COGNITIVE DIAGNOSTIC ASSESSMENT FOR LEARNING

EDITED BY: Peida Zhan, Feiming Li and Hong Jiao
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-096-3
DOI 10.3389/978-2-88974-096-3

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

COGNITIVE DIAGNOSTIC ASSESSMENT FOR LEARNING

Topic Editors:
Peida Zhan, Zhejiang Normal University, China
Feiming Li, Zhejiang Normal University, China
Hong Jiao, University of Maryland, College Park, United States

Citation: Zhan, P., Li, F., Jiao, H., eds. (2022). Cognitive Diagnostic Assessment for Learning. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-096-3

Table of Contents

frontiers in Psychology

EDITORIAL
published: 30 November 2021
doi: 10.3389/fpsyg.2021.806636

Editorial: Cognitive Diagnostic Assessment for Learning

Peida Zhan[1,2]*, *Feiming Li*[1,2] *and Hong Jiao*[3]

[1] College of Teacher Education, Zhejiang Normal University, Jinhua, China, [2] Key Laboratory of Intelligent Education Technology and Application of Zhejiang, Zhejiang Normal University, Jinhua, China, [3] Measurement, Statistics and Evaluation, Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, United States

Keywords: cognitive diagnosis, longitudinal cognitive diagnostic assessment, assessment for learning, computerized adaptive test (CAT), cognitive diagnosis model (CDM)

Editorial on the Research Topic:

Cognitive Diagnostic Assessment for Learning

Measuring and improving individual development are actively tackled in psychological, educational, and behavioral sciences. In the past decades, cognitive diagnosis (Leighton and Gierl, 2007), which objectively quantifies students' current learning status and provides diagnostic feedback, has been increasingly needed in different settings to measure and improve individual development.

Although cognitive diagnosis aims to promote student learning based on diagnostic feedback and the corresponding remedial intervention, currently, only a few studies have focused on and evaluated the effectiveness of such feedback or remedial intervention (e.g., Wang et al., 2020; Tang and Zhan, 2021; Wang S. et al.). One of the main reasons is that most cognitive diagnoses adopt a cross-sectional design. This issue may also be reflected in the cognitive diagnosis models (CDMs) or diagnostic classification models (for review, see von Davier and Lee, 2019), the primary tools for data analysis in cognitive diagnosis. Although various CDMs have been proposed, they are only applicable to cross-sectional data analysis (see von Davier and Lee, 2019).

By contrast, longitudinal cognitive diagnosis evaluates students' knowledge and skills and identifies their strengths and weaknesses over a period of time. The data collected from longitudinal learning for diagnosis allow researchers to develop models for learning tracking, which can be used to track individual growth over time and evaluate the effectiveness of feedback. Compared to cross-sectional learning diagnosis, longitudinal cognitive diagnosis may provide an additional perspective to evaluate student learning when aiming to promote student learning.

Currently, longitudinal cognitive diagnosis (e.g., Li et al., 2016; Zhan et al., 2019) mainly stays in the model development stage and lacks practical applications and related research on issues such as missing data, measurement invariance, and linking methods. Moreover, although some longitudinal CDMs have been proposed, these models still have limitations that need further exploration and improvement.

This Research Topic intends to highlight issues, practices, and methodologies dealing with evaluating and improving individual growth in learning, especially using cognitive diagnosis. This Research Topic presents the cutting-edge research related to quantitative methods and applications related to student development (e.g., the development of longitudinal CDMs, the development of longitudinal diagnostic assessments, learning progression, and the impact of sample attrition), novel CDMs for specific test situations (e.g., random guessing behavior, rater effects, and mixed format assessments), theoretical issues in cognitive diagnosis (e.g., parameter estimation, Q-matrix specification, and non-parametric classification method), and application issues in adaptive testings (e.g., automated test assembly, item exposure control, online calibration, and attribute coverage). The contributions of this special topic are elaborated as follows.

