511.2021.1916395

Barker, L. E., and Shaw, K. M. (2015). Best (but oft-forgotten) practices: checking assumptions concerning regression residuals. *Am. J. Clin. Nutr.* 102, 533–539. doi: 10.3945/ajcn.115.113498

Betts, J. E., Appleton, J. J., Reschly, A. L., Christenson, S. L., and Huebner, E. S. (2010). A study of the factorial invariance of the student engagement instrument (SEI): results from middle and high school students. *Sch. Psychol. Q.* 25, 84–93. doi: 10.1037/a0020259

Bollen, K. A. (1989). *Structural equations with latent variables*. NJ: Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford publications.

Buehl, M. M., and Fives, H. (2009). Exploring teachers' beliefs about teaching knowledge: where does it come from? Does it change? *J. Exp. Educ.* 77, 367–408. doi: 10.3200/JEXE.77.4.367-408

Byrne, B. M., and van de Vijver, F. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies. *Int. J. Test.* 10, 107–132. doi: 10.1080/15305051003637306

Cattell, R. B., and Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educ. Psychol. Meas.* 24, 3–30. doi: 10.1177/001316446402400101

Chen, Y. C., and Techawitthayachinda, R. (2021). Developing deep learning in science classrooms: tactics to manage epistemic uncertainty during whole-class discussion. *J. Res. Sci. Teach.* 58, 1083–1116. doi: 10.1002/tea.21693

Chin, C. (2007). Teacher questioning in science classrooms: approaches that stimulate productive thinking. *J. Res. Sci. Teach.* 44, 815–843. doi: 10.1002/tea.20171

Deng, F., Yang, H. L., and Wang, L. J. (2019). Adaptive unscented Kalman filter based estimation and filtering for dynamic positioning with model uncertainties. *Int. J. Control. Autom. Syst.* 17, 667–678. doi: 10.1007/s12555-018-9503-4

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: toward better conceptualizations and measures. *Educ. Res.* 38, 181–199. doi: 10.3102/0013189X08331140

Ding, C., Lammert, C., and Fulmer, G. W. (2023). Refinement of an instrument measuring science teachers' knowledge of language through mixed method. *Disciplinary and Interdisciplinary Science Education Research* 5. doi: 10.1186/s43031-023-00080-7

Duschl, R. A., and Bybee, R. W. (2014). Planning and carrying out investigations: an entry to learning and to teacher professional development around NGSS science and engineering practices. *Int. J. STEM Educ.* 1, 1–9. doi: 10.1186/s40594-014-0012-6

Ferrando, P. J., and Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educ. Psychol. Meas.* 78, 762–780. doi: 10.1177/0013164417719308

Fulmer, G. W. (2008). *Successes and setbacks in collaboration for science instruction: connecting teacher practices to contextual pressures and student learning outcomes*. Riga, Latvia: VDM Publishing.

Fulmer, G. W., Hansen, W., Hwang, J., Ding, C., Ash, A., Hand, B., et al. (2023). "Development and application of a questionnaire on teachers' knowledge of argument as an epistemic tool" in *Advances in applications of Rasch measurement in science education*. eds. W. Liu and W. Boone (Cham: Springer)

Fulmer, G. W., Hwang, J., Ding, C., Hand, B., Suh, J. K., and Hansen, W. (2021). Development of a questionnaire on teachers' knowledge of language as an epistemic tool. *J. Res. Sci. Teach.* 58, 459–490. doi: 10.1002/tea.21666

Harris, L. R., Adie, L., and Wyatt-Smith, C. (2022). Learning progression–based assessments: a systematic review of student and teacher uses. *Rev. Educ. Res.* 92, 996–1040. doi: 10.3102/00346543221081552

Hashweh, M. Z. (1996). Effects of science teachers' epistemological beliefs in teaching. *J. Res. Sc. Teach.* 33, 47–63. doi: 10.1002/(SICI)1098-2736(199601)33:1<47::AID-TEA3>3.0.CO;2-PC

Hill, C. D., Edwards, M. C., Thissen, D., Langer, M. M., Wirth, R. J., and Burwinkle, T. M. (2007). Practical issues in the application of item response theory: a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales. *Med. Care* 45, S39–S47. doi: 10.1097/01.mlr.0000259879.05499.eb

Hirschfeld, G., and Von Brachel, R. (2014). Improving multiple-group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. *Pract. Assess. Res. Eval.* 19:7. doi: 10.7275/qazy-2946

Hofer, B. K., and Pintrich, P. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.3102/00346543067001088

Howard, B. C., McGee, S., Schwartz, N., and Purcell, S. (2000). The experience of constructivism: transforming teacher epistemology. *J. Res. Comput. Educ.* 32, 455–465. doi: 10.1080/08886504.2000.10782291

Kaufman, E. A., Xia, M., Fosco, G., Yaptangco, M., Skidmore, C. R., and Crowell, S. E. (2016). The difficulties in emotion regulation scale short form (DERS-SF): validation and replication in adolescent and adult samples. *J. Psychopathol. Behav. Assess.* 38, 443–455. doi: 10.1007/s10862-015-9529-3

Lammert, C., Hand, B., Suh, J. K., and Fulmer, G. (2022a). "It's all in the moment": a mixed-methods study of elementary science teacher adaptiveness following professional development on knowledge generation approaches. *Discip. Interdiscip. Sci. Educ. Res.* 4:12. doi: 10.1186/s43031-022-00052-3

Lammert, C., Sharma, R., and Hand, B. (2023). Beyond pedagogy: the role of epistemic orientation and knowledge generation environments in early childhood science teaching. *Int. J. Sci. Educ.* 45, 431–450. doi: 10.1080/09500693.2022.2164474

Lammert, C., Suh, J. K., Hand, B., and Fulmer, G. (2022b). Is epistemic orientation the chicken or the egg in professional development for knowledge generation approaches? *Teach. Teach. Educ.* 116:103747. doi: 10.1016/j.tate.2022.103747

Lei, L., Huang, X., Zhang, S., Yang, J., Yang, L., and Xu, M. (2020). Comparison of prevalence and associated factors of anxiety and depression among people affected by versus people unaffected by quarantine during the COVID-19 epidemic in Southwestern China. *Medical science monitor: international medical journal of experimental and clinical research*, 26:e924609-1. doi: 10.12659/MSM.924609

Liang, X., and Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *Int. J. Quan. Res. Educ.* 2, 17–38. doi: 10.1504/IJQRE.2014.060972

Liu, X., and Fulmer, G. (2008). Alignment between the science curriculum and assessment in selected NY state regents exams. *J. Sci. Educ. Technol.* 17, 373–383. doi: 10.1007/s10956-008-9107-5

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., and Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychol. Methods* 22, 486–506. doi: 10.1037/met0000075

Lorenzo-Seva, U., and Ferrando, P. J. (2021). MSA: the forgotten index for identifying inappropriate items before computing exploratory item factor analysis. *Methodology* 17, 296–306. doi: 10.5964/meth.7185

Maggioni, L., and Parkinson, M. M. (2008). The role of teacher epistemic cognition, epistemic beliefs, and calibration in instruction. *Educ. Psychol. Rev.* 20, 445–461. doi: 10.1007/s10648-008-9081-8

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement* 11, 71–101. doi: 10.1080/15366367.2013.831680

McGrath, R. E. (2015). Measurement invariance in translations of the VIA inventory of strengths. *Eur. J. Psychol. Assess.* 32, 187–194. doi: 10.1027/1015-5759/a000248

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychol. Methods* 23:412. doi: 10.1037/met0000144

Milanzi, E., Molenberghs, G., Alonso, A., Verbeke, G., and De Boeck, P. (2015). Reliability measures in item response theory: manifest versus latent correlation functions. *Br. J. Math. Stat. Psychol.* 68, 43–64. doi: 10.1111/bmsp.12033

Milfont, T. L., and Fischer, R. (2010). Testing measurement invariance across groups: applications in cross-cultural research. *Int. J. Psychol. Res.* 3, 111–130. doi: 10.21500/20112084.857

Miller, E., Manz, E., Russ, R., Stroupe, D., and Berland, L. (2018). Addressing the epistemic elephant in the room: epistemic agency and the next generation science standards. *J. Res. Sci. Teach.* 55, 1053–1075. doi: 10.1002/tea.21459

Millsap, R. E., and Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivar. Behav. Res.* 39, 479–515. doi: 10.1207/S15327906MBR3903_4

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Oxfordshire, England, UK: Routledge.

Morandi, S., Hagan, C., Granger, E. M., Schellinger, J., and Southerland, S. A. (2022). Exploration of epistemic orientation towards teaching science in a longitudinal professional development study. Paper presented at the annual meeting of NARST (Vancouver, BC). Available at: https://par.nsf.gov/servlets/purl/10330968

Muis, K. R. (2007). The role of epistemic beliefs in self-regulated learning. *Educ. Psychol.* 42, 173–190. doi: 10.1080/00461520701416306

Muthén, B., and Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: multiple-group and growth modeling in Mplus. *Mplus Web Notes* 4, 1–22. Available at: https://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, L. K., and Muthén, B. (2021). *Mplus. The comprehensive modelling program for applied researchers: User's guide*, 8 Available at: https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf.

National Research Council (2012). *A framework for K-12 science education: practices, crosscutting concepts, and Core ideas* The National Academies Press.

NGSS Lead States (2013). *Next generation science standards: for states, by states*. NW Washington, DC: The National Academies Press.

Palma, E. M. S., Gondim, S. M. G., and Aguiar, C. V. N. (2018). Epistemic orientation short scale: development and validity evidence in a sample of psychotherapists. *Paidéia (Ribeirão Preto)* 28:e2817. doi: 10.1590/1982-4327e2817

Ruiz-Primo, M. A., and Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: an update. *J. Res. Sci. Teach.* 33, 1045–1063. doi: 10.1002/(SICI)1098-2736(199612)33:10<1045::AID-TEA1>3.0.CO;2-S

Seung, E., Park, S., Kite, V., and Choi, A. (2023). Elementary preservice teachers' understandings and task values of the science practices advocated in the NGSS in the US. *Educ. Sci.* 13:371. doi: 10.3390/educsci13040371

Shi, D., Maydeu-Olivares, A., and DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivar. Behav. Res.* 53, 676–694. doi: 10.1080/00273171.2018.1476221

Singh, A. S., and Masuku, M. B. (2014). Sampling techniques & determination of sample size in applied statistics research: an overview. *Int. J. Econ. Com. Manag.* 2, 1–22.

Stroupe, D. (2014). Examining classroom science practice communities: how teachers and students negotiate epistemic agency and learn science-as-practice. *Sci. Educ.* 98, 487–516. doi: 10.1002/sce.21112

Suh, J. K., Hand, B., Dursun, J. E., Lammert, C., and Fulmer, G. (2023). Characterizing adaptive teaching expertise: teacher profiles based on epistemic orientation and knowledge of epistemic tools. *Sci. Educ.* 107, 884–911. doi: 10.1002/sce.21796

Suh, J. K., Hwang, J., Park, S., and Hand, B. (2022). Epistemic orientation toward teaching science for knowledge generation: conceptualization and validation of the construct. *J. Res. Sci. Teach.* 59, 1651–1691. doi: 10.1002/tea.21769

Weiss, K. A., McDermott, M. A., and Hand, B. (2022). Characterising immersive argument-based inquiry learning environments in school-based education: a systematic literature review. *Stud. Sci. Educ.* 58, 15–47. doi: 10.1080/03057267.2021.1897931

Widaman, K. F., Ferrer, E., and Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: measuring the same construct across time. *Child Dev. Perspect.* 4, 10–18. doi: 10.1111/j.1750-8606.2009.00110.x

Windschitl, M. (2002). Framing constructivism as the negotiation of dilemmas: an analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Rev. Educ. Res.* 72, 131–175. doi: 10.3102/00346543072002131

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 19 October 2023
DOI 10.3389/feduc.2023.1162499

Check for updates

OPEN ACCESS

EDITED BY
Hendrik Lohse-Bossenz,
University of Education Heidelberg, Germany

REVIEWED BY
Ferman Konukman,
Qatar University, Qatar
Matthias Baumgartner,
University of Teacher Education St. Gallen, Switzerland

*CORRESPONDENCE
Christian Leukel
✉ christian.leukel@ph-freiburg.de

RECEIVED 09 February 2023
ACCEPTED 15 September 2023
PUBLISHED 19 October 2023

CITATION
Leukel C, Leuders T, Bessi F and Loibl K (2023) Unveiling cognitive aspects and accuracy of diagnostic judgments in physical education teachers assessing different tasks in gymnastics.
Front. Educ. 8:1162499.
doi: 10.3389/feduc.2023.1162499

COPYRIGHT
© 2023 Leukel, Leuders, Bessi and Loibl. 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.

Unveiling cognitive aspects and accuracy of diagnostic judgments in physical education teachers assessing different tasks in gymnastics

Christian Leukel[1,2]*, Timo Leuders[1], Flavio Bessi[3] and Katharina Loibl[1]

[1]University of Education Freiburg, Freiburg, Germany, [2]Bernstein Center Freiburg, University of Freiburg, Freiburg, Germany, [3]Department of Sport Science, University of Freiburg, Freiburg, Germany

Introduction: Diagnostics is an essential part of teachers' profession. We investigated judgment accuracy and cognitive processes underlying judgment formation in physical education teachers who observed tasks in gymnastics, and compared teachers with gymnastics trainers as a reference group.

Methods: Teachers and trainers judged performance of prepuberal students in gymnastics, namely students exercising squat vault, underswing, and handstand. To investigate cognitive processes of judgment formation, participants were asked to structure the movements via event segmentation as well as to explain their judgments. All teachers and trainers had experience in working with prepuberal children similar to those they observed in this experiment, and the teachers completed a gymnastics class during their studies.

Results: Judgment accuracy (with reference to judgments made by expert trainers) was found to be significantly lower in teachers compared to trainers ($p < 0.001$). Moreover, agreement on the ratings among teachers was lower than among trainers. Agreement about the temporal structuring of the tasks from event segmentation was lower among teachers than among trainers ($p < 0.05$). When explaining their ratings, trainers referred more often than teachers to kinematic features of the task that were relevant to the judgments.

Discussion: We discuss these findings in context of the teachers' task to perform accurate judgments. For suggestions on teacher training, we particularly emphasize the relevance of implementing knowledge about kinematic features of the tasks and student errors into real-life scenarios resembling the complex skill of making accurate judgments in the physical education classroom.

KEYWORDS

movement, middle school, competencies, feedback, learning, assessment, teacher education, sports

Highlights

- Teacher's judgment accuracy (with reference to judgments made by expert trainers) about movement errors in tasks in gymnastics is lower compared to the judgment accuracy of trainers.
- Agreement on the judgments among teachers is significantly lower than among trainers.
- Agreement on the temporal structure of the tasks is significantly lower in teachers than in trainers.
- Trainers referred more often than teachers to kinematic features of the task that were relevant to the judgments.

1. Introduction

Learning depends on accurate feedback. In school, diagnostic judgments by teachers are the primary source of information to generate feedback. Accordingly, accuracy of teachers' diagnostic judgments has been recognized and studied in various contexts (Südkamp et al., 2012; Loibl et al., 2020; Urhahne and Wijnia, 2021), such as in reading (Bates and Nettelbeck, 2001), mathematics (Leuders et al., 2022) and physical education (PE) (Ward et al., 2020; Moura et al., 2021). With regard to diagnostic judgments in PE, O'Brien et al., 2023 emphasized that "Physical education as a subject has evolved beyond the idea of 'busy, happy, good' (Placek, 1983), constituting a successful learning experience. Students developing and illustrating their capabilities across the cognitive and psychomotor domains are now at the forefront of physical education assessment (Hay, 2006)." The primary goal of assessment in PE classes is not on testing and grading (testing culture), but on the promotion of learning and teaching (i.e., assessment for learning) (López-Pastor et al., 2013; Tolgfors, 2018).

Development of motor competences is considered one of the pillars of PE (Sacko et al., 2021; Dudley et al., 2022), and external feedback about the students' performance is necessary for improving these competences (Magill, 2001; Leukel and Lundbye-Jensen, 2012). Feedback for learning requires that teachers make accurate judgments on characteristics critical to the performance of an intended skill (Sacko et al., 2021). One way to achieve this is by standardized testing (Seidel and Bös, 2012; Herrmann et al., 2016). However, standardized testing has significant limitations, namely: (i) the focus is typically narrow, meaning that tests capture only a small subset of motor abilities and skills. Thus, (ii) for many skills of the PE curriculum no test exists. Furthermore, (iii) ceiling and floor effects of tests are problematic because students at the extremes of the spectrum (high-ability and low-ability students) often cannot be reliably judged (Rink, 2013). The central deficiency of formal tests is that (iv) though they can produce accurate diagnostic judgments, they do not improve teachers' ability to judge students' performance. However, teachers' diagnostic judgments are a prerequisite for individual and adaptive feedback which supports learning (Swinnen, 1996; Magill, 2001; Leukel and Lundbye-Jensen, 2012). Therefore, diagnostic judgments as a basis for feedback should be integrated in PE lessons throughout and whenever possible. The ability to judge students on a continuous basis, which is considered a core component of teacher knowledge (Baumert and Kunter, 2013; Urhahne and Wijnia, 2021), cannot be substituted by a plethora of formal tests. Hence, besides formal testing teachers are required to make accurate judgments on characteristics critical to the performance of (complex) motor skills which are part of the PE curriculum, by focussing on key movement features that are relevant for performance of an intended task (Sacko et al., 2021). Analysing these characteristics is not trivial but requires both extensive practice in observation and profound knowledge of the intended task (Barrett, 1983; Ward et al., 2020).

There is evidence that PE teachers do not make accurate judgments (Lorente-Catalán and Kirk, 2016; van der Mars et al., 2018; Sacko et al., 2021), and reasons have been put forward trying to explain why this is the case, e.g., that criteria used for assessment are quite subjective and not based on evidence (Tolgfors, 2018). In fact, there is still very little information about the cognitive processes underlying judgment formation in PE teachers. This knowledge is crucial because it can substantially aid further research and interventions in teacher education (Loibl et al., 2020). It allows explaining the diagnostic skills of teachers (Chernikova et al., 2020; Loibl et al., 2020; Leuders and Loibl, 2021) and designing instructions by which teachers' diagnostic competences can be enhanced (Chernikova et al., 2020; Leuders et al., 2022). To exemplify this point, Niederkofler et al. (2018) found deficient judgment accuracy in teachers assessing fundamental motor skills and argued that they should be trained to become more competent in diagnostics. However, it remains unclear which aspects of the teachers' reasoning should be targeted, because the cognitive processes that lead to decisions were not evaluated in the study. Like Niederkofler et al. (2018), Ferrari et al. (2022) also found that teachers generally overrated the capabilities of their students on the whole-class level. They found significant correlations of judgment accuracy with class size, but not with the number of weekly lessons spent in the class, with experience, or self-reported competence. This may be attributed to the fact that judging the whole class requires the integration of individual judgments, which is more difficult for larger classes. However, for making such assumptions it is essential to gain more detailed knowledge about the information processing when making decisions.

Decision making, in PE and other domains, occurs through information processing which has recently been explicated within a theoretical framework (Loibl et al., 2020). In this framework, diagnostic thinking is conceptualized as three steps, namely perception, interpretation, and decision making. Perception is primarily visual when making judgments in PE. According to the mentioned theoretical framework (Loibl et al., 2020), it is crucial to consider (a) what kind of (visual) information the teacher perceives (referring to situational cues), and (b) how this relates to the decision (referring to the interpretation of the cues). In gymnastics, there is empirical evidence that judgments are based on visual perception and evaluation of salient kinematic features, like the height of an athlete over the ground when jumping over the vault, or the time airborne

(Takei, 1998; Farana and Vaverka, 2012; Luis del Campo and Espada Gracia, 2018; Mack, 2020). In general, and not constrained to gymnastics, experts were shown to be better in picking up and evaluating relevant perceptual cues compared to novices (Abernethy et al., 2001; Mann et al., 2007). According to these empirical findings, studying cognitive processes underlying judgment formation in teachers should therefore be concerned with kinematic features that teachers perceive and process. Experimentally, this can be achieved with eye-tracking (Kredel et al., 2017; Mack, 2020) or through event segmentation (Zacks and Swallow, 2007; Kurby and Zacks, 2008). Event segmentation is concerned with the idea of how people automatically and unintentionally compartmentalize perceptual experience into temporally defined phases that are segregated by event boundaries. A central postulate is that boundaries and phases are used by humans to make predictions and inferences (Zacks and Swallow, 2007). When observing movements, it has been shown that event boundaries relate to salient kinematic features (Zacks et al., 2009; Newberry et al., 2021). Event segmentation even works for movements characterized by rapid kinematic changes like tasks in gymnastics (Bläsing, 2015; Newberry et al., 2021; Stadler et al., 2021), in contrast to eye-tracking where short fixations of the eyes with a rapidly changing visual scene is difficult to track (Mack, 2020). Therefore, and according to the theoretical grounding about the relevance of kinematic features for decision making in gymnastics, event segmentation was used in the current study.

Diagnostic judgments in PE are volatile and often based on normative descriptions of correctness (Mechling and Munzert, 2004; Hong and Bartlett, 2008). Research on judgment accuracy in PE requires benchmarking teachers' achievements according to a reference source. Referencing can be done by comparing teachers' results to standardized test outcomes like in the study of Niederkofler et al. (2018). As mentioned, a disadvantage of test outcomes as reference is that for many tasks in PE no tests exist. An alternative approach, therefore, is to utilize trainers who hold a sufficient level of judgment accuracy with regard to the studied task as reference (Bläsing, 2015; Mack, 2020; Newberry et al., 2021; Stadler et al., 2021).

In summary, accurate diagnostic judgments on motor skills from teachers are important in PE. Empirically substantiated knowledge about cognitive processes of judgment formation in PE teachers is scarce but a requirement for understanding deficiencies in teachers' judgments and for furthering teacher education in this area. Therefore, in the present study, two aspects of diagnostic judgments, namely judgment accuracy and cognitive processes of judgment formation, were theoretically defined and empirically studied.

1.1. The present study

In the present study, we investigated PE teachers' diagnostic judgments of volatile and short-lasting tasks in gymnastics. This is typical for situations in the PE classroom, in that students sequentially perform the task, teachers observe this performance, and subsequently provide feedback on that performance. It is also typical for such a situation that teachers have to generate judgments and feedback under time pressure while teaching the whole class. The selected tasks, namely squat vault, underswing, and handstand, are part of the PE curriculum in Germany, and suitable for investigating information processing because the visual cues essential for estimating movement errors and performance are well-defined (Heinen, 2015). In order to draw meaningful comparisons, we included regular teachers and trainers in gymnastics who both had worked with prepuberal students before.

The study had two aims. The first aim was to explore to what extent teachers are able to form accurate judgments in the mentioned situations. We therefore investigated teachers' ratings on the severity of movement errors from watching video vignettes of prepuberal students performing squat vault, underswing, and handstand, and we compared these ratings to the ratings made by trainers. The video vignettes were played in real-time and pictured the students from the side, which resembles the situation that the teachers typically face in the PE classroom.

The second aim of the study was to investigate cognitive processes underlying judgment formation, which in our case relates to the processing of kinematic features linked to the decision about movement errors. This was achieved on the basis of event segmentation on the one hand and verbal reasoning on the other. With the former we assessed aspects of the visual information the subjects focussed on, and this was measured by spontaneous reactions of the teachers and trainers during the observation of the video vignettes. With verbal reasoning, we assessed the explicit reasons of a particular judgment, revealing the consciously driven process of the decisions. Accordingly, event segmentation informs about what kind of sensory cues the teachers and trainers actively focussed on, and the comparison between event segmentation and verbal reasoning reveals if these aspects were considered as meaningful for the decision by the teachers and trainers.

Although teachers received training on these three tasks (which are part of the basic repertoire in gymnastics) during their studies, and studied the characteristics of the movements before entering the experiment, we expected that they would have difficulties in identifying and interpreting relevant sensory cues for performance because of limited or absent practical experiences. In particular, the students gained mostly theoretical knowledge during their study but were not trained to apply this knowledge in a real-live scenario resembling the complex task of making accurate diagnostic judgments in the PE classroom. These difficulties should become apparent when processing the information (event segmentation and written explanations of the judgments) and also manifest in the accuracy of the judgments when compared to trainers.

2. Methodology and methods

2.1. Subjects

Forty subjects (aged between 21 and 60 years) participated in this study (Table 1). Half of the subjects (20 subjects, aged 29 ± 3 years, 13 males, 7 female) were trainee teachers in their final year of an 18 months induction phase ("Referendariat"). All of them were PE teachers, and all had worked for 1 year as PE teachers in secondary schools at the time of the experiment. Gymnastics had been part of their study. They all had taken a gymnastics class for one semester, and were trained on the tasks they had to judge in this study in terms of self-performance and teaching methods including movement characteristics and movement errors. Importantly, the knowledge they gained during their study had not been applied in a real-live task, thus

TABLE 1 Subjects' characteristics.

Trainers						
ID	Sex	Age	Licence as trainer	Licence as judge	Years active	Main sport(s)
1	m	52	A	A	45	Gymnastics
2	f	27	C		10	Gymnastics
3	m	27	A	B	15	Gymnastics
4	m	29	C		7	Gymnastics
5	m	59	C	B	30	Gymnastics
6	m	23	C		3	Gymnastics
7	m	54	B	B	20	Gymnastics
8	f	25	C		5	Gymnastics
9	f	25	C		4	Gymnastics
10	f	26	C	B	15	Gymnastics
11	m	29	C	B	12	Gymnastics
12	m	34	C		15	Gymnastics
13	m	49	B	A	40	Gymnastics
14	f	43	B	B	36	Gymnastics
15	f	25	C	C	17	Gymnastics
16	m	28	B	B	20	Gymnastics
17	m	22	C		15	Gymnastics
18	m	60	B	A	50	Gymnastics
19	m	21	B	B	13	Gymnastics
20	f	52	C		30	Gymnastics

Teachers						
ID	Sex	Age	Licence as trainer	Licence as judge	Years active	Main sport(s)
1	m	30				Handball, kitesurfing
2	m	26			15	Soccer, gymnastics
3	m	28				Karate, soccer
4	m	28				Snowboarding, soccer
5	m	34				Soccer
6	f	27				Athletics
7	m	27			17	Gymnastics
8	m	32				Volleyball, tennis
9	f	37				Volleyball
10	m	28				Handball, wakeboarding
11	m	28				Climbing, basketball
12	f	27				Running, swimming
13	f	28				Volleyball
14	f	28				Skiing, volleyball
15	m	28				Climbing
16	m	28				Soccer
17	m	36				Tennis, athletics
18	m	32				Handball
19	f	26				Volleyball, skiing
20	f	26				Soccer

The terms "licence as trainer" and "licence as judge" refers to the licence obtained as trainer and/or judge in gymnastics. m, male; f, female. Trainers marked in green were taken for creating the reference trainer (i.e., median ratings of the 10 best trainers). The reference trainer was used for calculating judgment accuracy of the teachers and the remaining 10 trainers.

they were not trained in performing accurate diagnostic judgments in situations similar to the situation in the classroom. Approximately 2 weeks before the experiment, the teachers were informed about the types of motor tasks and type of students they had to judge. When asked after the experiment, the teachers stated that they had prepared for it, by rehearsing the movement characteristics of the tasks by reading the scripts from their gymnastics class and/or reading about movement characteristics and movement errors of these tasks in (a) gymnastics book(s). The other half of the subjects (20 subjects, aged 36 ± 13 years, 13 males, 7 females) were gymnastics trainers holding a C-licence in gymnastics at minimum at the time of the experiment. All of the trainers spent between 3 and 8 h weekly in the gym and also trained children who were beginners in gymnastics. Gymnastics was their main sport. They had been active in gymnastics between 2 years and 15 years. Importantly, all subjects (teachers and trainers) worked with prepuberal children (teachers: school, trainers: gym) who achieved performance levels similar to the students they judged in the current study. All subjects (trainers and teachers) provided written informed consent before participation. The study was conducted according to the guidelines set in the Declaration of Helsinki (latest revision in Fortaleza) and approved by the local ethics committee. All subjects received a book voucher of 10 Euros to compensate for the time they spent in the laboratory.

2.2. Experimental design

Thirty video vignettes showing prepuberal female students performing gymnastics tasks were presented to the subjects. In each vignette, a single student was shown who performed a single task, namely a squat vault, an underswing, or a handstand with subsequent roll-out and stance (Figure 1). All three tasks are part of the curriculum for secondary schools in the State of Baden-Württemberg in Germany.

Subjects had to complete three different tasks in a consecutive order when watching the vignettes: first, they had to segment the video into meaningful, temporally defined phases. Second, the subjects had to rate the severity of movement errors of the overall performance of the student. Third, the subjects had to explain the main reasons for their rating in written form. These three tasks are explained in detail below.

Psychopy 3.0 (Open Science Tools Ltd.) running on a 13-inch Macbook Pro computer (5th Generation, Apple Inc., California) was used to control the execution and the timing of the tasks and to record the data. The laptop was connected to an external keyboard (Wireless Keyboard, 3rd Generation, Apple Inc., California) used by the subjects, and a 24-inch external LED screen (refresh rate: 60 Hz, LG, Seoul). The external screen was placed at a distance of 80 cm in front of the subjects.

2.3. Videos

Videos were recorded with a Sony 4 K camcorder (AX100 E, Sony, Tokio) from 10 female students aged between 10 and 12 years. The camera angle was chosen so that the shots captured the whole body of the students performing the tasks. There was no panning and zooming during the recordings. For handstand, the camera was placed at a distance of 3.5 m from and orthogonal to the mat on which the

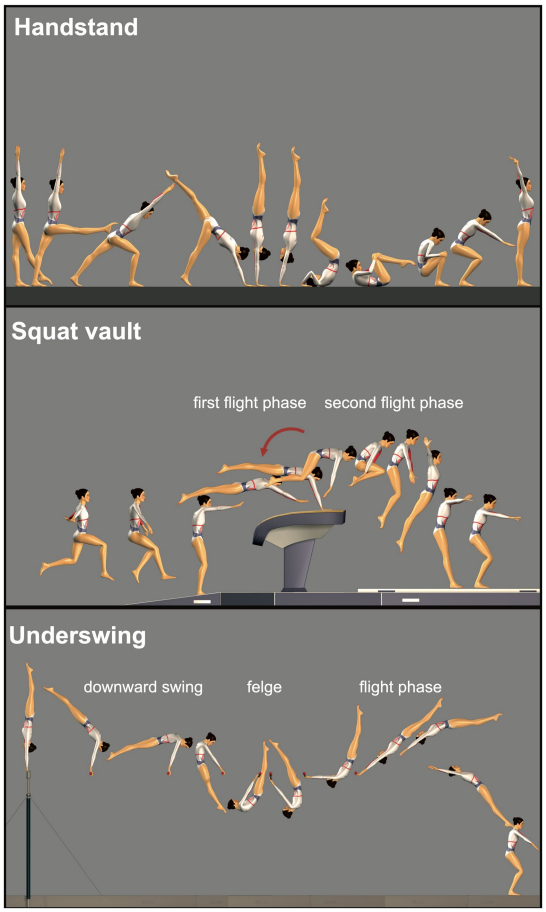


FIGURE 1
Illustrates movement phases for squat vault, underswing, and handstand. Note that for underswing, the athletes displayed in the videos in the current study did start with a support-position and not with a handstand, like it is displayed in the figure.

handstand was performed. For squat vault, the camera was placed at a distance of 5 m from and orthogonal to the vault. For underswing, the camera was placed at a distance of 4.8 m from the centre of the horizontal bar and orthogonal to the bar. For handstand and underswing, the camera captured the students from the beginning to the end of the task. For squat vault, the camera captured the students before jumping onto the springboard until landing on the mat after the vault. Thus, for squat vault, the camera did not capture the run-up. The contrast between the student and the background was increased by choosing light colors as background colors, i.e., yellow-colored vertically-placed mats when filming the underswing and the ivory-colored wall when filming the handstand and squat vault, respectively.

The raw video data were cut as follows: for handstand, the video started 1.5 s before the subjects initiated the movement, and ended after subjects reached stance. For squat vault, the video started 1.5 s before the subjects appeared in the picture (i.e., final step before the springboard), and ended after students reached stance (or crashed after landing). For underswing, the video started 1.5 s before the students started with the task, and ended after students reached stance (or crashed after landing). The final duration of each video was in between 4.2 and 6.8 s. The videos were presented in random order (i.e., no block design) to the subjects.

The students performing the tasks (age: 11–12 years) were recruited from a local gymnastics club and at the time of the recordings practiced gymnastics 1 to 2 times a week for a total of 2 h per week. The students had practiced this sport for half a year and up to 6 years at the time of the recordings. Thus, the performance level ranged between the children, which was intended, to cover the performance levels teachers typically see in the classroom (i.e., beginners to advanced). Each of the children performed all of the three mentioned tasks. All students wore the same clothes (namely black tight sports pants and a grey t-shirt) to reduce biased judgments relating to personal characteristics other than movement performance.

2.4. Event segmentation

Subjects had to segment the videos into temporally defined phases that were meaningful and seemed natural to them. Therefore, while watching the videos, they were instructed to press the spacebar on the keyboard whenever, in their opinion, a meaningful phase ended and a new one began. They had to place the index finger on the spacebar when performing the segmentation to reduce movement times affecting the timing of the presses. The time stamps marked by subjects are called event boundaries (Kurby and Zacks, 2008). Subjects viewed the entire video once before performing the segmentation. The segmentation had to be performed twice in consecutive order. Thus, subjects watched each of the videos three times (twice while also performing segmentation). When performing the segmentation the second time, we instructed the subjects to repeat what they did in the first run. Repetition was included because a previous study showed a systematic temporal shift of event boundaries with repeated exposure (Michelmann et al., 2021).

2.5. Judgment

Subjects had to rate the severity of movement errors they observed from viewing a task. They had to rate on a 5-point Likert scale: 1. no movement error; 2. task performed with minor movement errors; 3. task performed with medium movement errors; 4: task performed with major movement errors; or 5. task performed with very large movement errors. Note that subjects were asked about the severity of the errors, not the quantity (number) of errors they observed. The quantity indeed plays a role in competition but is not so important in a school setting. Here, effective performance-enhancive feedback rather addresses the severity of the error curtailing performance. Ratings had to be performed immediately after performing the segmentation. For each video, subjects were given 15 s to finalize their judgment, by pressing a number key on the keyboard referring to their rating (i.e., between 1 and 5).

2.6. Written explanations

After completing the judgment, subjects had to explain the main reasons for their rating in written form. They used the keyboard to write down the explanation in form of a text log. The subjects were instructed to explain the main reasons for the rating but were not constrained about what to include in and how to write the explanation (e.g., positive and negative aspects of the performance). They were

told to not pay attention to grammatical and language errors because spelling corrections were performed before analysing the data. For each video, subjects were given 50 s to write down the main reasons.

2.7. Choice reaction task

Subjects performed a choice-reaction task at the beginning of the experiment for assessing potential between-group differences in reaction times which could affect segmentation, namely how quickly subjects are able to press the button and set an event boundary. For the choice-reaction task, subjects viewed 5 different symbols (triangle, circle, cross, square, rectangle) on the computer screen (width: 10 cm, height: 10 cm, fill: light blue, background: white), which appeared in random order every 3.6 to 4 s for a duration of 300 ms. Each symbol was repeated 6 times; thus, 30 symbols were presented in total. A warning sign ("Get ready!") was presented on the screen for 4 s before showing the first symbol. Subjects were instructed to press the spacebar on the keyboard as quickly as possible as soon as a triangle appeared on the screen.

2.8. Experimental procedure

After having provided written informed consent, subjects were first tested in the choice-reaction task. Thereafter, they viewed a short video of 5 min in which the procedures of the main part of the experiment (i.e., event segmentation, rating, and explanation of the reasons for the rating) were explained. The subjects were allowed to ask questions concerning these procedures after having watched the instruction video. After the questions were answered by the experimenter the subjects executed three test trials and performed event segmentation, rating, and explanation of the reasons for the rating. The behavior of the subjects in these test trials was not recorded. The videos used for these test trials captured elementary school children from a local elementary school (fourth-graders) performing handstand, underswing, and squat vault, respectively. These videos were recorded and cut in the same way as the videos used in the main experiment. After finishing the test trials, subjects conducted the main experiment consisting of 30 videos. The duration of a single trial (event segmentation, rating, and explanation of the reasons for the rating for a single video) was 2 min and 10 s. The pause between two successive trials was 10 s. Thus, the overall duration of the main experiment was 1 h and 10 min. Subjects were given a break of 5 min after completing 15 videos to avoid fatigue. After completing 30 videos, at the end of the experimental session, subjects had to segment a final video showing a 10-year-old girl rising from a chair and leaving the room. This video showing a daily activity served as a control condition for the segmentation behavior. We expected that segmentation behavior would be different between trainers and teachers when viewing gymnastic tasks but not when viewing the daily activity.

2.9. Data analysis and statistics

2.9.1. Judgment

Interrater reliability within groups was assessed by calculating Krippendorff's alpha. Between-group differences (teachers versus trainers) of alpha were estimated by computing 95% confidence intervals from bootstrapping of the sample (2,000 sweeps) (Krippendorff, 2016). A significant difference between groups was assumed in case the confidence intervals of the groups did not overlap (Stolarova et al., 2014).

Rating accuracy was calculated with Spearman's rank correlations: the ratings (of 10 videos per task) of individual subjects were correlated with the ratings of a reference trainer. This approach of quantifying diagnostic accuracy via a correlation between the rank orders resulting from teachers' judgments on the one hand and from a reference order (often actual achievement in a test, but also expert judgments) is common in research concerned with diagnostic judgments. It goes back to a suggestion by Cronbach (1955), has been applied to teacher judgments by Helmke and Schrader (1987) and since then profusely and successfully used (Hoge and Coladarci, 1959; Südkamp et al., 2012; Urhahne and Wijnia, 2021). The reference trainer was created by calculating the median of the ratings per video of the 10 best trainers according to their experience level, namely their licence degree as trainer and judge in gymnastics (see Table 1). These 10 best trainers are assumed to perform the most accurate ratings. We compared (a) correlations between teachers and reference trainer (judgment accuracy of teachers) with (b) correlations between the remaining 10 trainers (those who were not taken for creating the reference trainer) and reference trainer (judgment accuracy of trainers). Statistical differences in judgment accuracy between groups (teachers against trainers) were assessed with unpaired Student's *t*-tests (for handstand, squat vault, and underswing, respectively) based on Fisher *z*-transformed correlation coefficients.

2.9.2. Segmentation behavior

Data from the segmentation task were aggregated in response vectors at a resolution of 60 Hz, corresponding to the screen refresh rate. Response vectors were set to 1 if a given participant had pressed the space bar within 200 milliseconds surrounding the time point and were set to 0 otherwise. To test for consensus between participants and the remaining members of the group (teachers and trainers) cosine similarity was computed between a participant's response vector and the average response vector across all other participants of the group. Accordingly, the average similarity to others' response was assessed per video. Differences in cosine similarity between groups were analysed with unpaired Student's t-tests.

The number of button presses was counted for each subject and video. Differences in button presses between groups were assessed with unpaired Student's *t*-tests.

Differences in the timings of button presses made between the first and the second run of segmentation were analysed by contrasting the instants of each of the button presses (with respect to video onset in deci-seconds) for the first and the second run. Values were discarded if the delay between run 1 and run 2 exceeded 500 ms, which was in most cases due to the fact that the subject did not perform a button press in run 1 or run 2, respectively. Linear regression was calculated from all button presses, for groups and tasks separately.

2.9.3. Segmentation behavior and movement characteristics

Movement characteristics linking to button presses were identified through analysing segmentation behavior in combination

with video analyses. The instants of the individual button presses (in deci-seconds from video onset) were marked in the videos, and movement characteristics occurring at these instants were identified. For squat vault and underswing, there were a small number of movement characteristics corresponding to these instants that were shared by members of the group and across videos, like the jump-off from the springboard for squat vault. According to the peaks of averaged response vectors from segmentation, 3 movement characteristics were selected for squat vault, and 4 movement characteristics were selected for underswing. The instants at which the movement characteristics occurred were determined for each video (in deci-seconds from video onset), and grand means were calculated from all of the 10 videos of each task.

In contrast to squat vault and underswing, for handstand the instants of button presses referred to a much larger number of movement characteristics, with larger inter-individual differences between subjects and videos. We therefore decided to select the top 5 characteristics shared by the trainers. Like for squat vault and underswing, the instants (in deci-seconds from video onset) at which these movement characteristics occurred were determined for each video, and grand means were calculated from all of the 10 videos.

2.9.4. Explanations

The 10 most frequently used nouns in explanations of the judgments were assessed separately for the two groups and the three tasks. The selection was performed according to the following procedure: first, the spelling of the written explanations was corrected. Second, the explanations of the subjects were tokenised and parsed into words. Capital letters were replaced by lower-case letters. Third, stop words were removed. Fourth, words with identical strings were counted. Fifth and finally, nouns were ranked according to the total count, and the 10 most frequently occurring nouns were listed.

2.9.5. Reaction times

Reaction times were recorded at a resolution of 60 Hz (screen refresh rate). Averaged reaction times were calculated for the 6 trials in each subject. Between-group differences in reaction times were analysed with an unpaired Student's t-test.

The level of significance was set to $p < 0.05$ for all tests. p-values from multiple comparisons were corrected according to Benjamini and Hochberg (1995). Data analyses were performed and graphs plotted using R programming language and R studio software (RStudio Inc., Boston).

3. Results

3.1. Judgment

This section of the results addresses the first aim of the study, which was to explore to what extent teachers are able to form accurate judgments. Therefore, teachers' ratings on the severity of movement errors were compared to the ratings made by trainers.

Single subject values of the ratings are depicted in Figure 2A. As it can be seen from Figure 2A, trainers declared movement errors to be more severe than the teachers across all tasks. The severity of movement errors was reported to be largest for underswing across groups, followed by squat vault and handstand.