Edited and reviewed by:
Alexander Robitzsch, IPN—Leibniz Institute for Science and Mathematics Education, Germany

***Correspondence:**
Peida Zhan
pdzhan@gmail.com

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

Received: *01 November 2021*
Accepted: *15 November 2021*
Published: *30 November 2021*

Citation:
Zhan P, Li F and Jiao H (2021) Editorial: Cognitive Diagnostic Assessment for Learning. Front. Psychol. 12:806636. doi: 10.3389/fpsyg.2021.806636

First, new quantitative methods and applications related to student development were proposed. Wen et al. proposed the HMM/ANN longitudinal CDM, in which the artificial neural network (ANN) was used as the measurement model of the hidden Markov model (HMM) to realize longitudinal tracking of students' cognitive skills. Pan et al. proposed a multivariate longitudinal CDM, in which the log-linear cognitive diagnostic model as the measurement model component evaluates the mastery status of attributes at each measurement occasion, and a generalized multivariate growth curve model that describes the growth of each attribute over time. Lin et al. proposed longitudinal CDMs that incorporate latent growth curve modeling and covariate extensions to measure the growth of skills mastery and evaluate attribute-level intervention effects over time. Tian et al. proposed a longitudinal CDM for hierarchical attributes by imposing model constraints on the transition CDM. In addition, Wang S. et al. reported developing and evaluating a learning program that integrated a longitudinal diagnostic assessment with two different learning interventions to diagnose and improve mental rotation skills. Furthermore, Bai and Wu et al. showed how to use CDMs to explore students' learning progression. Moreover, Pan and Zhan examined the impact of a common type of sample attrition, namely individual-level random attrition, on longitudinal cognitive diagnosis through a simulation study.

Second, novel CDMs for specific test situations were proposed. Choi et al. presented an approach in which the CDM was used with a statistical topic model to analyze item responses in mixed format assessments (i.e., multiple-choice and constructed-response items). Further, to estimate rater effects on constructed response times, Li X. et al. proposed CDMs within the frameworks of facets models and hierarchical rater models, using the log-linear cognitive diagnosis model as a template. Moreover, considering some students may engage in rapid guessing without thoughtful consideration on some items, Hsu et al. proposed a CDM with item response and response time to model rapid guessing behavior and enhance cognitive diagnosis.

Third, some theoretical details of cognitive diagnosis have also been concerned. Zhang et al. proposed a highly effective Pólya-Gamma Gibbs sampling algorithm to estimate the DINA model based on auxiliary variables. Furthermore, Wang W. et al. proposed a semi-supervised learning approach and an optimal design for examinee sampling for Q-matrix specification under the conjunctive and disjunctive model with an independent structure. In addition to parametric models, non-parametric diagnostic methods are also an essential method in cognitive diagnosis. Guo et al. introduced a non-parametric spectral clustering algorithm to cluster students according to their responses.

Fourth, although classification accuracy is critical in cognitive diagnostic computerized adaptive testing (CAT), attention has increasingly shifted to item exposure control to ensure test security and attribute balance/coverage to ensure test fairness. In such cases, Sun et al. developed the binary restrictive threshold method to balance measurement accuracy and item exposure. Wang Y. et al. proposed the attribute discrimination index-based method to balance the attribute coverage. Furthermore, online calibration is a technique to calibrate the parameters of new items in CAT, which seeds new items in answering operation items and estimates the parameters of new items through the response data of examinees on new items. Xiong et al. extended the two most popular calibration methods, one- and multiple EM cycle methods, to the graded response model for polytomous data. Moreover, Li G. et al. explored the automated test assembly in cognitive diagnostic multistage adaptive testing that can be seen as a combination of the paper and pencil-based test and CAT.

Finally, Zhan reviewed the current status and possible future research directions of longitudinal cognitive diagnosis. He pointed out that there are still many issues related to longitudinal cognitive diagnosis worthy of discussion. For example, (a) only binary attributes (e.g., "1" means mastery and "0" means non-mastery) were considered in most current studies. In the future, the polytomous attributes (Karelitz, 2004) or probabilistic attributes (Zhan et al., 2018) can be incorporated into longitudinal CDMs to track students' refined development (e.g., Zhan, 2021); (b) only item response accuracy data were considered in most current studies. In the future, utilizing multimodal data (e.g., item response times and eye-tracking indices) can evaluate the growth of students in multiple aspects, which is conducive to a more comprehensive understanding of the development of students (e.g., Wang et al., 2018); (c) most current studies assumed that attributes are structurally independent. However, when attribute hierarchy (Leighton et al., 2004) exists, the development trajectory of students is not arbitrary and should be developed in such hierarchical order. Therefore, incorporating the attribute hierarchy into current longitudinal CDMs is worth exploring (e.g., Zhan and He, 2021), and (d) adaptive learning and testing system involving longitudinal CDMs is also worthy of further study.

With 19 papers from 62 authors, this topic enhances interdisciplinary research fields such as psychometrics, pedagogy, psychology, statistics, computer science, educational technology, to name a few. The categorization focused on each paper's core contribution though some papers can be cross-classified. The papers' key findings and advancements well-represent the current state-of-the-art in the field of longitudinal cognitive diagnosis in educational and psychological assessments. As topic editors, we are happy to receive such a great collection of papers with various foci and make these publications right when the concept of assessment for/as learning is rapidly gaining popularity. We hope these papers fill some gaps in the literature related to longitudinal cognitive diagnosis modeling and applications. It is expected that the methodological papers will inspire more researchers to explore new frontiers in models and methods for longitudinal cognitive diagnosis; in the meantime, the methodological innovation will guide practitioners to improve their practices.

AUTHOR CONTRIBUTIONS

PZ contributed to manuscript drafting and revising. FL and HJ contributed to manuscript revising. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No. 19YJC190025) and the National Natural Science Foundation of China (Grant No. 31900795).

REFERENCES

Karelitz, T. M. (2004). *Ordered Category Attribute Coding Framework for Cognitive Assessments* (Unpublished doctoral dissertation). Champaign, IL: The University of Illinois at Urbana Champaign.

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

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

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

Tang, F., and Zhan, P. (2021). *Does Diagnostic Feedback Promote Learning? Evidence From a Longitudinal Cognitive Diagnostic Assessment*. AERA Open.

von Davier, M., and Lee, Y.-S. (2019). *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. New York, NY: Springer.

Wang, L., Tang, F., and Zhan, P. (2020). Effect analysis of individualized remedial teaching based on cognitive diagnostic assessment: taking "linear equation with one unknown" as an example. *J. Psychol. Sci.* 43, 1490–1497. doi: 10.16719/j.cnki.1671-6981.20200630

Wang, S., Zhang, S., Douglas, J., and Culpepper, S. A. (2018). Using response times to assess learning progress: a joint model for responses and response times. *Meas. Interdiscipl. Res. Perspect.* 16, 45–58. doi: 10.1080/15366367.2018.1435105

Zhan, P. (2021). Refined learning tracking with a longitudinal probabilistic diagnostic model. *Educ. Meas. Issues Pract.* 40, 44–58. doi: 10.1111/emip.12397

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

Zhan, P., Wang, W.-C., Jiao, H., and Bian, Y. (2018). Probabilistic-input, noisy conjunctive models for cognitive diagnosis. *Front. Psychol.* 9:997. doi: 10.3389/fpsyg.2018.00997

Zhan,. P., and He, K. (2021). A longitudinal diagnostic model with hierarchical learning trajectories. *Educ. Meas. Issues Pract.* 40, 18–30. doi: 10.1111/emip.12422

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 January 2020
doi: 10.3389/fpsyg.2019.03085

Online Calibration of Polytomous Items Under the Graded Response Model

Jianhua Xiong[1,2]*, *Shuliang Ding*[2], *Fen Luo*[1,2] *and Zhaosheng Luo*[1]*

[1] School of Psychology, Jiangxi Normal University, Nanchang, China, [2] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China

Computerized adaptive testing (CAT) is an efficient testing mode, which allows each examinee to answer appropriate items according his or her latent trait level. The implementation of CAT requires a large-scale item pool, and item pool needs to be frequently replenished with new items to ensure test validity and security. Online calibration is a technique to calibrate the parameters of new items in CAT, which seeds new items in the process of answering operational items, and estimates the parameters of new items through the response data of examinees on new items. The most popular estimation methods include one EM cycle method (OEM) and multiple EM cycle method (MEM) under dichotomous item response theory models. This paper extends OEM and MEM to the graded response model (GRM), a popular model for polytomous data with ordered categories. Two simulation studies were carried out to explore online calibration under a variety of conditions, including calibration design, initial item parameter calculation methods, calibration methods, calibration sample size and the number of categories. Results show that the calibration accuracy of new items were acceptable, and which were affected by the interaction of some factors, therefore some conclusions were given.