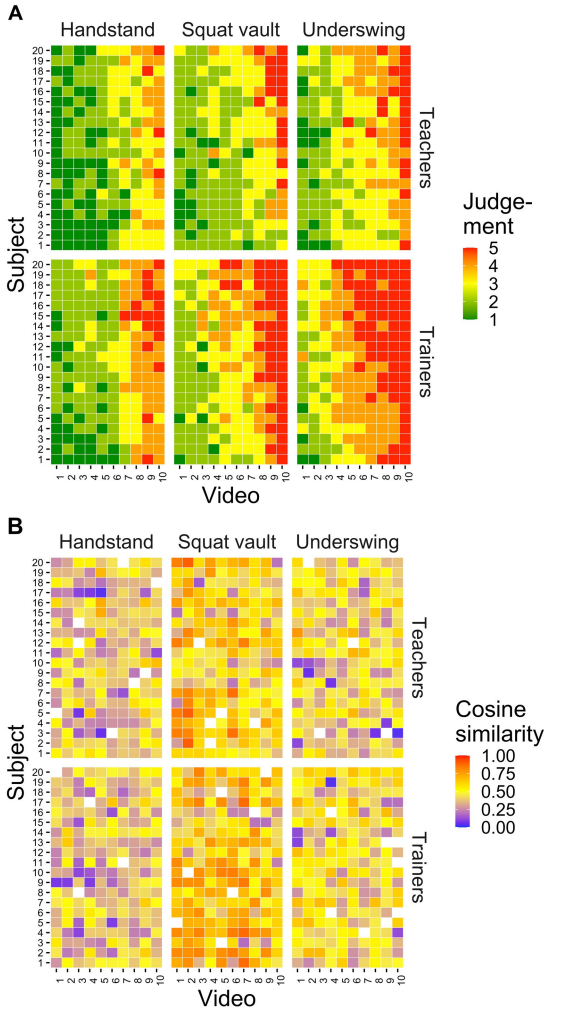


FIGURE 2
(A) Displays individual judgments. The dark green color indicates that, according to the subject's opinion, the task was performed with no movement error. The dark red color indicates that the task was performed with very large movement errors. The subjects were sorted according to the judgment mean (bottom: lowest value; top: highest value). Sorting was performed for each of the tasks separately. The same applies to the sorting of the videos. Videos were sorted according to the judgment mean from trainers (left: lowest value; right: highest value). **(B)** Displays cosine similarity calculated from segmentation data. Cosine similarity was computed between a participant's response vector (i.e., button presses of a single trial) and the average response vector (of the same trial) across all other participants of the group. Thus, the values display agreement between individual segmentation behavior and the segmentation behavior of others corresponding to the same group. A value of 1 indicates total agreement. Note that the sorting of the videos and subject IDs correspond to the sorting in part A of this figure. White spots: missing values, i.e., the subject did not perform segmentation.

Attributed performance levels clearly differed between students performing the tasks, ranging from students performing with no or few movement errors to students performing with very large movement errors. This was true for all three tasks.

Interrater reliability of the two groups and for the three tasks is displayed in Figure 3A. Krippendorff's alpha, expressing interrater reliability, was higher for trainers than for teachers for each of the

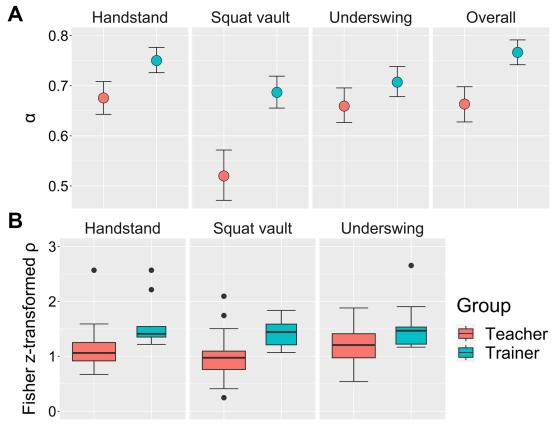


FIGURE 3
(A) Depicts Krippendorff's alpha calculated from judgments. "Overall" refers to the pooling of all data from the three tasks. Black bars display 95% confidence intervals which were calculated from bootstrapping of the sample (2,000 sweeps). **(B)** Depicts boxplots of Spearman's correlation coefficients (Fisher *z*-transformed values). Correlations were calculated between individual subjects and a reference trainer. Lower and upper hinges correspond to the first and third quartile, and the thick horizontal line of each boxplot represents the median. The upper whisker displays the largest value within 1.5 times the interquartile range above the third quartile. The lower whisker displays the smallest value within 1.5 times the interquartile range below the first quartile. Black dots display outliers. Note that, for reasons of clarity, one outlier in the trainers' group for handstand (*z*-score at 18.7) is not plotted.

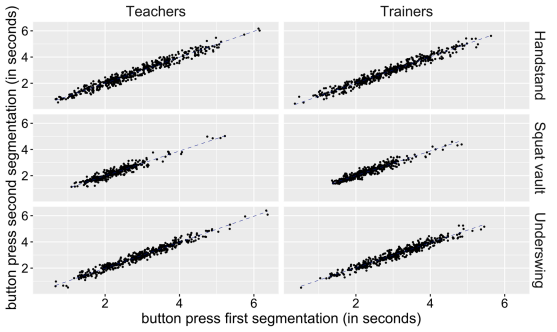


FIGURE 4
Shows correlations of individual button presses between the first and the second run of segmentation. The black dashed lines display the linear regression of data points.

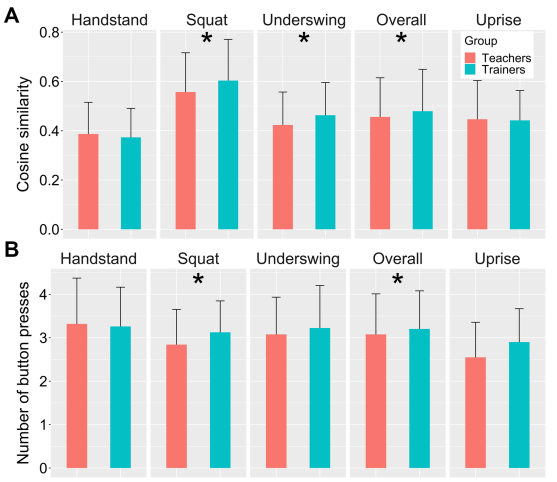


FIGURE 5
(A) Depicts grand mean values of cosine similarity from segmentation data. "Overall" refers to the pooling of all data from the three tasks. "Uprise" refers to the control condition (i.e., child stands up from a chair and leaves the room). Asterisks indicate significant differences between teachers and trainers. **(B)** Depicts grand mean values of the number of button-presses from segmentation. Asterisks indicate significant differences between the two groups.

three tasks and for the pooled data from all three tasks (Overall). 95% confidence intervals of alpha did not overlap between trainers and teachers for handstand, squat vault, and pooled data. This indicates significantly higher interrater agreement among trainers than among teachers.

The accuracy of the teachers' judgments was estimated with Fisher z-transformed Spearman's rank correlation coefficients that were compared between teachers and trainers (Figure 3B). Single subject values are depicted in Supplementary Figure S1. On the group level, results from the Student's *t*-tests yielded significant lower judgment accuracy in teachers compared to trainers for squat vault ($p < 0.01$, $t = -3.2$, Cohen's $d = 1.13$), underswing ($p < 0.05$, $t = -1.96$, Cohen's $d = 0.79$), but not for handstand ($p = 0.24$). Note that the result for handstand did not reach significance because of an outlier in the trainers' group (subject number 2) with a z-score at 18.7. Removing this outlier yielded a *p*-value of <0.05 ($t = -2.46$, Cohen's $d = 1.01$).

3.2. Segmentation

This section of the results addresses the second aim of the study, which was to investigate cognitive processes underlying judgment formation. Spontaneous reactions concerning the segmentation of the video vignettes were assessed and compared between trainers and teachers.

Timings of button presses in the first and second run of segmentation are displayed in Figure 4. Timings were not different between the two runs, as indicated by linear regression of data points. Slopes of all regressions were close to 1, and the regressions fitted the data points sufficiently well. Slopes (S) and coefficients of determination (R^2) were as follows: handstand: teachers ($S = 0.98$, $R^2 = 0.96$), trainers ($S = 1.0$, $R^2 = 0.96$); squat vault: teachers ($S = 0.95$, $R^2 = 0.92$), trainers ($S = 0.94$, $R^2 = 0.92$); underswing: teachers ($S = 0.99$, $R^2 = 0.96$), trainers ($S = 0.96$, $R^2 = 0.95$). According to these results, further analyses were conducted with data from the first run.

Single subject values are depicted in Figure 2B. As can be seen from Figure 2B, the biggest similarity of segmentation behavior among members of a group was observed for squat vault, followed by underswing and handstand, respectively.

There were no clear associations between cosine similarity and judgment. Declared severity of movement errors did not relate to the level of cosine similarity (comparison between Figures 2A,B). Otherwise, the coloring of the tiles in Figure 2B would follow the pattern of the coloring in Figure 2A. This is clearly not the case.

Cosine similarity was compared between the two groups for the three tasks, for the pooled data (overall), and for the control condition (uprise), respectively (Figure 5A). Unpaired Student's *t*-tests yielded

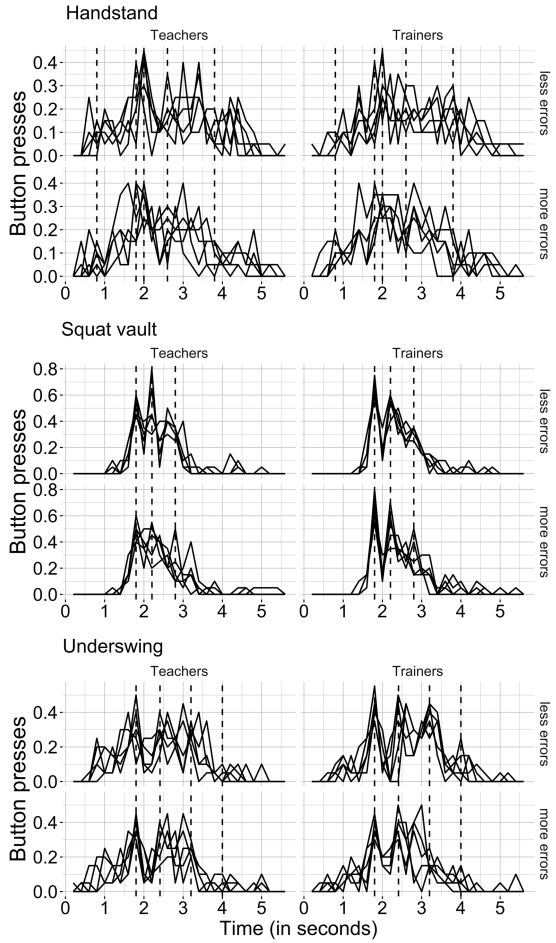


FIGURE 6
Displays averaged response vectors from segmentation over time, beginning at video onset. The 10 videos corresponding to each task were divided equally according to the judgment mean from trainers' ratings. Movement features that link to button presses from segmentation are displayed by dashed vertical lines. For handstand, the lines refer to: (i) subjects start moving, (ii) foot of supporting leg touches floor, (iii) hands touch floor (beginning of handstand), (iv) start of roll-out, (v) beginning of stance. For squat vault, the lines refer to: (i) feet leave springboard, (ii) hands leave vault, (iii) feet touch floor. For underswing, the lines refer to: (i) transition between backswing and forward swing, (ii) upper body crossing bar during forward swing, (iii) felge, (iv) feet touch floor. See also the text for a more detailed description of these movement features.

significant differences between teachers and trainers, for squat vault ($p < 0.01$, $t = -2.79$, Cohen's d = 0.28) and underswing ($p < 0.01$, $t = -2.92$, Cohen's d = 0.29). There were no differences for handstand ($p = 0.28$) and the control condition Uprise (i.e., child stands up from a chair and leaves the room) ($p = 0.92$). This indicates that, for squat vault and underswing, trainers' agreement about the temporal structuring of the tasks was significantly higher than the agreement among teachers.

Differences in the number of button-presses between groups are depicted in Figure 5B. For squat vault, the number of button presses was higher in trainers than in teachers (squat vault: $p < 0.01$, $t = -3.6$, Cohen's $d = 0.36$). There were no significant differences for handstand ($p = 0.56$), underswing ($p = 0.11$), and the control condition Uprise ($p = 0.17$). According to these results the higher number of button presses observed for the pooled data is likely due to the significant effect in squat vault.

For squat vault and underswing, button presses referred to distinct movement features marked in the average response vectors displayed in Figure 6. For handstand, in contrast to squat vault and underswing, there were no clear associations between response vectors and movement features across subjects and videos (see Figure 6). For squat vault and underswing, the movement features marked by the subjects refer to kinematic features that often occurred at transitions between subsequent movement phases. For squat vault, the first feature (in chronological order) marks a rapid change in acceleration of the body on the springboard, at the transition between the jump-off and the first flight phase. The second characteristic marks a rapid change in acceleration on the vault, at the transition between the first and the second flight phase (when the hands push off at the vault). The third feature marks the deceleration of the body at the transition between the second flight phase and the landing. For underswing, the first feature marks a rapid change in acceleration of the body at the transition between backswing and downward swing. The second feature marks the lifting of the body in the upward direction occurring between downward and upward swing. The third feature marks the maximum acceleration induced by the felge at the transition between the felge action and the flight phase. The fourth feature marks a deceleration of the body at the transition between the flight phase and the landing. Interestingly, for the underswing averaged response vectors of trainers indicated the presence/absence of the felge (cf. lower right part of Figure 6: the third dashed vertical line marks the transition between felge and flight phase). All subjects with fewer errors in Figure 6 (upper part) showed the felge, whereas all subjects with more movement errors (lower part) did not. Thus, trainers' segmentation behavior was sensitive to a critical feature for movement performance of the underswing.

3.3. Written explanations

In addition to segmentation, written explanations were analysed to investigate if the movement features marked when segmenting the videos were important for the judgment.

The 10 most frequently mentioned nouns of written explanations by trainers and teachers are listed in Figure 7. For squat vault, 5 out of the 10 nouns referred to movement features the trainers focussed on during segmentation, compared to 3 out of 10 in teachers. For underswing, 3 out of 10 nouns referred to features the trainers focussed on during segmentation, compared to 1 out of 10 in teachers. In total, 493 nouns referred to movement features the trainers focussed on during segmentation, compared to 236 nouns in teachers. This indicates that trainers referred to these features more often in their judgments (twice as much) compared to teachers.

4. Discussion

This study had two aims. The first aim was to elucidate to which extent teachers are able to perform accurate judgments. Therefore, we compared teachers' ratings to the ratings of trainers. The second aim was to investigate cognitive processes underlying judgment

Trainers

Handstand		Squat vault		Underswing	
handstand	153	**push-off (hands, vault)**	85	bar	84
roll-out	113	**start second flight phase**	67	**start forward swing (after backswing)**	75
arms/hands	86	**landing**	64	**felge/felgposition**	70
uprise	59	**start first flight phase**	61	hip	62
stiffness (body)	43	arms / hands	61	undershoot	49
back	30	vault	50	legs	40
head	28	**take-off (springboard)**	39	**landing**	32
feet	18	legs	37	shoulder	32
step (before handstand)	18	crash	34	height	30
upswing	17	touch down (on springboard)	29	tension	25

Teachers

Handstand		Squat vault		Underswing	
handstand	116	flight phase	96	sway	121
roll-out	111	**landing**	96	bar	76
arms/hands	80	arms/hands	87	legs	65
sway	64	**take-off (springboard)**	73	flight	50
stiffness (body)	51	vault	64	hip	47
uprise	41	legs	50	**landing**	47
back	41	sway	39	undershoot	46
legs	41	run-up	36	stiffness (body)	33
movement	33	spring board	22	feet	30
head	27	**stance**	20	body	29

FIGURE 7
Displays the most frequently mentioned nouns in written explanations of the judgments by trainers and teachers. The numbers indicate the total amount of entries in written explanations. Words in bold are expressions that refer to significant movement features from segmentation (see also Figure 2).

formation on the basis of event segmentation and written explanations of the judgments.

Concerning the first aim, we found significantly lower judgment accuracy in teachers compared to trainers. While this finding is in line with research on teachers' judgment accuracy in other domains (Südkamp et al., 2012), it is not trivial, given the fact that teachers received training on the tasks they judged and additionally prepared for the task they were required to accomplish, and the tasks are not difficult but part of the basic repertoire in gymnastics. Agreement on the ratings was significantly lower among teachers than among trainers. In general, teachers declared movement errors to be less severe compared to trainers.

Concerning the second aim, agreement about the temporal structuring of the tasks (squat vault and underswing) from event segmentation was significantly lower among teachers than among trainers. Trainers' segmentation behavior (i.e., button presses) referred to kinematic features that mostly indicated transitions between movement phases. Written responses from trainers, explaining the judgments, referred to these features more often than responses from trainers.

In the following, we discuss these results separately for each of the two aims, then discuss limitations of the study, and finally summarize the findings and their implications with regard to teacher education.

4.1. Teachers' ability to perform accurate judgments

There are two types of feedback students utilise when learning movements (Magill, 2001): intrinsic feedback refers to feedback that originates from the body's own sensors. Extrinsic feedback refers to feedback that originates from an external source, typically the teacher in a school setting. Accurate extrinsic feedback is necessary for motor learning of (complex) motor skills and for the consolidation of learned behavior (Leukel and Lundbye-Jensen, 2012; Leukel and Gollhofer, 2023). Accurate judgment is a prerequisite for providing accurate extrinsic feedback. In the current study, we observed lower judgment accuracy in teachers compared to trainers. Furthermore, ratings were less consistent among teachers than trainers. Indeed, Figure 7 visualizes the higher variability of teachers' ratings on individual students. What does this mean for PE? Are teachers' ratings deficient concerning their task to form accurate judgments in the classroom, and is this relevant? We argue that this is the case, and that it is relevant, and particularly refer to the inconsistency of judgments among teachers. Judgments on individual students did substantially vary between teachers, and because accurate judgment is a prerequisite for supportive feedback, a students would receive quite different feedback from different teachers. This raises concerns about the quality of extrinsic feedback necessary for learning (Magill, 2001; Leukel and Lundbye-Jensen, 2012; O'Brien et al., 2023). Furthermore, considering that the severity of movement errors may also be relevant for the grading, students would be graded quite differently from different teachers. Admittedly, PE covers a variety of tasks and not only tasks in gymnastics, and the goals are not just constrained to the development of physical competences. This means that not all judgments in PE have to be of high quality, nor can this be expected from teachers who typically spend only a fraction during their studies on gymnastics. However, we argue that if a task is considered important in PE, accurate judgment is a necessary ingredient for learning (Swinnen, 1996; Wolpert et al., 2011). We specifically consider the squat vault such a task because it is part of the basic repertoire in gymnastics, and considered relevant in PE and present in curricula from primary school to secondary school in many States in Germany.

Teachers' ratings are surely not as accurate as standardized kinematic measures. However, we refrain do conclude that teacher judgments should be replaced by standardized assessments tool (Seidel and Bös, 2012; Herrmann et al., 2016). We rather advocate that

teachers' judgment competences should be improved by training. As argued before, assessment tools are limited, with regard to focus (i.e., rather narrow, only a subset of motor abilities and skills are accessible), adaptability to individual needs, and ceiling and floor effects of the tests. Diagnostic competences of teachers are necessary because they overcome these limitations. Our results regarding the processes of judgment formation provide a starting point for elaborations on teacher training as will be discussed below.

We observed teachers to be milder in their judgments than trainers. Research on judgment accuracy in other subjects also shows that teachers usually overestimate their students (Ostermann et al., 2018; Oudman et al., 2018). This could be due to teachers' level of experience and/or due a personality trait. Teachers are less experienced than trainers with the tasks and thus might overlook movement errors. Concerning the personality trait, teachers might be generally milder in their judgments than trainers', because they not only focus on performance but also on social skills.

4.2. Cognitive processes of judgment formation

Investigating cognitive aspects of judgment formation is considered necessary for building a theoretical understanding of diagnostic judgments and deriving measures to improve diagnostic competences in teachers (Loibl et al., 2020). An important finding in the present study in this respect was that teachers had a less consistent concept of the temporal structuring of the tasks than trainers. Event boundaries, setting the temporal structure, have recently been shown to be important for memory formation and information retrieval (Michelmann et al., 2021). In the study of Michelmann et al. (2021), the time surrounding event boundaries was linked to information flows between cortex and hippocampus. The hippocampus is known for its role in memory formation and consolidation, and the area that is described as cortex in this study is linked to aspects of sensory (visual and auditory) integration and processing. According to these recently published findings, at event boundaries sensory inputs are likely compared to stored information. Thus, when judging tasks in gymnastics, at event boundaries the brain may perform comparisons of the target and the actual performance, separately for fundamental building blocks (i.e., different phases) of the observed act. This means that the brain may compare stored information of desired values of kinematic features with actual sensory (visual) values for each of the phases of the task separately, and the differences between desired and actual values indicate movement errors. It makes sense that this process does not cover the whole task but rather segregated parts, because this limits the amount of information that needs to be processed at once. Importantly, when trainers in the present study explained their ratings, the most frequently used nouns referenced to the kinematic features they focussed on during segmentation. This was also the case in teachers, but to a much lesser degree (half as much). This indicates that these features constitute the grounding of the ratings. Thus, according to these explanations, teachers may have more difficulties in (i) identifying kinematic features relevant for performance, and (ii) judging them appropriately.

Differences between trainers and teachers that concern the agreement of the temporal structuring were observed for squat vault and underswing, but not handstand. This could be due to the ambiguity of kinematic markers for handstand. For squat vault and underswing there are clear markers indicating salient changes in kinematics, like the feet leaving the springboard in squat vault indicating the beginning of the first flight phase. This might be the reason why, for handstand, event boundaries were set slightly differently from different subjects.

4.3. Limitations

This study has several methodological limitations: firstly, the button presses from segmentation do not clearly indicate to which aspects of the performance the subjects referred to. This is also due to a movement delay between the instant of the decision and the pressing of the button, and this delay varies between subjects (Norman and Komi, 1979; Kurz et al., 2019). The events in the videos that are referred to by the subjects are therefore not exactly traceable from the button presses.

Secondly, when analysing written explanations, we counted nouns/conjunct nouns and did not look for other word classes and combinations of other words except nouns. We did this on purpose because we were interested if subjects referred to movement features from event segmentation. In German language, movement features are expressed by (conjunct) nouns in combination with attributes (e.g., high *take-off*: hoher *Absprung*). It could well be that a thorough linguistic analysis would have brought up additional findings about the explanation of judgments. However, this was beyond the purpose of this study.

Thirdly, there is a risk that we might have overlooked semantically shared expressions pointing to movement features. This is particularly true for teachers. Teachers typically do not use technical lingo, in contrast to trainers (e.g., second flight phase in squat vault). They thus might have used various expressions referring to the same semantic content. We tried to account for this by searching for words with similar meanings, but there is still the risk that teachers used lingo we did not recognize as being similar.

Furthermore, there are some restrictions connected to more fundamental considerations: firstly, we cannot clearly define the level at which teachers' judgments are regarded as sufficient for students' development and grading. Statements about diagnostic competences in teachers were derived from comparisons to a reference group. Further empirical research is required to determine how the quality of a teachers' diagnostic feedback is coupled to students' learning.

Secondly, it is important to remember that accurate judgments are necessary for providing accurate feedback, but they are not sufficient. Improving teachers' diagnostic competences in the PE classroom does not necessarily mean that the students receive feedback that promotes learning. Providing feedback that promotes learning requires additional knowledge components (pedagogical content knowledge (Shulman, 1986), content knowledge, pedagogical knowledge) in addition to diagnostic information, which need to be flexibly applied according to the range of performance levels of the students (Hattie and Yates, 2014; Richartz et al., 2022). Indeed, a student who has difficulties to jump off the springboard requires very different feedback than a student who almost masters the squat vault.

Thirdly, the results of the present study are constrained to a limited number of tasks in PE, so findings only apply to a small portion within a broader range of tasks in sports. Hence, it cannot be concluded that PE teachers are poor judges in general.

Fourthly and finally, we did not measure pre-existing (declarative) knowledge about movement characteristics relevant for accurate

diagnoses in teachers. The extent of this knowledge might partly explain the interindividual differences we observed in the teachers.

4.4. Implications for teacher training

There is a broad consensus that diagnostics should be an integral part of PE with a strong aim to promote student learning and also teaching (Hay et al., 2015; O'Brien et al., 2023). Yet, teachers' practice does not meet this goal (Lorente-Catalán and Kirk, 2016; Moura et al., 2021). As a consequence, awareness of diagnostics in PE and promotion of diagnostic competences should be a fundamental part of teacher education (Ward et al., 2020; Moura et al., 2021; O'Brien et al., 2023). As mentioned before in this article, diagnostic judgments should not be constrained to standardized testing, but also need to be concerned with the assessment of (complex) motor skills which are part of the PE curriculum (Sacko et al., 2021). Accurate judgments on complex motor skills require knowledge about movement characteristics of the intended motor task, knowledge on typical student errors with respect to these characteristics, and extensive experience in observation (i.e., practical training in a real-life scenario) (Ward et al., 2020; Sacko et al., 2021). These aspects relate to the situational (cues) and person characteristics described in the framework of Loibl et al. (2020). According to the outcome of the present study, knowledge about movement characteristics and typical movement errors of the students are likely not sufficient for performing accurate judgments. Participants gained this knowledge and rehearsed it prior to the experiment but still showed insufficient performance. Performing judgments on motor skills in PE is a complex and a practical task involving many knowledge components which are declarative (e.g., knowledge about movement characteristics and typical movement errors) and procedural (e.g., knowledge about which position to take in the classroom for observing the students' performance, knowledge about where and when to look at the students while they perform the task) in nature. Accordingly, learning environments should acknowledge this complex nature of the task, including the procedural knowledge components. Instructional designs like the four-component instructional design model for training complex skills (4CID) (Van Merriënboer et al., 1992, 1997) do acknowledge these components, when they position the real-life scenario in the centre of the instructional design, and provide individual support in terms of knowledge and part-task practice. Hence, future studies may want to investigate how features of such an instructional design can effectively support the development of diagnostic competences in PE teachers.

5. Conclusion

In summary, in this study we found significant differences between trainers' and teachers' ratings on movement errors of tasks in gymnastics, with teachers' ratings being less accurate and consistent. The segmentation data indicated that the temporal structuring of the tasks was less consistent in teachers than trainers, and referred to kinematic features that are mostly linked to transitions between movement phases. In trainers, written explanations of the judgments contained these kinematic features from segmentation more often compared to teachers. We conclude from these results that diagnostic competences in teachers are insufficient and should be improved. According to the results from segmentation and written explanations, a preferable strategy for teacher education would be to focus on kinematic features relevant for performance in a practical teacher-training setting that resembles the real-life-scenario in the classroom, in which the different knowledge components (declarative and procedural) for making accurate judgments are integrated. Future studies might want to investigate the effectiveness of these learning environments on diagnostic competences in teachers.

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 humans were approved by the Ethics committee at the University of Freiburg, Germany. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

CL, TL, FB, and KL: conceptualization, project administration, writing – review, and editing. CL: data acquisition. CL, TL, and KL: data analysis, interpretation of the data, and writing – original draft. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank Lars Breuning, Sabine Karoß, and Fabian Vogt, who helped with the recruitment of the subjects.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1162499/full#supplementary-material

References

Abernethy, B., Gill, D. P., Parks, S. L., and Packer, S. T. (2001). Expertise and the perception of kinematic and situational probability information. *Perception* 30, 233–252. doi: 10.1068/p2872

Barrett, K. R. (1983). A hypothetical model of observing as a teaching Skill1. *J. Teach. Phys. Educ.* 3, 22–31. doi: 10.1123/jtpe.3.1.22

Bates, C., and Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educ. Psychol.* 21, 177–187. doi: 10.1080/01443410020043878

Baumert, J., and Kunter, M. (2013). "The COACTIV model of teachers' professional competence" in *Cognitive activation in the mathematics classroom and professional competence of teachers*. (New York: Springer), 25–48.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bläsing, B. E. (2015). Segmentation of dance movement: effects of expertise, visual familiarity, motor experience and music. *Front. Psychol.* 5:1500. doi: 10.3389/fpsyg.2014.01500

Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., and Fischer, F. (2020). Facilitating diagnostic competences in higher education—a meta-analysis in medical and teacher education. *Educ. Psychol. Rev.* 32, 157–196. doi: 10.1007/s10648-019-09492-2

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychol. Bull.* 52, 177–193. doi: 10.1037/h0044919

Dudley, D., Mackenzie, E., Van Bergen, P., Cairney, J., and Barnett, L. (2022). What drives quality physical education? A systematic review and meta-analysis of learning and development effects from physical education-based interventions. *Front. Psychol.* 13:799330. doi: 10.3389/fpsyg.2022.799330

Farana, R., and Vaverka, F. (2012). The effect of biomechanical variables on the assessment of vaulting in top-level artistic female gymnasts in world cup competitions. *Acta Gymnica* 42, 49–57. doi: 10.5507/ag.2012.012

Ferrari, I., Kühnis, J., Bretz, K., and Herrmann, C. (2022). "Diagnostische Kompetenz der Lehrpersonen und deren Bedeutung für die Förderung motorischer Basiskompetenzen", in *Narrative Zwischen Wissen Und Können*, Eds. Roland M. and Claus K (Baden-Baden: Academia), 175–194.

Hattie, J. A. C., and Yates, G. C. R. (2014). "Using feedback to promote learning", in *Applying science of learning in education: Infusing psychological science into the curriculum*. eds. V. A. Benassi, C. E. Overson and C. M. Hakala, Society for the Teaching of Psychology, 45–58., 45–58.

Hay, P. (2006). "Assessment for learning in physical education", in *The handbook of physical education*. eds. D. Kirk, D. Macdonald and M. O'Sullivan, (London, UK: SAGE), 312–25.

Hay, P., Tinning, R., and Engstrom, C. (2015). Assessment as pedagogy: a consideration of pedagogical work and the preparation of kinesiology professionals. *Phys. Educ. Sport Pedagog.* 20, 31–44. doi: 10.1080/17408989.2013.788145

Heinen, T. (2015). *Advances in visual perception research*. Milton Park, UKMilton Park Abingdon, Oxfordshire United Kingdom: Nova Science Publishers Inc.

Helmke, A., and Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teach. Teach. Educ.* 3, 91–98. doi: 10.1016/0742-051X(87)90010-2

Herrmann, C., Gerlach, E., and Seelig, H. (2016). Motorische Basiskompetenzen in der Grundschule. *Sportwissenschaft* 46, 60–73. doi: 10.1007/s12662-015-0378-8

Hoge, R. D., and Coladarci, T. (1959). Teacher-based judgments of academic achievement: a review of literature. *Rev. Educ. Res.* 59, 297–313. doi: 10.3102/00346543059003297

Hong, Y., and Bartlett, R. (2008). *Routledge handbook of biomechanics and human movement science* Routledge.

Kredel, R., Vater, C., Klostermann, A., and Hossner, E.-J. (2017). Eye-tracking technology and the dynamics of natural gaze behavior in sports: a systematic review of 40 years of research. *Front. Psychol.* 8:1845. doi: 10.3389/fpsyg.2017.01845

Krippendorff, K. (2016). *Bootstrapping distributions for Krippendorff's alpha*. Available at: https://www.asc.upenn.edu/sites/default/files/2021-03/Algorithm%20for%20Bootstrapping%20a%20Distribution%20of%20Alpha.pdf

Kurby, C. A., and Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends Cogn. Sci.* 12, 72–79. doi: 10.1016/j.tics.2007.11.004

Kurz, A., Xu, W., Wiegel, P., Leukel, C., and Baker, S. N. (2019). Non-invasive assessment of superficial and deep layer circuits in human motor cortex. *J. Physiol.* 597, 2975–2991. doi: 10.1113/JP277849

Leuders, T., and Loibl, K. (2021). Beyond subject specificity—student and teacher thinking as sources of specificity in teacher diagnostic judgments. *RISTAL* 4, 60–70. doi: 10.23770/RT1842

Leuders, T., Loibl, K., Sommerhoff, D., Herppich, S., and Praetorius, A.-K. (2022). Toward an overarching framework for systematizing research perspectives on diagnostic thinking and practice. *J. Math.-Didakt.* 43, 13–38. doi: 10.1007/s13138-022-00199-6

Leukel, C., and Gollhofer, A. (2023). Applying augmented feedback in basketball training facilitates improvements in jumping performance: augmented feedback in basketball. *Eur. J. Sport Sci.*, 1–16. doi: 10.1080/17461391.2022.2041732

Leukel, C., and Lundbye-Jensen, J. (2012). "The role of augmented feedback in human motor learning" in *Routledge handbook of motor control and motor learning*. eds. A. Gollhofer, W. Taube and J. B. Nielsen (Milton Park, Abingdon, Oxfordshire, United Kingdom: Routledge).

Loibl, K., Leuders, T., and Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (Dia CoM). *Teach. Teach. Educ.* 91:103059. doi: 10.1016/j.tate.2020.103059

López-Pastor, V. M., Kirk, D., Lorente-Catalán, E., Mac Phail, A., and Macdonald, D. (2013). Alternative assessment in physical education: a review of international literature. *Sport Educ. Soc.* 18, 57–76. doi: 10.1080/13573322.2012.713860

Lorente-Catalán, E., and Kirk, D. (2016). Student teachers' understanding and application of assessment for learning during a physical education teacher education course. *Eur. Phys. Educ. Rev.* 22, 65–81. doi: 10.1177/1356336X15590352

Luis del Campo, V., and Espada Gracia, I. (2018). Exploring visual patterns and judgments predicated on role specificity: case studies of expertise in gymnastics. *Curr. Psychol.* 37, 934–941. doi: 10.1007/s12144-017-9572-1

Mack, M. (2020). Exploring cognitive and perceptual judgment processes in gymnastics using essential kinematics information. *Adv. Cogn. Psychol.* 16, 34–44. doi: 10.5709/acp-0282-7

Magill, R. A. (2001). "Augmented feedback in motor skill acquisition" in *Handbook of sport psychology*. eds. R. N. Singer, H. A. Hausenblas and C. M. Janelle (Hoboken, New Jersey: John Wiley & Sons)

Mann, D. T., Williams, A. M., Ward, P., and Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: a meta-analysis. *J. Sport Exerc. Psychol.* 29, 457–478. doi: 10.1123/jsep.29.4.457

Mechling, H., and Munzert, J. (2004). Handbuch Bewegungswissenschaft—Bewegungslehre. *Hofmann* 34, 231–236. doi: 10.1007/BF03176404

Michelmann, S., Price, A. R., Aubrey, B., Strauss, C. K., Doyle, W. K., Friedman, D., et al. (2021). Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nat. Commun.* 12, 1–15. doi: 10.1038/s41467-021-25376-y

Moura, A., Graça, A., Mac Phail, A., and Batista, P. (2021). Aligning the principles of assessment for learning to learning in physical education: a review of literature. *Phys. Educ. Sport Pedagog.* 26, 388–401. doi: 10.1080/17408989.2020.1834528

Newberry, K. M., Feller, D. P., and Bailey, H. R. (2021). Influences of domain knowledge on segmentation and memory. *Mem. Cogn.* 49, 660–674. doi: 10.3758/s13421-020-01118-1

Niederkofler, B., Herrmann, C., and Amesberger, G. (2018). Diagnosekompetenz von Sportlehrkräften—Semiformelle Diagnose von motorischen Basiskompetenzen. *Zeitschrift für sportpädagogische Forschung* 6, 72–96. doi: 10.5771/2196-5218-2018-2-72

Norman, R. W., and Komi, P. V. (1979). Electromechanical delay in skeletal muscle under normal movement conditions. *Acta Physiol. Scand.* 106, 241–248. doi: 10.1111/j.1748-1716.1979.tb06394.x

O'Brien, W., Philpott, C., Lester, D., Belton, S., Duncan, M. J., Donovan, B., et al. (2023). Motor competence assessment in physical education–convergent validity between fundamental movement skills and functional movement assessments in adolescence. *Phys. Educ. Sport Pedagog.* 28, 306–319. doi: 10.1080/17408989.2021.1990241

Ostermann, A., Leuders, T., and Nückles, M. (2018). Improving the judgment of task difficulties: prospective teachers' diagnostic competence in the area of functions and graphs. *J. Math. Teach. Educ.* 21, 579–605. doi: 10.1007/s10857-017-9369-z

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., and van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teach. Teach. Educ.* 76, 214–226. doi: 10.1016/j.tate.2018.02.007

Placek, J. H. (1983). Conceptions of success in teaching: busy, happy and good. *Teach. Phys. Educ.* 14, 46–56.

Richartz, A., Kohake, K., and Maier, J. (2022). "Pädagogische Qualität des Trainings im Kinder-und Jugendsport" in *Materialien für die Lehre im Gerätturnen: Bd. 3. Gerätturnen 2.0 (1. Aufl.). e-gymnastics*. ed. F. Bessi Germany: Freiburg.

Rink, J. E. (2013). Measuring teacher effectiveness in physical education. *Res. Q. Exerc. Sport* 84, 407–418. doi: 10.1080/02701367.2013.844018

Sacko, R. S., Utesch, T., Cordovil, R., De Meester, A., Ferkel, R., True, L., et al. (2021). Developmental sequences for observing and assessing forceful kicking. *Eur. Phys. Educ. Rev.* 27, 493–511. doi: 10.1177/1356336X20962134

Seidel, I., and Bös, K. (2012). Chancen und Nutzen motorischer Diagnostik im Schulsport am Beispiel des DMT 6–18 Sportunterricht, Schorndorf, 61, 228–233.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educ. Res.* 15, 4–14. doi: 10.3102/0013189X015002004

Stadler, W., Kraft, V. S., Beer, R., Hermsdörfer, J., and Ishihara, M. (2021). Shared representations in athletes: segmenting action sequences from taekwondo reveals implicit agreement. *Front. Psychol.* 12, –733896. doi: 10.3389/fpsyg.2021.733896

Stolarova, M., Wolf, C., Rinker, T., and Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Front. Psychol.* 5:509. doi: 10.3389/fpsyg.2014.00509

Südkamp, A., Kaiser, J., and Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: a meta-analysis. *J. Educ. Psychol.* 104, 743–762. doi: 10.1037/a0027627

Swinnen, S. P. (1996). "Information feedback for motor skill learning: a review" in *Advances in motor learning and control*. ed. N. Zelaznik (Champaign, IL 61820: Human Kinetics), 37–66.

Takei, Y. (1998). Three-dimensional analysis of handspring with full turn vault: deterministic model, coaches' beliefs, and judges' scores. *J. Appl. Biomech.* 14, 190–210. doi: 10.1123/jab.14.2.190

Tolgfors, B. (2018). Different versions of assessment for learning in the subject of physical education. *Phys. Educ. Sport Pedagog.* 23, 311–327. doi: 10.1080/17408989.2018.1429589

Urhahne, D., and Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educ. Res. Rev.* 32:100374. doi: 10.1016/j.edurev.2020.100374

van der Mars, H., McNamee, J., and Timken, G. (2018). Physical education meets teacher evaluation: supporting physical educators in formal assessment of student learning outcomes. *Phys. Educ.* 75, 582–616. doi: 10.18666/TPE-2018-V75-I4-8471

Van Merriënboer, J. J. (1997). *Training complex cognitive skills: a four-component instructional design model for technical training*. Englewood Cliffs, NJ: Educational Technology.

Van Merriënboer, J. J., Jelsma, O., and Paas, F. G. (1992). Training for reflective expertise: a four-component instructional design model for complex cognitive skills. *Educ. Technol. Res. Dev.* 40, 23–43. doi: 10.1007/BF02297047

Ward, P., Ayvazo, S., Dervent, F., Iserbyt, P., Kim, I., and Li, W. (2020). Skill analysis for teachers: considerations for physical education teacher education. *J. Phys. Educ. Recr. Dance* 92, 15–21. doi: 10.1080/07303084.2020.1853635

Wolpert, D. M., Diedrichsen, J., and Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nat. Rev. Neurosci.* 12, 739–751. doi: 10.1038/nrn3112

Zacks, J. M., Kumar, S., Abrams, R. A., and Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition* 112, 201–216. doi: 10.1016/j.cognition.2009.03.007

Zacks, J. M., and Swallow, K. M. (2007). Event segmentation. *Curr. Dir. Psychol. Sci.* 16, 80–84. doi: 10.1111/j.1467-8721.2007.00480.x

frontiers | Frontiers in Education

TYPE Methods
PUBLISHED 20 November 2023
DOI 10.3389/feduc.2023.1271335

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Dennis Arias-Chávez,
Continental University, Peru
Luis Gómez,
University of the Bío Bío, Chile

*CORRESPONDENCE
Marcela Romero Jeldres
✉ marcela.romero@umce.cl

RECEIVED 02 August 2023
ACCEPTED 31 October 2023
PUBLISHED 20 November 2023

CITATION
Romero Jeldres M, Díaz Costa E and Faouzi Nadim T (2023) A review of Lawshe's method for calculating content validity in the social sciences.
Front. Educ. 8:1271335.
doi: 10.3389/feduc.2023.1271335

COPYRIGHT
© 2023 Romero Jeldres, Díaz Costa and Faouzi Nadim. 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.

A review of Lawshe's method for calculating content validity in the social sciences

Marcela Romero Jeldres[1]*, Elisabet Díaz Costa[2] and Tarik Faouzi Nadim[3]

[1]Department of Pedagogical Training, Metropolitan University of Educational Sciences, Santiago, Chile, [2]Department of Early Childhood Education, Metropolitan University of Educational Sciences, Santiago, Chile, [3]Department of Mathematics and Computer Science, Faculty of Science, University of Santiago de Chile, Santiago, Chile

This study aimed to show the usefulness of Lawshe's method (1975) in investigating the content validity of measurement instruments under the strategy of expert judgment. The research reviewed the historical use of Lawshe's method in the social sciences and analyzed the main criticisms of this method using mathematical hypotheses. Subsequently, we experimented with an instrument designed to determine the pedagogical skills possessed by students undertaking initial teacher training in Chile. The results showed that in Lawshe's proposal, it is essential to highlight the need to reconsider the sum of all Content Validity Ratio (CVR) indices for the calculation of the context validity index and not only the indices that exceed the critical Content Validity Index (CVI) given the greater power of discrimination presented by the latter alternative.

KEYWORDS

instrument development, content validity index, content validity ratio, panel of experts, Bayesian methods

1 Introduction

In research, especially in education, measurement instruments are valuable tools to generate and structure complex constructs and convert them into analyzable parameters (Juárez-Hernández and Tobón, 2018). Implementing these instruments entails using statistical techniques that allow their validity to be demonstrated in the measurement field. In this sense, scientific organizations recommend considering the following sources of validity: content, response processes, internal structure, relations with other variables, and evaluation consequences (Eignor, 2013; Plake and Wise, 2014). Measuring the correspondence between the content of the items and the evaluated content is very relevant for evaluating content validity.

The literature identifies two approaches for addressing the content analysis of a measurement instrument: the first is related to the methods based on expert judgment; the second is related to the methods derived from the application of measuring instruments (Pedrosa et al., 2014; Urrutia et al., 2014). These methods aim to collect evidence on two sources of validity: the definition of the items' domain (representativeness) and the adequacy of the content (relevance). This study opted to focus on methods based on expert judgment.

According to Urrutia et al. (2014), before carrying out content analysis under the expert judgment strategy, the researcher must resolve two critical issues: first, determine what can be measured, and second, define the number and characteristics of the experts who will participate in evaluating the relevance of the instrument items. Moreover, the variability of the

number of participants during the expert judgment, their suitability concerning the study's objective, their work activity, and their geographical area of origin. However, the literature on measurement instruments indicates that when defining the number of experts for a panel, the researcher must predict the type of statistical analysis that will be carried out with the responses obtained in such a way that the number of selected participants is equal to or greater than the minimum number to the minimum number of judges for the statistical test to be valid (Tristán-López, 2008).