Keywords: online calibration, computerized adaptive testing, graded response model, squeezing average method, one EM cycle method, multiple EM cycle method

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Yi Zheng,
Arizona State University, United States
Yinhong He,
Nanjing University of Information Science and Technology, China

***Correspondence:**
Jianhua Xiong
002279@jxnu.edu.cn
Zhaosheng Luo
luozs@126.com

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

Received: *03 September 2019*
Accepted: *30 December 2019*
Published: *23 January 2020*

Citation:
Xiong J, Ding S, Luo F and Luo Z (2020) Online Calibration of Polytomous Items Under the Graded Response Model. Front. Psychol. 10:3085. doi: 10.3389/fpsyg.2019.03085

INTRODUCTION

Computerized adaptive testing (CAT), which is considered to be one of the most important applications of item response theory (IRT; Lord, 1980), is a tailored test mode (e.g., Chang and Zhang, 2002; Chang, 2015). The goal of CAT is to construct an optimal test for each examinee (Meijer and Nering, 1999). Compared with the traditional paper-pencil test (PandP), CAT has many advantages such as more flexible testing time, more diverse items, shorter test length, more accurate ability estimation, and more timely score reporting (e.g., Weiss, 1982; Meijer and Nering, 1999; Cheng and Chang, 2009; Wang and Chang, 2011; Wang et al., 2013). Therefore, many large-scale evaluation programs such as the Graduate Management Admission Test (GMAT) and the Armed Services Vocational Aptitude Battery (ASVAB; Sands et al., 1997) adopted the CAT test mode (Chang and Ying, 2009).

The implementation of CAT requires a large-scale item pool, and the maintenance and management of item pool is critical to ensure the validity and security of CAT. After a period of time, some operational items may be no longer suitable for use due to overexposure, obsoleteness, or flaw, thus it is necessary to replace unsuitable items by new ones (Wainer and Mislevy, 1990; Zheng, 2014; Zheng and Chang, 2017). The new items should be precisely calibrated

before being put into the item pool for use formally. Moreover, the calibration accuracy of the new items has great influence on the estimation accuracy of the examinees' latent trait in the ensuing CAT sessions (e.g., van der Linden and Glas, 2000; Chang and Lu, 2010).

Wainer and Mislevy (1990) proposed two strategies for calibrating new items based on CAT in the literature. The first strategy is traditional offline calibration with anchor-item design. Namely, set some anchor items between the new and the operational items, and do equating transformation through the collected responses to ensure the item parameters of the new items and those of the operational items on the same scale. Because the traditional calibration method needs to organize P and P test in advance, there are some shortages, such as the consumption of manpower and material resources, the easy exposure of new items and so on. The second strategy is online calibration, which refers to the process of assigning the new items to examinees during the course of their adaptive tests and then estimating the item parameters of new items based on the collected responses. In the online calibration framework, new items can be embedded inconspicuously within the operational tests, and be pretested and calibrated in the same testing environment as the operational items. Compared with the traditional calibration, online calibration is not only time-saving but also cost-effective. It places new items on the same scale as the operational items without *post hoc* scaling.

Online calibration design and online calibration method are two crucial aspects of online calibration (Chen and Xin, 2014). Online calibration design refers to the way which the new items are assigned to examinees during the CAT process, and collects the responses of the new items. Online calibration design mainly includes two types. One is random design, and the other is adaptive design. Random design randomly selects a new item and then stochastically seeds it in the current examinee's adaptive test (Wainer and Mislevy, 1990). Adaptive design selects the most suitable new item according to some criterion when he or she reaches a seeding location (He and Chen, 2019). The online calibration method uses the responses collected during the online calibration design phase to estimate the item parameters of new items. The most popular estimation methods proposed for online calibration include one EM cycle method (OEM; Wainer and Mislevy, 1990) and multiple EM cycle method (MEM; Ban et al., 2001).