The importance of defining statistical parameters (measures) in statistics (Statistical Measure)lies in the fact that the qualitative assessment of items is not sufficient —to determine the degree of agreement of evaluators on the suitability of an item (Sireci, 1998).

1.1 Critical analysis of Lawshe's proposal

In the search for indices to calculate inter-judge agreement, we reviewed different methods used in the social sciences to calculate content validity. In this review, the proposal of Lawshe (1975). This strategy considers a panel of judges with expertise in the construct of the instrument, and each specialist individually evaluates the items associated with said construct. Lawshe (1975) suggested that under sociological principles, the minimum inter-judge agreement should be 50%, and the use of two indices: the content validity ratio (CVR), which measures the agreement of the panelists on an item, and the CVI, which presents the average of the CVR that constitutes the final instrument. The CVR can be presented as follows:

$$CVR = \frac{n_e - n_{ne}}{N} \quad (1)$$

where ne is the number of panelists in agreement and n_{ne}, is the number of panelists in disagreement.

A CVR index is considered acceptable depending on the level of agreement of the panelists regarding an item. Lawshe (1975) presented a table of the critical values of the CVR index according to the number of panelists. For the research exemplified below, it was necessary to find an index that applied to seven judges who participated in the process. For Lawshe, the critical CVR of a 7-member panel of experts should be equal to or greater than 0.75, a value that would allow the Interjudge agreement to be considered statistically valid.

Polit et al. (2007) suggested a critical CVR value equal to.78 for three or more panelists. It represents an advance concerning Lawshe, who proposed a critical CVR calculation with at least five panelists. In any case, the proposal by Polit et al. does not apply to the social sciences because, although it supports a smaller number of panelists, the recommended critical CVR is too low for panels of five or fewer experts, thus losing discriminant validity. Likewise, in panels with six or more members, it is objectionable for the critical CVR value to remain very close to that defined for three panelists, as it would contradict studies on the indirectly proportional relation between the number of rating experts on a panel and the critical value of CVR (Tristán-López, 2008; Wilson et al., 2012; Ayre and Scally, 2014).

Tristán-López (2008) pointed out that Lawshe's formula does not apply to panels with fewer than five members and further generated a critical CVR table, revealing, unlike Lawshe, the statistical test he used (Chi-squared = $\chi^2(\alpha = 0.1, gl = 1)$), while Lawshe had used the value of α =0. 05 as significance level without mentioning the test used, A question that, to date, represents the most significant criticism of the work of the latter author. Although Tristán-López's proposal represents an advance concerning the previous ones, and the use of the statistic used in the critical CVR table for each number of evaluating experts is made explicit, the increase of the admissible error to the value 1 (α =0.1) is not justified.

As an effect of the above, we observe that the higher the value of α, the better the probability of calculating the critical values for a smaller number of panelists, which is beneficial from the point of view of reducing the number of panelists but not from the perspective of the error of permissible significance and, therefore, of the level of exigency of the proof. On the other hand, according to Ayre and Scally (2014), using the Tristan-López chi-square test is inadequate, considering that the data with which the RCV is calculated would have a binomial distribution. Consequently, the statistical modalization of Tristán-López (2008) to generate the critical CVR table is inappropriate.

Wilson et al. (2012) proposed a new critical CVR table using the normal approximation for a binomial distribution under the assumption of the central limit theorem, in which the rating expert's answers do not interact with each other (each rating expert is an independent variable). The complexity of this interesting proposal for its application in social sciences lies in the fact that this approach is valid only for a considerable size of panelists, which corresponds to another of the assumptions of the central limit theorem and represents a great difficulty for the search for expert judges.

Ayre and Scally (2014) proposed a new critical CVR table using the exact binomial test (EBT). Unlike the chi-squared test used by Tristán-López (2008) and the normal approximation of Wilson et al. (2012), the EBT is an appropriate alternative to the model proposed by Lawshe for the following reasons. Lawshe proposed two response categories (essential and non-essential) such that the data distribution of each expert rater corresponds to a Bernoulli distribution with parameter $p = 0.5$. If, in addition, each expert rater's responses are considered independent variables. The criteria of an exact binomial distribution are met, with parameter $p = 0.05$ and N, the number of expert raters.

Suppose we consider an instrument composed of "n" items evaluated by "N" panelists with a dichotomous scale ("essential" or "non-essential"). Then, modeling the problem statistically and finding the critical value of panelists in agreement that allows an item to be considered essential using the CVR_{cr} index calculated using equation (1), we considered the null hypothesis.

$$H_0: ne \leq \frac{N}{2'} \quad (2)$$

where n_e is the number of panelists in agreement, with a significance level equal to 0.05.

The null hypothesis is rejected if the probability of the "number of panelists in agreement is greater than the minimum required number of panelists in agreement with the essential modality ncr" is greater than 0.95.

$$P(n_e > n_{cr}) > 0.95,$$

where n_{cr} is the minimum required number of panelists who agree with the "essential" modality.

Ayre and Scally (2014) calculated the critical values table for each number of panelists using STATA. Using the R program, we compared the three results (Lawshe, 1975; Tristán-López, 2008; Ayre and Scally, 2014).

As shown in Figure 1, the three methods for nine or fewer panelists coincide with the value calculated for critical CVR, but the values differ for ten or more panelists. Given that the method based on the binomial test is more precise than the chi-squared test, proposing an alternative to the CVR index would not change the decision-making under the exact binomial method. However, the Tristán-López (2008) method can be used as an alternative to calculating the critical CVR index when the number of panelists is equal to four (see Table 1).

Baghestani et al. (2019) proposed a new method to calculate the critical CVR based on Bayesian statistics. They replaced the null hypothesis given in (1) with.

$$H_0: p \leq 0.5, vs\, H_1 : p > 0.5,$$

where p is the realization of a random variable X, whose distribution is unknown. The prior misinformation on the distribution of X was studied by Jeffreys (1935) and Berger (2013), who considered that the previous distribution of X is beta with parameters α *and* β, denoted by $X \sim Beta(\alpha,\beta)$. Given the above, the posterior density function of X is

$$\pi(X|Ne) \propto f(X) * g(Ne|X),$$

where $f(X; \alpha, \beta) = \text{dBeta}(\alpha, \beta)$ is the density function of the variable X, and $g(\text{Ne}|\text{X}) = \text{dbinom}(p, N)$ is the density function of Ne given the variable X.

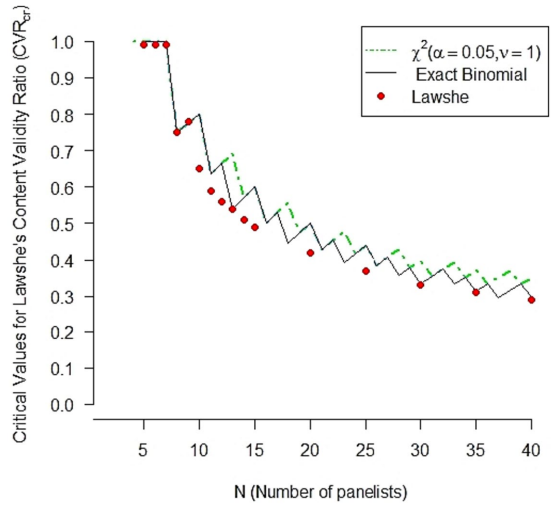


FIGURE 1
Comparison of Critical CVR Between Chi-squared and Exact Binomial Methods The graph represents the critical values for the CVR index and their variation based on the number of panelists comparing the proposals of Lawshe (1975), Tristán-López (2008), and Ayre and Scally (2014).

Thus, the posterior probability of hypothesis H_0 is

$$P(X \leq .5|Ne) = \int_0^{.5} f(p) * g(Ne|p)\, dp$$

$$= \int_0^{.5} f(p, Ne+\alpha, N-Ne+\beta)\, dp.$$

$$= \int_0^{.5} p^{Ne+\alpha}(1-p)^{N-Ne+\beta}\, dp.$$

At a significance level of X, the null hypothesis H_0 is rejected if $\int_0^{.5} p^{Ne+\alpha}(1-p)^{N-Ne+\beta}\, dp < .05$. Then, the parameter Ne is determined as the minimum value to reject the null hypothesis considering the model of Berger (2013), with $\alpha = \beta = 1$.

Figure 2 shows that with the method developed by Baghestani et al. (2019), the number of judges can be reduced while still obtaining acceptable critical CVR results. This fact is a significant advantage, given the difficulty in obtaining the required judges. However, this conclusion is only possible after reviewing the calculations carried out by Baghestani et al. (2019), which originated in discovering discrepancies regarding using the "pbinom" function of the program's stats package.

Compared with the exact binomial method, using Bayesian statistics to calculate the critical CVR helps reduce the number of panelists. This issue became evident after correcting the calculation error in Baghestani et al. (2019). In addition, these alternative shares the advantage of the chi-squared test method to calculate critical CVR (Tristán-López, 2008), where four panelists can evaluate an instrument.

Lawshe (1975) asserted that the CVI value is dependent on the number of panelists and that the following equation would represent it:

$$CVI_{accepted} = \sum_{j=0}^{m} \frac{CVR_j}{m},$$

Where m is the number of items whose CVR values exceed the critical CVR values, and CVR_j is the value of the CVR index of the accepted item "j": the number of questions that obtained a CVR greater than the critical CVR.

Lawshe's criterion can be considered an exaggeration in the sense that, according to the Lawshe model, the value of the CVI of the complete instrument (with accepted and non-accepted items) will not exceed, in any case, the CVI_{accepted} (calculated from the average of the CVR values for each accepted article). For this reason, we consider the critical value CVI_{cr} de Tilden et al. (1990), which suggests a value greater than 0.7 for the research outlined here.

However, unlike the CVR index, the critical CVI value, which allows accepting or rejecting the content validity of the total instrument, differs between authors. Tilden et al. (1990) suggested that this index is satisfactory, starting at 0.7, whereas Davis (1992) proposed a value of 0.8. The critical CVI value of 0.5 was proposed by

TABLE 1 Comparison of the Critical Content Validity Ratio (CVR) according to proposals from different authors.

N	CVR_{cr}	CVR_{cr}	CVR_{cr}	CVR_{cr}
	Exact binomial	Chi square	Lawshe	Bayesian statistics
2	–	–	–	–
3	–	–	–	–
4	–	1.00	–	1.00
5	1.00	1.00	0.99	1.00
6	1.00	1.00	0.99	1.00
7	1.00	1.00	0.99	0.71
8	0.75	0.75	0.75	0.75
9	0.78	0.78	0.78	0.78
10	0.80	0.80	0.65	0.60
11	0.64	0.64	0.59	0.64
12	0.67	0.67	0.56	0.50
13	0.54	0.69	0.54	0.54
14	0.57	0.57	0.51	0.57
15	0.60	0.60	0.49	0.47
16	0.50	0.50	–	0.50
17	0.53	0.53	–	0.41
18	0.44	0.56	–	0.44
19	0.47	0.47	–	0.47
20	0.50	0.50	0.42	0.40
21	0.43	0.43	–	0.43
22	0.45	0.45	–	0.36
23	0.39	0.48	–	0.39
24	0.42	0.42	–	0.42
25	0.44	0.44	0.37	0.36
26	0.38	0.38	–	0.38
27	0.41	0.41	–	0.33
28	0.36	0.43	–	0.36
29	0.38	0.38	–	0.31
30	0.33	0.40	0.33	0.33
31	0.35	0.35	–	0.35
32	0.38	0.38	–	0.31
33	0.33	0.39	–	0.33
34	0.35	0.35	–	0.29
35	0.31	0.37	0.31	0.31
36	0.33	0.33	–	0.28
37	0.30	0.35	–	0.30
38	0.32	0.37	–	0.32
39	0.33	0.33	–	0.28
40	0.30	0.35	0.29	0.30

The table represents the calculation of the critical values of the CVR index at a confidence level of.05 and their variation based on the number of panelists comparing different methods: exact binomial (Ayre and Scally, 2014), chi-squared (Tristán-López, 2008), Lawshe's (1975), and Bayesian statistics (Baghestani et al., 2019).

Tristán-López (2008). The key question is which of the two indices proposed by Lawshe —CVR and CVI— provides better and more information on the content validity of an instrument? Gilbert and

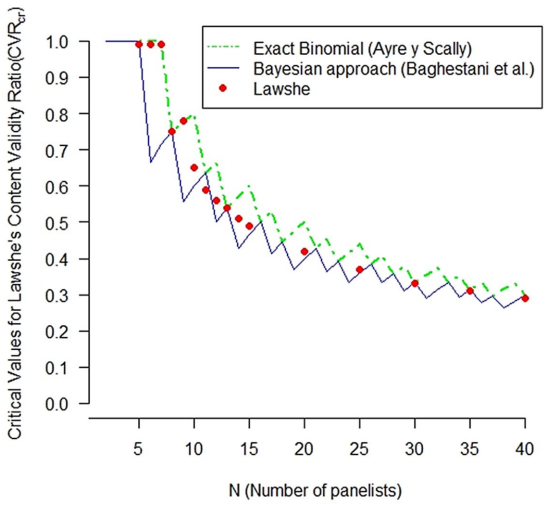


FIGURE 2
Comparison of Critical CVR Between the Exact Binomial and Bayesian Methods The graph represents the critical values for the CVR index and their variation based on the number of panelists comparing the proposals of Ayre and Scally (2014) and Baghestani et al. (2019).

Prion (2016) pointed out that the choice of one index over another depends on the orientation of the study. Based on the above and given that the general objective of our study was to analyze inter-judge agreement against a set of items to interpret their theoretical perspective and improve the instrument, we opted to use the CVI index.

The index used to calculate the CVI (Lawshe, 1975), although less demanding because it considers only the accepted values of the CVR, allows the maintenance of a vision of the relevance and representativeness of the items, mainly because this stage was combined with the analysis of the validity of the metric characteristics of the instrument, called construct validity, where the reliability of the items and the unidimensionality of the factors formed by the items are statistically verified through confirmatory factor analysis (CFA) (Díaz Costa et al., 2015; Romero-Jeldres and Faouzi, 2018). Consequently, and based on the fact that the CVI calculated for the instrument was more significant than 0.7, it was decided to maintain the total of the items elaborated for each dimension of the theoretical construct to which they referred.

Romero-Jeldres and Faouzi (2018) demonstrated that inter-judge agreement is complex, given the cultural elements that mediate the object of study. In the case of pedagogical skills for the exercise of teaching, each country develops different socio-political constructs for their understanding and evaluation, for which an essential adjustment of the Chilean evaluators to the regulatory framework of that country has been observed, in comparison with German evaluators who have shown differences regarding the theoretical frameworks supporting their vision of pedagogical competences for professional practice.

In other words, the CFA considered all the instrument items to verify instrument dimensionality, analyzing the data through various indices, namely, the Joreskög coefficient (Joreskög's rho), variance extracted index, factorial contribution, and chi-squared statistic divided by the degree of freedom. By combining the strategies for

calculating the CVI with the confirmatory factor analysis (CFA) strategy, we reduced the bias of the item evaluators (Juárez-Hernández and Tobón, 2018; Ventura-León, 2019).

2 Methodology

The present study was preceded by another research that defined the items associated with the construct to be measured (see variable operationalization Appendices A in Supplementary material) by Romero-Jeldres and Faouzi (2018), The conditions for performing a content validity analysis described by Guion (1977) were partially assured:

1. The domain content had to be rooted in behavior with a generally accepted meaning.
2. The domain content had to be relevant to the measurement objectives.
3. The domain content had to be adequately sampled.
4. Qualified evaluators must have agreed that the domain had been adequately sampled.
5. The content of the responses must have been reliably observed and evaluated.

Therefore, in addition to deciding the type of index to use for content validation, we defined the number of evaluators necessary to perform the validation. This question always turns out to be complex owing to the laboriousness of the task for the evaluators and the low recognition of this type of contribution to the academe. With these difficulties, recruiting a certain number of evaluators with expertise in the research topic is always a challenge.

For this purpose, we adopted expert judgment, defined as an informed opinion of people with experience in the subject and recognized as qualified experts able to provide information, evidence, judgments, and assessments (Cabero et al., 2013). Several expert judgment methods are available, varying in whether the evaluation is done individually or in a group. In all cases, the research problem determines the profile of the specialists. Therefore, defining the attributes of the possible expert persons is a prerequisite, considering as basic requirements having a background, experience, and disposition toward the topic, as well as being willing to review their initial judgment in the development of the study (López-Gómez (2018); Moreno López et al., 2022).

For the present study, we accounted for two additional criteria: the specificity of the content of the object of study and available resources. Akins et al. (2005) suggested that small panels must have at least seven experts to maintain the representation of the information obtained. However, to maintain fairness at the national level, we opted to identify 14 evaluators who met the study criteria in the regions where the sample was collected. We contacted the experts personally and then by email to provide them with all the information and documents needed for the evaluation. The recruitment of the participants was sustained until eight qualified panelists confirmed their willingness to participate. The characterization of the group of experts was highly relevant to the categorical development of the instrument. Table 2 shows the attributes considered.

2.1 Analysis

Based on the Bayesian statistical strategy Baghestani et al., 2019, we calculated the CVR index for each instrument item to define their relevance concerning the purposes of our study. We calculated the CVI for the instrument using Lawshe's model. This value was compared with the critical value Tilden et al. (1990) defined. This criterion was chosen for the convenience of maintaining all the items. For this purpose, the following equation was used:

$$CVI = \frac{1}{n}\sum_{j=1}^{n} CVR_j$$

where n is the total number of items.

According to Pedrosa et al. (2014), the CVR calculation represents a problem when half of the experts indicate an item as relevant and the other half as irrelevant. Concerning this criticism, we showed that having a certain number of experts makes it possible to define a critical value of CVR = 0, with which it is possible to affirm that even a value close to 0 can be admissible under certain conditions.

3 Results

About the content validation of the items, we affirm that based on the CVR index of the Lawshe (1975) method and using Tables 1, 56 of the 71 items exceeded 0.71. The calculated CVI value, including all instrument items, was 0.77, which is considered acceptable by Tilden et al. (1990), according to their CVI critical value proposition of 0.7. To verify that from a certain number of experts, a critical value of CVR = 0 can be defined, we implement the following:

$$CVR_{cr} = 0. \tag{4}$$

Using Table 1 and Equation (1), we found that CVR_{cr} was a decreasing positive function concerning N; thus, critical CVR must be strictly positive for any n_e greater than n_{ne}. Equation (4) was obtained using the following (Armitage et al., 2008; Wilson et al., 2012; Ayre and Scally, 2014).

$$CVR_{cr} = \frac{1 + z\sqrt{N}}{N}$$

N is extensive, $N = n_e + n_{ne}$, and the variable z has a standard normal distribution.

Thus, we could formulate the following:

$$\lim_{n_e \to \infty} CVR_{cr} = \lim_{n_e \to \infty}\left(\frac{1}{N} + \frac{z}{\sqrt{N}}\right) = 0.$$

Subsequently, when $n_e = \frac{N}{2} + \varepsilon$, where ε is a fixed integer value greater than 0, the value of CVR approaches 0 as N approaches infinity. Therefore, an item with a CVR value close to 0 would be acceptable if there are many panelists.

TABLE 2 Characterization of panel judges.

Judges	Age (years)	Sex	Title	Postgraduate		Participants	Territorial extension that reaches
				Master	Doctorate		
J1	65	Female	Basic general education teacher	Special education	Didactics and organization of educational institutions	Associate professor, researcher in teaching performance and initial teacher training. CRUCH University	Santiago
J2	34	Female	Secondary education state teacher in history and social sciences	Education, Curriculum, and Evaluation	Education	Coordinator of the Curriculum Design and Updating Area at CUECH University. Academic management researcher in university education Pedagogical excellence in H and G Award.	Santiago
J3	43	Female	State teacher in Spanish	Education, Curriculum, and Evaluation	Education	Executive Secretary Educational Foundation	National
J4	53	Female	Biology Teacher	Educational Administration	Education and society	Associate professor researcher in initial teacher training, reflective practice, professional development, and learning. Council of Rectors of Chilean Universities, CRUCH,	Concepción
J5	35	Male	Secondary education teacher in history, geography, and social sciences	Ethnographic research, anthropological theory, and intercultural relations	Social anthropology	Associate Professor-Researcher Coordinator of Curriculum Design and Continuous Improvement. Expert in Interculturality. Council of Rectors of Chilean Universities, CRUCH.	Antofagasta
J6	48	Male	Religion teacher in basic and middle education	Educational Sciences, m/c Curriculum	© Doctor of Educational Sciences.	Associate Professor- Researcher in Strategic Diagnosis for the strengthening of pedagogies. Council of Rectors of Chilean Universities, CRUCH,	Antofagasta
J7	54	Male	State teacher in music education	Educational Sciences, m/c Curriculum	None	Academic and academic director Master's Program in Management and Leadership Curriculum Development and Management Coordinator. Private university	Santiago

Source: Own elaboration.

4 Discussion and conclusions

The concept of content validity has undergone multiple transformations, but its essence has remained stable since its origin. Pedrosa et al. (2014) reviewed the conceptions developed about content validity, going through perspectives that considered it only for one type of test and others that considered it the basis of construct validity.

Although Lawshe's (1975) method for analyzing the content validity of a measurement instrument presents numerous advantages over alternatives for these purposes (Divayana et al., 2020), it needs more critical values for the reliability of the CVR calculation and the acceptance of the CVI. Thus, researchers using this method have been compelled to decide between several proposals from different authors.

The present work opted for the method of Baghestani et al. (2019) as an appropriate proposal for variables of unknown distribution. Meanwhile, and about Lawshe's proposal, it is essential to highlight the need to reconsider the sum of all the CVR indices for calculating the CVI—and not only the indices that exceed the critical CVR—given the greater power of discrimination presented by this last alternative.

We also noted the need to consider mixed methods to provide greater veracity to the content validation process. In this sense, it is helpful to add spaces to the analysis matrix offered to experts to enable them to provide contributions and insights regarding the questions

(Urrutia et al., 2014). This fact allows the collection of a broader spectrum of information beyond pertinence and relevance. This type of opening makes it possible to interpret the experts' responses and understand their frame of reference for their responses, both at the theoretical and representational levels.

Concerning the possible sources of interpretation error (Pedrosa et al., 2014), it is arguable to consider the CVR index equal to 0 as a difficulty when half of the experts indicate an item as relevant and the other half as irrelevant. Lawshe identified the need for an agreement of at least 50%, with which the previously exposed problem is overcome. In addition, we demonstrated the possibility of considering the critical value of CVR = 0 when the number of experts tends toward infinity, allowing for straightforward interpretation.

Although different authors justify statistical indices, such as Lawshe's CVR, there is a strong tendency to use other strategies that complement these indices to circumvent their limitations. In this sense, it is currently suggested that "once the relevant items have been defined, they should be applied to a set of participants in order to apply the GT [Generalization Theory] to their answers (...), so that it would be possible to quantify the effect of possible sources of error" (Pedrosa et al., 2014, p. 15).

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

MR: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. ED: Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. TF: Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – review & editing, Software.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The results of this publication have been possible thanks to the financing awarded by DIUMCE of the Metropolitan University of Educational Sciences through the PMI-EXA-PNNI 01-17 Project.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1271335/full#supplementary-material

References

Akins, R. B., Tolson, H., and Cole, B. R. (2005). Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. *BMC Med. Res. Methodol.* 5, 1–12. doi: 10.1186/1471-2288-5-37

Armitage, P, Berry, J., and Matthews, N.S. *Statistical methods in medical research*. New York: John Wiley & Sons (2008).

Ayre, C., and Scally, A. J. (2014). Critical values for Lawshe's content validity ratio. *Meas. Eval. Couns. Dev.* 47, 79–86. doi: 10.1177/0748175613513808

Baghestani, A. R., Ahmadi, A., and Meshkat, M. (2019). Bayesian critical values for Lawshe's content validity ratio. *Meas. Eval. Couns. Dev.* 52, 69–73. doi: 10.1080/07481756.2017.1308227

Berger, J O. *Statistical decision theory and Bayesian analysis*. Berlin: Springer Science & Business Media (2013).

Cabero, J., del Carmen, M., and Llorente, F. (2013). La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información y comunicación (TIC). *Revista Eduweb* 7, 11–22.

Davis, L. (1992). Instrument review: getting the most from a panel of experts. *Appl. Nurs. Res.* 5, 194–197. doi: 10.1016/S0897-1897(05)80008-4

Díaz Costa, E., Fernández-Cano, A., Faouzi, T., and Henríquez, C. (2015). Validación del constructo subyacente en una escala de evaluación del impacto de la investigación educativa sobre la práctica docente mediante análisis factorial confirmatorio. *Revista de Investigación Educativa* 33, 47–63. doi: 10.6018/rie.33.1.193521

Divayana, D. G. H., Suyasa, P. W. A., and Adiarta, A. (2020). Content validity determination of the countenance-tri kaya parisudha model evaluation instruments using Lawshe's CVR formula. *J. Phys. Conf. Ser.* 1516:012047. doi: 10.1088/1742-6596/1516/1/012047

Eignor, D. R. (2013). "The standards for educational and psychological testing," in *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology*. Eds. K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez (American Psychological Association), 245–250.

Gilbert, G. E., and Prion, S. (2016). Making sense of methods and measurement: Lawshe's content validity index. *Clin. Simulation Nursing* 12, 530–531.

Guion, R. M. (1977). Content Validity. *Am. Psychol.* 5, 522–517.

Jeffreys, H. (1935). Some tests of significance treated by the theory of probability. *Math. Proc. Camb. Philos. Soc.* 31, 203–222. doi: 10.1017/S030500410001330X

Juárez-Hernández, L. G., and Tobón, S. (2018). Análisis de los elementos implícitos en la validación de contenido de un instrumento de investigación. *Revista Espacios* 39.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Pers. Psychol.* 28, 563–575. doi: 10.1111/j.1744-6570.1975.tb01393.x

Moreno López, N. M., and Monroy González, J. D. (2022). *Validez de contenido por juicio de expertos del inventario de respuestas de afrontamiento CRI-A dirigido a víctimas del conflicto*

armado colombiano. Cátedra Villarreal-Psicología, 1. Recuperado a partir de. Available at: https://revistas.unfv.edu.pe/CVFP/article/view/137810.3389/feduc.2023.1271335

López-Gómez, E. (2018). El método Delphi en la investigación actual en educación: Una revisión teórica y metodológica. *Educación XXI* 1, 17–40.

Pedrosa, I., Suárez-Álvarez, J., and García-Cueto, E. (2014). Evidencias sobre la validez de contenido: Avances teóricos y métodos para su estimación. *Revista Acción Psicológica* 10, 3–20. doi: 10.5944/ap.10.2.11820

Plake, B. S., and Wise, L. L. (2014). "what is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing?" educational. *Measurement* 33, 4–12. doi: 10.1111/emip.12045

Polit, D. F., Tatano, C. B., and Owen, S. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res. Nurs. Health* 30, 459–467. doi: 10.1002/nur.20199

Romero-Jeldres, M., and Faouzi, T. (2018). Validación de un modelo de competencias pedagógicas para docentes de Educación Media Técnica. *Educación y Educadores* 21, 114–132. doi: 10.5294/edu.2018.21.1.6

Sireci, S. G. (1998). The construct of content validity. *Soc. Indic. Res.* 45, 83–117. doi: 10.1023/A:1006985528729

Tilden, V. P., Nelson, C., and May, B. A. (1990). Use of qualitative methods to enhance content validity. *Nurs. Res.* 39, 172–175. doi: 10.1097/00006199-199005000-00015

Tristán-López, A. (2008). Modificación al modelo de Lawshe para el dictamen cuantitativo de la validez de contenido de un instrumento objetivo. *Avances en medición* 6, 37–48.

Urrutia, E., Marcela, S. B., Araya, M. G., Núñez, , and Camus, M. M. (2014). Métodos óptimos para determinar validez de contenido. *Educación Médica Superior* 28, 547–558.

Ventura-León, J. (2019). De regreso a la validez basada en el contenido. *Adicciones* 34, 323–326. doi: 10.20882/adicciones.1213

Wilson, F. R., Pan, W., and Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Meas. Eval. Couns. Dev.* 45, 197–210. doi: 10.1177/0748175612440286

frontiers | Frontiers in Education

TYPE Perspective
PUBLISHED 23 November 2023
DOI 10.3389/feduc.2023.1270700

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Syamsul Nor Azlan Mohamad,
MARA University of Technology, Malaysia
Jason M. Lodge,
The University of Queensland, Australia
Kim Schildkamp,
University of Twente, Netherlands

*CORRESPONDENCE
Therese N. Hopfenbeck
✉ Therese.hopfenbeck@unimelb.edu.au

RECEIVED 01 August 2023
ACCEPTED 02 November 2023
PUBLISHED 23 November 2023

CITATION
Hopfenbeck TN, Zhang Z, Sun SZ, Robertson P and McGrane JA (2023) Challenges and opportunities for classroom-based formative assessment and AI: a perspective article.
Front. Educ. 8:1270700.
doi: 10.3389/feduc.2023.1270700

COPYRIGHT
© 2023 Hopfenbeck, Zhang, Sun, Robertson and McGrane. 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.

Challenges and opportunities for classroom-based formative assessment and AI: a perspective article

Therese N. Hopfenbeck[1,2,3]*, Zhonghua Zhang[1], Sundance Zhihong Sun[1], Pam Robertson[1] and Joshua A. McGrane[1,2]

[1]Assessment and Evaluation Research Centre, Graduate School of Education, The University of Melbourne, Parkville, VIC, Australia, [2]Kellogg College, University of Oxford, Oxford, England, [3]The University of Science and Technology, Trondheim, Norway

The integration of artificial intelligence (AI) into educational contexts may give rise to both positive and negative ramifications for teachers' uses of formative assessment within their classrooms. Drawing on our diverse experiences as academics, researchers, psychometricians, teachers, and teacher educators specializing in formative assessment, we examine the pedagogical practices in which teachers provide feedback, facilitate peer- and self-assessments, and support students' learning, and discuss how existing challenges to each of these may be affected by applications of AI. Firstly, we overview the challenges in the practice of formative assessment independently of the influence of AI. Moreover, based on the authors' varied experience in formative assessment, we discuss the opportunities that AI brings to address the challenges in formative assessment as well as the new challenges introduced by the application of AI in formative assessment. Finally, we argue for the ongoing importance of self-regulated learning and a renewed emphasis on critical thinking for more effective implementation of formative assessment in this new AI-driven digital age.

KEYWORDS

artificial intelligence, formative assessment, self-regulation, critical thinking, classroom based assessment

Introduction

In an era marked by rapid technological advancements, artificial intelligence (AI) is now increasingly used in diverse sectors of our society, fundamentally transforming the way we live, work, and learn. Within the field of educational assessment, the introduction of AI has raised both concerns and optimisms, particularly with respect to the dynamics around AI and formative assessment classroom practices. In the current paper, we explore the opportunities and challenges AI offers and underscore the continued significance of self-regulated learning and critical thinking as essential skills in this AI -driven digital age.

A brief background of classroom-based formative assessment

Classroom-based assessment has been internationally researched for decades, both with respect to summative assessments that typically occur at the end of a learning process (e.g., McMillan, 2013; Brookhart, 2016), as well as formative assessments that

involve feedback processes promoting students' learning as it happens (e.g., Brown, 2018; Lipnevich and Smith, 2018). Black and Wiliam (1998) emphasized the pivotal role of formative assessment in providing valuable information not only to teachers but also to students, guiding improvements in teaching and learning to optimize student outcomes. Since the publication of their classic work, they have continued to refine their model through subsequent theoretical papers (e.g., Black and Wiliam, 1998). Additionally, they have supported their theoretical insights with empirical studies, documenting the tangible impact of formative assessment practices on students' learning within classroom settings (e.g., Wiliam et al., 2004).

While there is a consensus among researchers regarding the positive effects of formative assessment on students' learning (Hattie 2009; Lipnevich and Smith, 2018), the term "formative assessment" itself has faced critique for lacking a cohesive definition. Instead, it has been argued to be a collection of varied definitions and practices, making it challenging to conduct rigorous evaluations of its effects (Bennett 2011; Stobart and Hopfenbeck, 2014). We aim to navigate these complexities by adopting Black and Wiliam's (2009) definition of formative assessment: "Practice in a classroom is formative to the extent that the evidence about student achievement is elicited, interpreted, and used by teachers, learners, and their peers to make decisions about their next steps in instruction that are likely to be better, or better founded than the decisions they would have taken in the absence of the elicited evidence" (p.9).

In other words, for a teacher to conduct formative assessment, they will need to know each student, their learning progress, and how to support them to achieve their learning goals. In traditional classrooms, formative assessment challenges teachers, as it requires them to find ways of following up a whole class or classes of students and provide individualized feedback to everyone, either through teacher-assessment, peer assessment, self-assessment, group-assessment, or by other means (Double et al., 2020). As we will discuss in the next section, research has shown that these practices are difficult to implement at scale and in ways that are sustainable over time (Hopfenbeck and Stobart, 2015).

Challenges to implementing formative assessment

Several challenges to the implementation of formative assessment have been documented by researchers as presented in the January 2015 Special Issue of the Assessment in Education journal. Wylie and Lyon (2015) found substantial variation in the quality of implementing formative assessments among 202 Mathematics and Science teachers in the US context. They suggest that more targeted professional development is needed to secure high-quality implementations of formative assessment practices. Further, since formative assessment requires teachers to have high competency across a range of knowledge and skills (e.g., domain content knowledge, pedagogical content knowledge, assessment and data literacy, and knowledge of measurement fundamentals), such professional development needs to be wide in scope.

Challenges have also been found when stakeholders involved in the assessment do not share a mutual understanding of its purpose(s). For example, in the same Special Issue, Hopfenbeck et al. (2015) conducted an evaluation of a large-scale implementation of an assessment for learning program in Norway and found that implementation was weaker in schools where the assessment was perceived as part of an accountability system, while in schools with a high degree of trust between teachers, head teachers and the school owners at the municipality level, the quality of the implementation was better. Similarly, a study of school-based assessment in Singapore found that their high-stakes examination-focused education system created tensions when trying to implement formative assessment processes, thus demonstrating how context matters in terms of the challenges that arise between different stakeholders in the interaction between formative assessment and accountability systems (Ratnam-Lim and Tan, 2015). These findings indicate how teachers' formative assessment implementations are influenced by accountability structures, educational leadership, resources, workload and social pressures within their context.

Thirdly, formative assessment practices have primarily been researched and developed in contexts where students and teachers have access to a wealth of resources, and, thus, do not necessarily generalize to more challenging contexts. Halai et al. (2022) evaluated an implementation of assessment for learning practices in six schools in Dar es Salaam, Tanzania, and documented the challenges created by very large class sizes, with one teacher responsible for up to 180 students at a time, as well as under-resourced classrooms. Furthermore, it was found that the cultural assumptions of the student role assumed in most of the Western, English-speaking literature did not fit what is seen as a good student in Tanzania. Formative assessment practices expect students to be self-regulated and proactive so they can participate in peer-discussions and assessment, and as part of this, they are supposed to engage in dialog in groups and with the teacher and be able to ask critical questions. In contrast, a good student in Tanzania is expected to listen to the teacher, not ask questions or be too critical, and overall follow instructions and do what the teacher tells them to do. This is enforced by the parents' expectations of how schools and teachers need to help raise the child. Thus, the interactive dialog between teachers, students and peers that are at the heart of formative assessment can be culturally and contextually sensitive, which poses challenges for implementing a 'one size fits all' formative assessment practice across different contexts.

Finally, studies have reported that teachers find it challenging to provide enough feedback to students, particularly at crucial times in the learning process, as well as with the quality of feedback required to further each student's learning, due to time and other resource constraints (Brooks et al., 2019; Gamlem and Vattøy, 2023). As a result, formative assessment theory suggests that teachers need to design classrooms where students can provide feedback to each other to help reduce this workload (Wiliam, 2011). However, teachers still report that they struggle to manage classrooms where these peer assessment practices are established (Dignath et al., 2008; Halai et al., 2022). So, even in well-resourced contexts where teachers endeavor to engage best-practices in the implementation of formative assessment in their classrooms, the high workload such practices engender continues to be a barrier.

Evolving and revolutionizing formative assessment: what can AI bring?

Given the challenges to implementing formative assessment in the classroom that are outlined in the previous section, the advancement of AI and relevant products (e.g., ChatGPT) may provide opportunities to overcome some of these challenges, such as high number of students, and only one teacher to provide feedback. Indeed, researchers have identified assessment as one of the most significant areas of opportunity that AI and related technologies offer in education (Cope et al., 2021; Swiecki et al., 2022; Zhai and Nehm, 2023). However, new challenges may also be introduced through the widespread use of AI, including practical and ethical challenges (Milano et al., 2023). This section focuses on the changing landscape of the practice of formative assessment, especially under the influence of AI, and discusses how AI can help support teachers to provide formative assessment for students on a large and sustainable scale.

What do we mean by AI?

In our discussion, AI refers to the application of sophisticated algorithms that allow computers and machines to simulate human intelligence for successfully completing tasks (Murphy, 2019). Although different technical approaches and methods, e.g., supervised learning versus deep learning, have been used to develop AI systems, the essence of AI is to use data to teach machines to make classifications, diagnoses, decisions, predictions, and/or recommendations (Gardner et al., 2021). More specifically, the application of AI typically involves collecting large multivariate data relevant to the task of interest, applying statistical methods and sophisticated algorithms to process sets of input data to build a model(s) that identifies and weights features and/or patterns of the input variables relevant to the task, then using a different pre-collected dataset to validate the model(s) for the task where the correct output is known in advance, and then applying the model(s) to generate a task output(s) (e.g., classification, prediction, decision, or recommendation) in a context where the correct output is unknown (Murphy, 2019; Gardner et al., 2021).

Opportunities for implementing formative assessment with AI

As mentioned above, one of the challenges in formative assessment is to provide individualized, high-quality feedback to students. It is highly resource intensive for teachers to personally give or find other ways (e.g., peer assessment, self-assessment, group assessment) to provide individualized feedback to each student on a large scale. However, AI can make some of the assessment procedures fully or partly automated, making the assessment practices more feasible to maintain, which can then reduce the time burden on teachers (Swiecki et al., 2022). A typical example that has been widely discussed is automated essay scoring systems (Ke and Ng, 2019; Gardner et al., 2021). The application of AI in automated essay scoring frees teachers from the labor-intensive grading process and allows them to assign more extended writing tasks to students, automate the grading process, and, more importantly, with the integration of natural language processing-based AI, provide timely formative feedback to help students revise and improve their writing (Murphy, 2019).

Liu et al. (2016) showed that a machine learning enabled automated scoring tool, c-rater-ML, could produce scores that were comparable to human raters in scoring students' responses to constructed response questions about science inquiry, offering a promising solution to improving the efficiency of not only obtaining the summative scores but also generating instant formative feedback (Linn et al., 2014). Another example of how AI can help is by using computers to support the management and delivery of formative assessments (e.g., Webb et al., 2013; Tomasik et al., 2018). These systems have the capacity to discern distinct learning pathways in students' progress, enabling the identification of the most suitable tasks or questions for each student at different points in time. In addition, computerized formative assessment systems can optimize the administration of formative assessments by determining their frequencies and schedules customized for every individual student (Shin et al., 2022). These findings demonstrate how AI can improve the efficiency and flexibility of formative assessment practices at the individual student level.

Another significant opportunity that AI offers for formative assessment is the improvement of feedback both in quantity and quality (Gardner et al., 2021). The main goals of formative assessment are to provide constructive feedback based on students' responses and to help teachers design differentiated instructional strategies and sustain students in self-regulating their learning. AI can delve into the data to identify the patterns on which dynamic, customized, individualized, and visualized feedback can be automatically generated (Verma, 2018; Tashu and Horvath, 2019; Lee, 2021, 2023). For example, the adaptive nature of some computerized formative assessment systems and intelligent tutoring systems enables every student's attainment to be individually and more precisely assessed, which facilitates more appropriate and targeted feedback based on their individual learning stage and trajectory (Ma et al., 2014; Tomasik et al., 2018; Mousavinasab et al., 2021). Adaptive multi-strategy feedback models, based on AI methods, have been applied in the context of such systems to automatically adapt the feedback generating strategy to individual students, which have, in turn, been found to generate more effective feedback than the traditional feedback generation methods (Gutierrez and Atkinson, 2011).

In addition, AI can improve the quality and effectiveness of peer assessment in classrooms with large class sizes. Peer assessment can be supported with prompts from language models (Er et al., 2021). This approach to peer assessment supports students in not only providing feedback to peers but also reflecting on and justifying their judgments, providing further opportunities for them to develop their self-regulated learning skills (Liu and Carless, 2006), and has been found to provide useful peer feedback to students (Luaces et al., 2018). In addition, the peer assessment reviews can both help the teachers to better understand the performance of the students in their classroom and also provide additional data (e.g., the review text) that can be analyzed using AI-based techniques (e.g., semantic, lexical, and psycho-linguistic analyses; Vincent-Lamarre and Larivière, 2021) to further enhance teachers' understanding of the performance of their students.

AI can also aid teachers in collecting and analysing longitudinal formative assessment data, and in generating learner profiles to trace

students' learning progression over time (Swiecki et al., 2022). This application of AI enables the scalable implementation of formative assessment in both cross-sectional and longitudinal contexts, which makes it more sustainable, allowing teachers to efficiently monitor the growth of student learning and identify the knowledge and skill gaps in their learning over time (Barthakur et al., 2023). Another contribution of AI to facilitating longitudinal formative assessment lies in its ability to analyse the large-scale longitudinal formative assessment data to trace the learning trajectories of the students and predict their future learning states. For example, some of the widely applied statistical methods in AI (e.g., hidden-Markov models, artificial neural networks) have been combined with traditional cognitive diagnostic models (CDM) to analyse longitudinal formative assessment data to track the changes of students' learning over time (e.g., Chen et al., 2018; Wen et al., 2020).

The application of AI also allows teachers to get an in-depth understanding of students' learning processes based on the analysis of large volumes of 'process data' rather than just the assessment artefacts (e.g., responses to questions, items or tasks) produced by the students. With modern technologies, the processes leading to assessment responses can be captured in time-stamped log stream data (Cope et al., 2021). For example, students' actions (e.g., keystrokes, editing, chat history, video watching) during an assessment can be captured. Which potentially contain additional information for understanding how students produce their responses. With the support of AI, the process data can be analyzed to investigate their strategies (e.g., identifying misconceptions), which can then provide invaluable information for individualized feedback. In addition, taking advantage of AI's increasing capacity to deal with complex, multi-media data, more authentic assessment tasks (e.g., multimedia, game-based problem solving, essay writing, performance-based tasks) can be effectively used in undertaking formative assessments (Swiecki et al., 2022).