There are many studies on online calibration based on dichotomously scored models (e.g., You et al., 2010; Chen et al., 2012; van der Linden and Ren, 2015; He et al., 2017, 2019). One purpose of modern item response theory research is to exhaust all types of models to cover test data from any "natural" form (van der Linden and Hambleton, 1997). And compared with dichotomously scored items, polytomously scored items have many advantages, such as measuring more complex knowledge structure and providing higher item and test information. Therefore, examinees' ability can be estimated with greater precision by the same number of items, or the same level of precision can be obtained with fewer items. More and more tests involving polytomously scored items have emerged. However, online calibration of polytomously scored model is reported rarely. Zheng (2016) extends the formula, procedure and algorithm of online calibration under dichotomously scored models to the generalized partial credit model (GPCM). The extended formulas and algorithms are studied by simulation method, and some constructive conclusions are obtained. The graded response model (GRM; Samejima, 1969, 1996), like GPCM, is a polytomously scored model. But they have many differences. First, the ideas of model construction are different, GPCM is a division model, that is, the proportion of part to whole. In contrast, GRM is a deviation model, that is, the difference between adjacent categories. Second, the meanings of difficulty parameters in GRM and GPCM models are different, GPCM emphasizes the difficulty of each step on an item, and the difficulty value does not necessarily increase monotonously, GRM emphasizes the difficulty of getting different scores on an item, and the difficulty value increases monotonously. Therefore, it is necessary to discuss online calibration based on GRM, it is of great significance to the expansion of the item pool with GRM items.

The structure of this article is as follows. First, the GRM, an IRT model used in this research is introduced. Second, online calibration method (OEM and MEM method) based on GRM is introduced. Two methods for calculating initial item parameters are given in detail. Third, two simulation studies are designed, and the research results are presented. Fourth, a batch of real data are used to verify the validity of the method. The last part involves conclusions, a supplementary study, discussions, and directions for future research.

METHODOLOGY

The GRM

The GRM is an IRT model suitable for polytomous data with ordered categories. It is an extension of two parameters logistic model (2PLM). In GRM, an examinee's likelihood of responding in a particular response category is obtained by two steps. First, category boundary response functions (CBRFs) are calculated to determine boundary decision probabilities of t response categories for each item. The equation for a CBRF is similar to 2PLM for dichotomous data:

$$p^*_{ijt} = \frac{1}{1 + \exp(-D \cdot a_j(\theta_i - b_{jt}))} \tag{1}$$

In Equation (1), p^*_{ijt} is the probability that an examinee with ability level θ_i will respond positively at the boundary of category t for item j where $t = 1, 2, \cdots f_j$, θ_i represents the ith examinee's ability; a_j represents the item discrimination parameter or slope for item j; b_{jt} represents the item difficulty parameter or category location. Importantly, the values of b_{jt} should satisfy monotonically increasing, that is $b_{j1} < b_{j2} < \cdots b_{jt} < \cdots b_{j,f_j}$.

In the second step of GRM, the probability of responding in a particular category is determined by CBRF, which are derived by subtracting p^*_{ijt} from the following category. The process is illustrated in Equation (2) (adapted from Embretson and Reise, 2000).

$$p_{ijt} = p^*_{ijt} - p^*_{ij,t+1} \tag{2}$$

Further, make the following constraints, $p^*_{ij0} = 1$, namely, the probability of scoring more than 0 must be 1; $P^*_{ij,fj+1} = 0$, that is, the probability of scoring more than the item's full score is naturally 0.

Extend OEM and MEM Methods to GRM

Under the dichotomous model, OEM (Wainer and Mislevy, 1990) and MEM (Ban et al., 2001) are based on the framework of MMLE with the EM algorithm. Their main difference is the number of EM cycles. The OEM method takes just one E step using the posterior distribution of ability, which is estimated based on item responses only from the operational CAT items, and just one M step to estimate the new item parameters, involving response data from only the new items. The MEM method is similar mathematically to the OEM method. The first EM cycle of the MEM method is the same as the OEM method. The parameter estimates of new items obtained from the first EM cycle is regarded as the initial values of the new items for the second EM cycle. However, from the second E step, the MEM method uses item responses on both the operational items and new items to obtain the posterior distribution. For each M step iteration, the item parameter estimates for the operational items are fixed, whereas parameter estimates for the new items are updated until the new item parameter estimates converge. The principles of OEM and MEM under GRM are basically the same as those under the dichotomous model, but there are some differences in implementation details. The details of OEM and MEM implementation under GRM are described below.