AI also offers opportunities for assessing some hard-to-measure constructs. Students' non-scholastic attributes, including social–emotional traits (e.g., classroom engagement, self-efficacy, motivation, resilience) and social-cognitive skills (e.g., metacognition, collaborative problem-solving, critical thinking, digital literacy, self-regulated learning), have been attracting more attention and are increasingly recognized as equally important as their scholastic achievement (Durlak and Weissberg, 2011). Advances in AI and related technologies allow for these constructs to be more validly assessed, instead of purely relying on students' self-reported beliefs and behavior through questionnaires. Now, the data collected through different channels (e.g., time-stamped process data, eye contact, feedback, facial expression, eye movements, body posture and gesture) can be mined to develop indicators for assessing these different aspects of student learning. For example, MOOC data has been used to design indicators through a thorough analysis of students learning behaviors in online courses to measure students' self-regulated learning (e.g., Milligan and Griffin, 2016) and leadership development in workplace learning in an online environment (e.g., Barthakur et al., 2022). Another example is the measurement of collaborative problem-solving skills through process data that captured the actions and chats of pairs of team members collaboratively solving tasks (Griffin and Care, 2014). AI-based Large Language Models provide further promise for mining chat history data to support assessing how team members explore, define, plan, execute and solve tasks in a collaborative way.

Challenges arising from using AI in formative assessment

AI introduces not only opportunities but also challenges to formative assessment practices (Swiecki et al., 2022). A primary challenge that needs to be addressed before teachers can apply AI in their formative assessment practices is their lack of knowledge and skills relevant to AI techniques as well as their limited access to big data. Thus, although AI can potentially ease the workload of teachers by automating some aspects of formative assessment (e.g., automatising scoring and tracing students' learning progress), it adds further burden through the need for professional development in its use (Engeness, 2021). Moreover, despite the promising future for formative assessment brought by big data, with the possibility of collecting the process data through students' learning, a new challenge arises in identifying which part of the collected data is most helpful and relevant to improve student learning. In addition, the unique features of current big data (e.g., time-stamped process data, sparse data) are significantly different from that of the traditional assessment data and pose a variety of challenges to the psychometric methods for analyzing the data. To deal with this, scholars have been endeavoring to introduce new methods to integrate data science and machine learning into psychometrics (e.g., von Davier et al., 2022).

Another challenge arising from using AI in formative assessment is to tackle relevant issues about investigating the best way to use AI in formative assessment practices. One of the hotly debated issues is whether AI will replace teachers. We argue that AI should not replace but facilitate teachers' formative assessment practices and promote the role of formative assessments in supporting instruction and learning. As stated by Murphy (2019), "the best use of AI in education is to augment teacher capacity by helping teachers deliver more effective classroom instruction" (p. 14). Teachers need to understand the limitations of the AI techniques when they review the assessment results. For example, automated scoring systems have long been criticized for their inability to measure higher-order aspects of writing (e.g., creativity, argumentation, reasoning) (Gardner et al., 2021). One of the primary aims of formative assessment is to diagnose gaps in students' learning based on the well-established interpretability of the measurement scales. However, many approaches based on machine learning are designed for prediction involving complicated models for improving accuracy but sacrificing the ease of interpretation. Therefore, any inferences from the results of formative assessments involving the integration of AI techniques should only be made after having examined the assessments' validity and interpretability (Bejar et al., 2016; Scalise et al., 2021). Teachers need to critically review how the assessment results are reached and identify any sources of bias introduced by the application of AI techniques in assessment, which in turn adds to their workloads (Murphy, 2019). Finally, but not at least, the introduction of AI in the classroom cannot happen without ethical considerations for the use and application of it. Scholars have emphasized the importance of having conversations with students on the productive, ethical and critical relationship around the use of AI and future technologies (Bearman et al., 2023) and improving knowledge on data privacy for children (Johnston, 2023). With these considerations in mind, we will now turn to one example of formative assessment practices and AI.

The formative uses of rubrics and the opportunities and challenges of using AI

The formative use of rubrics, i.e., the scoring guides that are used to make judgments about the quality of students' constructed responses, such as writing, performances or products (Popham, 1997), has been shown to have a positive influence on learning. Specifically, Panadero and Jönsson (2013) argued that the use of quality rubrics plays a key role in enhancing the effectiveness of formative assessment practices. There are two main ways that rubrics are thought to improve formative assessment. They make assessment expectations explicit, thereby assisting with understanding and feedback, and they support self-regulated learning by supporting learners to monitor their own learning and make decisions about areas for improvement (Jönsson and Panadero, 2016). While the use of rubrics for formative assessment purposes has a positive effect on learning, this effect is amplified when it is accompanied by teacher-given rubric feedback that addresses the three feedback questions (Wollenschläger et al., 2016). Consequently, we now examine the formative use of rubrics as a specific example of how AI can address existing challenges, provide additional opportunities, and present new challenges to this formative assessment practice.

Firstly, AI can support the formative uses of rubrics by helping teachers to overcome some of the time needed to construct rubrics and to teach students how to use them, as these have been found to be a constraint to rubric use (English et al., 2022). Generative AI can speed up rubric design, as teachers can use it to provide samples of rubrics to assess specified constructs, and a teacher can choose to directly use these rubrics or to use them as a source of ideas for designing their own. AI also has the potential to assist students as they learn to use rubrics by providing work samples matching different levels on a rubric, by assessing student-generated work samples against a rubric so a student can check the accuracy of self-assessments, and by providing written feedback to accompany a rubric assessment. These possible AI-augmented rubric uses by students help build agency, as the students can have more control over the timing and style of feedback they receive. Therefore, AI has the potential to help teachers overcome some of the common challenges to using rubrics in the classroom.

Nonetheless, the challenges presented by potential biases in training data are also applicable to rubrics generated with AI (Li et al., 2023). Rubrics for constructs with a greater cultural influence, such as communication, are likely to be more affected than those for constructs where the subject matter is more consistent irrespective of culture, like chemistry. In addition, while there is an acknowledged need for more research on rubric design (e.g., English et al., 2022), the findings of such research often fail to be commonly adopted by teachers. One example of this is that most rubrics have structurally aligned categories, e.g., all parts of the rubric have five levels of quality. Unfortunately, there is empirical evidence that this common structure is ineffective because it increases construct-irrelevant variance by facilitating scoring based on a halo effect where the assessor makes a global judgment of quality and simply aligns the ratings on different criteria of the rubric to match rather than making independent decisions for each (Humphry and Heldsinger, 2014). Rubrics, thus, support more accurate judgments when the number of levels of quality are tailored to the specific criterion being evaluated where some criterion (e.g., quality of argumentation) have more levels than others (e.g., use of paragraphs). Consequently, without careful curation of training data sets to ensure they meet best practice in rubric design, AI generated rubrics will likely propagate common design flaws. Moreover, exploration by researchers of the different ways AI is already augmenting rubric use in classrooms is warranted, especially in cultures and contexts that are not well represented in training data sets.

The role of self-regulated learning and critical thinking in formative uses of AI

Based on the formative assessment cycle in Ruiz-Primo and Brookhart (2018), there is a natural bridge between self-regulated learning and formative assessment, as formative assessment can be considered as a self- and co-regulated process of improving learning, which starts with defining and sharing learning goals and then through a process of gathering or eliciting information, analysing and interpreting the collected information, and finally using the collected information to make a reflective judgment on whether the pre-defined learning goal has been achieved or not. There has been a call for linking the research into self-regulation with formative assessment, as it is recognized that self-regulation will enhance students' ability to act as peer-assessors, do self-assessment and take on the proactive role needed for formative assessment practices (Brandmo et al., 2020). Despite decades of educational research into what improves students' learning, few researchers have tried to combine the two fields of formative assessment and self-regulation, although exceptions include Allal (2010), Andrade and Brookhart (2016), Brown (2018), Butler and Winne (1995), Nicol and MacFarlane-Dick (2006), and Panadero et al. (2018). Moreover, more recent research has demonstrated how students can benefit more from formative assessment practices if they are self-regulated learners (Allal, 2020; Andrade and Brookhart, 2020; Perry et al., 2020). With the rise of AI, students' ability to self-regulate will be even more important, as it opens opportunities but also challenges in how we plan, use strategies, and evaluate our learning processes.

Furthermore, we argue for the importance of critical thinking for both teachers and learners to navigate the principled use of AI and leverage the effectiveness of formative assessment as part of their process of self-regulated learning, particularly when confronted with the novel challenges of AI. Although highlighting the importance of critical thinking may seem like an already labored point in educational settings, as it has been acknowledged as a fundamental generic skill necessary for individuals to live and thrive in the 21st century (e.g., Davies and Barnett, 2015), none of the existing research has yet built a connection between critical thinking, formative assessment and self-regulated learning under the impact of AI. Before getting into the specific argument on the role of critical thinking in formative assessment and self-regulated learning, it is worth clarifying that self-regulated learning is used in a broader way in this section, extending beyond learners to encompass teachers who also need to apply their self-regulated learning skills to effectively acquire new knowledge and skills to harness the potential of AI in their teaching practice effectively. In the following part, we will briefly explain our understanding of critical thinking, depicting the role of critical thinking when facing new challenges brought by AI, and then

describing the role of critical thinking in formative assessment and self-regulated learning.

When facing the uncertain, complex issues brought by the advancement of AI, critical thinking, defined as "reasonable and reflective thinking focused on deciding what to believe or do" (Ennis et al., 2005: 1), becomes increasingly pronounced. As a complex competency, good critical thinking practitioners are expected to have a sophisticated level of epistemic beliefs (i.e., attitudes to knowledge and knowing), which are essential to recognizing the uncertainty and complexity of such controversial issues as the ethical use of AI in education. These beliefs lay a foundation for the engagement of their thinking skills (e.g., understanding, applying, analysing, synthesizing, evaluating) to be well-informed of the issue and navigate through a vast amount of potentially conflicting information (King and Kitchener, 1994; Kuhn and Weinstock, 2004; Wertz, 2019; Sun, 2021, 2023). As theorized by Dewey (1910), suspending judgments may be the most effective course of action prior to acquiring a comprehensive understanding of a relevant issue.

For different responses on the issue of whether AI should be used in educational settings, it is not surprising to witness resistance toward emerging technologies because there is a natural fear and unease that often accompanies the introduction of new technologies (Ball and Holland, 2009). However, if critical thinking is engaged before deciding what to believe or do, this natural tendency can be challenged. AI is far from a novel concept and has been an academic discipline since the 1950s (Haenlein and Kaplan, 2019; Gillani et al., 2023), and various AI technologies (e.g., image recognition, smart speakers, self-driving cars) and models (e.g., AlphaGo, Deep Blue and ELIZA) have already significantly impacted our ways of living and working (Haenlein and Kaplan, 2019). Yet, limited transformation has happened in education, with 20th-century traditions and practices still dominating our schools (Luckin and Holmes, 2016). Despite the recognition of the enormous benefits and potential of AI in transforming education, scholars' impatience is mounting because many of these promising ideas remain confined to the lab or lecture halls with few practical breakthroughs (Luckin and Holmes, 2016).

Specific to the context of formative assessment and self-regulated learning, the role of critical thinking is also pivotal, which equips both teachers and learners to effectively address emerging challenges. For teachers, as the landscape evolves with a growing array of AI products and assessment data, being flooded by abundant online learning resources can be overwhelming (Schwartz, 2020). It would be increasingly important for teachers to critically evaluate what, when, and how to utilize these resources to enhance their teaching methodologies and bolster student learning. When facing an increasing amount of data that has been collected or needs to be collected, teachers need to critically discern how assessment data can best inform their pedagogical strategies rather than have data to dictate their teaching. Additionally, teachers should exercise discernment in determining the level of trust they can place in specific AI models when making judgments about student learning outcomes. This becomes especially crucial for teachers who should critically assess the potential biases that AI models might carry due to the use of training data (Li et al., 2022, 2023). Furthermore, as AI advancements have the potential to liberate teachers from routine and time-consuming tasks like assignment grading and rubric development, they must engage in critical reflection. They need to consider which skills they should prioritize for their professional development, such as data literacy, and what skills should remain at the core of their teaching, notably critical and creative thinking. This critical assessment of their evolving role is essential in navigating the transformative impact of AI in education.

Regarding individual learners, engaging critical thinking can have positive contribution to the effectiveness of formative assessment and self-regulated learning when facing the opportunities and challenges introduced by the advancement of AI. For instance, the advancement of AI, indeed, can certainly be used to generate text to pass the assessment of a subject, but if learners are practicing their critical thinking and self-regulated learning skills, they may ask themselves some reflective questions, such as what is the purpose of learning? Will a certain way of using AI contribute to achieving their learning goals? When specific solutions have not been produced to address the new challenges brought by AI, individual learner's practice of their critical thinking and self-regulated learning may contribute to the ethical use of new technologies. Despite some instances of learners exploiting AI to evade plagiarism detection systems, it is encouraging to learn from recent empirical research that many students genuinely benefit from the timely feedback and companionship provided by AI (Skeat and Ziebell, 2023). Moreover, these students display ethical awareness, being cautious and mindful of their AI usage even in the absence of well-developed regulations governing AI in education.

While some scholars have suggested that assessment is holding us back from transforming our education systems (Luckin and Holmes, 2016, p.35), the advancement of AI may catalyze a "Renaissance in Assessment" (Hill and Barber, 2014). Although the acceptance of AI may encounter some resistance, the power of new technologies, if unleashed with principled and research-driven use, may significantly change and improve ways of teaching and learning. In this vein, Australian educational policymakers made a significant shift by granting permission for the use of ChatGPT and generative AI in all government schools. This change followed the release of the Australian Framework for Generative Artificial Intelligence in Schools. It is encouraging to observe the transition from a policy that limited the use of ChatGPT across every Australian state and territory, except South Australia, to a more welcoming and adaptable stance.

Conclusion

As we have outlined in this article, despite decades of research on formative assessment practices, teachers still face several challenges in implementing these practices on a large scale. The use of AI in classrooms has the potential for supporting formative assessment practices, although we will argue, it will require some careful considerations. Based upon what we have outlined in this article, we will conclude with the following suggestions on how to integrate AI into formative assessment:

1. Utilize AI for feedback assistance, particularly in large classes where teachers struggle to give timely feedback to all students.
2. Promote self-regulating skills as students will need to take even more responsibility for their own learning, when using AI. This

includes goal-setting, monitoring progress, and adjusting their study strategies based upon AI feedback.
3. Emphasize the ethical use of AI in formative assessment and discuss the importance of integrity and responsible use of AI tools, to avoid inappropriate uses.
4. Emphasize the role of teachers in guiding the students' use of AI. Teacher can help students interpret AI feedback, set learning goals, and make informed decisions based upon AI recommendations.
5. Encourage collaborative research between educators and researchers to explore the effectiveness of AI in formative assessment. Co-design studies with teachers and students to assess how AI impacts learning outcomes and student engagement.
6. Recognize the evolving role of teachers and facilitators of AI-enhanced learning.

In the changing times of AI, students more than ever need teachers to guide them using AI, and as researchers, we would encourage colleagues to take part in co-designing studies with teachers and students, where we together examine how to improve students learning through formative assessment practices, critical thinking, self-regulated learning and AI.

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

TH: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft. ZZ: Conceptualization, Writing – original draft. SS: Conceptualization, Writing – original draft, Writing – review & editing. PR: Writing – original draft, Writing – review & editing. JM: Conceptualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

References

Allal, L. (2010). "Assessment and the regulation of learning" in *International encyclopedia of education*. eds. P. Peterson, E. Baker and B. McGaw (Elsevier), 348–352.

Allal, L. (2020). Assessment and the co-regulation of learning in the classroom. *Assess. Educ.* 27, 332–349. doi: 10.1080/0969594X.2019.1609411

Andrade, H. L., and Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assess. Educ.* 27, 350–372. doi: 10.1080/0969594X.2019.1571992

Ball, W., and Holland, S. (2009). The fear of new technology: A naturally occurring phenomenon. *The American Journal of Bioethics*, 9, 14–16. doi: 10.1080/15265160802617977

Barthakur, A., Kovanovic, V., Joksimovic, S., Zhang, Z., Richey, M., and Pardo, A. (2022). Measuring leadership development in workplace learning using automated assessments: learning analytics and measurement theory approach. *Br. J. Educ. Technol.* 53, 1842–1863. doi: 10.1111/bjet.13218

Barthakur, A., Dawson, S., and Kovanovic, V. (2023). Advancing leaner profiles with learning analytics: a scoping review of current trends and challenges. In LAK23: 13th international learning analytics and knowledge conference, 606–612. doi: 10.1145/3576050.3576083

Bearman, M., Ajjawi, R., Boud, D., Tai, J., and Dawson, P. (2023) *CRADLE suggest assessment and gen AI*. Centre for Digital Learning, Deakin University, Melbourne, Australia.

Bejar, I. I., Mislevy, R. J., and Zhang, M. (2016). "Automated scoring with validity in mind" in *The Wiley handbook of cognition and assessment: frameworks, methodologies, and applications*. eds. A. A. Rupp and J. P. Leighton (Wiley), 226–246.

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ.* 18, 5–25. doi: 10.1080/0969594X.2010.513678

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assess. Educ.* 5, 7–74. doi: 10.1080/0969595980050102

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. *Educ. Assess.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Black, P., and Wiliam, D. (2018). Classroom assessment and pedagogy. *Assess. Educ.* 25, 551–575. doi: 10.1080/0969594X.2018.1441807

Boud, D. (2005). *Enhancing learning through self-assessment* Routledge.

Brandmo, C., Panadero, E., and Hopfenbeck, T. N. (2020). Bridging classroom assessment and self-regulated learning. *Assess. Educ.* 27, 319–331. doi: 10.1080/0969594X.2020.1803589

Brookhart, S. M. (2016). "The use of teacher judgement for summative assessment in the USA" in *International teacher judgement practices*. ed. V. Klenowski (Oxon, New York: Routledge), 69–90.

Brooks, C., Carroll, A., Gillies, R. M., and Hattie, J. (2019). A matrix of feedback for learning. *Australian J. Teach. Educ.* 44, 14–32. doi: 10.14221/ajte.2018v44n4.2

Brown, G. T. L. (2018). *Assessment of student achievement*. Oxon, New York: Routledge.

Butler, D. L., and Winne, P. H. (1995). Feedback and self-regulated learning: a theoretical synthesis. *Rev. Educ. Res.* 65, 245–281. doi: 10.3102/00346543065003245

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

Cope, B., Kalantzis, M., and Searsmith, D. (2021). Artificial intelligence for education: knowledge and its assessment in AI-enabled learning ecologies. *Educ. Philos. Theory* 53, 1229–1245. doi: 10.1080/00131857.2020.1728732

Davier, A.A.von, Mislevy, R.J., and Hao, J. eds. (2021). Introduction to computational psychometrics: towards a principled integration of data science and machine learning techniques into psychometrics. *Computational psychometrics: new methodologies for a new generation of digital learning and assessment*. Boston: Springer.

Davies, M., and Barnett, R. (2015). *The Palgrave handbook of critical thinking in higher education*. Springer.

Dewey, J. (1910). *How we think*. Boston: Dover Publications.

Dignath, C., Büttner, G., and Langfeldt, H. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educ. Res. Rev.* 3, 101–129. doi: 10.1016/j.edurev.2008.02.003

Double, K. S., McGrane, J. A., and Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: a meta-analysis of control group studies. *Educ. Psychol. Rev.* 32, 481–509. doi: 10.1007/s10648-019-09510-3

Durlak, J. A., and Weissberg, R. P. (2011). Promoting social and emotional development is an essential part of students' education. *Hum. Dev.*, 54, 1–3. doi: 10.1159/000324337

Engeness, I. (2021). Developing teachers' digital identity: towards the pedagogic design principles of digital environments to enhance students' learning in the 21st century. *Eur. J. Teach. Educ.* 44, 96–114. doi: 10.1080/02619768.2020.1849129

English, N., Robertson, P., Gillis, S., and Graham, L. (2022). Rubrics and formative assessment in K-12 education: a scoping review of literature. *Int. J. Educ. Res.* 113:101964. doi: 10.1016/j.ijer.2022.101964

Ennis, R., Millman, J., and Tomko, T. (2005). *Cornell critical thinking tests level X & Level Z Manual*. The Critical Thinking Co.

Er, E., Dimitriadis, Y., and Gašević, D. (2021). A collaborative learning approach to dialogic peer feedback: a theoretical framework. *Assess. Eval. High. Educ.* 46, 586–600. doi: 10.1080/02602938.2020.1786497

Gamlem, S. M., and Vattøy, K.-D. (2023). "Feedback and classroom practice" in *International encyclopedia of education*. eds. R. J. Tierney, F. Rizvi and K. Ercikan, vol. *13. 4th* ed (Elsevier), 89–95.

Gardner, J., O'Leary, M., and Yuan, L. (2021). Artificial intelligence in educational assessment: 'breakthrough? Or buncombe and ballyhoo?'. *J. Comput. Assist. Learn.* 37, 1207–1216. doi: 10.1111/jcal.12577

Gillani, N., Eynon, R., Chiabaut, C., and Finkel, K. (2023). Unpacking the "Black Box" of AI in Education. *J Educ Techno Soc*, 26, 99–111.

Griffin, P., and Care, E. (2014). *Assessment and teaching of 21st century skills: methods and approach*. Springer Dordrecht: Springer.

Gutierrez, F., and Atkinson, J. (2011). Adaptive feedback selection for intelligent tutoring systems. *Expert Syst. Appl.* 38, 6146–6152. doi: 10.1016/j.eswa.2010.11.058

Haenlein, M., and Kaplan, A. (2019). A brief history of artificial intelligence: on the past, present, and future of artificial intelligence. *Calif. Manag. Rev.* 61, 5–14. doi: 10.1177/0008125619864925

Halai, A., Sarungi, V., and Hopfenbeck, T. N. (2022). Teachers' perspectives and practice of assessment for learning in classrooms in Tanzania. *Int. Encycl. Educ.* 63-72. doi: 10.1016/B978-0-12-818630-5.09039-4

Hattie, J. A. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. London and New York: Routledge.

Hill, P. W., and Barber, M. (2014). *Preparing for a renaissance in assessment* Pearson.

Hopfenbeck, T. N., and Stobart, G. (2015). Large-scale implementation of assessment for learning. *Assess. Educ.* 22, 1–2. doi: 10.1080/0969594X.2014.1001566

Hopfenbeck, T. N., Flórez Petour, M. T., and Tolo, A. (2015). Balancing tensions in educational policy reforms: large-scale implementation of assessment for learning in Norway. *Assess. Educ.* 22, 44–60. doi: 10.1080/0969594X.2014.996524

Humphry, S. M., and Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educ. Res.* 43, 253–263. doi: 10.3102/0013189X14542154

Johnston, S. -K. (2023). *Privacy considerations of using social robots in education: policy recommendations for learning environments*. United Nations, Department of Economics and Social Affairs, Sustainable Development.

Jönsson, A., and Panadero, E. (2016). "The use and Design of Rubrics to support assessment for learning" in *Scaling up assessment for learning in higher education*. eds. D. Carless, S. M. Bridges, C. K. Y. Chan and R. Golfcheski (York: Springer), 99–111. (https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/PreparingforaRenaissanceinAssessment.pdf)

Ke, Z., and Ng, V. (2019). Automated Essay Scoring: A Survey of the State of the Art. *Paper presented at the 28th International Joint Conference on Artificial Intelligence*.

King, P. M., and Kitchener, K. S. (1994). *Developing reflective judgment: understanding and promoting intellectual growth and critical thinking in adolescents and adults*. San Francisco: Jossey-Bass.

Kuhn, D., and Weinstock, M. (2004). "What is epistemological thinking and why does it matter?" in *Personal epistemology: the psychology of beliefs about knowledge and knowing*. eds. B. K. Hofer and P. R. Pintrich (New York: Routledge), 121–144.

Lee, A. V. Y. (2021). Determining quality and distribution of ideas in online classroom talk using learning analytics and machine learning. *Educ. Technol. Soc.* 24, 236–249.

Lee, A. V. Y. (2023). Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation. *Stud. Educ. Eval.* 77:101250. doi: 10.1016/j.stueduc.2023.101250

Li, C., Xing, W., and Leite, W. (2022). Using fair AI to predict students' math learning outcomes in an online platform. *Interact. Learn. Environ.* 1-20, 1–20. doi: 10.1080/10494820.2022.2115076

Li, T., Reigh, E., He, P., and Adah Miller, E. (2023). Can we and should we use artificial intelligence for formative assessment in science? *J. Res. Sci. Teach.* 60, 1385–1389. doi: 10.1002/tea.21867

van der Linden, W. J., and Glas, C. A. (2010). *Elements of adaptive testing*. New York: Springer.

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., and Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science* 344, 155–156. doi: 10.1126/science.1245980

Lipnevich, A. A., and Smith, J. K. (2018). *The Cambridge handbook on instructional feedback* Cambridge University Press.

Liu, N. F., and Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teach. High. Educ.* 11, 279–290. doi: 10.1080/13562510600680582

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., and Linn, M. C. (2016). Validation of automated scoring of science assessments. *J. Res. Sci. Teach.* 53, 215–233. doi: 10.1002/tea.21299

Luaces, O., Díez, J., and Bahamonde, A. (2018). A peer assessment method to provide feedback, consistent grading and reduce students' burden in massive teaching settings. *Comput. Educ.* 126, 283–293. doi: 10.1016/j.compedu.2018.07.016

Luckin, R., and Holmes, W. (2016). *Intelligence unleashed: an argument for AI in education*. Available at: https://discovery.ucl.ac.uk/id/eprint/1475756

Ma, W., Adesope, O. O., Nesbit, J. C., and Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: a meta-analysis. *J. Educ. Psychol.* 106, 901–918. doi: 10.1037/A0037123

McMillan, J. H. (2013). *SAGE handbook of research on classroom assessment*. Los Angeles: Sage.

Milano, S., McGrane, J. A., and Leonelli, S. (2023). Large language models challenge the future of higher education. *Nat. Mach. Intell.* 5, 333–334. doi: 10.1038/s42256-023-00644-2

Milligan, S. K., and Griffin, P. (2016). Understanding learning and learning design in MOOCs: a measurement-based interpretation. *J. Learn. Analyt.* 3, 88–115. doi: 10.18608/jla.2016.32.5

Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, S. R., Rakhshan, M., Keikha, L., and Ghazi Saeedi, M. (2021). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interact. Learn. Environ.* 29, 142–163. doi: 10.1080/10494820.2018.1558257

Murphy, R. F. (2019). Artificial intelligence applications to support K-12 teachers and teaching. *Rand Corp.* 10, 1–20. doi: 10.7249/PE315

Nicol, D., and MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Stud. High. Educ.* 31, 199–218. doi: 10.1080/03075070600572090

Panadero, E., and Jönsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: a review. *Educ. Res. Rev.* 9, 129–144. doi: 10.1016/j.edurev.2013.01.002

Panadero, E., Andrade, H., and Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: a roadmap of where we are, how we got here, and where we are going. *Aust. Educ. Res.* 45, 13–31. doi: 10.1007/s13384-018-0258-y

Perry, N., Lisaingo, S., Yee, N., Parent, N., Wan, X., and Muis, K. (2020). Collaborating with teachres to design and implement assessments for self-regulated learning in the context of authentic classroom writing tasks. *Assess. Educ.* 27, 416–443. doi: 10.1080/0969594X.2020.1801576

Popham, W. J. (1997). What's wrong - and what's right - with rubrics. *Educ. Leadersh.* 55, 72–75.

Ratnam-Lim, C. T. L., and Tan, K. H. K. (2015). Large-scale implementation of formative assessment practices in an examination-oriented culture. *Assess. Educ.* 22, 61–78. doi: 10.1080/0969594X.2014.1001319

Ruiz-Primo, M., and Brookhart, S. (2018). *Using feedback to improve learning*. Routledge.

Scalise, K., Wilson, M., and Gochyyev, P. (2021). A taxonomy of critical dimensions at the intersection of learning analytics and educational measurement. *Front. Educ.* 6:656525. doi: 10.3389/feduc.2021.656525

Schwartz, S. (2020). Flood of online learning resources overwhelms teachers. *Educ. Week*. Available at: March 25, 2020: https://www.edweek.org/teaching-learning/flood-of-online-learning-resources-overwhelms-teachers/2020/03

Shin, J., Chen, F., Lu, C., and Bulut, O. (2022). Analyzing students' performance in computerized formative assessments to optimize teachers' test administration decisions using deep learning frameworks. *Journal of Computers in Education* 9, 71–91. doi: 10.1007/s40692-021-00196-7

Skeat, J., and Ziebell, N. (2023). University students are using AI, but not how you think. Available at: https://pursuit.unimelb.edu.au/articles/university-students-are-using-ai-but-not-how-you-think

Stobart, G., and Hopfenbeck, T. (2014). "Assessment for learning and formative assessment" in *State of the field review assessment and learning*. eds. J.-A. Baird, T. Hopfenbeck, P. Newton, G. Stobart and A. Steen-Utheim (Oxford: Norwegian Knowledge Centre for Education).

Sun, S. Z. (2021). Epistemological beliefs: the key to enhance critical thinking for higher education students in the east. Paper presented at *the 2021 American Educational Research Association (AERA) annual meeting*.

Sun, Z. S. (2023). *Developing and validating an operationalisable model for critical thinking assessment in different cultures*. The University of Melbourne.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., et al. (2022). Assessment in the age of artificial intelligence. *Comp. Educ.* 3:100075. doi: 10.1016/j.caeai.2022.100075

Tashu, T. M., and Horvath, T. (2019). Semantic-based feedback recommendation for automatic essay evaluation. Proceedings of SAI Intelligent Systems Conference 334–346. London: Springer.

Tomasik, M. J., Berger, S., and Moser, U. (2018). On the development of a computer-based tool for formative student assessment: epistemological, methodological, and practical issues. *Front. Psychol.* 9:2245. doi: 10.3389/fpsyg.2018.02245

Verma, M. (2018). Artificial intelligence and its scope in different areas with special reference to the field of education. *Int. J. Adv. Educ. Res.* 3, 5–10.

Vincent-Lamarre, P., and Larivière, V. (2021). Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome. *Quant. Sci. Stud.* 2, 662–677. doi: 10.1162/qss_a_00125

von Davier, A. A., Mislevy, R. J., and Hao, J. (2022). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python*. Springer.

Webb, M., Gibson, D., and Forkosh-Baruch, A. (2013). Challenges for information technology supporting educational assessment. *J. Comput. Assist. Learn*, 29, 451–462. doi: 10.1111/jcal.12033

Wen, H., Liu, Y., and Zhao, N. (2020). Longitudinal cognitive diagnostic assessment based on the HMM/ANN model. *Front. Psychol.* 11:2145. doi: 10.3389/fpsyg.2020.02145

Wertz, M. H. (2019). *Epistemological developmental level and critical thinking skill level in undergraduate university students*. University of South Florida.

Wiliam, D. (2011). What is assessment for learning? *Stud. Educ. Eval.* 37, 3–14. doi: 10.1016/j.stueduc.2011.03.001

Wiliam, D., Lee, C., Harrison, C., and Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. *Assess. Educ.* 11, 49–65. doi: 10.1080/0969594042000208994

Wylie, C. E., and Lyon, C. J. (2015). The fidelity of formative assessment implementation: issues of breadth and quality. *Assess. Educ.* 22, 140–160. doi: 10.1080/0969594X.2014.990416

Wollenschläger, M., Hattie, J., Machts, N., Möller, J., and Harms, U. (2016). What makes rubrics effective in teacher-feedback? Transparency of learning goals is not enough. *Contemp. Educ. Psychol.* 44-45, 1–11. doi: 10.1016/j.cedpsych.2015.11.003

Zhai, X., and Nehm, R. H. (2023). AI and formative assessment: the train has left the station. *J. Res. Sci. Teach.* 60, 1390–1398. doi: 10.1002/tea.21885

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 29 November 2023
DOI 10.3389/feduc.2023.1253926

Check for updates

OPEN ACCESS

EDITED BY
Marcel Mierwald,
Georg Eckert Institute (GEI), Germany

REVIEWED BY
Mario Carretero,
Autonomous University of Madrid, Spain
Arja Virta,
University of Turku, Finland
Martin Nitsche,
University of Applied Sciences and Arts Northwestern Switzerland, Switzerland

*CORRESPONDENCE
David Rosenlund
david.rosenlund@mau.se

RECEIVED 06 July 2023
ACCEPTED 14 November 2023
PUBLISHED 29 November 2023

CITATION
Rosenlund D (2023) The nature of historical knowledge in large-scale assessments – a study of the relationship between item formats and offerings of epistemic cognition in the Swedish national test in history.
Front. Educ. 8:1253926.
doi: 10.3389/feduc.2023.1253926

COPYRIGHT
© 2023 Rosenlund. 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.

The nature of historical knowledge in large-scale assessments – a study of the relationship between item formats and offerings of epistemic cognition in the Swedish national test in history

David Rosenlund*

Faculty of Learning and Society, Malmö University, Malmö, Sweden

Issues of validity and reliability have an impact on the construction of tests. Since the 2010s, there has been increasing emphasis in Sweden on enhancing reliability in the large-scale test system to combat grade inflation. This study aims to examine how this increased focus on reliability has affected how the nature of historical knowledge is presented in the national test in history. Accordingly, it addresses the following research question: what kinds of epistemic cognition does the test communicate to students? The concept of *epistemic cognition* builds on Kuhn et al.'s discussion on epistemic understanding, regarding the balance between the objective and subjective dimensions of knowledge. Furthermore, the concept of *companion meanings* is used to establish a connection between the items in the test and students' epistemic cognition. The findings show that the selected-response tasks predominantly communicate an objective dimension of historical knowledge, while the constructed-response tasks communicate both subjective and objective dimensions of historical knowledge. The findings regarding the offerings of epistemic cognition are discussed in relation to validity, reliability, item formats and classroom practices.

KEYWORDS

epistemic cognition, epistemic beliefs, epistemic understanding, reliability, validity, history education, item formats

1 Introduction

In test construction, the balance between construct validity and assessment reliability is important. There are two aspects of validity that is addressed in the study presented here. First, the concept *construct validity* refers to the degree to which an assessment construct captures the knowledge dimensions in the intended construct. Second, *cognitive validity* addresses the degree to which an assessment instrument can be argued to elicit the intended cognitive processes (Kaliski et al., 2015). Reliability addresses the degree to which test items enable assessors to make equal evaluations of student responses with similar quality. Reliability concerns both the agreement between different teachers (inter-rater reliability) as well as

agreement between individual teachers' assessments of different student responses (intra-rater reliability). There is an ongoing debate in history education on the relation between validity, reliability, and item formats in test construction and assessment. One central aspect in this discussion is how appropriate the item formats used in the tests are to address the construct that is to be assessed. When it comes to reliability in assessment, there is a tendency to promote selected-response (SR) items, namely items in which select the correct answer from several predefined alternatives. These items are suitable for ensuring reliability because they reduce the room for different interpretations of student responses, so assessors are more likely to make equal evaluations of responses of the same quality, resulting in higher degrees of inter-rater reliability (Rodriguez, 2015). On the other hand, the literature tends to ascribe the ability to address more complex knowledge to open-ended questions, also known as constructed-response (CR) items, where pupils construct their answer on an item. This is because complex knowledge is often characterized by variety and issues of interpretation, which are more appropriate to express in open-ended contexts (Koretz, 2008). Another relevant aspect in the debate about validity, reliability, and item formats is that large-scale tests tend to influence what teachers include in their teaching, a phenomenon labeled as a 'washback effect' (Au, 2007; Hardy, 2015).

Existing studies on item formats in history tests have investigated challenges in the use of both SR and CR items in assessing historical knowledge. When constructing a test to tap students' ability to handle three historical thinking concepts— evidence, historical perspectives, and the ethical dimension— Seixas et al. (2015) chose to use a combination of both SR and CR items. The CR items were considered necessary to elicit information about the students' perceptions of the intended construct, which are fundamental to the discipline of history (Seixas, 2015). Similarly, US history tests in the National Assessment of Educational Progress (Lazer, 2015) and the Advanced Placement (Charap, 2015) programs use a combination of the two item formats. In these studies, the challenge that comes with CR items regarding reliability was also raised, stressing the need for assessment criteria. Such criteria are needed because students have to formulate their own answers to CR questions, the assessment of which leaves room for examiners' interpretations and risks lowering inter-rater reliability (Shemilt, 2018). In the three examples above, SR items are complemented with CR items because of the assumption that the latter are better suited to address more complex types of knowledge. This implies that SR items are seen as more appropriate for assessing the less complex knowledge type, factual, or content, knowledge. This position is problematized by Shemilt (2018), who argues that although using SR items to assess factual knowledge is practical, the format is still beset with challenges because its application is based on the questionable supposition that the items are interchangeable and thus of equal difficulty. Meanwhile, there are attempts to use SR items to address more complex knowledge of the history subject. Körber and Meyer-Hamme (2015) used SR items in a study to examine students' ability to handle historical accounts. The items included the content knowledge needed, and students were asked to apply more complex concepts and provide different answers to receive points. Other studies have problematized the use of SR items to assess complex knowledge in history because the format has proven to be difficult to use for addressing the interpretative nature of historical knowledge (Reich, 2009; Smith et al., 2019).

Based on the studies presented above, test developers face a challenge in the context of history, where the construct that is to be assessed consists of complex knowledge. In such a context, the need for reducing assessors' room for interpretations (to ensure reliability) has to be balanced with the need for measuring students' proficiency regarding knowledge with higher degrees of complexity. Such a situation is at hand in Sweden, where the balance between validity and reliability has become more prominent because of a persistent grade inflation. This inflation is characterized by an increasing tendency among teachers to hand out higher grades than are motivated by students' actual levels of knowledge. The main factor behind this process is the marketization of education in Sweden, initiated in the 1992, resulting in a competition between schools (Wennström, 2020). The Swedish National Agency for Education wants to handle this inflation by improving the national tests (Skolverket, 2021). This means that reliability is likely to be given more weight in the construction of national tests.

In the Swedish national test in history, administered annually to students in Grade 9 (15–16 years), the construct that is to be assessed is formulated in the official history curriculum. This curriculum is largely recontextualized from the academic discipline of history, and the complexity of this disciplinary knowledge is also transferred to the curricula (Samuelsson, 2014; Eliasson et al., 2015). Conversely, the share of SR items in the national history test increased between the years 2013 and 2015, a fact that is noteworthy, considering that these items address knowledge with a lower degree of complexity than the knowledge prescribed by the curriculum (Rosenlund, 2022). This observed discrepancy in complexity between the history curriculum and the SR items in the national history test raises a question regarding what kind of history subject the test offers to the students taking it. Is it a subject that aligns with the constructivist complexity of the historical discipline (Zeleňák, 2015), as recontextualized in the curriculum, or is it a subject where historical knowledge is communicated as objective statements?

The study presented here aims to further the understanding of how the balance between validity and reliability affects the subject of history that is communicated to students in the national test. To address this aim, the study pursued the following research question: What stances of epistemic cognition does the national test in history communicate to students?

2 Epistemic cognition and companion meanings

Epistemic cognition is a concept that is used to address individuals' perceptions of knowledge, how knowledge is constructed, validated and the limits it is beset with (Kitchner, 1983). In this study, the operationalization of the concept is rooted in a discussion about a closely related concept, epistemic understanding (Kuhn et al., 2000), where it is characterized by the coordination of the objective and subjective dimensions of knowledge. Since this division implies that there are elements in a subject that can be treated as objective entities, and that how individuals organize and make meaning of these objective elements can be described as a subjective dimension of knowledge, it provides elements suitable to establish an analytical framework for this study. The concept is relevant to address in an educational context, since a more qualified understanding of epistemic

issues affects how individuals can utilize the knowledge that they have (Kuhn et al., 2000). Further, when students have more nuanced epistemic stances toward historical knowledge, their proficiency in other aspects of the subject increases (Van Boxtel and van Drie, 2017). However, research has shown that many adolescents have more simplistic epistemic understandings of history (Lee and Shemilt, 2004; Miguel-Revilla, 2022), indicating that history education needs to address issues of epistemic cognition (Seixas, 2015). Research has provided examples of how that can be done (Marczyk et al., 2022).

In history education research, the first two decades of the 21st century saw increased attention to the concept of epistemic cognition. It was built on ideas that took a coherent form in the United Kingdom in the 1970s about school history as a subject where strategies from the academic historical discipline are prominent (Shemilt, 1983). Similar ideas about history education took form in both the United States (Wineburg, 1991) and Canada (Seixas, 2015), partially building on the ideas from the United Kingdom. The UK-based research on the disciplinary strategies resulted in progression models that formulated suggestions regarding how students' knowledge about such strategies develop. These disciplinary strategies are recontextualized into so-called second-order concepts. These concepts are *evidence*, which concerns how information in historical sources can be addressed (Lee and Shemilt, 2003); *accounts* about the construction of historical narratives (Lee and Shemilt, 2004); and *causation*, which emphasizes how historians establish relations between historical phenomena and their causes and consequences (Lee and Shemilt, 2009).

The definition of epistemic cognition as the coordination between the objective and subjective dimensions of knowledge is generic, thereby making it applicable to several knowledge domains. It has already been applied in studies addressing history education. For instance, Maggioni et al., 2009 used this definition as one source of inspiration in an influential study on epistemic cognition in history. Their second source of inspiration was the aforementioned progression model related to evidence. In this progression model, the coordination between the objective and subjective dimension is described differently on each level. Maggioni et al., 2009 used the model and the generic model of epistemic cognition as scaffolds to define three epistemic stances—the copier, the borrower, and the criterialist stance—and construct an instrument to map individuals' epistemic cognition. The three stances describe a development in epistemic cognition from a more naïve view of knowledge that is characterized by acknowledging only the objective dimension of knowledge, via a focus on the subjective dimension, to a more nuanced view that coordinates both the objective and subjective dimension. In history, this could mean moving away from an understanding that there only is one answer to the question of whether the consequences of the industrial revolution were positive or negative; individuals holding this understanding cannot use different perspectives or address different interpretations of the consequences. As epistemic cognition progresses, individuals move from this stance to one where the subjective dimension of knowledge replaces the objective, resulting in knowledge being seen as mere opinions. In the history subject, in line with Lee and Shemilt (2003, 2004, 2009), this would mean that students see historical knowledge as dependent only on the historian's viewpoint, bias, and other personal attributes. On this level, the methodological strategies that are used by historians to bridge the objective and subjective dimensions are not yet acknowledged. On the most advanced level, students form an understanding where both the objective and subjective dimensions are acknowledged and coordinated. Here, historical knowledge is seen as the result of a subjective arrangement of objectively observable phenomena.

In research on students' epistemic cognition, both models mentioned above have been used as analytical frameworks. Stoel et al. (2017) revised the instrument constructed by Maggioni and VanSledright and surveyed 922 secondary school students. They found that students who can coordinate the two dimensions of knowledge also find history as a subject more interesting. Basing their analytical framework on Kuhn et al.'s model, Ní Cassaithe et al. (2022) conducted an interview study with 17 primary school students and found that the students' view on the nature of history and the concept of evidence affect their possibilities for progression in epistemic cognition. Similar results were reported in a survey study with 62 undergraduate students; based on the framework by Stoel et al. (2017), Sendur et al. (2022) found a strong correlation between epistemic beliefs and the quality of source-based argumentation.

Teachers are important as educators of epistemic cognition, and in the context of this study, their assessment practices are of extra interest and there are studies that highlights epistemic cognition within history tests. Van Nieuwenhuyse et al. (2015) examined four tests by each of 70 Flemish upper-secondary school teachers and found that 3 % of the questions in the tests made the constructed nature of historical knowledge visible for the students. In a similar study, Rosenlund (2016) examined all tests used by 23 upper-secondary school teachers during one academic year and found that 3.5 percent of 893 tasks communicated what is labeled as an integrated epistemic cognition in this study.

In this study, the assumptions underlying the model established by Kuhn et al. (2000) are used to formulate an analytical framework consisting of three categories. Each category describes one particular way in which historical knowledge can be presented in the tests, and thus offered to the students. The three categories comprise historical knowledge presented from (a) an objective perspective, (b) a subjective perspective, and (c) both an objective and subjective perspective. Although Kuhn et al.'s model is developmental—that is, the authors explicitly stated that it is meant to describe a progression between levels of cognitive cognition—this study does not use the model to address issues of development. Rather, it aims to employ it to identify what dimensions of historical knowledge are presented to the students; accordingly, the approach can be described as more dimensional (Nitsche et al., 2022). Furthermore, how the balance between the objective and subjective dimensions of knowledge is communicated in the national test in history will be used to indicate the companion meanings regarding epistemic cognition that are offered to the students encountering the tests.

The concept of *companion meanings* (Roberts, 1998) directs attention to the implicit learning that is present in educational contexts. Implicit learning is not just present in situations where learning is foregrounded; it is present in most educational situations, including testing. The national test in history, which is investigated in this study, carries with it companion meanings that offer students certain ways to understand historical knowledge. That the companion meanings are *offered* indicates that the meanings identified by the researcher do not have a one-to-one relationship with the meanings perceived by the students. Therefore, what this study examines is what stances to epistemic cognition that is *most likely* perceived by the

students taking the test. Since the concept companion meanings focuses on what is implicitly communicated, the intentions behind the items in the history test are of secondary importance in this study. Regardless of the complexity of the knowledge that the Swedish National Agency for Education is looking to assess, the items in the national tests communicate certain companion meanings regarding epistemic cognition to the students. Generally, tests communicate to students what aspects and dimensions of a subject are deemed as important (Black and Wiliam, 2009). Students are assumed to perceive the companion meanings in the national tests as important for three reasons: First, the tests are constructed externally from the schools; second, they are administered by an authoritative body, the National Agency of Education; and third, they are high-stake tests that affect the students' grades in history. These factors contribute to making the national tests appear authoritative to the students, which is why the tests' companion meanings regarding epistemic cognition are likely to have an impact on them.

In sum, three assumptions underlie the analytical framework (see Table 1) of this study: When an item in the test presents knowledge from an objective perspective, it offers students an *objectivist* epistemic stance concerning historical knowledge. In the same way, items presenting historical knowledge from a solely subjective perspective are assumed to offer students a *subjectivist* epistemic stance concerning historical knowledge. Finally, items containing both the objective and the subjective dimensions of historical knowledge are assumed to offer an *integrative* epistemic stance concerning historical knowledge.

3 Context of the national test and its role in teachers grading practice

As mentioned above, the construct that is assessed in the Swedish national test in history is the official history curriculum. The curriculum that was in effect when the tests investigated in this study were administered aimed at enabling students to develop four abilities after participating in history education: (1) to use a historical frame of reference that incorporates different interpretations of time periods, events, notable figures, cultural meetings, and development trends; (2) to critically examine, interpret, and evaluate sources as a basis for creating historical knowledge; (3) to reflect over their own and others' use of history in different contexts and from different perspectives; and (4) to use historical concepts to analyze how historical knowledge is organized, created, and used (Skolverket, 2011). As mentioned in the introduction, these four abilities share similarities with the academic subject of history (Samuelsson, 2014). A common feature of the abilities in the curriculum is that they communicate that interpretations have a central role in the subject, something that historians also express when describing the subject (McCullagh, 2004; Berkhofer, 2008). The level descriptors in the curriculum, characterizing the differences between the grades E, C and A, are not incorporated from the disciplines (Rosenlund, 2019), instead the Agency of Education (2011) have used generic descriptions of increasing complexity to describe them. In one line of progression, addressing how students handle relationships between time periods, the levels are described with *simple* (for grade A) – *relatively complex* (for grade B) and *complex* (for grade C).

TABLE 1 Description of coding categories.

Coding category	Description
Objectivist stance	Historical knowledge is communicated as consisting of objective elements and statements about history can be true or false.
Subjectivist stance	Historical knowledge is communicated as consisting of subjective elements and statements about history can be true or false
Integrative stance	Historical knowledge is communicated as consisting of a coordination between objective and subjective elements and statements about history can vary in quality.

In Sweden, teachers have a large degree of autonomy when it comes to grading their students. However, the national tests are an instrument that infringes on this autonomy. Namely, teachers are obliged to take the test results into account when they grade their students—a feature of the tests that was further strengthened in 2018 (Skolverket, 2018). The students receive a grade on each item in the tests, and these item grades are combined into a test grade: F (fail), E (pass), C (pass with distinction), and A (pass with special distinction). There are national tests in primary school and in both lower- and upper-secondary school. The tests are most common in the subjects Swedish (mother tongue), Mathematics, and English, which are administered at all three stages. In the natural science subjects, the tests are administered in lower- and upper-secondary school, whereas tests in social science subjects are used solely in lower-secondary school. Students have to take the national test in one of the four social science subjects (geography, history, religion, and social science) and one of the three natural science subjects (biology, chemistry, and physics). This study examines the history test.

4 Materials and methods

The empirical material examined in this study consists of the items in large-scale tests, the Swedish national history tests for Grade 9 conducted 2016–2019. The four tests included in this study contained 89 graded items categorized into two types: selected-response items and constructed response items. In many cases, the items comprised subitems that were graded individually, and these subitem-grades were combined to make up a total item grade (F, E, C, and A). This was the case for both SR items and CR items. In this study, each subitem is considered to offer a companion meaning to the students taking the test. This is based on the assumption that the companion meanings are inherent in each of the stems the students encounter and that this is valid both when the item in itself results in a grade and when its result is combined with other items. Accordingly, although the tests contained 89 graded items, 507 items were analyzed in this study.

The methodological approach used in the study is content analysis (Krippendorff, 2004), which is a useful method when looking for implicit meanings in texts. Each item was examined in relation to the analytical framework in a three-stage process (Bowen, 2009). The first two stages involved an initial, superficial reading followed by a thorough examination, and the third stage entailed a final categorization and coding based on the analytical framework (Bowen, p. 32–33). In order to enhance intra-rater reliability a reexamination

of the material was conducted. The aim was to ensure that the understanding of the analytical framework shifted during the first round of analysis. Following this process, each item was provided with two codes, one that categorizes the epistemic cognition that is communicated in the item and one that describes whether it is an SR or a CR item. The corresponding researcher was responsible for the coding procedure and in order to enhance transparency, the principles of the coding are presented in detail below.

5 The balance between the objective and subjective dimensions of historical knowledge

The 507 items were categorized according to the analytical framework presented above. The analysis of the empirical material will be presented in the following order: first, two items coded as offering an objectivist epistemic stance concerning historical knowledge; second, two items coded as offering a subjectivist epistemic stance; and lastly, two items coded as offering an integrative stance, combining the subjective and the objective dimensions of historical knowledge. The items that serve as examples in each of the following subsections are all from the 2017 test, since it at the time of publication is the most recent test that is not confidential. These items were chosen as representative examples of the item formats in the tests, and they will also be used to explain how the analytical framework has been applied in relation to the empirical material. This is because they share important characteristics with items in the national tests in history of the same format. This goes both for the other items in test from 2017, but also for the items in the tests conducted in 2016, 2018 and 2019.

5.1 Items representing an objectivist stance concerning historical knowledge

As mentioned in the theoretical section, knowledge is understood to contain both an objective and a subjective dimension. This section addresses the 432 items in the national test that solely present the objective dimension of historical knowledge. In the four analyzed tests, only SR items present knowledge from solely an objective perspective.

The first example is item number 4, which has been selected as a representative SR example due to its large number of subitems; it consists of 12 subitems, four of which are presented in Figure 1. These subitems have been categorized as communicating an objectivist epistemic stance because for each of them, there is only one correct alternative. Consequently, the subjective dimension of historical knowledge (for example, sampling, interpretation, and representation) is not offered to the students. This is not an argument for a view of historical knowledge where there are no objective aspects; the tricolor was indeed an important artifact during the French Revolution, and the Soviet Communist Party did use the symbol of the hammer and sickle. However, there are historical themes in these subitems where the subjective dimension is present but not made visible in the test. For example, two subjective aspects can be argued to nuance the objectivist stance communicated in these two items. First, what criteria can be used to substantiate that the tricolor was introduced during the French Revolution, not in the Netherlands in the 15th century? Second, the use of a hammer and a sickle to address the unity between workers and peasants can be found in other contexts before the Russian Revolution, so linking its origin to the Russian Revolution can also be nuanced by adding on a subjective perspective.

Another example of an item consisting of multiple subitems is number 18 (see Figure 2). In this item, students are asked to fill in a table with letters provided in four lists, where each letter represents (a) dates, (b) people, (c) events, and (d) countries/regions. The information from the lists is to be matched with the periods that are found in the table's leftmost column. In this item, there are 19 pieces of information that the students have to match to one of the historical periods. This type of item is present in several of the tests. It has been categorized as offering an objectivist epistemic stance for two reasons. First, each of the 19 pieces of information and the five historical periods are presented as objective entities; namely, there are no indications of a subjective aspect that can be related to them. Second, there can be only one correct response in each cell, meaning that students cannot take their subjective perspectives on history into account when responding to the item, which is also a factor for coding this kind of item as objective.

The four tests administered between 2016 and 2019 contained 432 items that have been coded as presenting historical knowledge from an objective perspective. That is, students responding to these items are being offered an objectivist epistemic stance concerning historical knowledge.

Regarding the issue of reliability and validity, the items categorized into the objective category have one important thing in common: they leave no room for interpretation in the assessment process. This feature increases both the inter- and the intra-rater reliability of the tests containing such items. On the issue of validity, since the items categorized to the objective dimension are void of the interpretative aspects of history that are present in the history curriculum, the construct validity is compromised. Also, as the number of correct results on subitems are combined to an item grade without any differentiation between the subitems regarding difficulty, the cognitive validity is reduced. This is because the level descriptors in the curriculum for the grades E, C and A are characterized by increasing complexity and not by an accumulation of content knowledge – which is what is rewarded in this kind of tasks in the national test in history.

5.2 Items representing a subjectivist stance concerning historical knowledge

Six items in the tests were coded as showing only the subjective dimension of history. These items address two different aspects of the history curriculum, namely, the uses of history and the handling of historical sources.

An example where the subjective dimension is likely to be prominent for students taking the test is item number 21 (see Figure 3). In this item, the students are supposed to provide a reason why a company in an advertisement is using a reference to its origin in 1880. This item aims to tap into student knowledge regarding the use of history, an aspect that was introduced in the history curriculum in 2011 and that history teachers in Swedish lower-secondary schools tend to neglect in their teaching (Skolinspektionen, 2015; Eliasson and Nordgren, 2016). Due to this neglect, a large share of students have likely not met this aspect as

Connections between the present day and the American, French and Russian Revolutions

4 Complete each sentence by putting a cross against the American, French or Russian Revolutions. Put **one** cross for each task (1-12).

1. Many flags of the world consist of three stripes in different colours. Such flags are called tricolour. The tricolour comes from ...

The American Revolution	
The French Revolution	
The Russian Revolution	



An example of tricolor: the Belgian flag.

2. One country is governed by a Senate and a House of Representatives. The country decided on this form of government after ...

The American Revolution	
The French Revolution	
The Russian Revolution	

3. A symbol for a particular political ideology is "the hammer and sickle". It comes from ...

The American Revolution	
The French Revolution	
The Russian Revolution	



The hammer and sickle.

4. The Marseillaise is a famous national anthem. The Marseillaise was sung for the first time during ...

The American Revolution	
The French Revolution	
The Russian Revolution	

8

FIGURE 1
Item 4 in the Swedish national test in history, 2017.

a part of their history education. The item lacks scaffolds that could inform students about how to respond to it, which increases the likelihood that students would not be aware of the objective aspects in this item. Students who lack knowledge of strategies that are productive for analyzing the uses of history and who are confronted with these tasks without any scaffolds are likely to perceive this kind

Periods of history

18 Complete tasks 1-4 by writing the correct **letters** in the table.
You should only write **one letter** in each square.
We have already written in the letter for Great Britain in Column 4: Important areas/states.

Period of History	1: Dates	2: People	3: Events	4: Important regions/ states
Ancient Egypt				
European Colonialism				
The Industrial Revolution in Great Britain				F
The World Wars				
The Post-War Period				

1. Dates

Write **the letter** for the date the period of history **begins** in column 1.
Three letters will remain over.

- **A.** 3000 BC
- **B.** 13th century
- **C.** 14th century
- **D.** 15th century
- **E.** 18th century
- **F.** 1914
- **G.** 1946
- **H.** 1989

12

FIGURE 2
Item 18 in the Swedish national test in history, 2017.

of item as presenting history solely from a subjective perspective. This means that these students are likely to perceive their response to this item as merely a personal standpoint.

The second area where the subjective dimension is likely to speak most clearly to students concerns the historicity of individuals' actions and ideas. This area is addressed in item number 11, which references

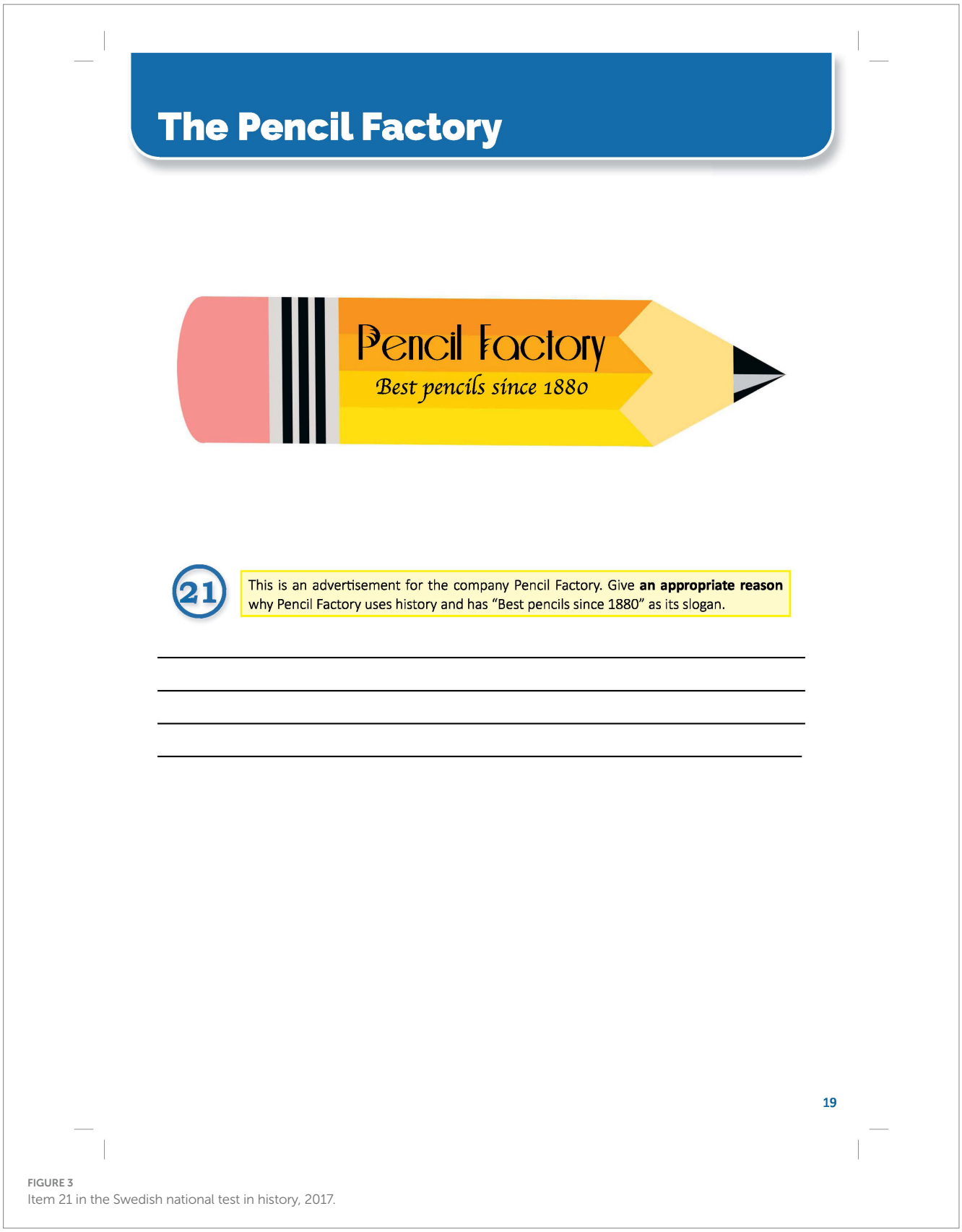


FIGURE 3
Item 21 in the Swedish national test in history, 2017.

a commercial for men's clothing (see Figure 4). In this item, students are supposed to provide reasons why the person in the commercial is smoking. To be able to do this, students need to have met the idea of historicity and learned how to apply it. Considering that historicity is a rather complex concept and a procedural aspect of history, an aspect that is seldom addressed in Swedish history education

Advertising then and now



Advertisement from 1953. The picture is advertising suits for men. The men in the picture are smoking on board an aeroplane.

11 Why did people smoke in advertising pictures in 1953? Give **one** reason.

22

FIGURE 4
Item 11 in the Swedish national test in history, 2017.

(Skolinspektionen, 2015; Eliasson and Nordgren, 2016), a large share of students are likely unfamiliar with the concept. As with item 21 above, the lack of scaffolds also contributes to the possibility that students fail to identify the objective aspects in the item and, thus, perceive the historical knowledge addressed in this kind of item as solely a subjective enterprise.

There are six items in the examined tests that have been categorized as addressing historical knowledge from a subjective perspective, all being in the CR format. The conclusions regarding these items should be considered as tentative since they are based on indications found in previous research. Nonetheless, these items likely offer students a subjectivist epistemic stance concerning historical knowledge, meaning that knowledge in the subject is more akin to opinions than based on criteria and related to objective reality.

The fact that these items lack scaffolds and address aspects that teachers tend to neglect in their teaching is likely to reduce inter-rater reliability. Regarding construct validity, the items tap into content of the history curriculum, providing students with the opportunity to show their proficiency in relation to a relevant historical content. The lack of scaffolds, however and the indications of scarce teaching in these areas risks compromising the cognitive validity. Reasons for this is that many students are likely to be unaware of what to include in their responses, something that might result in a disconnect between the level descriptors and the responses.

5.3 Items representing an integrative stance concerning historical knowledge

In the analyzed tests, 69 items have been characterized as addressing historical knowledge from a perspective where both the subjective and objective dimensions are visible. The example in Figure 5 concerns one of the CR items that have been coded as part of this integrative category. In this item, students are supposed to respond by providing examples of consequences of the industrial revolution, and they have one and a half pages at their disposal for this task. The grading rubric contains descriptions of the criteria that a response has to meet for each of the three passing grades (E, C and A). The objective dimension is present in this item in that it directs students' attention to a historical phenomenon and asks for its consequences. This is likely to communicate to the students that there are events that have happened in the past and that these have real consequences. The subjective dimension is first communicated to the students through the term *discuss* in the stem, indicating that there is not one correct way to address this item. *Discuss* is a frequently used term in the Swedish school system; it tells students that there are several aspects to be considered and that an evaluation of some sort can be included in the response. Thus, both the objective and subjective dimensions of historical knowledge are present in this item, demonstrating that it offers students an integrative epistemic stance concerning historical knowledge.

In item number 22, the SR-subitems are meant to address students' understanding of the uses of history (see Figure 6). The students are provided with two contrasting standpoints on Swedish membership in NATO through fictive persons who use historical arguments to substantiate the two standpoints. Subitems one, two, and six (shown in Figure 6) have been coded as addressing both the objective and the subjective dimension of historical knowledge because they apply to both the positive and the negative standpoints. They communicate that how the objective dimension (i.e., the historical examples referenced in the item) is used can be dependent on the subjective dimension. In this case, the subjective dimension relates to attitude toward NATO membership.

The tests contain 69 items where both the objective and the subjective dimensions have been identified. These items offer students an integrative epistemic stance, where both dimensions of knowledge are present. Among these items, 61 are in the CR format and 9 are subitems in the SR format. The SR-items present in the tests from 2016 and 2017 have high inter rater reliability since there are no room for interpretation in the assessment of student responses, and that both the objective and subjective dimensions are present increases construct validity. However, as the case is with the items addressing only the objective dimension discussed above, the cognitive validity of these SR-items is compromised because there is no differentiation regarding difficulty between them.

One aspect included in the CR-items that can help strengthen both inter- and intra-rater reliability is assessment rubrics, reducing the room for interpretations in assessments of the student responses. These rubrics also have the potential to strengthen the cognitive validity because they make it transparent what differences there are between the three grading levels. Finally, that the two dimensions of knowledge are combined strengthens the construct validity of these items.

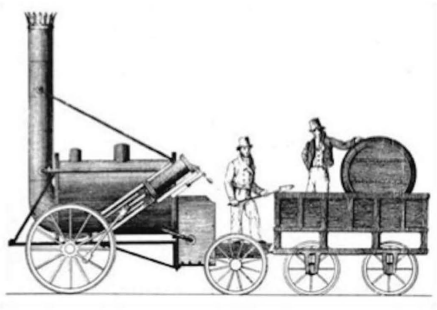
6 Conclusions and discussion

There is one consistent pattern in the four tests examined in this study: a vast majority of the items address the knowledge dimensions separately. In each of the four test years, between 84 and 90% of the items address one of the dimensions in isolation. Among this majority of items, only a small share addresses the subjective dimension, whereas between 82 and 87% of the items address the objective dimension. Moreover, only 10 to 16% of the items present historical knowledge as consisting of both the objective and the subjective dimensions. These numbers are presented in Table 2. For each year, the gray rows show the number of items that correspond with each of the three categories of epistemic cognition. The non-shaded rows show the number and percentage of SR and CR items each year.

A majority of the 507 items offer the students in Grade 9 a one-dimensional and objectivist epistemic stance concerning historical knowledge. I will in the following discuss this finding from two perspectives: first, in relation to students' proficiency in history and, second, in relation to test construction and issues of reliability and validity. Regarding students' proficiency in history, there is a risk that an objectivist approach in large-scale testing strengthens the objectivist preconceptions of historical knowledge already held by many adolescents (Lee and Shemilt, 2004; Miguel-Revilla, 2022). This is likely to impede the endeavors of history teachers to educate their students so that they can acquire an integrated epistemic cognition of historical knowledge. Such a nuanced epistemic cognition is necessary for developing other relevant aspects of the history subject, such as historical reasoning (Van Boxtel and van Drie, 2017) and source based argumentation (Sendur et al., 2022). Importantly, a more nuanced understanding of historical knowledge also increases individuals' ability to handle conflicting accounts (Nokes, 2014), an ability of great importance for active participation in society (Lee, 2011).

Regarding test construction, the analysis of the relation between the item formats and the epistemic cognition that is offered to students shows two significant correlations. The first is that the objectivist epistemic stance is exclusively offered in SR items and the second is

Consequences of the industrial revolution



Drawing of Stephenson's steam engine Rocket in an American technical journal from 1829.

7 Write about and discuss the **consequences** of the industrial revolution using **historical examples**.

Your discussion should include		
E	C	A
Two consequences	**Three** consequences. **At least two different types** of consequence, for example social, economic, political etc.	The same as for C, but also one consequence that **occurred then** and a consequence that **had a long-term affect**. Discuss this by **describing in what way** it was a consequence that lasted for a long time.

14

FIGURE 5
Item 7 in the Swedish national test in history, 2017.

that the integrative epistemic stance is offered mostly in CR items. This is in line with previous research where the use of SR-items for assessing complex constructs have been questioned (Reich, 2009; Smith et al., 2019). This study adds two aspects to this this research. First, that history assessments communicate perceptions about the nature of historical knowledge. Second, that SR-items, the way they



FIGURE 6
Item 22 in the Swedish national test in history, 2017.

are used in the context of this study, communicate an objectivist stance concerning historical knowledge. This finding strongly indicates that increasing the share of SR items in tests is likely to result in a more persistent offering of an objectivist epistemic stance. Since SR items are efficient in achieving high inter-rater reliability (Rodriguez, 2015), the share of SR items in large-scale tests is likely to

TABLE 2 Number of items for each category, divided by item type.

	Epistemic stances							
Year	Objectivist		Subjectivist		Integrative		Total	
2016	109 (87%)		0 (0%)		16 (13%)		125	
	SR: 109 (100%)	CR: 0	SR: 0	CR: 0	SR: 4 (25%)	CR: 12 (75%)	SR: 109 (87%)	CR: 16 (13%)
2017	95 (82%)		2 (2%)		19 (16%)		116 (100%)	
	SR: 95 (100%)	CR: 0	SR: 0	CR: 2 (100%)	SR: 4 (21%)	CR: 15 (79%)	SR: 99 (85%)	CR: 17 (15%)
2018	122 (87%)		4 (3%)		14 (10%)		140 (100%)	
	SR: 122 (100%)	CR: 0	SR: 0	CR: 4 (100%)	SR: 0	CR: 14 (100%)	SR: 122 (87%)	CR: 18 (13%)
2019	106 (84%)		0 (0%)		20 (16%)		126 (100%)	
	SR: 106 (100%)	CR: 0	SR: 0	CR: 0	SR: 0	CR: 20 (100%)	SR: 106 (84%)	CR: 20 (16%)

For each year, the gray rows show the number of items that correspond with each of the three categories of epistemic cognition. The non-shaded rows show the number and percentage of SR and CR items each year.

be substantial when reliability is a prioritized objective, as is the case in Sweden. Also, such a prioritization would risk compromising both the construct- and the cognitive validity of the national test in history. Regarding construct validity, the results presented in this study show that the objective and subjective dimensions of knowledge rarely are combined in SR-items. The risk for cognitive validity comes from the fact that the SR-items in the test are interchangeable and do thus not mirror the progression of complexity described in the curriculum.

Considering the possibility of a washback effect (Au, 2007), it is not a farfetched concern that if the objectivist epistemic stance will predominate in the large-scale tests, it is likely that will also be so in the history classrooms. If students are to be acquainted with an integrative epistemic stance, teachers have to promote such an understanding in their own teaching despite what is emphasized in the tests. History education researchers have suggested concept such as evidence, causation, and significance (Lee and Shemilt, 2004; Seixas, 2015) as disciplinary strategies that an education promoting an integrative epistemic stance can focus on. That said, research on teachers' assessment practices indicates that these findings from history education research have yet to be applied in history classrooms (Van Nieuwenhuyse et al., 2015; Rosenlund, 2016).

Bearing in mind that the four social science subjects examined through national tests in Sweden (i.e., geography, history, civics, and religious education) share many similarities, both regarding the complexity of curriculum content (Samuelsson, 2014) and the construction of the national tests, endeavors to ensure reliability in assessment are likely to result in similar consequences regarding the epistemic cognition offered in all four subjects. Furthermore, an approach to ascribe items the same value with no differentiations based on difficulty is likely to strengthen the image of the subjects as consisting of non-related pieces of information (Shemilt, 2018), increasing the obstacles for students to construct coherent images of the subject at hand (Shemilt, 2009).

Returning to the balance between reliability and validity, this study cannot provide an answer regarding how a reasonable balance can be achieved. However, the study can offer one recommendation that may be useful when seeking such a balance. Test developers are encouraged to consider how the balance between validity and reliability impacts how the subject is perceived by the test takers. If, for example, the share of SR items are increased in an attempt to increase inter- and intra-rater reliability, they should be mindful of the effects this strategy can have on the epistemic cognition offered to students. This is a crucial point because companion meanings in the test items impact students' perception of what is important in a subject (Black and Wiliam, 2009). In addition, offering to students a one-dimensional cognition of knowledge hinders their ability to use their knowledge in constructive ways (Kuhn et al., 2000) when participating in society.

Finally, it is important to mention two limitations of his study. First, this study is not a reception study, meaning that the relation between the epistemic stances communicated in the items and the statements made here about how students perceive the nature of historical knowledge is theoretical. Second, this study does not take into account other aspects of history education where epistemic stances are communicated, history teaching and teacher-made tests, being two examples. This means that this study addresses one of several aspects that influence students' epistemic cognition. However, the consistency of results presented above, regarding the stances that are communicated in the tests, calls for further research on the washback effects of high-stakes testing on students' epistemic cognition.

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

DR: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares 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.

References

Agency of Education. (2011). *Curriculum for the compulsory school system, the pre-school class and the leisure-time Centre 2011*. Stockholm: Swedish National Agency for Education Skolverket: Fritze distributör.

Au, W. (2007). High-stakes testing and curricular control: a qualitative metasynthesis. *Educ. Res.* 36, 258–267. doi: 10.3102/0013189X07306523

Berkhofer, R. F. (2008). *Fashioning history*. New York: Palgrave, Macmillan.

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. *Educ. Assess. Eval. Account.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qual. Res. J.* 9, 27–40. doi: 10.3316/QRJ0902027

Charap, L. G. (2015). "Assessing historical thinking in the redesigned advanced placement United States history course and exam" in *New directions in assessing historical thinking*. eds. K. Ercikan and P. Seixas (New York: Routledge), 159–170. doi: 10.4324/9781315779539

Eliasson, P., Alvén, F., Yngvéus, C.A. and Rosenlund, D. (2015). "Historical consciousness and historical thinking reflected in large-scale assessment in Sweden," in *New directions in assessing historical thinking*. eds. K. Ercikan and P. Seixas. Routledge, 171.

Eliasson, P., and Nordgren, K. (2016). Vilka är förutsättningarna i svensk grundskola för en interkulturell historieundervisning? [what are the conditions in Swedish lower secondary school for intercultural history teaching?]. *Nordidicatica-J Hum Soc Sci Educ* 2, 47–68.

Hardy, I. (2015). A logic of enumeration: the nature and effects of national literacy and numeracy testing in Australia. *J. Educ. Policy* 30, 335–362. doi: 10.1080/02680939.2014.945964

Kaliski, P., Smith, K., and Huff, H. (2015). "The importance of construct validity evidence in history assessment: what is often overlooked or misunderstood?" in *New directions in assessing historical thinking*. eds. K. Ercikan and P. Seixas (New York: Routledge)

Kitchner, K. S. (1983). Cognition, metacognition, and epistemic cognition: a three-level model of cognitive processing. *Hum. Dev.* 26, 222–232. doi: 10.1159/000272885

Körber, A., and Meyer-Hamme, J. (2015). "Historical thinking, competencies, and their measurement: challenges and approaches" in *New directions in assessing historical thinking*. eds. K. Ercikan and P. Seixas (New York: Routledge), 89–101.

Koretz, D. M. (2008). *Measuring up*. Cambridge, Mass: Harvard University Press.

Krippendorff, K.H. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks: Sage Publications.

Kuhn, D., Cheney, R., and Weinstock, M. (2000). The development of epistemological understanding. *Cogn. Dev.* 15, 309–328. doi: 10.1016/S0885-2014(00)00030-7

Lazer, S. (2015). "A large-scale assessment of historical knowledge and reasoning: NAEP US history assessment" in *New directions in assessing historical thinking*. eds. K. Ercikan and P. Seixas (New York: Routledge), 145–158.

Lee, P. (2011). "History education and historical literacy," in *Debates in history teaching*, ed. I. Davies. Routledge London).

Lee, P., and Shemilt, D. (2003). A scaffold, not a cage: progression and progression models in history. *Teach. Hist.* 113, 13–23.

Lee, P., and Shemilt, D. (2004). 'I just wish we could go back in the past and find out what really happened': progression in understanding about historical accounts. *Teach. Hist.* 117, 25–31.

Lee, P., and Shemilt, D. (2009). Is any explanation better than none? Over-determined narratives, senseless agencies and one-way streets in student's learning about cause and consequence in history. *Teach. Hist.* 137, 42–49.

Maggioni, L., VanSledright, B., and Alexander, P. A. (2009). Walking on the borders: a measure of epistemic cognition in history. *J. Exp. Educ.* 77, 187–214. doi: 10.3200/JEXE.77.3.187-214

Marczyk, A. A., Jay, L., and Reisman, A. (2022). Entering the historiographic problem space: scaffolding student analysis and evaluation of historical interpretations in secondary source material. *Cogn. Instr.* 40, 517–539. doi: 10.1080/07370008.2022.2042301

McCullagh, C.B. (2004). *The logic of history: Putting postmodernism in perspective*. Routledge, New York

Miguel-Revilla, D. (2022). What makes a testimony believable? Spanish students' conceptions about historical interpretation and the aims of history in secondary education. *Historical Encounters* 9, 101–115. doi: 10.52289/hej9.106

Ní Cassaithe, C., Waldron, F., and Dooley, T. (2022). We can't really know cos we weren't really there: identifying Irish primary children's bottleneck beliefs about history. *Historical Encounters* 9, 78–100. doi: 10.52289/hej9.105

Nitsche, M., Mathis, C., and O'Neill, D. (2022). Epistemic cognition in history education, historical. *Encounters* 9, 1–10. doi: 10.52289/hej9.101

Nokes, J. D. (2014). Elementary students' roles and epistemic stances during document-based history lessons. *Theory & Research in Social Education* 42, 375–413. doi: 10.1080/00933104.2014.937546

Reich, G. A. (2009). Testing historical knowledge: standards, multiple-choice questions and student reasoning. *Theory & Research in Social Educ* 37, 325–360. doi: 10.1080/00933104.2009.10473401

Roberts, D. A. (1998). "Analyzing school science courses: the concept of companion meaning" in *Problems of meaning in science curriculum*. eds. D. A. Roberts and L. Ostman (New York: Teachers College Press), 5–12.

Rodriguez, M. C. (2015). "Selected-response item development" in *Handbook of test development*. eds. S. Lane, M. R. Raymond and T. M. Haladyna (New York: Routledge).

Rosenlund, D. (2016). "*History education as content, methods or orientation? A study of curriculum prescriptions, teacher-made tasks and student strategies*". (Frankfurt: Peter Lang).

Rosenlund, D. (2019). Internal consistency in a Swedish history curriculum—a study of vertical knowledge discourses in aims, content and level descriptors. *J. Curric. Stud.*, 51, 1, 84–99. doi: 10.1080/00220272.2018.151356

Rosenlund, D. (2022). Två olika berättelser om historieämnets komplexitet?: En jämförelse mellan grundskolans kursplan och det nationella provet i historia, [two different narratives about the complexity of the subject of history?: a comparison between the primary school curriculum and the national exam in history], in *Cross-sections: Historical Perspectives from Malmö University* ed. J.H. Glaser, Julia; Lund, Martin; Lundin, Emma (Malmö: Malmö university).

Samuelsson, J. (2014). Ämnesintegrering och ämnesspecialisering: SO-undervisning i Sverige 1980-2014. [subject integration and subject specialization: social-studies teaching in Sweden 1980-2014]. *Nordidactica: J Human Soc Sci Educ* 2014, 85–118.

Seixas, P. (2015). *"looking for history," in joined up history: New directions in history education research, eds. A. Chapman & a. Wilschut*. (Charlotte: Information Age Publishing).

Seixas, P., Gibson, L., and Ercikan, K. (2015). "A design process for assessing historical thinking" in *New directions in assessing historical thinking*. eds. K. Ercikan and P. Seixas (New York: Routledge), 102.

Sendur, K., van Drie, J., and van Boxtel, C. (2022). Epistemic beliefs and written historical reasoning: exploring their relationship. *Historical Encounters* 9, 141–158. doi: 10.52289/hej9.108

Shemilt, D. (1983). The devil's locomotive. *History and theory* 22, 1–18. doi: 10.2307/2505213

Shemilt, D. (2009). "Drinking an ocean and pissing a cupful" in *National history standards: The problem of the canon and the future of teaching history*. eds. L. Symcox and A. Wilschut (Charlotte, NC: Information Age Pub)

Shemilt, D. (2018). "Assessment of learning in history education: past, present, and possible futures" in *The Wiley international handbook of history teaching and learning*. eds. S. A. Metzger and L. M. Harris (New York: Wiley), 449–471.

Skolinspektionen (2015). *Undervisningen i historia*.

Skolverket (2011). *Curriculum for the compulsory school system, the pre-school class and the leisure-time centre 2011*. Swedish National Agency for Education.

Skolverket. (2018). *Genomföra och bedöma nationella prov i grundskolan[conduct and assess national tests in primary school]*.

Skolverket (2021). *Betygen måste bli mer likvärdiga [The grades must become more equal]*. Stockholm: Skolverket.

Smith, M., Breakstone, J., and Wineburg, S. (2019). History assessments of thinking: a validity study. *Cogn. Instr.* 37, 118–144. doi: 10.1080/07370008.2018.1499646

Stoel, G., Logtenberg, A., Wansink, B., Huijgen, T., van Boxtel, C., and van Drie, J. (2017). Measuring epistemological beliefs in history education: an exploration of naïve and nuanced beliefs. *Int. J. Educ. Res.* 83, 120–134. doi: 10.1016/j.ijer.2017.03.003

Van Boxtel, C., and van Drie, J. (2017). "Engaging students in historical reasoning: the need for dialogic history education" in *Palgrave handbook of research in historical culture*

and education. eds. M. Carretero, S. Berger and M. Grever (London: Palgrave), 573–589. doi: 10.1057/978-1-137-52908-4_30

Van Nieuwenhuyse, K., Wils, K., Clarebout, G., Draye, G., and Verschaffel, L. (2015). "Making the constructed nature of history visible. Flemish secondary history education through the lens of written exams" in *Joined-up history: New directions in history education research*. eds. A. Chapman and A. Wilschut (Charlotte, NC: Information Age Publishing), 231–253.

Wennström, J. (2020). Marketized education: how regulatory failure undermined the Swedish school system. *J. Educ. Policy* 35, 665–691. doi: 10.1080/02680939.2019.1608472

Wineburg, S. (1991). Historical problem solving: a study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *J. Educ. Psychol.* 83, 73–87. doi: 10.1037/0022-0663.83.1.73

Zeleňák, E. (2015). Two versions of a constructivist view of historical work. *History and Theory* 54, 209–225. doi: 10.1111/hith.10754

frontiers | Frontiers in Education

TYPE Perspective
PUBLISHED 21 June 2024
DOI 10.3389/feduc.2024.1216481

Check for updates

OPEN ACCESS

EDITED BY
Gautam Biswas,
Vanderbilt University, United States

REVIEWED BY
Stephanie M. Gardner,
Purdue University, United States

*CORRESPONDENCE
Jesse R. Sparks
✉ jsparks@ets.org

RECEIVED 03 May 2023
ACCEPTED 23 April 2024
PUBLISHED 21 June 2024

CITATION
Sparks JR, Lehman B and Zapata-Rivera D (2024) Caring assessments: challenges and opportunities.
Front. Educ. 9:1216481.
doi: 10.3389/feduc.2024.1216481

COPYRIGHT
This work is authored by Jesse R. Sparks, Blair Lehman and Diego Zapata-Rivera. © 2024 Educational Testing Service. 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 authors and the copyright owner are credited and 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.

Caring assessments: challenges and opportunities

Jesse R. Sparks*, Blair Lehman and Diego Zapata-Rivera

ETS, Princeton, NJ, United States

Caring assessments is an assessment design framework that considers the learner as a whole and can be used to design assessment opportunities that learners find engaging and appropriate for demonstrating what they know and can do. This framework considers learners' cognitive, meta-cognitive, intra- and inter-personal skills, aspects of the learning context, and cultural and linguistic backgrounds as ways to adapt assessments. Extending previous work on intelligent tutoring systems that "care" from the field of artificial intelligence in education (AIEd), this framework can inform research and development of personalized and socioculturally responsive assessments that support students' needs. In this article, we (a) describe the caring assessment framework and its unique contributions to the field, (b) summarize current and emerging research on caring assessments related to students' emotions, individual differences, and cultural contexts, and (c) discuss challenges and opportunities for future research on caring assessments in the service of developing and implementing personalized and socioculturally responsive interactive digital assessments.

KEYWORDS

caring assessments, formative assessment, interactive digital assessment, caring systems, adaptive assessment and learning, socioculturally responsive assessment

1 Introduction

Personalization in the assessment context is an umbrella term that can include many different approaches. Most prior research and development has focused on adaptations based on students' prior knowledge or performance during the assessment (e.g., Shemshack et al., 2021). However, personalization may sometimes consider other intra-or interpersonal aspects of students' experience (Du Boulay, 2018). For example, student engagement has been utilized in effort-monitoring computer-based tests (Wise et al., 2006, 2019) and a wider range of student emotions have been used to enhance performance-based adaptation in several personalized learning systems (D'Mello et al., 2011; Forbes-Riley and Litman, 2011). Research in the field of artificial intelligence in education (AIEd) has increasingly emphasized a more holistic picture of learners which takes into account cognitive, metacognitive, and affective aspects of the learner to explain their behavior in learning environments (Grafsgaard et al., 2012; Kizilcec et al., 2017; Yadegaridehkordi et al., 2019), reflecting growing interest in integrating positive psychology into research within the AIEd community (Bittencourt et al., 2023).

The caring assessments (CA) framework provides an approach for designing adaptive assessments that learners find engaging and appropriate for demonstrating their knowledge, skills, and abilities (KSAs; Zapata-Rivera, 2017). This conceptual framework considers cognitive aspects of the learner as well as metacognitive, intra-and interpersonal skills, aspects of the learning context, cultural and linguistic backgrounds, and interaction behaviors within an integrated learner model and uses this model to personalize assessment to students' needs

(Zapata-Rivera et al., 2023). Multiple lines of research must be conducted to bring this vision for caring assessment to fruition. This Perspective article describes the CA framework and its unique contributions to the field (Section 2) and summarizes current and emerging research on the CA framework emphasizing students' emotions (Section 3), individual differences (Section 4), and cultural contexts (Section 5). Challenges and opportunities emerging from this literature are also discussed (Section 6), highlighting gaps and future directions for AIEd research that is most promising to advance the vision of CA.