OEM

OEM has only one EM cycle. For each examinee $i = 1, 2, \cdots N_j$ who takes item j, q_i denotes his/her responses to the operational items, η_{op} is a vector of the known item parameters of the operational items. The E-step of the OEM method marginalizes the log-likelihood of new item j using q_i and η_{op}. Based on the common assumption that examinees are independent from each other, the log-likelihood of item j from the N_j examinees are summed up as the final marginalized log-likelihood of item j to be taken to the subsequent M-step. The M-step seeks the item parameter vector $\hat{\eta}_j$ that maximizes the final marginalized log-likelihood of item j.

These two steps are adapted from described in Muraki (1990) of item parameter estimation. The difference between the algorithms here for online calibration and Muraki's algorithm is in the computation of the two quantities: $\bar{r}_{jtk}$ and $\bar{f}_k$, where $\bar{r}_{jtk}$ is the temporary expected frequency of the tth category response of item j at the kth quadrature point; $\bar{f}_k$ is the temporary expected sample size at quadrature point k. In his original EM algorithm, every examinee receives the same set of items. In the online calibration setting, as described earlier in this article, each new item j is administered to a different sample of examinees; and each examinee who takes new item j takes a different set of operational items. To adapt these variations, the formulae for $\bar{r}_{jtk}$ and $\bar{f}_k$ in the EM algorithm are modified into as follows:

$$\bar{r}_{jtk} = \sum_{i=1}^{N_j} u_{ijt} h(X_k) \tag{3}$$

$$\bar{f}_k = \sum_{i=1}^{N_j} h(X_k) \tag{4}$$

$$h(X_k) = \frac{L_i(X_k)A(X_k)}{\sum_{k=1}^{K} L_i(X_k)A(X_k)} \tag{5}$$

$$L_i(X_k) = \prod_{h=1}^{m_i} \prod_{t=1}^{f_h} [p_{ht}(X_k)]^{q_{iht}} \tag{6}$$

Where $i = 1, 2, \cdots, N_j$ denote the N_j examinees who received new item j; X_k is the quadrature point; $A(X_k)$ is the corresponding weight, which is approximately the standard normal probability density at the point X_k, assuming there are K quadrature points, such that $\sum_{k=1}^{K} A(X_k) = 1$.$U_{ijt}$ is an indicator variable expressed in a binary format; $U_{ijt} = 1$ represents examinee i scored exactly t on new item j; otherwise $U_{ijt} = 0$.$L_i(X_k)$ is the likelihood of examinee i's response to all operational items given quadrature point X_k; h denotes the hth operational items answered by examinee i; f_h is the number of categories of hth operational item, $p_{ht}(X_k)$ is the probability of correct response to the tth category of item h at given quadrature point X_k, q_{iht} is an indicator variable too, which denotes the examinee i's responses to operational item h in a binary format to category t.

With the one EM cycle in the OEM method, the revised $\bar{r}_{jtk}$ and $\bar{f}_k$ are inserted into the Newton-Raphon iteration in the single EM cycle to get a set of parameter estimates.

MEM

The MEM method allows multiple EM cycles. The first cycle is the same as OEM. Beginning with the second cycle, response data from both the operational items and the new items are used to update the posterior ability distribution in the E-step. Specifically, the only change in computation from OEM is that beginning with the second cycle of MEM, $L_i(X_k)$ is replaced by:

$$L_i(X_k) = \left(\prod_{h=1}^{m_i} \prod_{t=1}^{f_h} [p_{ht}(X_k)]^{q_{iht}} \right) \left(\prod_{t=1}^{f_j} [p_{jt}(X_k)]^{x_{ijt}} \right) \tag{7}$$

Where x_{ijt} denotes examinee i's response to new item j in the binary format for category t.

The E-step and the M-step iterate until a certain convergence criterion is met, for example the maximum absolute change in the item parameters between two consecutive EM cycles are less than a small threshold.

Calculate the Initial Value of OEM and MEM

OEM and MEM are both iterative algorithms, the initial item parameters have a great influence on the calibration accuracy. However, there are few reports on the calculation of initial iteration values. In the dichotomous model, a squeezing average method is given to compute the initial value of difficulty parameter and a biserial correlation method

is used to compute the initial value of discrimination parameter (You et al., 2010). Under GRM, Xiong et al. (2018) also proposed a methods for calculating the initial item parameters, namely, deleting extremum and squeezing average method and polyserial correlation coefficient method. They had better calibration results under the experimental conditions given in these literatures (You et al., 2010; Xiong et al., 2018). Their theories and implementation details are as follows.