2 The caring assessments framework

The CA framework (see Figure 1) is a *conceptual framework for adaptive assessment design* which proposes that assessments can provide a more engaging student experience while collecting more precise information about their KSAs by better understanding who students are and how they interact with the assessment (Zapata-Rivera, 2017). This better understanding of students can be leveraged to provide "caring" in terms of *adaptations* before, *caring support* during, and *feedback* after the assessment (Lehman et al., 2018).

Caring support before the assessment involves the development of *student profiles* that include a variety of information about the student, from their *personal characteristics* (e.g., interests, beliefs, linguistic background) to *contextual information* such as prior learning opportunities (Zapata-Rivera et al., 2020). These profiles can then be leveraged to provide students with an adapted version of the assessment that affords them the best opportunity to engage with the assessment and demonstrate what they know and can do. Alternative versions of the assessment could vary from the assessment format (e.g., multiple-choice items or game-based) to the language (e.g., toggle between English/Spanish) to the context of the assessment (e.g., using different texts while measuring the same underlying reading skills).

The student profiles that enable caring support before the assessment also serve as the start for providing caring support during the assessment. Caring support during the assessment will require an *integrated learner model* (ILM) that considers both student and contextual characteristics (from the student profile) and the *interaction behaviors* students demonstrate during the assessment. This ILM is a more complex learner model than is typically employed in personalized learning and assessment tasks but draws on prior research on various types of learner models (Zapata-Rivera and Arslan, 2021; Bellarhmouch et al., 2023). This ILM can leverage information from the student profile and interactions to provide on-demand support. For example, a student might become disengaged during the assessment and the ILM could deploy a motivational message that has been personalized based on the student's interests or prior opportunity to learn within the domain (Kay et al., 2022).

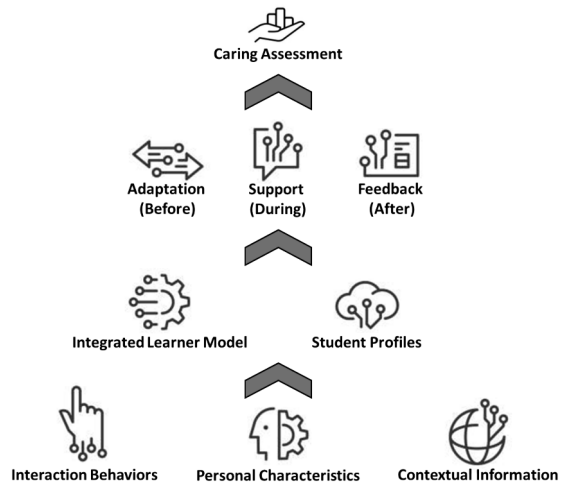


FIGURE 1
Caring assessment framework.

Caring support after the assessment is primarily provided in the feedback report. The goal is to provide feedback to the student that will be easy to understand and motivate them to continue their learning journey. This necessitates feedback reports that utilize asset-based language (Gay, 2013; Ramasubramanian et al., 2021) and provide context for performance on the assessment by leveraging the information in the ILM (e.g., identifying learners' relevant prior knowledge and lived experiences and the strengths they demonstrated on the assessment along with areas for improvement). This contextualized reporting could, for example, identify if student responses were connected to specific behavioral patterns or could connect current performance to students' prior experiences or opportunities to learn to highlight progress. This contextualized reporting can also be utilized when providing feedback to teachers, which can then support teacher decision-making on the next appropriate steps to support student learning and continue caring support outside of the assessment.

The CA framework builds on several areas of prior research. The notion of an adaptive "caring" assessment system (Zapata-Rivera, 2017) builds on AIED research on adaptive intelligent tutoring systems that "care" as they support learning (Self, 1999; Kay and McCalla, 2003; Du Boulay et al., 2010; Weitekamp and Koedinger, 2023). Attending to a broader set of student characteristics, contexts, and behaviors also allows the CA framework to leverage findings from multiple learning theories when developing "caring" supports. Emphasis on using intra-and interpersonal characteristics and other contextual information to drive assessment adaptation is consistent with and can leverage models of self-regulated learning (e.g., Winne and Hadwin, 1998; Pintrich, 2000; Kay et al., 2022). The inclusion of a broader set of characteristics, contexts, and behaviors also extends the idea of "conditional fairness" in assessments that use contextual information about students' backgrounds to adapt assessment designs and scoring rules (Mislevy, 2018) and extends typical research on computer adaptive assessments driven by performance and item difficulty (van der Linden and Glas, 2010; Shemshack et al., 2021).

While the CA framework has relevance to both large-scale summative and classroom formative assessment contexts, there is greater potential flexibility in applying this framework to the design of tools to be used in formative contexts, due to the emphasis on providing on-demand "caring" support to help learners maximize their learning and engagement during assessment tasks (Zapata-Rivera, 2017). Efforts toward realizing this framework have investigated how students' emotions, individual differences, and cultural contexts can best be leveraged to provide personalized assessment experiences. Next, we summarize this current and emerging research.

3 Student emotions

As anyone who has completed an assessment knows, it can be an emotional experience. However, very few assessments support students to remain in a productive emotional state (see Wise et al., 2006, 2019 for exceptions) or consider students' emotions when determining assessment outcomes (see Wise and DeMars, 2006 for an exception). Most research on student emotions during test taking has focused on documenting those experiences after test completion – and have shown that the experience of different emotions are differentially related to assessment outcomes (Spangler et al., 2002; Pekrun et al., 2004, 2011; Pekrun, 2006). Research on the impact of student emotions during learning activities has received far greater attention (see D'Mello, 2013 for a review) and there are multiple examples of personalized learning systems that leverage both student cognition and emotions to provide feedback and guide instructional decisions (e.g., D'Mello et al., 2011; Forbes-Riley and Litman, 2011).

In our own research on the emotional experiences of students during interactions with conversation-based assessments we build on prior work in both assessment and learning contexts by focusing on the intensity of discrete emotions (Lehman and Zapata-Rivera, 2018). When intensity was considered, we found the same pattern across boredom, frustration, and confusion: low intensity was positively correlated, medium intensity was not correlated, and high intensity was negatively correlated with performance, despite no overall relationship with performance. While it has been found that confusion has a more positive relationship with learning than boredom and frustration (e.g., Craig et al., 2004; D'Mello and Graesser, 2011, 2012, 2014; D'Mello et al., 2014) and frustration a more positive relationship than boredom (Baker et al., 2010), in assessment context it appears that the three emotions have a similar relationship with outcomes. However, the intensity findings for confusion, specifically, may relate to prior findings that the partial (Lee et al., 2011; Liu et al., 2013) or complete resolution of confusion (D'Mello and Graesser, 2014; Lehman and Graesser, 2015) is necessary for learning. Real-time tracking of students' emotional experiences (states and intensity) can be leveraged to provide caring support during the assessment as has been successfully implemented in personalized digital learning systems. However, integration of emotion detectors into the ILM will require going beyond prior research as both the experience of emotions and the ways in which those experiences are supported to promote learning will need to consider more factors (e.g., student interest, cultural background). In the assessment context, the use of student emotions can be expanded to provide caring support after the assessment by providing context for a student's performance to both the teacher and the student (e.g., student was confused while responding to items 2, 5, and 7), which can allow for more informed instructional decisions. In the CA framework, the ways in which student emotions are leveraged to support student learning will build upon prior learning research and will require new research efforts to ensure that emotions are productively integrated with other individual differences.

4 Individual differences

Students enter into test-taking experiences with a wide variety of interests, prior knowledge, experiences, attitudes, motivations, dispositions, or other intra-or interpersonal qualities that can affect their engagement with and performance on educational assessments and other academic outcomes (Braun et al., 2009; Lipnevich et al., 2013; Duckworth and Yaeger, 2015; West et al., 2016; Abrahams et al., 2019). For example, self-efficacy beliefs are strongly linked to academic achievement across domains (Guthrie and Wigfield, 2000; Richardson et al., 2012; Schneider and Preckel, 2017). Understanding how individual differences influence performance in interactive learning environments suggests directions for interventions or dynamic supports (Self, 1999) based on cognitive or motivational variables (Du Boulay et al., 2010) or prior knowledge (Khayi and Rus, 2019) that can be applied in assessments.

In previous work, we investigated student characteristics that predict performance on innovative conversation-based assessments of science inquiry and mathematical argumentation (Sparks et al., 2019, 2022). Students' science self-efficacy, growth mindset, cognitive flexibility, and test anxiety (with a negative coefficient) predicted performance on a science assessment (Sparks et al., 2019), while cognitive flexibility and perseverance (with a negative coefficient) predicted performance on mathematical argumentation (Sparks et al., 2022), controlling for student demographics and domain skills. Cluster analyses resulted in interpretable profiles with distinct relationships to student characteristics and performance, suggesting distinct paths for caring support within the CA framework (Sparks et al., 2020). For example, one profile represented students with average domain ability but relatively low cognitive flexibility, while another reflected motivated but test-anxious students. We hypothesize that these profiles would benefit from different supports (i.e., motivational messages vs. anxiety-reduction strategies; Arslan and Finn, 2023). However, the profiles and associated supports must be developed and validated in future research with students and teachers to ensure that the profiles reflect, and the adaptations address, the aspects most meaningful for instruction.

5 Cultural contexts

The prominence of social justice and anti-racist movements has resulted in increasing or renewed interest in (socio-)culturally responsive assessment (SCRA) practices (Hood, 1998; Lee, 1998; Qualls, 1998; Sireci, 2020; Bennett, 2022, 2023; Randall, 2021) which are themselves grounded in culturally relevant, responsive, and sustaining pedagogies (Paris, 2012; Gay, 2013; Ladson-Billings, 2014). Recent research reflects increasing attention to students' cultural characteristics when designing and evaluating AI-enabled instructional systems (Blanchard and Frasson, 2005; Mohammed and Watson, 2019; Talandron-Felipe, 2021); we can apply lessons from this work toward digital assessment design. As the K-12 student population becomes increasingly demographically, culturally, and linguistically diverse (National Center for Education Statistics, 2022), educational assessments must account for such variation, enabling test-takers to demonstrate their knowledge, skills, and abilities in ways that are most appropriate considering their cultural, linguistic, and social contexts (Mislevy, 2018; Sireci and Randall, 2021). Test items can include content reflective of situations, contexts, and practices students encounter in their lives (Randall, 2021), which can tap into students' home and community funds of knowledge (Moll et al., 1992; González et al., 2005) in ways that foster deeper student learning through meaningful connections to familiar, interesting contexts (Walkington and Bernacki,

2018). Math problems assessing knowledge of fractions within a recipe context could vary the context to align with students' cultural background (e.g., beans and cornbread vs. peanut butter sandwich). Positive effects have been shown for African American students interacting with pedagogical agents that employ dialects similar to their own in personalized learning systems (Finkelstein et al., 2013).

Emerging work is exploring cultural responsiveness in the context of scenario-based assessments (SBAs). SBAs are a useful context for exploring cultural factors in assessment performance and potential for implementing personalization within the CA framework (Sparks et al., 2023a,b). SBAs intentionally situate students in meaningful contexts for problem solving, providing a purpose and goal for responding to items (Sabatini et al., 2019). SBA developers have emphasized how scenarios can be made relevant to students from diverse racial, ethnic, and cultural backgrounds by intentionally incorporating contexts and content that celebrate students' cultural identities and integrate funds of knowledge from an asset-based perspective (O'Dwyer et al., 2023). Similar work has been conducted in designing robots for educational purposes in which students serve as co-creators to enable cultural relevance and responsiveness (Li et al., 2023). For example, SBA topics with greater cultural relevance to Black students (i.e., the Harlem Renaissance) show comparable reliability and validity but smaller group differences in performance versus more general topics (Ecosystems, Immigration), potentially due to Black students' greater engagement (Wang et al., 2023). Our current research (Sparks et al., 2023a,b) involves measuring students' self-identified cultural characteristics to examine relationships among their engagement and performance on SBAs, their racial, ethnic, and cultural identities, as well as their emotions, interests, motivations, prior knowledge, and experiences (i.e., home and community experiences, values, and practices related to assessment topics; Lave and Wenger, 1991; Gutiérrez and Rogoff, 2003; González et al., 2005). In future research, we aim to incorporate these characteristics into student profiles and evaluate how the profiles can be leveraged to provide a personalized assessment experience. This combination of cultural responsiveness and personalization has been explored in the learning context (Blanchard, 2010); however, additional research is needed to understand these dynamics to provide caring support within assessments.

6 Challenges and opportunities for caring assessment

Personalization within a CA framework introduces several challenges as well as opportunities when considering implementation of this framework within a digital learning system. The holistic view of students reflected in the ILM – going beyond measures of cognitive skill or performance to incorporate emotions, motivations, knowledge, interest, and other characteristics – requires access to data that is not typically collected during educational assessments (Zapata-Rivera, 2017). Contextual variables are often collected via survey methods (e.g., Braun et al., 2009; Abrahams et al., 2019) but could increasingly be collected by other means such as embedded assessment (Zapata-Rivera and Bauer, 2012; Zapata-Rivera, 2012; Rausch et al., 2019), and *stealth* assessment (Shute et al., 2009, 2015; Shute and Ventura, 2013) approaches which use logfile data from the assessment interaction and are less intrusive. For example, student interest could be measured by utilizing time-on-task and clickstream behaviors, versus a survey. Such approaches may collect multimodal interaction data (e.g., audio, or visual data) and leverage this information in an ILM. Collection of such multimodal data introduces the potential for privacy concerns regarding what is being collected, where data is stored, and who has access, especially to the extent that Personally Identifiable Information (PII) may be collected. Policy prohibitions may prevent collection and storage of certain data types (Council of the European Union, 2023). The importance of ethical and secure data handling and transparency with users about what and how data will be collected, retained, and used, is paramount, especially for K-12 students. Thus, implementation of the CA framework will require innovative measurement and modeling methodologies as well as close collaboration with students and teachers to build trust. Much like the ILM, it will be critical to integrate these independent lines of work in new research efforts that apply the CA framework in practice. Such integrated research is being actively explored in the INVITE institute[1] toward development of "caring" STEM learning environments for K-12 students.

A further challenge relates to the inherent tradeoffs in selecting the key student characteristics and behaviors that should be used to implement personalization. Variable selection requires care to ensure that measures are reliable and appropriate, so that personalization can be implemented along the dimensions that are most pertinent to students' needs. However, this challenge also inspires new research opportunities – particularly ones that focus on students that have been historically underrepresented in both research and educational technology to determine what characteristics and behaviors are most relevant for different student groups. Research that is more inclusive and aware of the diverse experiences that students bring to personalized digital assessment and learning experiences can support effective variable selection. Open learner modeling approaches (Bull and Kay, 2016; Bull, 2020; Zapata-Rivera, 2020) introduce an opportunity to further refine CAs while building user trust by giving teachers and students the chance to inspect and reflect on the ILM, highlighting where the model and its interpretations should be revised or qualified. Development of the infrastructure needed to collect variables, classify behaviors, deploy adaptations, and continually update a caring system requires computational modeling, machine learning, and artificial intelligence expertise to help develop, test, and iterate on the learner models. ILMs can be leveraged toward effective decision cycles within the caring system that, for example, provide necessary supports, route students to appropriate versions of subsequent tasks, and provide tailored, asset-based feedback.

A related issue concerns teachers' perceptions of personalization and whether they prioritize mastery of content or embrace a more holistic view and a need to personalize based on a broader set of emotional, motivational, or cultural aspects. The effectiveness of CAs will rest on their ability to effectively integrate with teacher practice by supporting students with different constellations of strengths and challenges, detecting for teachers the students who are most in need of their additional attention and support. Again, this challenge offers an opportunity for new research that incorporates teachers into the research and development process to bring CAs into practice that are reflective of

1 https://invite.illinois.edu

current best practices and work with teachers in achieving the shared goal of student learning in a caring and supportive environment.

Integrating cultural responsiveness into the CA framework introduces additional challenges. While personalization implies treating students as individuals, culturally situated perspectives emphasize how individual students are positioned as members of socially-and historically-defined racial, ethnic, and cultural groups (Gutiérrez and Rogoff, 2003). Such views acknowledge that groups are not monolithic and that identification with the racial, ethnic, and cultural contexts individual students experience also varies (Tatum, 2017). Adapting at the group level necessitates acknowledgment of this individual variation as well as the potential for individuals to identify in ways that may (not) be congruent with demographic group membership. Demographics may intersect in meaningful ways that impact students' lived experiences (Crenshaw, 1989). However, culture is embodied in participation in practices with shared meaning and significance (Lave and Wenger, 1991; Gutiérrez and Rogoff, 2003; Nasir et al., 2014). This implies that CA should enable student self-identification of demographic characteristics, cultural group memberships, and engagement in home and community practices (i.e., in terms of their funds of knowledge). Further research is needed to best understand how the complexity of student identities interact and impact their learning experiences.

Intersections among students' cultural backgrounds, knowledge, and experiences might be leveraged to increase the relevance and responsiveness of assessments (Walkington and Bernacki, 2018). Meaningful co-design activities in which the knowledge, interests, values, and experiences of students and teachers from historically marginalized groups can be centered, celebrated, and prioritized has the potential to result in more engaging, relevant, and valid assessments and would support more responsive personalized designs (O'Dwyer et al., 2023; Ober et al., 2023). Open learner models that can be interrogated and critiqued by students and teachers will be essential for a culturally responsive CA framework, so that student profiles and ILMs do not reflect biases or stereotypes, that misclassifications are appropriately corrected, and that contextual factors are considered when interpreting students' performance. Continued partnerships with teachers and students are needed to maximize the benefits for learning through connections to students' funds of knowledge while also minimizing unintended consequences.

7 Discussion

The CA framework can be leveraged toward personalized and culturally responsive assessments designed to support K-12 teaching and learning. This article outlines the current state of CA research on student emotions, individual differences, and cultural contexts, and highlights key challenges and opportunities for future research. Critical issues for future research include collection and handling of student data (characteristics, behavioral, multimodal) and associated privacy and security concerns, selection of characteristics for learner modeling, teacher perceptions of personalization, individual variation and self-identification of students' cultural identities and contexts, and engaging students and teachers in co-design of personalized ILMs and responsive adaptations. Research that integrates these independent areas is needed to bring the CA conceptual framework into practice in personalized digital assessments.

Whether the primary aim is individual personalization or responsiveness to students' cultural contexts, it is imperative that researchers engage in deep, sustained co-design partnerships with teachers and students to ensure validity and utility for those most in need of support (Penuel, 2019). It is also important to consider the assessment context (e.g., formative vs. summative, group-vs. individual-level reporting) and implications for measurement (e.g., comparability, scoring, interpretation) when determining how best to apply CA in practice. CA introduces opportunities to enhance students' assessment experiences and to advance use of assessment outcomes to further individuals' educational opportunities and wellbeing (Bittencourt et al., 2023). However, effective design and implementation of personalized assessments is a complex endeavor, which may necessitate new processes for designing assessments (O'Dwyer et al., 2023). We invite other scholars to conduct research addressing these challenges, advancing the field's ability to provide personalized, culturally responsive assessments.

Author contributions

JRS, BL, and DZ-R contributed to manuscript conceptualization, writing, reviewing, and editing prior to submission. All authors contributed to the article and approved the submitted version.

Funding

This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

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.

Author disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

Abrahams, L., Pancorbo, G., Santos, D., John, O. P., Primi, R., Kyllonen, P., et al. (2019). Social-emotional skill assessment in children and adolescents: advances and challenges in personality, clinical, and education contexts. *Psychol. Assess.* 31, 460–473. doi: 10.1037/pas0000591

Arslan, B., and Finn, B. (2023). The effects of nudges on cognitively disengaged student behavior in low-stakes assessments. *J. Intelligence* 11:204. doi: 10.3390/jintelligence11110204

Baker, R., D'Mello, S. K., Rodrigo, M. M. T., and Graesser, A. C. (2010). Better to be frustrated than bored: the incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *Int. J. Hum. Comput. Stud.* 68, 223–241. doi: 10.1016/j.ijhcs.2009.12.003

Bellarhmouch, Y., Jeghal, A., Tairi, H., and Benjelloun, N. (2023). A proposed architectural learner model for a personalized learning environment. *Educ. Inf. Technol.* 28, 4243–4263. doi: 10.1007/s10639-022-11392-y

Bennett, R. E. (2022). *The good side of COVID-19. Educational Measurement: Issues and Practice*. 41, 61–63. doi: 10.1111/emip.12496

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educ. Assess.* 28, 83–104. doi: 10.1080/10627197.2023.2202312

Bittencourt, I. I., Chalco, G., Santos, J., Fernandes, S., Silva, J., Batista, N., et al. (2023). Positive artificial intelligence in education (P-AIED): a roadmap. *Int. J. Artif. Intell. Educ.* doi: 10.1007/s40593-023-00357-y

Blanchard, E. G. (2010). "Adaptation-oriented culturally-aware tutoring systems: when adaptive instructional technologies meet intercultural education" in *Handbook of research on human performance and instructional technology*. eds. H. Song and T. Kidd. IGI Global. 413–430. doi: 10.4018/978-1-60566-782-9.ch025

Blanchard, E. G. M., and Frasson, C. (2005). Making intelligent tutoring systems culturally aware: the use of Hofstede's cultural dimensions. In: *Proceedings of the 2005 International Conference on Artificial Intelligence (ICAI 2005)*. eds. H. R. Arabnia and R. Joshua. Las Vegas, Nevada: CSREA Press. 2, 644–649.

Braun, H., Coley, R., Jia, Y., and Trapani, C. (2009). "Exploring what works in science instruction: a look at the eighth-grade science classroom" in *ETS policy information report* (Princeton, NJ: ETS). Available at: https://www.ets.org/Media/Research/pdf/PICSCIENCE.pdf

Bull, S. (2020). There are open learner models about! *IEEE Trans. Learn. Technol.* 13, 425–448. doi: 10.1109/TLT.2020.2978473

Bull, S., and Kay, J. (2016). SMILI☺: a framework for interfaces to learning data in open learner models learning analytics and related fields. *Int. J. Artif. Intell. Educ.* 26, 293–331. doi: 10.1007/s40593-015-0090-8

Council of the European Union. (2023). Artificial intelligence act: council and parliament strike a deal on the first rules for AI in the world. Press release. Available at: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

Craig, S., Graesser, A., Sullins, J., and Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning. *J. Educ. Media* 29, 241–250. doi: 10.1080/1358165042000283101

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Univ. Chic. Leg. Forum* 1989, Available at: http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

D'Mello, S. K. (2013). A selective Meta-analysis on the relative incidence of discrete affective states during learning with technology. *J. Educ. Psychol.* 105, 1082–1099. doi: 10.1037/a0032674

D'Mello, S., and Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognit. Emot.* 25, 1299–1308. doi: 10.1080/02699931.2011.613668

D'Mello, S., and Graesser, A. (2012). Dynamics of affective states during complex learning. *Learn. Instr.* 22, 145–157. doi: 10.1016/j.learninstruc.2011.10.001

D'Mello, S., and Graesser, A. (2014). Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychol.* 151, 106–116. doi: 10.1016/j.actpsy.2014.06.005

D'Mello, S. K., Lehman, B. A., and Graesser, A. C. (2011). "A motivationally supportive affect-sensitive AutoTutor" in *New perspectives on affect and learning technologies*. eds. R. A. Calvo and S. K. D'Mello (New York: Springer), 113–126.

D'Mello, S., Lehman, B., Pekrun, R., and Graesser, A. (2014). Confusion can be beneficial for learning. *Learn. Instruct.* 29, 153–170. doi: 10.1016/j.learninstruc.2012.05.003

Du Boulay, B. (2018). Intelligent tutoring systems that adapt to learner motivation. In *Tutor. Intell. Tutor. Syst.*, Ed. S. D. Craig. New York: Nova Science Publishers, Inc. 103–128.

Du Boulay, B., Avramides, K., Luckin, R., Martinuz-Miron, E., Rebolledo Mendez, G., and Carr, A. (2010). Towards systems that care: a conceptual framework based on motivation, metacognition and affect. *Int. J. Artif. Intell. Educ.* 20, 197–229. doi: 10.5555/1971783.1971784

Duckworth, A. L., and Yaeger, D. S. (2015). Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. *Educ. Res.* 44, 237–251. doi: 10.3102/0013189X15584327

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., and Cassell, J. (2013). "The effects of culturally congruent educational technologies on student achievement" in *Proceedings of 16th international conference on artificial intelligence in education (AIED2013)*. eds. K. Yacef, C. Lane, J. Mostow and P. Pavlik (Berlin Heidelberg: Springer-Verlag), 493–502.

Forbes-Riley, K., and Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Comm.* 53, 1115–1136. doi: 10.1016/j.specom.2011.02.006

Gay, G. (2013). Teaching to and through cultural diversity. *Curric. Inq.* 43, 48–70. doi: 10.1111/curi.12002

González, N., Moll, L. C., and Amanti, C. (Eds.) (2005). *Funds of knowledge: theorizing practices in households, communities, and classrooms*. Lawrence Erlbaum associates publishers.

Grafsgaard, J. F., Boyer, K. E., Wiebe, E. N., and Lester, J. C. (2012). Analyzing posture and affect in task-oriented tutoring. In *Proceedings of the Intelligent Tutoring Systems Track of the Twenty-Fifth International Conference of the Florida Artificial Intelligence Research Society (FLAIRS)*. Marco Island, FL. Washington, DC: AAAI. 438–443.

Guthrie, J. T., and Wigfield, A. (2000). "Engagement and motivation in reading" in *Handbook of reading research*. eds. M. L. Kamil, P. B. Mosenthal, P. D. Pearson and R. Barr (Mahwah, NJ: Erlbaum).

Gutiérrez, K. D., and Rogoff, B. (2003). Cultural ways of learning: individual traits or repertoires of practice. *Educ. Res.* 32, 19–25. doi: 10.3102/0013189X032005019

Hood, S. (1998). Culturally responsive performance-based assessment: conceptual and psychometric considerations. *J. Negro Educ.* 67, 187–196. doi: 10.2307/2668188

Kay, J., Bartimote, K., Kitto, K., Kummerfeld, B., Liu, D., and Reimann, P. (2022). Enhancing learning by open learner model (OLM) driven data design. *Comput. Educ. Artif. Intell.* 3:100069. doi: 10.1016/j.caeai.2022.100069

Kay, J., and McCalla, G. (2003). The careful double vision of self. *Int. J. Artif. Intell. Educ.* 13, 1–18,

Khayi, N. A., and Rus, V. (2019). Clustering students based on their prior knowledge. In M. Desmarais, C. F. Lynch, A. Merceron and R. Nkambou (Eds.), Proceedings of the 12th international conference on educational data mining.

Kizilcec, R. F., Pérez-Sanagustín, M., and Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Comput. Educ.* 104, 18–33. doi: 10.1016/j.compedu.2016.10.001

Ladson-Billings, G. (2014). Culturally relevant pedagogy 2.0: a.k.a. the remix. *Harv. Educ. Rev.* 84, 74–84. doi: 10.17763/haer.84.1.p2rj131485484751

Lave, J., and Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lee, C. D. (1998). Culturally responsive pedagogy and performance-based assessment. *J. Negro Educ.* 67, 268–279. doi: 10.2307/2668195

Lee, D., Rodrigo, M., Baker, R., Sugay, J., and Coronel, A. (2011). "Exploring the relationship between novice programmer confusion and achievement" in *ACII 2011*. eds. S. D'Mello, A. Graesser, B. Schuller and J. Martin (Berlin/Heidelberg: Springer), 175–184.

Lehman, B., and Graesser, A. (2015). "To resolve or not to resolve? That is the big question about confusion" in *Proceedings of the international conference on artificial intelligence in education*. eds. C. Conati, N. Heffernan, A. Mitrovic and M. F. Verdejo (Springer: Verlag), 216–225.

Lehman, B., Sparks, J. R., and Zapata-Rivera, D. (2018). "When should an adaptive assessment care?" in *Proceedings of ITS 2018: Intelligent tutoring systems 14th international conference, workshop on exploring opportunities for caring assessments*. eds. N. Guin and A. Kumar (Montreal, Canada: ITS), 87–94.

Lehman, B., and Zapata-Rivera, D. (2018). "Frequency, intensity, and mixed emotions, oh my! Investigating control-value theory in conversation-based assessments" *Paper presented at the annual meeting of the American Educational Research Association*. New York, NY.

Li, Y., Nwogu, J., Buddemeyer, A., Soloist, J., Lee, J., Walker, E., et al. (2023). "I want to be unique from other robots": positioning girls as co-creators of social robots in culturally-responsive computing education. In *Proceedings of the 2023 CHI conference on human factors in computing systems (CHI '23), Article No. 441*. eds. A. Schmidt, K. Vaananen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, et al. New York, NY: ACM, 1–14.

Lipnevich, A. A., MacCann, C., and Roberts, R. D. (2013). Assessing non-cognitive constructs in education: a review of traditional and innovative approaches. In *Oxford Handbook Child Psychol. Assess.* eds. D. H. Saklofske, C. R. Reynolds and V. Schwean. New York: Oxford University Press. 750–772. doi: 10.1093/oxfordhb/9780199796304.013.0033

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., and Baker, R. (2013). "Sequences of frustration and confusion, and learning" in *EDM 2013*. eds. S. D'Mello, R. Calvo and A. Olney. *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*. International Educational Data Mining Society. 114–120.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge. doi: 10.4324/9781315871691

Mohammed, P. S., and Watson, E. N. (2019). "Towards inclusive education in the age of artificial intelligence: perspectives, challenges, and opportunities" in *Artificial intelligence and inclusive education. Perspectives on rethinking and reforming education.* Book series on perspectives on rethinking and reforming education. eds. J. Knox, Y. Wang and M. Gallagher (Singapore: Springer) 17–37. doi: 10.1007/978-981-13-8161-4_2

Moll, L. C., Amanti, C., Neff, D., and Gonzalez, N. (1992). Funds of knowledge for teaching: using a qualitative approach to connect homes and classrooms. *Theory Pract.* 31, 132–141.

Nasir, N. S., Rosebery, A. S., Warren, B., and Lee, C. D. (2014). "Learning as a cultural process: achieving equity through diversity" in *The Cambridge handbook of the learning sciences*. 2 Edition ed. R. K. Sawyer (Cambridge, UK: Cambridge University Press), 686–706. doi: 10.1017/CBO9781139519526.041

National Center for Education Statistics (2022). *Racial/ethnic enrollment in public schools. Condition of education*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

O'Dwyer, E., Sparks, J. R., and Nabors Oláh, L. (2023). Enacting a process for developing culturally relevant classroom assessments. *Appl. Meas. Educ.* 36, 286–303. doi: 10.1080/08957347.2023.2214652

Ober, T., Lehman, B., Gooch, R., Oluwalana, O., Solyst, J., Garcia, A., et al. (2023). Development of a framework for culturally responsive personalized learning. Paper presented at the annual meeting of the International Society for Technology in Education (ISTE). Philadelphia, PA.

Paris, D. (2012). Culturally sustaining pedagogy: a needed change in stance, terminology, and practice. *Educ. Res.* 41, 93–97. doi: 10.3102/0013189X12441244

Pekrun, R. (2006). The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* 18, 315–341. doi: 10.1007/s10648-006-9029-9

Pekrun, R., Goetz, T., Frenzel, A., Barchfield, P., and Perry, R. (2011). Measuring emotions in students' learning and performance: the achievement emotions questionnaire (AEQ). *Contemp. Educ. Psychol.* 36, 36–48. doi: 10.1016/j.cedpsych.2010.10.002

Pekrun, R., Goetz, T., Perry, R., Kramer, K., Hochstadt, M., and Molfenter, S. (2004). Beyond test anxiety: development and validation of the test emotions questionnaire (TEQ). *Anxiety Stress Coping* 17, 287–316. doi: 10.1080/10615800412331303847

Penuel, W. R. (2019). "Co-design as infrastructuring with attention to power: building collective capacity for equitable teaching and learning through design-based implementation research" in *Collaborative curriculum Design for Sustainable Innovation and Teacher Learning*. eds. J. Pieters, J. Voogt and N. P. Roblin (Cham, Switzerland: SpringerOpen), 387–401.

Pintrich, P. R. (2000). "The role of goal orientation in self-regulated learning" in *Handbook of Self-regulation*. eds. M. Boekaerts, P. R. Pintrich and M. Zeidner (San Diego, CA: Academic Press), 452–502.

Qualls, A. L. (1998). Culturally responsive assessment: development strategies and validity issues. *J. Negro Educ.* 67, 296–301. doi: 10.2307/2668197

Ramasubramanian, S., Riewestahl, E., and Landmark, S. (2021). The trauma-informed equity-minded asset-based model (TEAM): the six R's for social justice-oriented educators. *J. Media Lit. Educ.* 13, 29–42. doi: 10.23860/JMLE-2021-13-2-3

Randall, J. (2021). Color-neutral is not a thing: redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educ. Meas. Issues Pract.* 40, 82–90. doi: 10.1111/emip.12429

Rausch, A., Kögler, K., and Siegfried, J. (2019). Validation of embedded experience sampling (EES) for measuring non-cognitive facets of problem-solving competence in scenario-based assessments. *Front. Psychol.* 10, 1–16. doi: 10.3389/fpsyg.2019.01200

Richardson, M., Abraham, C., and Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychol. Bull.* 138, 353–387. doi: 10.1037/a0026838

Sabatini, J., O'Reilly, T., Weeks, J., and Wang, Z. (2019). Engineering a 21st century reading comprehension assessment system utilizing scenario-based assessment techniques. *Int. J. Test.* 20, 1–23. doi: 10.1080/15305058.2018.1551224

Schneider, M., and Preckel, F. (2017). Variables associated with achievement in higher education: a systematic review of meta-analyses. *Psychol. Bull.* 143, 565–600. doi: 10.1037/bul0000098

Self, J. A. (1999). The distinctive characteristics of intelligent tutoring systems research: ITSs care, precisely. *Int. J. Artif. Intell. Educ.* 10, 350–364,

Shemshack, A., Kinshuk, , and Spector, J. M. (2021). A comprehensive analysis of personalized learning components. *J. Comput. Educ.* 8, 485–503. doi: 10.1007/s40692-021-00188-7

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., et al. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Comput. Educ.* 86, 224–235. doi: 10.1016/j.compedu.2015.08.001

Shute, V. J., and Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. Cambridge, MA: The MIT Press.

Shute, V. J., Ventura, M., Bauer, M. I., and Zapata-Rivera, D. (2009). "Melding the power of serious games and embedded assessment to monitor and foster learning: flow and grow" in *Serious games: mechanisms and effects*. eds. U. Ritterfeld, M. Cody and P. Vorderer (Mahwah, NJ: Routledge, Taylor and Francis), 295–321.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educ. Meas. Issues Pract.* 39, 100–105. doi: 10.1111/emip.12377

Sireci, S. G., and Randall, J. (2021). "Evolving notions of fairness in testing in the United States" in *The history of educational measurement: Key advancements in theory, policy, & practice*. eds. B. E. Clauser and M. B. Bunch (New York: Routledge), 111–135.

Spangler, G., Pekrun, R., Kramer, K., and Hofmann, H. (2002). Students' emotions, physiological reactions, and coping in academic exams. *Anxiety Stress Coping* 15, 413–432. doi: 10.1080/1061580021000056555

Sparks, J. R., Lehman, B., Ober, T., Zapata-Rivera, D., Steinberg, J., and McCulla, L. (2023b). *Linking students' funds of knowledge to task engagement and performance: toward culturally responsive "caring assessments"*. Paper presented in coordinated symposium session at the 7th Culturally Responsive Evaluation and Assessment Conference, Chicago, IL.

Sparks, J.R., Ober, T., Steinberg, J., Lehman, B., McCulla, L., and Zapata-Rivera, D. (2023a). *PK and IDK are OK, IMO: correlating students' knowledge, metacognition, and perceptions of scenario-based assessments*. Poster presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Sparks, J. R., Peters, S., Steinberg, J., James, K., Lehman, B. A., and Zapata-Rivera, D. (2019). *Individual difference measures that predict performance on conversation-based assessments of science inquiry skills*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Sparks, J. R., Steinberg, J., Castellano, K., Lehman, B., and Zapata-Rivera, D. (2020). *Generating individual difference profiles via cluster analysis: toward caring assessments for science*. Poster accepted at the annual meeting of the National Council for Measurement in Education, San Francisco, CA. (Conference canceled).

Sparks, J. R., Steinberg, J., Lehman, B., and Zapata-Rivera, D. (2022). *Leveraging students' background characteristics to predict performance on conversation-based assessments of mathematics*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Talandron-Felipe, M. M. P. (2021). "Considerations towards culturally-adaptive instructional systems" in *Adaptive instructional systems. Design and evaluation. HCII 2021*. eds. R. A. Sottilare and J. Schwarz, Lecture Notes in Computer Science (Cham: Springer), 12792.

Tatum, B. D. (2017). *Why are all the black kids sitting together in the cafeteria? And other conversations about race. Revised* Edn. New York: Basic Books.

van der Linden, W. J., and Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. New York: Springer.

Walkington, C., and Bernacki, M. L. (2018). Personalization of instruction: design dimensions and implications for cognition. *J. Exp. Educ.* 86, 50–68. doi: 10.1080/00220973.2017.1380590

Wang, Z., Bruce, K. M., O'Reilly, T. P., Sparks, J. R., and Walker, M. (2023). *Group differences across scenario-based reading assessments: examining the effects of culturally relevant test content*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Weitekamp, D., and Koedinger, K. (2023). Computational models of learning: deepening care and carefulness in AI in education. In: N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda and O.C Santos. (Eds.), Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky. AIED 2023. Communications in Computer and Information Science, vol 1831. Cham, Switzerland: Springer. doi: 10.1007/978-3-031-36336-8_2

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., et al. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educ. Eval. Policy Analysis* 38, 148–170,

Winne, P. H., and Hadwin, A. F. (1998). "Studying as self-regulated engagement in learning" in *Metacognition in educational theory and practice*. eds. D. Hacker, J. Dunlosky and A. Graesser (Hillsdale, NJ: Erlbaum), 277–304.

Wise, S. L., Bhola, D., and Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: the effort-monitoring CBT. *Educ. Meas. Issues Pract.* 25, 21–30. doi: 10.1111/j.1745-3992.2006.00054.x

Wise, S. L., and DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *J. Educ. Meas.* 43, 19–38. doi: 10.1111/j.1745-3984.2006.00002.x

Wise, S. L., Kuhfeld, M. R., and Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Appl. Meas. Educ.* 32, 183–192. doi: 10.1080/08957347.2019.1577248

Yadegaridehkordi, E., Noor, N. F. B. M., Ayub, M. N. B., Affal, H. B., and Hussin, N. B. (2019). Affective computing in education: a systematic review and future research. *Comput. Educ.* 142:103649. doi: 10.1016/j.compedu.2019.103649

Zapata-Rivera, D., and Bauer, M. (2012). "Exploring the role of games in educational assessment" in *Technology-based assessments for twenty-first-century skills: theoretical and practical implications from modern research*. eds. J. Clarke-Midura, M. Mayrath, D. Robinson and G. Schraw. Charlotte, NC: Information Age. 147–169.

Zapata-Rivera, D., Forsyth, C., Sparks, J. R., and Lehman, B. (2023). "Conversation-based assessment: current findings and future work" in *International encyclopedia of education*. eds. R. Tierney, F. Rizvi and K. Ercikan. *4th* ed (Amsterdam, Netherlands: Elsevier Science) 504–518.

Zapata-Rivera, D. (2012) Embedded assessment of informal and afterschool science learning. Summit on Assessment of Informal and After-School Science Learning Available at: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072564.pdf

Zapata-Rivera, D. (2017). Toward caring assessment systems. In *Adjunct publication of the 25th conference on user modeling, adaptation and personalization (UMAP '17)*. eds. M. Tkalcic, D. Thakker, P. Germanakos, K. Yacef, C. Paris and O. Santos. New York: ACM. 97–100.

Zapata-Rivera, D. (2020). Open student modeling research and its connections to educational assessment. *Int. J. Artif. Intell. Educ.* 31, 380–396. doi: 10.1007/s40593-020-00206-2

Zapata-Rivera, D., and Arslan, B. (2021). "Enhancing personalization by integrating top-down and bottom-up approaches to learner modeling" in *Proceedings of HCII 2021*. eds. R. A. Sottilare and J. Schwarz (Switzerland: Springer Nature), 234–246.

Zapata-Rivera, D., Lehman, B., and Sparks, J. R. (2020). "Learner modeling in the context of caring assessments" in *Proceedings of the second international conference on adaptive instructional systems, held as part of HCI international conference 2020, LNCS 12214*. eds. R. A. Sottilare and J. Schwarz (Cham, Switzerland: Springer Nature), 422–431.

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 30 October 2024
DOI 10.3389/feduc.2024.1248770

Check for updates

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
University of Auckland, New Zealand

REVIEWED BY
Yi-Fang Wu,
Cambium Assessment, Inc., United States
Preeta Hinduja,
Iqra University, Pakistan

*CORRESPONDENCE
Jari Metsämuuronen
✉ jari.metsamuuronen@gmail.com

RECEIVED 27 June 2023
ACCEPTED 09 May 2024
PUBLISHED 30 October 2024

CITATION
Metsämuuronen J (2024) Note on the radical inflation in the estimates of error variance in measurement models.
Front. Educ. 9:1248770.
doi: 10.3389/feduc.2024.1248770

COPYRIGHT
© 2024 Metsämuuronen. 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.

Note on the radical inflation in the estimates of error variance in measurement models

Jari Metsämuuronen[1,2]*

[1]Finnish Education Evaluation Center (FINEEC), Helsinki, Finland, [2]Turku Research Institute for Learning Analytics (TRILA), University of Turku, Turku, Finland

This note discusses the radical technical inflation in error variance and the related standard error of test scores from both conceptual and empirical viewpoints. This technical inflation arises as a direct consequence of the technical underestimation of item-score correlation by the product–moment coefficient of correlation (PMC), which is embedded in the traditional estimators of reliability such as coefficients alpha, theta, omega, or rho (maximal reliability). Specifically, in educational settings where compilations usually include both easy and difficult items, the estimate by PMC may substantially deviate from the true association between an item and the score. Consequently, the use of traditional estimators of reliability leads to technically inflated estimates of standard errors, as the error variance related to these traditional measurement models is significantly inflated, resulting in deflated reliability estimates. In educational testing, employing deflation-corrected standard errors, calculated using deflation-corrected reliability estimators, would provide a more accurate measure of the test score's true precision.

KEYWORDS

item-score correlation, error variance, standard errors, deflation-corrected error variance, deflation-corrected standard errors

1 Introduction

This note focuses on a consequential outcome concerning significant deflation observed in the primary reliability estimators used within classical test theory, namely, coefficients alpha, theta, omega, and rho (maximal reliability), as previously discussed by researchers such as Zumbo and colleagues (e.g., Zumbo et al., 2007; Gadermann et al., 2012) and more recently by Metsämuuronen (2022a,b,c,d,e,f, 2023). The reader is led to the concepts and literature from four perspectives. Section 1.1 discusses the general phenomena of deflation in reliability and inflation in error variance. Section 1.2 explores the phenomenon where correlation estimates serve as the primary cause of deflation in reliability estimates and inflation in error variance. Section 1.3 briefly examines conceptual aspects related to error variance inflation. Section 1.4 provides a hypothetical example illustrating the magnitude of error variance, inflation, and standard error. The empirical section investigates the extent of error variance, inflation, and standard errors, aiming to elucidate the circumstances under which notable effects are expected.

1.1 Deflation in reliability and inflation in error variance as phenomena

In certain kinds of tests, which typically include items of extreme difficulty levels, as is common in educational testing settings (see discussion in Metsämuuronen, 2023), the technical deflation in the estimates of reliability has been reported to range from 0.40 to 0.70 units of the reliability coefficient. In these types of tests, the standard errors related to the score are significantly inflated. For extremely easy or difficult tests, standard errors can be more than ten times higher when using traditional reliability estimators compared to deflation-corrected reliability estimators (DCER) (Metsämuuronen and Ukkola, 2019; Metsämuuronen, 2022b; for DCER details, see Metsämuuronen, 2022c,d,e). When tests include easy, medium, and difficult items, the standard errors can be two to three times higher with traditional estimators (Metsämuuronen, 2022f). This indicates that the estimated error variance related to the measurement model is radically inflated.

The deflation of 0.40–0.70 units of reliability discussed above related to the artificial technical or mechanical errors in the estimation of correlation needs to be separated from attenuation related to the violations against the measurement model. The attenuation related to estimators of reliability and, consequently, in the estimated standard errors has been discussed widely, especially the challenges related to coefficient alpha (Kuder and Richardson, 1937; Jackson and Ferguson, 1941; Guttman, 1945; Cronbach, 1951), which are well known (see discussions and literature in, e.g., Sijtsma, 2009; Cho and Kim, 2015; Hoekstra et al., 2019; Metsämuuronen, 2022b,d).

Guttman (1945) was the first to show that the coefficient we know today as coefficient alpha always gives estimates that are lower in magnitude than true population reliability. The magnitude of the attenuation related to the violations against the assumption related to coefficient alpha has been reported to vary from 1% (Raykov, 1997) to 11% (Green and Yang, 2009). However, it is commonly accepted that if the assumptions for the coefficient alpha are met, the items are (essentially) tau-equivalent, the phenomenon is unidimensional, and the measurement errors related to test items do not correlate. Alpha would give unattenuated estimates (see Novick and Lewis, 1967; Raykov and Marcoulides, 2017; see the discussion also in, e.g., Green and Yang, 2009, 2015; Davenport et al., 2015, 2016; Trizano-Hermosilla and Alvarado, 2016; McNeish, 2017). However, there is an ongoing debate among scholars about whether we could continue to use coefficient alpha as one of the lower boundaries of reliability or not at all (see a positive view in, e.g., Bentler, 2009; Falk and Savalei, 2011; Raykov et al., 2015; Metsämuuronen, 2017; Raykov and Marcoulides, 2017; and a negative view in, e.g., Sijtsma, 2009; Yang and Green, 2011; Dunn et al., 2013; Trizano-Hermosilla and Alvarado, 2016; McNeish, 2017).

Notably, the underestimation in the estimates of reliability is only partly related to attenuation, as discussed above. Metsämuuronen (2016) shows algebraically that the radical deflation in the estimates of reliability is directly related to technical or mechanical errors in the estimates of correlation by item-score correlation (*Rit*). This issue affects not only coefficient alpha but also other reliability coefficients such as theta, omega, and rho, which also incorporate item-score correlation in some form (see Metsämuuronen, 2022b,c,d). Metsämuuronen (2022b,d) identifies several other estimators of reliability with the same challenge. The role of *Rit* in the deflation is discussed later.

Deflation in reliability has a direct effect on the traditional standard error of measurement (*S.E.m*) related to the test score (see Metsämuuronen, 2023). The standard error is a concept used in quantifying the average amount of random measurement error in a score variable generated by a compilation of multiple test items; the technicalities are discussed in Section 1.3. Notably, in large-scale testing settings such as Program of International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), the focus is mainly on the standard errors across different parts of the ability scale, referred to as conditional standard errors, rather than the average *S.E.m.* (see, e.g., Schult and Sparfeldt, 2016; Foy and LaRoche, 2019). In this note, however, the traditional *S.E.m.* is discussed because it has a direct relationship with the traditional estimate of reliability (*REL*), that is, $S.E.m. = \sigma_E = \sigma_X\sqrt{1-REL}$ (e.g., Gulliksen, 1950), based on the classical test theory definition of reliability:

$$REL = \sigma_T^2 / \sigma_X^2 = 1 - \sigma_E^2 / \sigma_X^2 \quad (1)$$

where σ_T^2, σ_X^2, and σ_E^2 refer to the variances of the observed score variable (X), unobserved true score (T), and error element (E) related to the profound idea in measurement modeling: $X = T + E$.

Because of the simplicity of the definition of reliability in Equation (1), the technical reason for the observed radical deflation in the estimates of reliability can be traced to two sources: either the population variance (σ_X^2) is deflated, or the error variance (σ_E^2) is inflated—or both may happen at the same time. Metsämuuronen (2022h) specifically studies the magnitude and limits of the deflation in the population variance. The deflation in the population variance is an obvious reason for the deflation in the estimates by coefficients alpha and theta and related coefficients from the extended family (see the discussion in, e.g., Metsämuuronen, 2016, 2022b,d). The reason is obvious because the element σ_X^2 embedded in the reliability formulae embeds the item–score correlation ($Rit = \rho_{iX}; \sigma_X^2 = \left(\sum_{i=1}^{k}\sigma_i \times \rho_{iX}\right)^2$; see also Equation 8), and *Rit* is known to give radical underestimates when the scales of two variables differ. This is always the case with an item and a score. This is deepened and illustrated in Section 1.2.

In the more advanced estimators of reliability, such as coefficients omega and rho, the reason for the deflation is partly the overestimated error variance observed in the form of $\left(1-\lambda_i^2\right)$ in the formulae (see later Equation 4), where λ_i refers to factor loading. Notably, factor loadings are essentially correlations between an item and a factor score (e.g., Cramer and Howitt, 2004; Yang, 2010). The consequences and magnitude of the deflation in reliability are discussed in the empirical part of this article.

1.2 Deflation in estimates of correlation due to inflation in error variance

Due to being the oldest estimator of association still in use—with over a century of research on and with the product–moment coefficient of correlation (PMC)—most of its weaknesses are well-known. General challenges are extensively covered in standard

textbooks (e.g., Salkind, 2010; Tabachnick and Fidell, 2021). Two specific challenges strictly related to the topic of the article are discussed here.

First, scholars have extensively discussed a particular challenge of the product–moment coefficient of correlation (PMC) under the topic of "restriction of range" or "range restriction" (RR) for over a century, starting from the works of Pearson (1903) and Spearman (1904) onward. More recent discussions are summarized by Sackett and Yang (2000), Sackett et al. (2007), Meade (2010), Walk and Rupp (2010) and Metsämuuronen (2022d). This phenomenon refers to situations where only a portion of the range of values of a variable is realized in the sample, leading to inaccurate correlation estimates by PMC. These estimates are attenuated, meaning they are lower than the true correlation (see various patterns of RR in Sackett and Yang, 2000). Pearson (1903) and Spearman (1904) proposed initial solutions to correct this attenuation, and numerous solutions have been suggested since then (see typologies in Mendoza and Mumford, 1987; Sackett et al., 2007). This characteristic of PMC has been investigated and addressed, particularly within meta-analytic studies (see, e.g., Schmidt and Hunter, 2003, 2015; Schmidt et al., 2008).

The other challenge the PMC poses, closely related to the inflation in error variance, is its inaccurate estimation in item analysis settings. This is considered the primary reason for the deflation of reliability because PMC is embedded in the most widely used reliability estimators (see compiled in Metsämuuronen, 2022d). Through simulations, Metsämuuronen (2021a, 2022a); also partly observed in simulations by Martin (1973, 1978) and Olsson (1980) has identified seven cumulative and partly interrelated conditions where deflation in estimates by PMC is anticipated.

Based on these simulations, the item–score correlation (*Rit*) tends to consistently and systematically underestimate the true association between an item and a score variable under the following conditions:

1 The deflation approximates 100% the greater the extremity of the item difficulty is.
2 Scale discrepancy: The greater the discrepancy between the item's scale and the score.
3 Fewer item categories: The fewer categories present in the item.
4 Fewer score categories: The fewer categories present in the score.
5 Number of items: The smaller the number of items comprising the score. This is closely linked to the number of categories in the score's scale.
6 Non-uniform tied cases: The greater the presence of non-uniformly distributed tied cases in the score. This a consequence of having a small number of items.
7 Distribution: If the distribution of the latent variable (and score) deviates from a uniform distribution.

Consequently, if the test contains items with extreme difficulty levels, a small number of items, and items with a narrow scale, resulting in a score with a narrow scale, we anticipate obtaining significantly deflated item-total correlations. This leads to markedly inflated measurement errors, substantially deflated reliability estimates, and inflated standard errors. The extent of this inflation is illustrated in Section 1.4 with a numerical example.

The phenomenon of technical or mechanical deflation in the estimates of correlation can be easily illustrated with two identical (latent) variables that have an obvious perfect correlation $\rho_{\theta\theta} = 1$. If one of these identical variables is dichotomized (item) and the other polytomized into several categories (score), *Rit* cannot reach the perfect (latent) correlation. This is unlike other measures, such as polychoric correlation (R_{PC}; Pearson, 1900, 1913), Goodman–Kruskal gamma (G; Goodman and Kruskal, 1954), dimension-corrected G (G_2; Metsämuuronen, 2021a), and attenuation-corrected *Rit* and *eta* (R_{AC} and E_{AC}; Metsämuuronen, 2022e,g) (see simulations in Metsämuuronen, 2021a, 2022a). Some estimators, such as r-bireg and r-polyreg correlation (R_{REG}; Livingston and Dorans, 2004; Moses, 2017), Somers delta directed so that "score dependent" (D; Somers, 1962), and dimension-corrected D (D_2; Metsämuuronen, 2020b, 2021a) come close to a deflation-free outcome.

As an example of the radical technical deflation in PMC, let us take the vector of $n = 1{,}000$ cases from a normally distributed population and double it. Of these identical variables, one (item g) is divided into a binary form [$df(g) = 1$] by using a cut-off of $p = 0.90$; that is, 90% of the hypothetical test-takers give the correct answer, and the other (score X) is divided into seven categories [$df(X) = 6$] with an average difficulty level of [$p(X) = 0.50$]; this could be a latent reflection of a short subtest (e.g., "geometry") amid a longer test ("mathematical achievement"). The difference between the latent correlation ($\rho_{\theta\theta} = 1$) and the observed correlation ($\rho_{iX} = \rho_{i.} = Rit$) indicates the magnitude of technical deflation in the estimates, even without attenuation, which may add some additional deflation to the outcome. Figure 1 illustrates the magnitudes of the technical deflation in selected estimators of association.

Notably, the estimates by such known estimators of item-score association based on the mechanics of PMC as Henrysson's item–rest correlation *Rir* (Henrysson, 1963), Spearman's rank-order correlation R_{Rank} (Spearman, 1904), *Rit*, and *eta* cannot detect the obvious perfect latent correlation, and the magnitude of deflation is notable (> 0.47 units of correlation).[1] Moreover, Kendall's *tau-b* (Kendall, 1948) gives a deflated estimate because the values are always lower than those by PMC (see the reasons in, e.g., Metsämuuronen, 2021b). Such estimators as R_{PC}, G, G_2, $R_{A,\,C,}$ and E_{AC} are found deflation-*free* in this kind of comparison. However, they may have some other challenges in fully reaching the true association (see Metsämuuronen, 2022a). R_{REG} is almost deflation-free, and, in D and D_2, the magnitude of deflation may be nominal, depending on the number of tied pairs in the items and score as well as the widths of the scales in item and score (see Metsämuuronen, 2021a). Hence, based on an analysis of 11 sources of deflation, Metsämuuronen (2022a) lifts coefficients R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, and E_{AC} as superior options for *Rit* to be used in estimators of reliability to reach deflation-corrected estimates of reliability. Some of these estimators are used as benchmarks in the numerical example and empirical section to assess the magnitude of the inflation in error variance and standard errors.

1 Notably, although coefficient *eta* uses different information in comparison with *Rit*, in the binary case, their formulae are identical (see Wherry and Taylor, 1946; see also Metsämuuronen, 2022g).

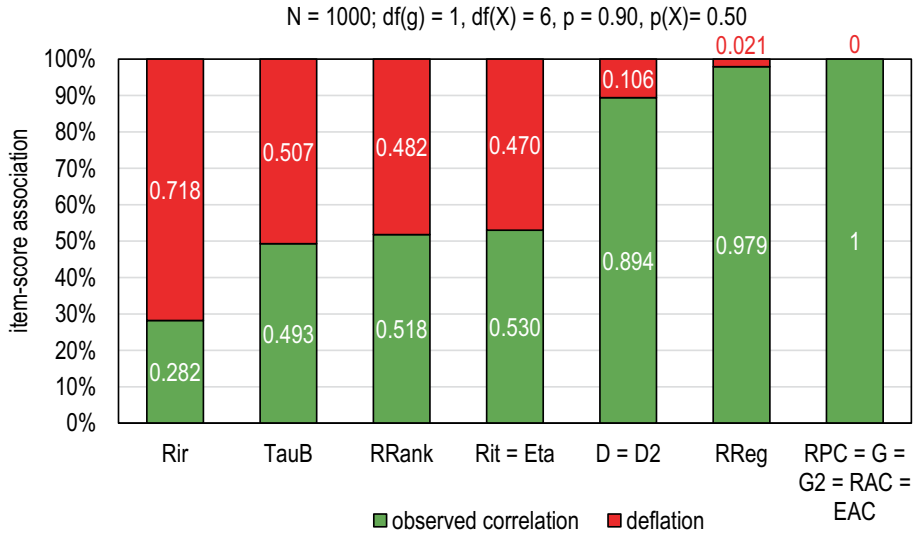


FIGURE 1
Magnitude of deflation in the selected estimators of association.
Rir, Henrysson item–rest correlation (= PMC); Tau-b, Kendall tau-b; RRank, Spearman rank-order correlation (= PMC); Rit, Item-total correlation (= PMC); eta, Coefficient eta (X dependent) (= PMC in the binary case); D, Somers delta (X dependent); D2, Dimension-corrected D; RReg, r-bireg correlation; RPC, Polychoric correlation; G, Goodman-Kruskal gamma; G2, Dimension-corrected G; RAC, Attenuation-corrected Rit; EAC, Attenuation-corrected eta.

1.3 Briefly on the basic concepts related to inflation in error variance

Section 1.4 gives a practical, hypothetical example of the phenomenon of inflation. Some concepts are needed to understand the notation in that section. However, the conceptual discussion is minimal in this section (see in detail in Supplementary Appendix 1).

Let us assume a congeneric measurement model with one latent variable (θ):

$$x_i = w_i\theta + e_i \tag{2}$$

where x_i denotes the observed values of an item g_i and w_i denotes a weight factor that links θ with x_i (e.g., Metsämuuronen, 2022a,c,d). This congeneric measurement model is generalized from the traditional model (e.g., McDonald, 1999; Cheng et al., 2012). In the traditional model, the weight factor w_i is usually assumed to be a factor loading (λi), and the factor score variable is assumed to reflect the most accurately the latent variable. In the general model, the weight factor w_i is a coefficient of association in *some* form, also including principal component and factor loadings. The unobservable θ may manifest as a varying type of relevantly formed compilation of items such as a raw score (θ_X), standardized raw score (θ_{XSDT}), principal component score (θ_{PC}), factor score (θ_{FA}), theta[2] score formed by the item response theory (IRT) or Rasch modeling (θ_{IRT}), or various non-linear combinations of the items ($\theta_{Non\text{-}Linear}$).

If we assume that errors in the individual items do not correlate with each other, the error variance related to the compilation of the items is as follows:

$$VAR\left(\sum_{i=1}^{k} e_i\right) = \sigma_E^2 = \sum_{i=1}^{k}\left(1 - w_i^2\right) = k - \sum_{i=1}^{k} w_i^2. \tag{3}$$

In practical terms, the traditional measurement model takes the factor loading as the weight factor, and this leads to the following error variance related to the score variable:

$$\sigma_E^2 = \sum_{i=1}^{k}\left(1 - \lambda_i^2\right) = k - \sum_{i=1}^{k} \lambda_i^2. \tag{4}$$

Notably, the traditional model assumes that the weight factor w_i, i.e., factor loading being a correlation coefficient, always gives accurate estimates. This assumption is too optimistic, as observed above, and the deflation in the estimate may be remarkable. However, if we select the correlation w wisely so that the magnitude of the mechanical error is as small as possible, that is, if we use some of the deflation-free or deflation-corrected estimators of correlation (w_{i_DC}), the outcome is deflation-free or near. The magnitude of the error component related to deflation may be near zero. This leads us to a deflation-corrected measurement model and, consequently, to deflation-corrected error variance as follows:

$$\sigma_{E_DC}^2 = \sum_{i=1}^{k}\left(1 - w_{i_DC}^2\right) = k - \sum_{i=1}^{k} w_{i_DC}^2. \tag{5}$$

In practical terms, if using R_{PC}, G, and D as the deflation-corrected estimators of association between an item i and the (undefined) latent

2 It may cause some confusion that the tradition within IRT and Rasch modeling uses "theta" as a general name for the observed score variable. While logically consistent, it creates a tension between the notation used within the article, where "theta" refers to the latent variable rather than the observed variable. To resolve this tension, "theta" is written with a subscript when referring to the manifestation of the latent variable (e.g., θ_X or θ_{IRT}), while the latent variable itself is denoted by the Greek letter θ.

variable θ, a theoretical deflation-corrected error variance based on R_{PC} is as follows (Metsämuuronen, 2023):

$$\sigma^2_{E_DC_R_{PC}\theta} = \sum_{i=1}^{k}\left(1 - R^2_{PCi\theta}\right) = k - \sum_{i=1}^{k} R^2_{PCi\theta} \quad (6)$$

and, based on *G*, it is as follows:

$$\sigma^2_{E_DC_G\theta} = \sum_{i=1}^{k}\left(1 - G^2_{i\theta}\right) = k - \sum_{i=1}^{k} G^2_{i\theta} \quad (7)$$

and based on *D*, it is as follows:

$$\sigma^2_{E_DC_D\theta} = \sum_{i=1}^{k}\left(1 - D^2_{i\theta}\right) = k - \sum_{i=1}^{k} D^2_{i\theta}. \quad (8)$$

These estimators are used later in the hypothetical example and in the empirical section—except that instead of *G* and *D*, G_2 and D_2 are used in Equations (7) and (8) because they suit better polytomous settings; their computing is discussed later. Of these better-behaving estimators of association, R_{PC} refers to a theoretical association in that it refers to theoretical (latent) items and scores that a researcher is not privy to (see the critique in Chalmers, 2017). *G* and *D* and the derivatives G_2 and D_2 refer to observed items and scores with a practical interpretation: the estimates strictly indicate the proportion of the test takers that are logically (ascending) ordered after they are ordered by the score variable; $p = 0.5 \times G + 0.5$ and $p = 0.5 \times D + 0.5$ (see Metsämuuronen, 2022i based on Metsämuuronen, 2021b). Of *G* and *D*, the estimates by *D* are more conservative in comparison with *G* because *G* omits the tied pairs in the computing proportions while *D* uses them (see Metsämuuronen, 2021b). In polytomous settings, the magnitude of the estimates by G_2 tends to follow close to those by R_{PC} and the estimates by D_2 close to those by R_{REG} (Metsämuuronen, 2022i).

1.4 A hypothetical numerical example of the inflation in estimates of error variance

Assume a hypothetical dataset, as in Table 1 with $k = 5$ items and incremental difficulty levels in items ($p = 0.083$–0.917) and $n = 12$ test takers. This could be a short subtest of "Sets" amid a larger mathematics achievement test given to a small group of students. Relevant indicators related to the traditional and deflation-corrected error variances are collected in Table 1. Four score variables are used: a raw score (θ_X), a standardized raw score (θ_{XSTD}), a factor score (θ_{FA}), and a theta score formed by the one-parameter logistic item response theory (1PL IRT) modeling or, factually, Rasch modeling (θ_{IRT}). The ML estimate is not optimal for the score variables because of the small sample size. However, it serves as an example of the computing process.

As an indicator of reliability, the coefficient omega total ($\rho_{\omega;}$ later, just omega) based on the works of Heise and Bohrnstedt (1970) and McDonald (1970, 1999) also known as McDonald's omega, is used. Omegas can be expressed as follows:

$$\rho_{\varpi} = \frac{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2 + \sum_{i=1}^{k}\left(1 - w^2_{i\theta}\right)}. \quad (9)$$

In the example of the possible outperforming estimators of correlation, *RPC*, *G*, and *D* and related deflation-corrected estimates of error variance are used as benchmarks for traditional factor loadings. Using *G* and *D* is justified because the items are binary (Metsämuuronen, 2020a,b, 2021a). From the viewpoint of the benchmarking estimators (Equations 6–8), using the raw score, standardized raw score, and 1PL model in IRT modeling leads to identical results because the order of the test takers does not change in the standardization and logistic transformation.

From the viewpoint of reliability estimates, the estimate by the traditional coefficient omega (Equation 8) is notably deflated as being $\hat{\rho}_{\omega} = (1.485)^2 / \left((1.485)^2 + 3.600\right) = 0.380$. It appears that the factor score is not the best reflection of the true ability in the case. Namely, the related deflation-corrected estimate is based on the form of omega, and using R_{PC} gives a deflated estimate of $\hat{\rho}_{\omega\underline{\grave{e}}R_{PC\ FA}} = (1.544)^2 / \left((1.544)^2 + 2.550\right) = 0.483$. In the hypothetical example, the estimates related to the raw score (and IRT score) appear more credible in comparison with the factor score, because items g_1, g_4, and g_5 can deterministically distinguish test takers from each other when tied cases are not considered ($R_{PC} \approx G = 1$). The estimates are quite close when the tied cases are considered ($D = 0.889$–0.909). Factor analysis can detect this phenomenon only in g_5 ($\lambda_{g_{5,\,FA}} = 0.999$) but fails notably in g_1 ($\lambda_{g_{1,\,FA}} = 0.091$) and g_4 ($\lambda_{g_{4,\,FA}} = 0.522$). Hence, we obtain the deflation in reliability by omega and inflation in the error variance.

If the raw score, standardized raw score, or IRT score are used as a justified reflection of the latent ability, the estimates of reliability would be notably higher by using R_{PC} as the weight factor, $\hat{\rho}_{\omega_R_{PC}X} = \hat{\rho}_{\omega_R_{PC}X_{STD}} = \hat{\rho}_{\omega_R_{PC}IRT} = (4.089)^2 / \left((4.089)^2 + 1.389\right) = 0.923$, mildly higher if *G* was used ($\hat{\rho}_{\omega_GX} = \hat{\rho}_{\omega_GX_{STD}} = \hat{\rho}_{\omega_GIRT} = 0.932$), and mildly lower if *D* was used ($\hat{\rho}_{\omega_DX} = \hat{\rho}_{\omega_DX_{STD}} = \hat{\rho}_{\omega_DIRT} = 0.869$).

When comparing the standardized score variables in the hypothetical example, the deflation in the estimate by the traditional omega is 56% [=(0.869–0.380)/0.869 × 100] if the conservative *D* is taken as the benchmarking weight factor and 59% if R_{PC} or *G* are taken as the benchmarks. The inflation in the traditional error variance based on the factor loadings is 30–41% when the factor score is considered and up to 76–182%, depending on the weight factor when the standardized raw score is considered. The magnitude of both deflation and inflation is notable and worth further investigation.

More in-depth analysis is discussed in Section 5, where a set of 1,440 real-life tests with various characteristics is used to explore the boundaries of the inflation in error variance.

TABLE 1 A hypothetic dataset related to inflation in the estimated error variance.

	Items					Scores			
Test taker	g1	g2	g3	g4	g5	θ_X	θ_{XSTD}	θ_{FA}	θ_{IRT}
1	1	0	0	0	0	1	−1.567	−0.28873	−1.976
2	0	1	0	0	0	1	−1.567	−0.28865	−1.976
3	1	1	0	0	0	2	−0.522	−0.28834	−0.642
4	1	0	1	0	0	2	−0.522	−0.28937	−0.642
5	1	1	0	0	0	2	−0.522	−0.28834	−0.642
6	1	1	0	0	0	2	−0.522	−0.28834	−0.642
7	1	0	1	1	0	3	0.522	−0.28778	0.642
8	1	1	1	0	0	3	0.522	−0.28897	0.642
9	1	1	1	0	0	3	0.522	−0.28897	0.642
10	1	1	1	0	0	3	0.522	−0.28897	0.642
11	1	1	0	1	1	4	1.567	3.17384	1.976
12	1	1	1	1	0	4	1.567	3.17384	1.976
p	0.917	0.750	0.500	0.250	0.083				
B (IRT)	−2.482	−1.238	0	1.238	2.482				
Score FA							SUM	Omega	
$\lambda_{i\theta_{FA}}$	0.091	0.174	−0.301	0.522	0.999		1.485	0.380	
$1-\lambda^2_{i\theta_{FA}}$	0.992	0.970	0.909	0.728	0.002		3.600		
$RPC_{i\theta_{FA}}$	−0.302	0.339	−0.493	1	1		1.544	0.483	
$1-RPC^2_{i\theta_{FA}}$	0.909	0.885	0.757	0	0		2.550		
$G_{i\theta_{FA}}$	0.091	0.259	−0.444	1	1		1.906	0.571	
$1-G^2_{i\theta_{FA}}$	0.992	0.933	0.803	0	0		2.728		
$D_{i\theta_{FA}}$	0.091	0.174	−0.444	1	1		1.821	0.545	
$1-D^2_{i\theta_{FA}}$	0.992	0.970	0.802	0	0		2.764		
Score X = Score XSTD = Score IRT									
$RPC_{i\theta_X}$	1	0.449	0.640	1	1		4.089	0.923	
$1-RPC^2_{i\theta_X}$	0	0.798	0.591	0	0		1.389		
$G_{i\theta_X}$	1	0.500	0.688	1	1		4.188	0.932	
$1-G^2_{i\theta_X}$	0	0.750	0.527	0	0		1.277		
$D_{i\theta_X}$	0.909	0.370	0.611	0.889	0.909		3.688	0.869	
$1-D^2_{i\theta_X}$	0.174	0.863	0.627	0.210	0.174		2.047		

(Continued)

TABLE 1 (Continued)

	Items					Scores			
Test taker	g1	g2	g3	g4	g5	θ_X	θ_{XSTD}	θ_{FA}	θ_{IRT}
Inflation in $1-\lambda^2_{i\theta_{FA}}$ when the factor score is considered							SUM	Deflation %[1]	
$dRPC_{i\theta_{FA}}$	0.083	0.085	0.153	0.728	0.002		1.050	41.2	
$dG_{i\theta_{FA}}$	0.000	0.037	0.107	0.728	0.002		0.873	32.0	
$dD_{i\theta_{FA}}$	0.000	0.000	0.107	0.728	0.002		0.836	30.3	
Inflation $1-\lambda^2_{i\theta_{FA}}$ when the (standardized) raw score is considered							SUM	Deflation %[1]	
$dRPC_{i\theta_X}$	0.992	0.171	0.319	0.728	0.002		2.211	159.2	
$dG_{i\theta_X}$	0.992	0.220	0.383	0.728	0.002		2.324	182.0	
$dD_{i\theta_X}$	0.818	0.107	0.283	0.518	−0.172		1.553	75.9	

[1] $\left(\left|\sum_{i=1}^{k}\left(1-w_{i,}^2\right)-\sum_{i=1}^{k}\left(1-\lambda^2_{i,FA}\right)\right| / \sum_{i=1}^{k}\left(1-w_{i,}^2\right)\right)\times 100$

1.5 Summary of the discussion by far

From earlier studies, it is known that the traditional estimates of reliability tend to be deflated. The deflation may be radical (up to 0.40–0.70 units of reliability), and the reason for this deflation is the poor behavior of the product–moment coefficient of correlation in the case that the widths of the scales of the variables are far from each other. This is always the case in measurement modeling settings, and it is often exacerbated in achievement testing, where we are willing to use both very easy, medium, and very demanding tasks to cover the full range of ability scales in one test. In these types of tests, the standard errors related to the score are radically inflated; in some extremely easy or difficult tests, the standard errors have been reported to be more than 10 times higher than they should be.

Because the relationship between reliability and error variance and the standard error of the score can be easily observed from the formulae, the technical reasons for the observed radical deflation in the estimates of reliability can be traced to three sources: either the population variance (σ^2_X) is deflated, or the error variance (σ^2_E) is inflated—or both may happen at the same time. This article focuses on error variance, which is strictly embedded in widely used reliability estimators such as omega and maximal reliability (see Supplementary Appendix 1). In some empirical settings, it has been noted that the estimates of reliability may be deflated by 0.40–0.70 units, and this can be directly connected to mechanical errors in the estimation of correlation, which needs to be separated from attenuation related to violations against the measurement model. From this viewpoint, it appears that the phenomenon of radical inflation in error variance and measurement error caused by technical error during the estimation process is discussed sparsely in literature, if at all, considering its possible consequences (see, however, discussion in Metsämuuronen, 2023 related to achievement testing, and Metsämuuronen, 2022b,f, related to inflation in conditional standard errors). Hence, it seems justified to further discuss the reasons, mechanisms, and consequences of the deflation observed in the estimates of reliability and the inflation in error variance.

2 Research questions

This note examines the magnitude and consequences of inflation in error variance estimates. The conceptual matters and reasons behind them are discussed in Section 1.3. The inflation in the error variance begs three key questions: (1) What is the magnitude of the inflation in the estimated error variance and the related standard errors in real-life testing settings? (2) How can the magnitude of inflation be predicted? and (3) How do deflation-corrected estimators of error variance and standard errors compare to traditional ones in real-life datasets? These questions are studied and discussed in the empirical section (Section 4) using a simulation dataset based on real-life settings.

3 Methods

3.1 Dataset

A dataset of 4,023 nationally representative test-takers of a mathematics test with 30 binary items (FINEEC, 2018) is used as the "population." From the original dataset, 10 samples with finite sample sizes of $n = 25$, 50, 100, and 200 test-takers were drawn. These samples imitate different real-life sample sizes, ranging from tests for a large student group ($n = 200$) to classroom settings ($n = 25$). In each of the 10×4 datasets, 36 tests were produced by varying the number of items in the tests, the difficulty levels of the items, and the length of the scales of the score [$df(X)$ = number of categories in the score scale − 1] and the item [$df(g)$ = number of categories in the item scale − 1].

Polytomous items were produced as a combination of binary items. In the final dataset, both the tests with the original items and the tests with fewer items but wider scales are mixed. Datasets comprising the traditional and deflation-corrected estimates of reliability, the estimates of error variance and standard errors and estimated population variances, and related derivatives and background information of the 1,440 tests are available at doi: 10.13140/RG.2.2.25390.79687 in CSV format and at doi: 10.13140/RG.2.2.33779.40481 in IBM SPSS format.

3.2 Estimators of association

Because we are using both binary and polytomous items, instead of G and D, their dimension-corrected modifications (G_2 and D_2) are used. It is known that when the number of categories in the item exceeds 3 (D) or 4 (G), G and D tend to underestimate the item-score association (see, e.g., Metsämuuronen, 2020a,b, 2021a). Hence, Metsämuuronen (2021a) suggests modifications specific to the measurement modeling settings as follows: $G_2 = G \times \left(1 + \left(1 - abs\left(G\right)\right) \times A\right)$ and $D_2 = D \times \left(1 + \left(1 - abs\left(D\right)\right) \times A\right)$, where G and D are the observed values of G and D and $A = \left(1 - \frac{1}{df\left(g\right)}\right)^3$. With binary items, $df(g) = 1$, and $A = 0$, and, hence, $G_2 = G$ and $D_2 = D$. Moreover, when $G = D = 1$, $G_2 = G$, and $D_2 = D$.

For the note, the estimates by G_2 and D_2 were computed manually from the values of G and D being standard outputs of a statistical software package (in the case of IBM SPSS; see syntaxes with some generally known packages in Supplementary Appendix 2).

3.3 Variables and statistics

In assessing the magnitude of the inflation in the estimates of error variance, a simple statistic is used: the difference between the traditional estimate and the deflation-corrected estimates. The traditional estimates are denoted $d\hat{\sigma}^2_{E_\lambda FA}$ or "VAR(*E*)_*LFA*" as an abbreviation of "error variance based on factor loadings as the linking factor and the factor score variable as the manifestation of the latent score estimated by using the maximum likelihood extraction method." Correspondingly, the deflation-corrected estimators are denoted $\hat{\sigma}^2_{E_R_{PC}X}$, $\hat{\sigma}^2_{E_G_2X}$, $\hat{\sigma}^2_{E_D_2X}$ or "*VAR(E)_*R_{PC}*X*," "*VAR(E)_*G_2*X*," and "*VAR(E)_*D_2*X*," respectively as abbreviations of "error variance based on R_{PC}/G_2/D_2 as the linking factor and the raw score as the manifestation of the latent score." The "written" version is seen, specifically in Figures to come. While traditional estimates are based on factor score variables, the latter estimates are based on raw scores. We may also note that the result would be equal if the standardized raw scores or IRT scores were used because the estimates of the item–score association by the deflation-corrected estimators of correlation are equal with the raw scores, standardized scores, and IRT scores because the order of the test takers does not change in these transformations.

A difference ("*d*") between the sample estimates of the traditional estimate of error variance and the deflation-corrected estimate reflects the magnitude of inflation. This difference is noted as follows: $d\hat{\sigma}^2_{E_R_{PC}X}$ or "*dVAR(E)_*R_{PC}*X*" refers to inflation in $\hat{\sigma}^2_{E_\lambda_{FA}}$ when $\hat{\sigma}^2_{E_R_{PC}X}$ is used as the benchmark. Similarly, "*dVAR(E)_G2X*" or "*dVAR(E)_D2X*" refers to the inflation in the case G_2 or D_2 have been used as the deflation-corrected estimator of weight factor w_i. Technically,

$$\begin{aligned} d\hat{\sigma}^2_{E_R_{PC}X_{STD}} &= \hat{\sigma}^2_{E_\lambda_{FA}} - \hat{\sigma}^2_{E_R_{PC}X} \\ &= \left(k - \sum_{i=1}^{k}\lambda^2_{i\theta_{FA}}\right) - \left(k - \sum_{i=1}^{k}R^2_{PCi\theta_{XSTD}}\right) \\ &= \sum_{i=1}^{k}R^2_{PCi\theta_{XSTD}} - \sum_{i=1}^{k}\lambda^2_{i\theta_{FA}} \end{aligned} \quad (10)$$

Then, if the magnitude of $\hat{\sigma}^2_{E_R_{PC}X_{STD}}$ is positive, the traditional estimated error variance of the score is overestimated. In some cases, this is expressed as percentages, which are notated using "*dp*" such as in $dp\hat{\sigma}^2_{E_R_{PC}X_{STD}}$. The percentages are computed so that the deflation-corrected estimate is the base; the percentage indicates the deflation in the traditional estimate, assuming that the deflation-corrected estimate represents the true value.

However, using the percentages is not necessarily wise to connect to the phenomenon because the magnitude of the error variance appears to vary radically depending on the number of items in the compilation. With two or three items, the magnitude of the error variances could be 0.2 and 0.6, leading to $dp\hat{\sigma}^2_{E_R_{PC}X_{STD}} = 200$; that is, the error variance seems to be inflated by 200%. If the difference is notably greater, such as 20 or 30, inflation would be only 50%.

3.4 Methods in analysis

The magnitude of the inflation is illustrated by using visual tools. The explaining factors are studied using standard linear regression modeling, and linear and non-linear graphical modeling is used with two variables. Decision tree analysis (DTA; IBM, 2017), a data mining tool with the CHAID algorithm (Chi-square Automatic Interaction Detector; Kass, 1980), explores the variables and groups the categories. In DTA, the outcome is a non-linear hierarchical model based on maximizing the F-test (or χ^2) statistics; all possible combinations of the explaining factors are computed, and the statistically best combination is selected. This tool is used when the number of items in the compilation explaining the error variance is of interest. A paired-sample t-test is used to compare mean differences in standard errors, and Cohen's d (Cohen, 1988) is used to indicate effect size.

4 Results

4.1 The magnitude of the error variance in real-life settings

Four lifts are made regarding the magnitude of the error variance in real-life settings. First, with the smallest sample size in the simulation (n = 25), it was not possible to produce all the factor models. While it *was* possible to produce 360 estimates of error variance related to the

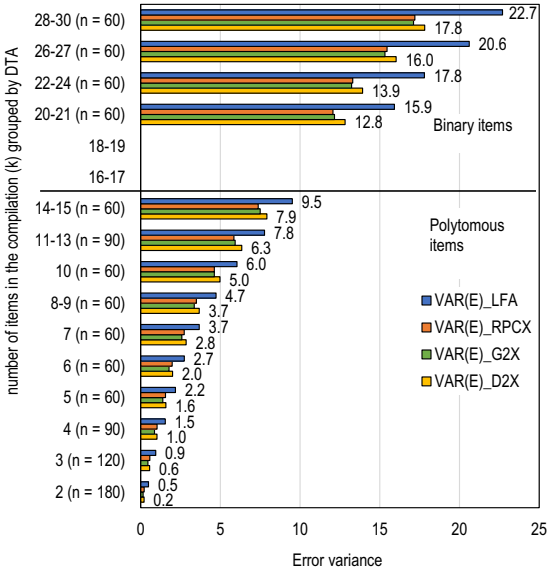


FIGURE 2
Magnitude of the error variance by the number of items in the compilation (sample sizes n = 50, 100, and 200) in groups suggested by DTA.
VAR*(E)_LFA* = error variance based on *factor loading (L) as the linking factor and a one-factor factor score variable (FA) estimated by using maximum likelihood estimation* as the manifestation of the latent score. *VAR(E)_RPCX, VAR(E)_G2X,* and *VAR(E)_D2X* = error variance based on $R_{PC}/G_2/D_2$ as the linking factor and the raw score (X) as the manifestation of the latent score.

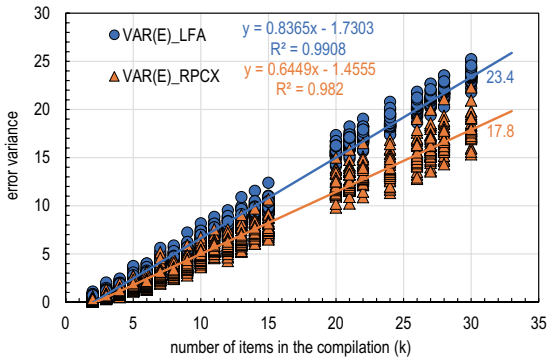


FIGURE 3
Models of the magnitude of the error variance by the number of items in the compilation (all tests).
VAR(E)_LFA=the traditional error variance based on *factor loading (L) as the weight factor and factor score variable (FA) from a one-factor solution estimated by using maximum likelihood estimation* as the manifestation of the latent ability. *VAR(E)_RPCX,* Deflation-corrected error variance based on R_{PC} as the weight factor and the raw score (X) as the manifestation of the latent ability.

deflation-corrected correlation estimators, the factor solution was found only in 314 out of 360 tests. This loss of 12.5% in the group of the smallest sample size is systematic in that the error variances were missing with binary items and tests with more than 24 items; that is, in the settings where the inflation was the greatest (see Figure 2). Hence, in Figure 2, only estimates with sample sizes of $n \geq 50$ (n = 1,080 tests out of 1,440) can be observed. Later, all possible estimates are used, that is, n = 1,394 for the traditional estimates and n = 1,440 for the deflation-corrected estimates. In the pairwise comparisons, only 1,394 pairs are available.

Second, the dataset used in the simulation did not include tests with 16–19 items or tests longer than 30 items. Technicalities in forming the dataset used in simulation led to practicalities such as the test with 20–30 items being based on binary items and the test with 2–15 items being polytomous items (Figure 2). Notably, some categories of k in Figure 2 are combined, as suggested by DTA with the CHAID algorithm; using these groups, the difference between the categories is the most statistically significant [for the traditional estimates, $F(13,\ 1{,}066) = 14{,}768.58$, $p < 0.001$]. In Figure 2, n refers to the number of tests; n = 60 indicates that the dataset consists of 60 tests compiled of 28–30 items.

Third, the magnitude of the error variance increases systematically with the number of items comprising the test (k). This is known from Eqs. (3) to (5), and it [i.e., the phenomenon in Eqs 3–5], is understandable because the error variance of the test is a cumulative sum of error variances of the single items. While the traditional error variance ("*VAR(E)_LFA*") is approximately 0.5–2.2 units with tests with few items (k = 2–5), with 20–30 binary items, it is 15.9–23.4 units. The number of items in the compilation explains almost all error variance variability; using the linear regression model, $R^2 > 0.99$ for the traditional error variance, and if using the deflation-corrected estimators of association, $R^2 > 0.98$ (Figure 3).

Fourth, all the estimators of deflation-corrected error variance give estimates that are systematically smaller in magnitude than traditional estimates. While, in the given dataset, the traditional estimates related to the factor score variable and factor loadings tend to range from 0.5 to 23.4 units, depending on the number of items in the compilation, the deflation-corrected estimates related to the raw score range from 0.2 to 17.8 units. It may be possible that the lower magnitude of the error variance related to the deflation-corrected estimates could be partly explained by the difference in the score variable; after all, the score variables differ between the estimators. However, traditionally, factor score variable has been taken as one of the "optimal linear combinations" discussed over years by, chronologically, e.g., Thompson (1940), Guttman (1941), Stouffer (1950), Lord (1958), and Bentler (1968) and later, for example, Li et al. (1996) and Li (1997); the "optimal" combination should be, logically, better than the raw score and, hence, it should include less error in comparison with the raw score. However, the studies with deflation-corrected estimators of reliability have shown that the reason for the deflation is mainly in estimates of the association between the item and score variable (see the discussion above) rather than in the difference between the score variables (see Metsämuuronen, 2022b of the effects of different sources of underestimation of reliability).

4.2 The magnitude of the inflation of the error variance in the real-life datasets

As shown in Figures 2 and 3, the inflation in the error variance tends to become greater the more items there are in the compilation. Figures 4A,B and 5 further exploit the same finding: Figures 4A,B use factual estimates, and Figure 5 uses the means of error variance in the compiled groups of the number of items in the compilation suggested by DTA. Three major points are highlighted.