Deleting Extremum and Squeezing Average Method

Under the dichotomous model, according to the characteristics of the item response curve, the correctness of the examinee's response to a certain item is related with the ratio of his/her ability to the difficulty parameter of the item. When the ratio is more than 1, the correct response probability is high; otherwise, the correct response probability is low. For the one-parameter logistic model (1PLM), when the examinee's ability value is equal to the difficulty of one item, his/her correct response probability on the item is 0.5. Therefore, as long as the number of responses is sufficiently large for one item, there must be some examinees whose abilities approach to the difficulty parameter of the item (You et al., 2010), and the abilities of these examinees can be used to estimate the difficulty parameter of the item. The method is called "squeezing average method." Under GRM, for a certain item, the difficulty of getting a high score is higher than that of getting a low score, so the initial parameters of different category can be squeezed out by the ability of the examinees who get the adjacent scores.

The steps of the squeezing average method (You et al., 2010) are described as follows. At first, put the ability values of all examinees who answered correctly on item j into the set correct(j), then sort correct(j) in ascending order; and put the ability values of all examinees who answered incorrectly on item j into the set wrong(j), then sort wrong(j) in descending order. Second, use the low part of correct(j) and the high part of wrong(j) to squeeze the difficulty of the item j. Because not all examinees' abilities in correct(j) or wrong(j) are used for squeezing, it is worth exploring how many examinees' abilities are used to squeeze item difficulty parameter. An empirical value of 18 is suggested by You et al. (2010).

Under the GRM model, GRM has multiple difficulty parameters, so multiple squeezing processes are required. For example, for the initial difficulty parameter of the tth category of the new item j, the ability of the examinees who scored t and t+1 on the item are used to squeeze. Pilot studies have shown that the result is unstable if the sample size for squeezing is still set to 18. A more flexible range of sample size for squeezing method, named "deleting extremum and squeezing average method," is proposed based on the original squeezing average method (Xiong et al., 2018). The ability of examinees who got t score in item j are put into one set, there are f_j sets for item j, and each set is sorted in ascending order by ability value. Then the top 5% and the bottom 5% of each set are deleted. The "deleting extremum and squeezing average method" can be formally expressed as:

$$b_{jt} = \left\{ \left(mean \left(\sum_{i=c(j,t)*5\%}^{c(j,t)*95\%} cap(t,i,j) \right) + mean \left(\sum_{i=c(j,t+1)*5\%}^{c(j,t+1)*95\%} cap(t+1,i,j) \right) \right) \right\} /2 \quad (8)$$

Where $cap(t, i, j)$ is the ability of the ith examinee's who got t score on item j, $c(j, t)$ is the number of examinees who scored t on item j, $cap(t + 1, i, j)$ and $c(j, t + 1)$ have the similar meaning.

In actual life, the evaluation of a contestant is generally based on a set of scores given by the experts. The highest and lowest score are removed, and then the average is taken, deleting extremum and squeezing average method takes this idea. The practice of choosing 5% as the extreme value in Equation (8) is derived from the way to obtain the initial value of the guess parameter under the three-parameter logistic model (3PLM). Pilot study also showed that the value had better results. It's easy to implement and guarantee the accuracy of parameter estimation.

Polyserial Correlation Coefficient Method

The polyserial correlation coefficient method is a common statistical method (Olsson et al., 1982), which is used to initialize the discrimination parameter and difficulty parameter of new items based on the examinee's responses. This method can be depicted by the following steps:

Step 1: For each new item, the pass rate of each category is calculated by using the responses of the examinees to the item, that is, $P^*_{jt} = n_{jt}/N$, where N is the total number of examinees, and n_{jt} is the number of examinees whose scores on the new item j are not lower than t.

Step 2: Convert P^*_{jt} to standard normal fraction Z_{jt}; then calculate the corresponding normal density function value $h(Z_{jt})$. The specific calculation formula is as follows:

$$y_j = -\ln(4P^*_{jt}(1 - P^*_{jt})) \quad (9)$$

$$Z_{jt} = sign(P^*_{jt} - \frac{1}{2}) \sqrt{y_j(2.0611786 - \frac{5.7262204}{y_j + 11.640595})} \quad (10)$$

$$h(Z_{jt}) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} Z^2_{jt}) \quad (11)$$

Step 3: Calculate the standard deviation (σ_j) of the score on the new item j, and the correlation coefficient (r_j) between the score of the new item j and the total score; then the point polyserial correlation coefficient is obtained via the following equation:

$$rpp_j = r_j * \sigma_j / \sum_{t=1}^{f_j} h(Z_{jt}) \quad (12)$$

Step 4: Transform the point polyserial correlation coefficient into polyserial correlation coefficient, that is:

$$rp_j = rpp_j * \sigma_j / \sum_{t=1}^{f_j} h(Z_{jt}) \tag{13}$$

Step 5: Calculate the initial value of the discrimination and difficulty of the new item j; the formula is:

$$a_j = rp_j / \sqrt{1 - rp_j^2} \quad b_{jt} = -Z_{j,t-1} / rp_j \tag{14}$$

Two methods of calculating the initial parameters of new items are given. The first method is called polyserial-initial method, abbreviated as Poly-Ini method, with this method, both a-parameter and b-parameters are calculated by polyserial correlation coefficient; the second method is called polyserial-squeezing-initial method, abbreviated as Poly-Sq-Ini method, with this method, a-parameter is calculated by polyserial correlation coefficient method and b-parameters are obtained by deleting extremum and squeezing average method.

SIMULATION STUDY

Research Objectives

Two simulation studies were conducted using programs written in Python 3.7. The program simulated the entire calibration workflow including the implementation of CAT and the calibration of the new items, and replicated 100 times in each circumstance. The main purpose of Study 1 is to explore the calibration results under a set of conditions fully crossed by two online calibration design methods (random design, adaptive design), two initial item parameter calculation methods (Poly-Ini method, Poly-Sq-Ini method), two calibration methods (OEM, MEM). There are 8 combinations, each combination takes 3-categories as an example.

The main purpose of Study 2 is to explore the calibration results under different calibration sample size and different number of categories. Two factors were manipulated: calibration sample size (300, 400, 500, 600, and 700) and the number of categories of new items (2, 3, 4, and 5). There are 20 combinations. Random design, Poly-Sq-Ini method and MEM are adopted in each combinations.

Generation of Items and Examinees

Suppose there are 1000 operational items with various categories (2–5 categories) in the CAT item pool, item parameters were randomly generated under GRM from the following distributions:

$$a_j \sim \log normal(0,1), \; b_{jt} \sim normal(0,1)$$

$j = 1, 2, \cdots 1000$, $t = 1, 2, \cdots f_j f_j$ is the number of categories. In addition, the generated a-parameter was truncated between 0.2 and 2.5, b-parameter was truncated between −3 and 3, and $b_{j1} < b_{j2} < \cdots b_{jt} < \cdots b_{j,f_j}$ in this paper.

A total number of 20 new items were generated in the same manner with the operational items.

3,000 examinees' ability values (θ) were randomly drawn from the standard normal distribution $\theta \sim normal(0,1)$, and θ was truncated between −3 and 3 too.

Simulation Details

The CAT test length is fixed 25 items, including 20 operational items and 5 new items. During the CAT test, the maximum Fisher information method (MFI; Lord, 1980) was chosen as the operational item selection method for its advantage of high accuracy. The Fisher information of an examinee i on a GRM item j was formulated as below:

$$I_j(\theta_i) = a_j^2 \sum_{t=1}^{f_j} p_{ijt}(1 - p_{ijt}^* - p_{ij,t+1}^*)^2 \tag{15}$$

During operational item selection, provisional θ estimates were used to replace the θ's in the formulae. After each operational item is administered, the examinee ability parameter $\hat{\theta}$ was updated by expected a posteriori (EAP) method (Baker and Kim, 2004).

The number of examinees who answer each new item must be sufficiently large to provide accurate item parameter estimates without placing an undue burden on examinees (Wainer and Mislevy, 1990). This paper investigates one sample size (3,000) and assumes that each examinee answers 5 new items, thus the number of examinees who answer each new item is approximately 750 [(3,000×5)/20] on average as in previous studies (e.g., Chen et al., 2012; Chen and Wang, 2016; He et al., 2017). In Study 1, the number of examinees to each new item is set 700. In addition