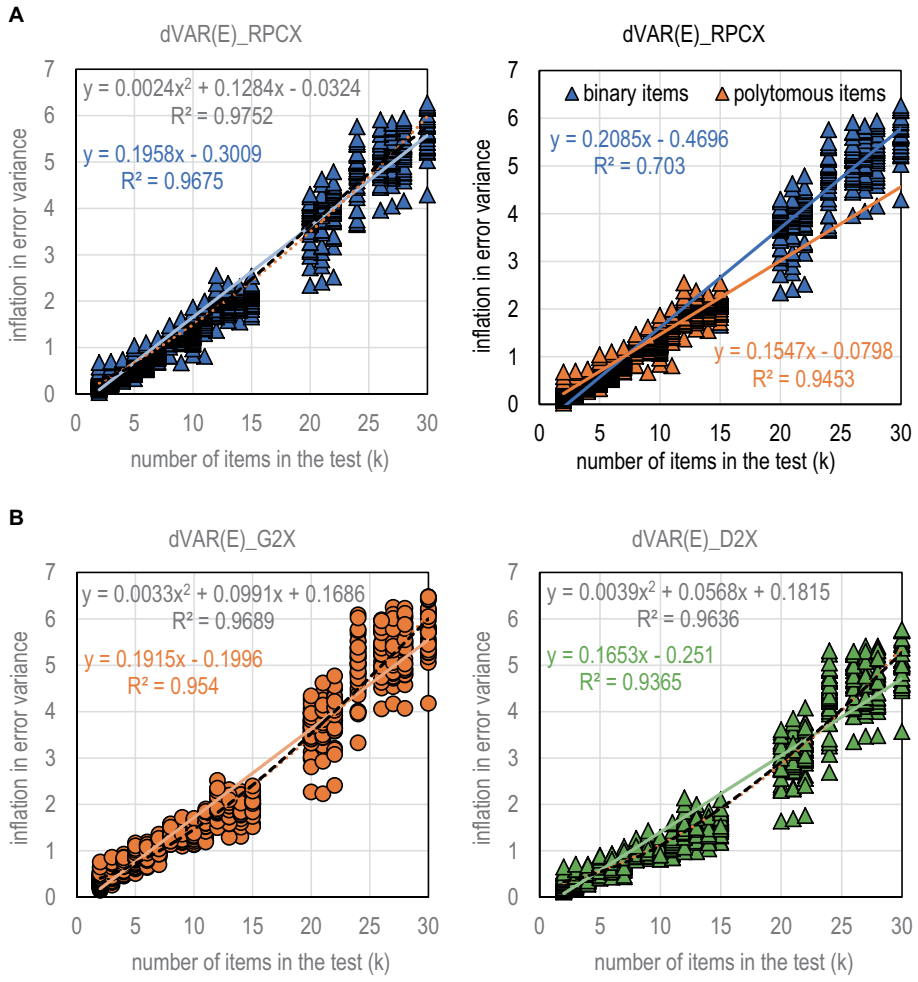


FIGURE 4
(A) Magnitude of the inflation in the error variance based on RPC (linear and non-linear models); all tests (left) and separated by the type of the tests (right). *dVAR(E)_RPCX=VAR(E)_RPCX−VAR(E)_LFA*, i.e., the difference (*d*) between the error variance based on the deflation-corrected and traditional estimates of error variance, that is, error variance based on R_{PC} as the weight factor and raw sum (*X*) as the manifestation of the latent ability and error variance based on the factor loading (*L*) as the weight factor and factor score variable (*FA*) from a one-factor solution as the manifestation of the latent ability. **(B)** The magnitude of the inflation in the error variance based on G2 and D2 (linear and non-linear models); all tests. *dVAR(E)_G2X=VAR(E)_G2X−VAR(E)_LFA*, i.e., the difference (*d*) between the error variance based on the deflation-corrected and traditional estimates of error variance, that is, error variance based on G_2 as the weight factor and raw sum (*X*) as the manifestation of the latent ability and error variance based on the factor loading (*L*) as the weight factor and factor score variable (*FA*) from a one-factor solution as the manifestation of the latent ability. Similarly, *dVAR(E)_D2X=VAR(E)_D2X−VAR(E)_LFA* based on D_2 as the weight factor.

First, the models of the magnitude of the inflation may be different for binary items and polytomous items. This is illustrated in Figure 4A with R_{PC}: the magnitude of the slope parameter with binary items is 0.209, and with the polytomous items, it is 0.155. In the dataset used in the simulation, the polytomous items were dependent on the binary items; after all, the polytomous items were formed as combinations of binary items. Systematic studies with independent polytomous and binary items would be valuable in confirming this phenomenon.

Second, in the simulation dataset, a linear model of $inflation = 0.2 \times k - 0.3$ explains well the magnitude of inflation when the benchmarking estimator is based on R_{PC} and $inflation = 0.2 \times k - 0.2$ when G_2 is the benchmark; that is, the estimates tend to be somewhat higher when using G_2 than R_{PC} (Figure 4B). The model for the conservative estimates by D_2 have a smaller magnitudes in the slope parameter and constant ($inflation = 0.17 \times k - 0.25$). In all cases, the explaining power for a linear model is high ($R^2 = 0.94$–0.97), although the models with a second power give slightly better explaining powers ($R^2 = 0.96$–0.98) (Figure 4B).

Third, not only is the error variance cumulative by the number of items (see Figures 4A,B), but the inflation in the error variance is also cumulative by the number of items. With 2–4 items, the error variance ranges from 0.3 to 0.5 regardless of the benchmarking deflation-corrected estimator, while with 30 binary items, the inflation in the error variance ranges 4.7–5.4 units depending on the benchmarking estimator (Figure 5). The technical reason for the phenomenon is that they tend to give estimates with a higher magnitude than PMC because of the better behavior of the deflation-corrected estimators of association. This is understood by the common characteristics of the deflation-corrected estimators of correlation, which give higher estimates than the traditional deflation-prone estimators. Because the error variance is cumulative, the more items we have in the compilation, the more cumulative error we obtain.

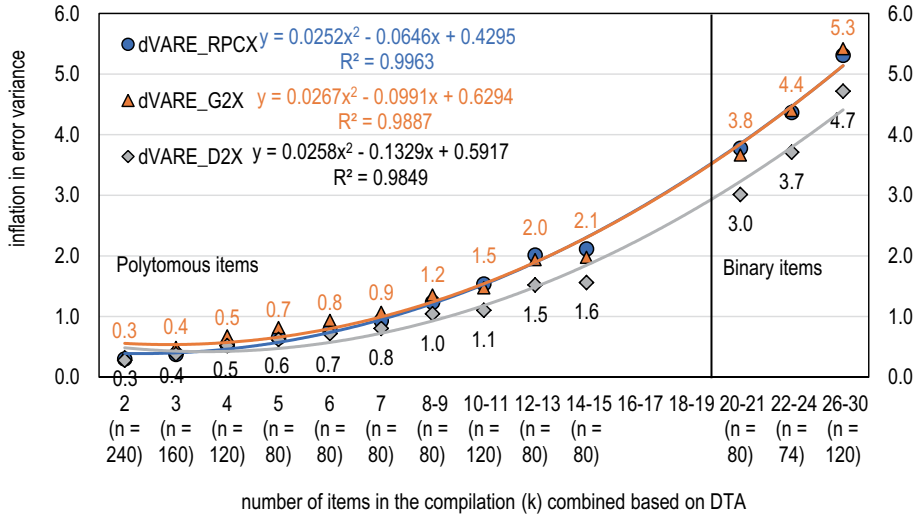


FIGURE 5
Magnitude of the inflation in the error variance by the number of items (*n* refers to the number of tests). *dVAR(E)_RPCX = VAR(E)_RPCX – VAR(E)_LFA*, *dVAR(E)_G2X = VAR(E)_G2X – VAR(E)_LFA*, and *dVAR(E)_D2X = VAR(E)_D2X – VAR(E)_LFA*, i.e., the difference (*d*) between the error variance based on the deflation-corrected and traditional estimates of error variance, that is, error variance based on $R_{PC}/G_2/D_2$ as the weight factor and raw sum (*X*) as the manifestation of the latent ability and error variance based on the factor loading (*L*) as the weight factor and factor score variable (*FA*) from a one-factor solution as the manifestation of the latent ability.

4.3 Note on the factors explaining the inflation in error variance

The magnitude of inflation was studied with linear regression analysis by using five factors related to the tests: the sample size (n), the test difficulty assessed by the average item difficulty ($\bar{p}$), the number of items (k), the number of categories in the score [$df(g)$], and the score [$df(X)$]. However, because the number of items alone explains 96.4–97.5% of the variability in inflation by the quadratic model (R^2 = 0.964–0.975), the other factors cannot add much information to the model. Factually, in all conjoint models, different combinations of other elements increase the explaining power statistically significantly, but the final explaining power of the more complicated linear model after Wherry's adjustment is *lower* (R^2_{Adj} = 0.952–0.974) than in the models with only one explaining factor without a need for the adjustment. Hence, these models are not included in this note. However, Table 2 condenses an example of the impact of different factors in a conjoint linear model where R_{PC} is used as a deflation-corrected estimator of association. By using R_{PC} in the correction, a number of items alone explain 97.7% (by quadratic model) or 96.8% (by linear model) of the variability in inflation. The whole linear model explains 97.4%. Notably, the relationship is not linear (see Figure 5).

4.4 Inflation in the standard errors

As discussed above, inflation in the error variance is strictly linked to the deflation of reliability. Another direct consequence is that the estimated standard errors are inflated. The more the item–score correlations are deflated, the more the reliability estimates are deflated, and, consequently, because of Equation (1), the more the standard errors are inflated (see the discussion in Metsämuuronen, 2023). The relation between the inflated error variance and inflated standard errors is somewhat more complicated than the inflation in error variance itself.

Taking the form of coefficient omega (Equation 9) as an example, the deflation in reliability depends not only on the inflated error variance $\sum_{i=1}^{k}\left(1-\lambda^2_{i\theta_{FA}}\right)$ but also on the other component related to "true variance" $\left(\sum_{i=1}^{k}\lambda_{i\theta_{FA}}\right)^2$, which is deflated when the traditional factor loadings are considered; these two elements are intertwined. If using the basic formula for the *S.E.m.* (based on Equation 1) with deflation-corrected estimators of correlation in estimation (e.g., $\sum_{i=1}^{k}\left(1-RPC^2_{i\theta}\right)$ and $\left(\sum_{i=1}^{k}RPC_{i\theta}\right)^2$) and related deflation-corrected estimators of reliability (DCER; Metsämuuronen, 2022c,d,e), we obtain deflation-corrected standard errors (*S.E.m._DC*).

The inflation in the standard errors is studied by using the coefficient omega as an example of an estimator of reliability. The traditional omega is used to estimate the reliability of a factor score, which is a standardized variable with $\sigma^2_X = 1$, and, hence, the traditional estimator of *S.E.m* using omega is $S.E.m. = \sqrt{(1-\rho_\omega)}$. The corresponding DCERs, "*OmegaR*$_{PC}$," "*OmegaG*$_2$," and "*OmegaD*$_2$," use

TABLE 2 Conjoint model of relevant factors explaining the inflation on error variance; dependent variable: dVAR(E)_RPCX.

	Unstandardized coefficients		Standardized coefficients		
Model	*B*	Std. Error	Beta	*t*	Sig.
Constant	−0.793	0.094		−8.471	< 0,001
Number of cases in the sample (*n*)	0.001	0	0.042	6.573	< 0,001
Test difficulty (mean of item difficulty)	0.635	0.141	0.02	4.513	< 0,001
Number of items (*k*)	0.216	0.001	1.083	148.834	< 0,001
(Average) number of categories in the item minus 1 (df(g))	0.057	0.003	0.122	16.851	< 0,001
Number of categories in the score minus 1 (df(X))	−0.027	0.003	−0.063	−9.356	< 0,001
R	R^2	R^2_{Adj}	Std. Error of the Estimate		
0.987	0.974	0.974	0.264		

the form of Equation (6) as the base and R_{PC}, G_2, and D_2 as the weight factors (see, e.g., Metsämuuronen, 2022d). However, the score variable in the datasets used in the simulation was originally the raw score. To compare estimated standard errors, without losing generalizability, we can assume that the raw scores were standardized; a correlation between an item and a raw score is identical to the correlation between an unstandardized item and a standardized raw score. Then, the deflation-corrected standard errors based on R_{PC} are computed as follows: $S.E.m._DC_R_{PC} = \sqrt{\left(1-\rho_{\omega_R_{PC}\theta_{XSTD}}\right)}$, which is abbreviated in the figures to come as "*SEM_RPC_STD*," referring to "standard error based on the formula of omega and using R_{PC} as the weight factor and standardized raw score (*STD*) as the manifestation of the latent ability."

Similarly, the deflation-corrected standard errors based on G_2 are computed as follows: $S.E.m._DC_G_2 = \sqrt{\left(1-\rho_{\omega_G_2\theta_{XSTD}}\right)}$. It is abbreviated as "*SEM_G2_STD*."

The deflation-corrected standard errors based on D_2 are computed as follows: $S.E.m._DC_D_2 = \sqrt{\left(1-\rho_{\omega_D_2\theta_{XSTD}}\right)}$. It is abbreviated as "*SEM_G2_STD*." The notations $\rho_{\omega_R_{PC}\theta_{XSTD}}$ ("*OmegaR*$_{PC}$*STD*"), $\rho_{\omega_G_2\theta_{XSTD}}$ ("*OmegaG*$_2$*STD*"), and $\rho_{\omega_D_2\theta_{XSTD}}$ ("*OmegaD*$_2$*STD*") indicate that the base of the estimator of reliability is omega (Equation 9), the weight factor w_i is operationalized as R_{PC}, G_2, or D_2, and the latent score variable is manifested as the standardized raw score (θ_{XSTD}). Hence, the standard errors related to the factor score variables and the standard errors of standardized raw scores are compared. Understandably, the outcome is not exact, but it gives us a rough idea of the magnitude of the inflation in standard errors.

In the datasets used in the simulation, the average *S.E.m* by using the traditional omega is 0.38 standard units, while the deflation-corrected standard errors using the deflation-corrected estimators of association with the formula of omega vary by 0.26–0.28, depending on the weight factor. Hence, on average, the traditional standard errors are inflated by 35–48% (Figure 6). The difference is statistically significant (paired-samples t-test, $t = 112.39–128.40$; $p < 0.001$) and remarkable or "huge" (Cohen's $d = 3.20–3.40$; see Sawilowsky, 2009). The modest inflation in comparison with the datasets by Metsämuuronen (2022b,f) is caused by the fact that the dataset used does not contain many items with extreme difficulty levels, and, hence, the deflation in the estimates of reliability is modest: $\rho_\omega = 0.85$ by using the traditional omega vs. $\rho_{\omega_wi\theta} = 0.92–0.93$ by using DCERs, that is, 7–8%. Notably, in the extremely easy dataset discussed by Metsämuuronen (2022b) (originally in Metsämuuronen and Ukkola, 2019), the deflation in the estimates by omega was 53–57% ($\rho_\omega = 0.42$ by omega vs. $\rho_{\omega_wi\theta} = 0.87–0.97$ by DCERs). In a real-life setting by Metsämuuronen (2022f), the deflation in reliability with easy items was 68–69% ($\rho_\omega = 0.29$ by omega vs. 0.86–0.90 by DCERs using *G* and *D*).

Even though the error variance by Equations (4, 6–8) is directly related to the number of items in the compilation, the magnitude of the standard error by Equation (1) is not systematically related to the number of items in the compilation, although it tends to become smaller the wider the scales of the item and score (Figure 7). In Figure 7, the abbreviations "*SEM L_FA*," "*SEM RPC_STD*," "*SEM G2_STD*," and "*SEM D2_STD*" refer to standard errors (SEM) estimated either by the traditional way by using coefficient omega with factor loadings (L_FA) or by using the formula of omega with deflation-corrected estimators of association (RPC/G2/D2) and standardized raw scores (STD). Formally, the DCERs are $\rho_{\omega_R_{PC}\theta_{XSTD}}$, $\rho_{\omega_G_2\theta_{XSTD}}$, and $\rho_{\omega_D_2\theta_{XSTD}}$, where the base of the estimator of reliability is omega (Equation 9), the weight factor w_i is operationalized as R_{PC}, G_2, or D_2, and the latent score variable is manifested as the standardized raw score (θ_{XSTD}).

4.5 Note on the standard errors and "standard errors"

We have presented two approaches to computing the average standard error. On the one hand, we have the traditional *S.E.m.* based on the definition of reliability of the score (Equation 1), that is,

$$\sigma_E = \sqrt{\sigma_X^2\left(1-REL\right)} \qquad (11)$$

This implies and determines that the standard error cannot exceed the magnitude of the standard deviation (σ_X) related to the score because $REL \leq 1$. With a standardized score with $\sigma_X^2 = 1$, according to Equation (11), the variance of the score can be divided into reliable variance (reliability, *REL*) and unreliable variance (σ_E^2), which together do not exceed the value 1, that is, $REL + \sigma_E^2 = 1$.[3]

3 Sincere thanks to PhD Christian Geiser from QuantFish LLC for reminding me of this in a private discussion concerning the matter.

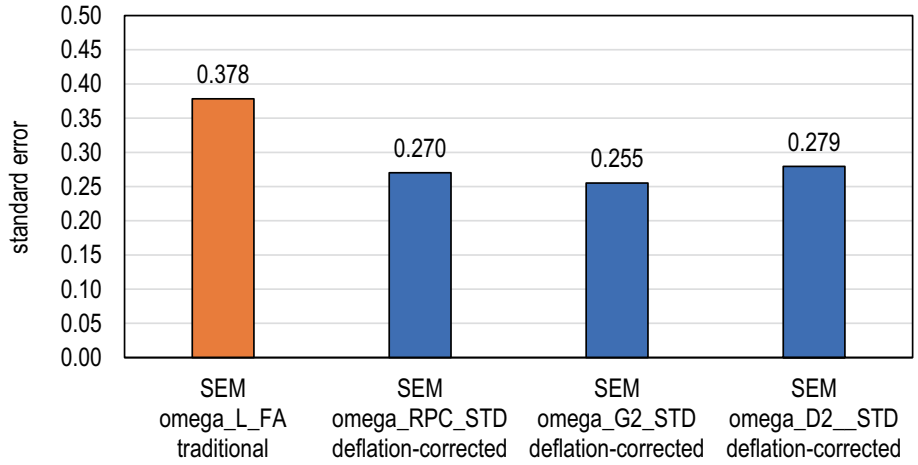


FIGURE 6
General tendencies of the traditional and deflation-corrected standard errors.
SEM omega_L_FA traditional, Traditional standard errors based on the estimates of reliability by coefficient omega using factor loadings (L) as weight factors in estimation. SEM omega_RPC_STD deflation-corrected, Deflation-corrected standard errors based on the estimates of reliability by coefficient omega using R_{PC} between items and the standardized raw score as weight factors in estimation; SEM omega_G2_STD deflation-corrected, Deflation-corrected standard errors based on the estimates of reliability by coefficient omega using G_2 between items and the standardized raw score as weight factors in estimation; SEM omega_D2_STD deflation-corrected, Deflation-corrected standard errors based on the estimates of reliability by coefficient omega using D_2 between items and the standardized raw score as weight factors in estimation.

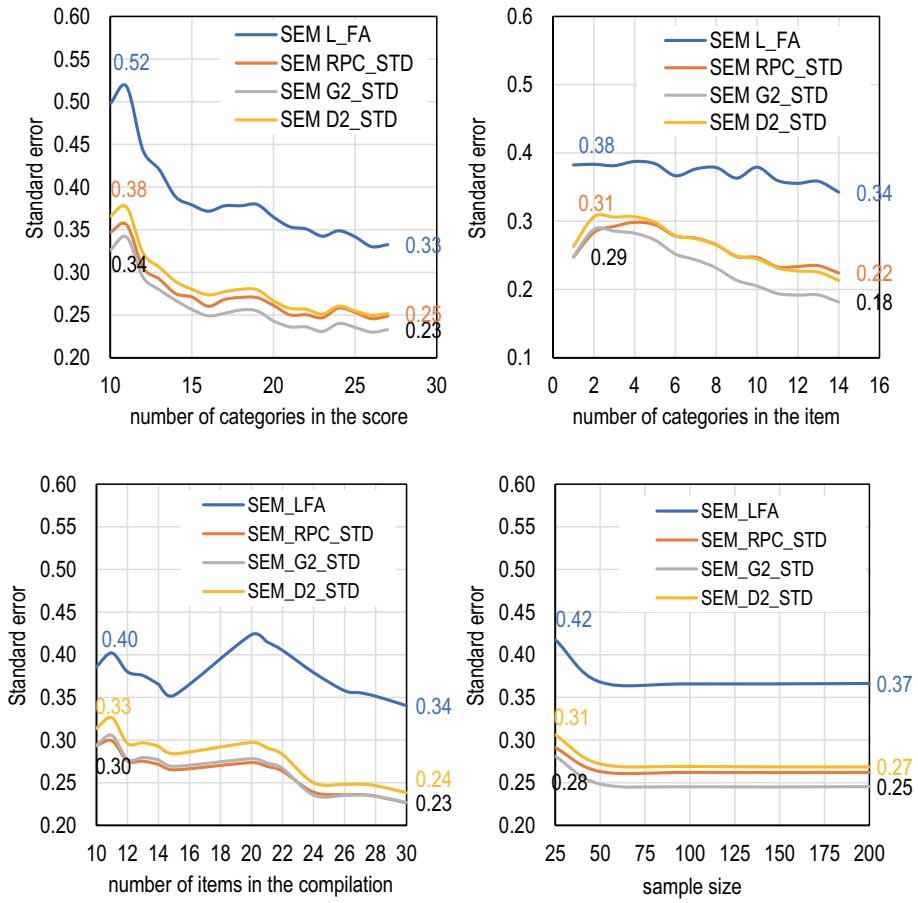


FIGURE 7
Selected factors explaining the inflation in standard errors.
SEM L_FA, Traditional standard errors (SEM) estimated using coefficient omega with factor loadings (L_FA). *SEM RPC_STD*, *SEM G2_STD*, *SEM D2_STD*, Deflation-corrected standard errors (SEM) estimated using the formula of omega with deflation-corrected estimators of association (RPC/G2/D2) and standardized raw scores (STD).

On the other hand, we can compute the "standard errors" based on the measurement model related to factor models by using Equation (3), that is,

$$\sigma_E = \sqrt{k - \sum_{i=1}^{k} w_i^2} \quad (12)$$

Even if this statistic is based on standardized score variables, it does not produce "standard errors" in the same metric as does the traditional formula (Equation 1), and the outcomes may differ radically from each other (see Figure 7). For example, with 100 items of $w_i = 0.4$ in each, the latter form leads to $\sigma_E = 7.7$ regardless of the reliability. The standard errors by Equation (11) and the "standard errors" by Equation (12) do not speak of the same thing.

5 Conclusion and limitations

5.1 Conclusion in a nutshell

The starting point of this note was the deflation in the reliability estimates. The term error variance related to the general one-factor measurement model ($\sigma_E^2 = k - \sum_{i=1}^{k} w_i^2 = \sum_{i=1}^{k}\left(1 - w_i^2\right)$) is embedded in classical reliability estimators such as coefficient omega and rho (maximal reliability). The traditional measurement model assumes that the weight factor w_i does not include technical or mechanical error. However, previous studies related to deflation in correlation estimates indicate that this is not true. If factor loadings are used as the weight coefficient w_i as they are with the traditional omega and rho, the error variance is always overestimated because factor loading is essentially a product–moment coefficient of correlation between the item and the score, and PMC is one of those estimators of correlation that are especially prone to deflation. Deflation-corrected estimators are obtained when, instead of PMC, some alternative, a better-behaving correlation estimator, such as polychoric correlation, Goodman–Kruskal gamma, or Somers delta, is used in the estimation.

Under the assumption of the one-factor measurement model, the error variance tends to be overestimated as the number of items on the test increases. This can also be derived from the error variance formula. Moreover, the inflation in the traditional error variance tends to grow by the number of items in relation to deflation-corrected estimators of error variance. The technical reason for the phenomenon is that, because of the better behavior of the deflation-corrected estimators of association, they tend to give estimates with a higher magnitude than PMC. The common characteristic of the deflation-corrected correlation estimators is that they give higher estimates than the traditional deflation-prone estimators. Because the error variance is cumulative, the more items we have in the compilation, the more cumulative error we obtain.

An obvious consequence of the inflated error variance is that the standard errors of the measurement are also inflated when the traditional reliability estimators are used. If the deflation-corrected reliability estimators are used, the consequent deflation-corrected standard errors may be notably lower. In the dataset used in the empirical section, the inflation was 35–48%, depending on the benchmarking coefficient of association. However, the deflation may be radically greater in magnitude if the difficulty levels of the items were extreme. This is typical in the tests within educational settings with achievement testing because, usually, the tests include both easy, medium, and difficult items.

5.2 Known limitations and suggestions for further studies

An obvious limitation in the empirical section is that the treatment was based on one real-world dataset with certain limitations: the latent reliability was not controlled, only small sample sizes were used, tests with more than 30 and less than 10 categories in the score were missing, and no tests with extreme difficulty levels or very short tests were included in the dataset used in the simulation. Systematic studies of the phenomenon would enrich our understanding of the nature of inflation in terms of error variance and standard error.

The theoretical basis for the deflation-corrected standard errors is somewhat underdeveloped. The estimators discussed in this article are mainly short-cuts where the poorly behaved *Rit* is replaced by better-behaving coefficients. However, these deflation-corrected estimators are theoretical because no such factor analysis routine currently exists that would yield some of the deflation-corrected estimators of association between an item and a score instead of the traditional product–moment coefficient of correlation (PMC). Of the alternative estimators of association, using R_{PC} and R_{REG} leads to theoretical standard errors because the outcome of deflation-corrected reliability by using R_{PC} or R_{REG} instead of the traditional estimator would lead us to infer something from the *theoretical score* that researchers do not have access to (see Chalmers, 2017; Metsämuuronen, 2022d). The other alternatives suggested by Metsämuuronen (2022a), G, G_2, D, D_2, R_{AC}, and E_{AC}, refer to observed scores and items.

This note is restricted to classical estimators of reliability. Consequently, we do not know much about how applicable the results would be with estimators of reliability within generalizability theory, confirmatory factor analysis (CFA), structural equation modeling (SEM), or IRT and Rasch modeling (see related discussion and literature in Metsämuuronen, 2022d).

Finally, this article aims to explore the reasons for, implications of, and factors related to the empirical finding discussed by Metsämuuronen (2023) that certain types of test settings, common in educational testing settings with widely varying levels of item difficulty, are prone to producing standard errors that may be vastly overestimated. The results of this note enrich our understanding of the factors associated with this phenomenon. Since inflation in the standard error tends to increase with the number of items and the traditional tenet in testing settings is that reliability increases with the number of items, there may be an apparent tension between these tendencies. Since the deflation-corrected estimates of reliability could be used to assess the magnitude

of inflation, it is strongly suggested that the estimates of the DCERs be reported alongside traditional reliability estimates for a more comprehensive evaluation.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: datasets comprising the traditional and deflation-corrected estimates of reliability, the estimates of error variance and standard errors and estimated population variances, and related derivatives and background information of the 1,440 tests are available at http://dx.doi.org/10.13140/RG.2.2.25390.79687 in CSV format and at http://dx.doi.org/10.13140/RG.2.2.33779.40481 in IBM SPSS format.

Ethics statement

Ethical approval was not required for the studies involving humans because according to the law in Finland, the national achievement test results can be used in research purposes by application. The datasets are always anonymized. Part of the simulation dataset is based on old dataset. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares 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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1248770/full#supplementary-material

References

Bentler, P. M. (1968). Alpha-maximized factor analysis (Alphamax): its relation to alpha and canonical factor analysis. *Psychometrika* 33, 335–345. doi: 10.1007/BF02289328

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika* 74, 137–143. doi: 10.1007/s11336-008-9100-1

Chalmers, R. P. (2017). On misconceptions and the limited usefulness of ordinal alpha. *Educ. Psychol. Meas.* 78, 1056–1071. doi: 10.1177/0013164417727036

Cheng, Y., Yuan, K.-H., and Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educ. Psychol. Meas.* 72, 52–67. doi: 10.1177/0013164411407315

Cho, E., and Kim, S. (2015). Cronbach's coefficient alpha: well known but poorly understood. *Organ. Res. Methods* 18, 207–230. doi: 10.1177/1094428114555994

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. *2nd* Edn. USA: Erlbaum.

Cramer, D., and Howitt, D. (2004). The Sage Dictionary of Statistics. A Practical Resource for Students. London: SAGE Publications, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Davenport, E. C., Davison, M. L., Liou, P.-Y., and Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: distinct albeit related concepts. *Educ. Meas. Issues Pract.* 34, 4–9. doi: 10.1111/emip.12095

Davenport, E. C., Davison, M. L., Liou, P.-Y., and Love, Q. U. (2016). Easier said than done: rejoinder on Sijtsma and on Green and Yang. *Educ. Meas. Issues Pract.* 35, 6–10. doi: 10.1111/emip.12106

Dunn, T. J., Baguley, T., and Brunsden, V. (2013). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br. J. Psychol.* 105, 399–412. doi: 10.1111/bjop.12046

Falk, C. F., and Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *J. Pers. Assess.* 93, 445–453. doi: 10.1080/00223891.2011.594129

FINEEC (2018). National assessment of learning outcomes in mathematics at grade 9 in 2002 (Unpublished dataset opened for the re-analysis 18.2.2018). Finnish education evaluation centre.

Foy, P., and LaRoche, S. (2019). Estimating standard errors in the TIMSS 2019 results. Ch. 14 in TIMSS 2019 Technical Report. eds. M. O. Martin, M. von Davier, & I.V.S. Mullis, Available at: https://timssandpirls.bc.edu/timss2019/methods/chapter-14.html (Accessed September 4, 2022).

Gadermann, A. M., Guhn, M., and Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Pract. Assess. Res. Eval.* 17, 1–13. doi: 10.7275/n560-j767

Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Am. Stat. Assoc.* 49, 732–764. doi: 10.1080/01621459.1954.10501231

Green, S. B., and Yang, Y. (2009). Commentary on coefficient alpha: a cautionary tale. *Psychometrika* 74, 121–135. doi: 10.1007/s11336-008-9098-4

Green, S. B., and Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educ. Meas. Issues Pract.* 34, 14–20. doi: 10.1111/emip.12100

Gulliksen, H. (1950). Theory of Mental Tests. New York, NY: Lawrence Erlbaum Associates Publishers.

Guttman, L. (1941). "The qualifications of a class of attributes: a theory and method of scale construction" in The prediction of personal adjustment. Social Science Research Council, Bulletin 48. (ed.) P. Horst, 321–345.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika* 10, 255–282. doi: 10.1007/BF02288892

Heise, D., and Bohrnstedt, G. (1970). Validity, invalidity, and reliability. *Sociol. Methodol.* 2, 104–129. doi: 10.2307/270785

Henrysson, S. (1963). Correction of item–total correlations in item analysis. *Psychometrika* 28, 211–218. doi: 10.1007/BF02289618

Hoekstra, R., Vugteveen, J., Warrens, M. J., and Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *Int. J. Soc. Res. Methodol.* 22, 351–364. doi: 10.1080/13645579.2018.1547523

IBM (2017). IBM SPSS Decision Trees 25. IBM. Available at: https://www.ibm.com/docs/en/SSLVMB_25.0.0/pdf/en/IBM_SPSS_Decision_Trees.pdf (Accessed September 4, 2022).

Jackson, R. W. B., and Ferguson, G. A. (1941). Studies on the reliability of tests. Department of Educational Research, University of Toronto.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Appl. Stat.* 29, 119–127. doi: 10.2307/2986296

Kendall, M. G. (1948). Rank Correlation Methods. *1st* Edn. New York: Charles Griffin & Co Ltd.

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika* 2, 151–160. doi: 10.1007/BF02288391

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika* 62, 245–249. doi: 10.1007/BF02295278

Li, H., Rosenthal, R., and Rubin, D. B. (1996). Reliability of measurement in psychology: from Spearman-Brown to maximal reliability. *Psychol. Methods* 1, 98–107. doi: 10.1037/1082-989X.1.1.98

Livingston, S. A., and Dorans, N. J. (2004). A graphical approach to item analysis. Research Report No. RR-04-10. Educational Testing Service. doi: 10.1002/j.2333-8504.2004.tb01937.x

Lord, F. M. (1958). Some relations between Guttman's principal component scale analysis and other psychometric theory. *Psychometrika* 23, 291–296. doi: 10.1007/BF02289779

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: a test of validity. *J. Mark. Res.* 10, 316–318. doi: 10.1177/002224377301000315

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: additional considerations. *J. Mark. Res.* 15, 304–308. doi: 10.1177/002224377801500219

McDonald, R. P. (1970). Theoretical canonical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *Br. J. Math. Stat. Psychol.* 23, 1–21. doi: 10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). Test Theory: A Unified Treatment. New York: Lawrence Erlbaum Associates.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychol. Methods* 23, 412–433. doi: 10.1037/met0000144

Meade, A. W. (2010). "Restriction of range" in Encyclopedia of Research Design. ed. N. J. Salkind (London: SAGE Publications, Inc.), 1278–1280.

Mendoza, J. L., and Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *J. Educ. Stat.* 12, 282–293. doi: 10.3102/10769986012003282

Metsämuuronen, J. (2016). Item–total correlation as the cause for the underestimation of the alpha estimate for the reliability of the scale. *Glob. J. Res. Analy.* 5, 471–477.

Metsämuuronen, J. (2017). Essentials of Research Methods in Human Sciences. New Delhi: SAGE Publications, Inc.

Metsämuuronen, J. (2020a). Somers' D as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. *Int. J. Educ. Methodol.* 6, 207–221. doi: 10.12973/ijem.6.1.207

Metsämuuronen, J. (2020b). Dimension-corrected Somers' D for the item analysis settings. *Int. J. Educ. Methodol.* 6, 297–317. doi: 10.12973/ijem.6.2.297

Metsämuuronen, J. (2021a). Goodman–Kruskal gamma and dimension-corrected gamma in educational measurement settings. *Int. J. Educ. Methodol.* 7, 95–118. doi: 10.12973/ijem.7.1.95

Metsämuuronen, J. (2021b). Directional nature of Goodman-Kruskal gamma and some consequences. Identity of Goodman-Kruskal gamma and Somers delta, and their connection to Jonckheere-Terpstra test statistic. *Behaviormetrika* 48:283–307. doi: 10.1007/s41237-021-00138-8

Metsämuuronen, J. (2022a). Effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability. Seeking the best options of correlation for deflation-corrected reliability. *Behaviormetrika* 49, 91–130. doi: 10.1007/s41237-022-00158-y

Metsämuuronen, J. (2022b). How to obtain the most error-free estimate of reliability? Eight sources of underestimation of reliability. *Pract. Assess. Res. Eval.* 27:1–27. doi: 10.7275/7nkb-j673

Metsämuuronen, J. (2022c). Deflation-corrected estimators of reliability. *Front. Psychol.* 12:748672. doi: 10.3389/fpsyg.2021.748672

Metsämuuronen, J. (2022d). Typology of deflation-corrected estimators of reliability. *Front. Psychol.* 13:891959. doi: 10.3389/fpsyg.2022.891959

Metsämuuronen, J. (2022e). Attenuation-corrected reliability and some other MEC-corrected estimators of reliability. *Appl. Psychol. Meas.* 46:720–737. doi: 10.1177/01466216221108131

Metsämuuronen, J. (2022f). Reliability for a score compiled from multiple booklets with equated scores. ResearchGate [Preprint]. doi: 10.13140/RG.2.2.20880.69120/2

Metsämuuronen, J. (2022g). Artificial systematic attenuation in eta squared and some related consequences. Attenuation-corrected eta and eta squared, negative values of eta, and their relation to Pearson correlation. *Behaviormetrika*, 50:27–61. doi: 10.1007/s41237-022-00162-2

Metsämuuronen, J. (2022h). Note on the deflation in population variance in the measurement modelling settings. ResearchGate [Preprint]. doi: 10.13140/RG.2.2.31887.87202

Metsämuuronen, J. (2022i). Rank–polyserial correlation: quest for a "missing" coefficient of correlation. *Front. Appl. Math. Stat.* 8:914932. doi: 10.3389/fams.2022.914932

Metsämuuronen, J. (2023). Seeking the real reliability. Why the traditional estimators of reliability usually fail in achievement testing and why the deflation-corrected coefficients could be better options. *Pract. Assess. Res. Eval.* 28:10. doi: 10.7275/pare.1264

Metsämuuronen, J., and Ukkola, A. (2019). Methodological solutions of zero level assessment. Publications 18: 2019. Finnish education evaluation Centre. [in Finnish, abstract in English] Available at: https://www.karvi.fi/sites/default/files/sites/default/files/documents/KARVI_1819.pdf

Moses, T. (2017). A review of developments and applications in item analysis in Advancing Human Assessment. The Methodological, Psychological and Policy Contributions of ETS. (eds.) R. Bennett and M. von Davier (USA: Springer Open), 19–46

Novick, M. R., and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika* 32, 1–13. doi: 10.1007/BF02289400

Olsson, U. (1980). Measuring correlation in ordered two-way contingency tables. *J. Mark. Res.* 17, 391–394. doi: 10.1177/002224378001700315

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 195, 1–47. doi: 10.1098/rsta.1900.0022

Pearson, K. (1903). I. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 200, 1–66. doi: 10.1098/rsta.1903.0001

Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika* 9, 116–139. doi: 10.1093/biomet/9.1-2.116

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence for fixed congeneric components. *Multivar. Behav. Res.* 32, 329–354. doi: 10.1207/s15327906mbr3204_2

Raykov, T., and Marcoulides, G. A. (2017). Thanks coefficient alpha, we still need you! *Educ. Psychol. Meas.* 79, 200–210. doi: 10.1177/0013164417725127

Raykov, T., West, B. T., and Traynor, A. (2015). Evaluation of coefficient alpha for multiple component measuring instruments in complex sample designs. *Struct. Equ. Model.* 22, 429–438. doi: 10.1080/10705511.2014.936081

Sackett, P. R., Lievens, F., Berry, C. M., and Landers, R. N. (2007). A cautionary note on the effect of range restriction on predictor intercorrelations. *J. Appl. Psychol.* 92, 538–544. doi: 10.1037/0021-9010.92.2.538

Sackett, P. R., and Yang, H. (2000). Correction for range restriction: an expanded typology. *J. Appl. Psychol.* 85, 112–118. doi: 10.1037/0021-9010.85.1.112

Salkind, N. J. (Ed.) (2010). Encyclopedia of Research Design. London: SAGE Publications, Inc.

Sawilowsky, S. (2009). New effect size rules of thumb. *J. Mod. Appl. Stat. Methods* 8, 467–474. doi: 10.22237/jmasm/1257035100

Schmidt, F. L., and Hunter, J. E. (2003). "History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975–2001" in Validity Generalization: A Critical Review. ed. K. R. Murphy (USA: Erlbaum), 31–66.

Schmidt, F. L., and Hunter, J. E. (2015). Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. *3rd* Edn: London: SAGE Publications, Inc.

Schmidt, F. L., Shaffer, J. A., and Oh, I.-S. (2008). Increased accuracy for range restriction corrections: implications for the role of personality and general mental ability in job and training performance. *Pers. Psychol.* 61, 827–868. doi: 10.1111/j.1744-6570.2008.00132.x

Schult, J., and Sparfeldt, J. R. (2016). Reliability and validity of PIRLS and TIMSS. *Eur. J. Psychol. Assess.* 34, 258–269. doi: 10.1027/1015-5759/a000338

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 107–120. doi: 10.1007/s11336-008-9101-0

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.* 27, 799–811. doi: 10.2307/2090408

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159

Stouffer, S. A., Princeton, N. J. (Ed.) (1950). "Measurement and prediction" in Studies in Social Psychology in World War II, vol. *IV* (Princeton, N.J: Princeton university press).

Tabachnick, B. G., and Fidell, L. S. (2021). Using Multivariate Statistics. *6th* Edn. India: Pearson Education.

Thompson, G. H. (1940). Weighting for battery reliability and prediction. *Br. J. Math. Stat. Psychol.* 30, 357–360. doi: 10.1111/j.2044-8295.1940.tb00968.x

Trizano-Hermosilla, I., and Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Front. Psychol.* 7:769. doi: 10.3389/fpsyg.2016.00769

Walk, M. J., and Rupp, A. A. (2010). "Pearson product-moment correlation coefficient" in Encyclopedia of Research Design. ed. N. J. Salkind (London: SAGE Publications, Inc.), 1022–1026.

Wherry, R. J., and Taylor, E. K. (1946). The relation of multiserial eta to other measures of correlation. *Psychometrika* 11, 155–161. doi: 10.1007/BF02289296

Yang, H. (2010). "Factor loadings" in Encyclopedia of Research Design. ed. N. J. Salkind (London: SAGE Publications, Inc.), 480–483.

Yang, Y., and Green, S. B. (2011). Coefficient alpha: a reliability coefficient for the 21st century? *J. Psychoeduc. Assess.* 29, 377–392. doi: 10.1177/0734282911406668

Zumbo, B. D., Gadermann, A. M., and Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *J. Mod. Appl. Stat. Methods* 6, 21–29. doi: 10.22237/jmasm/1177992180

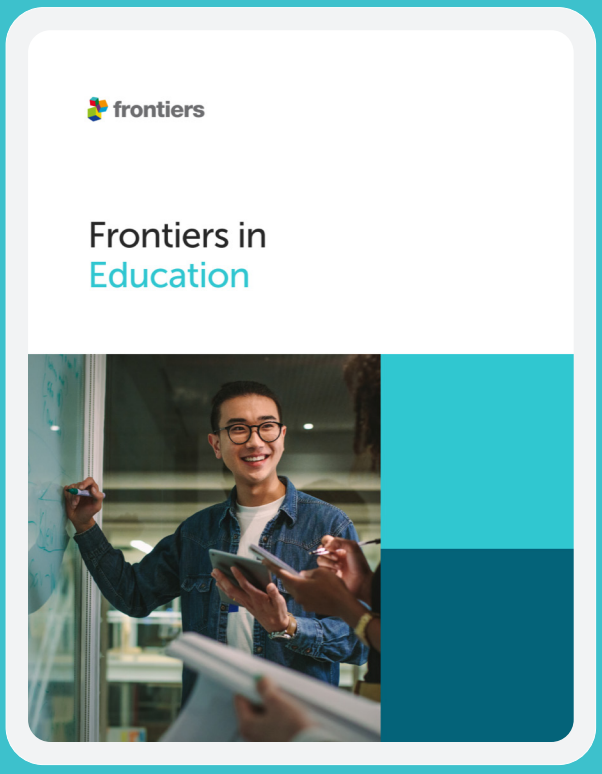
Frontiers in Education

Explores education and its importance for individuals and society

A multidisciplinary journal that explores research-based approaches to education for human development. It focuses on the global challenges and opportunities education faces, ultimately aiming to improve educational outcomes.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact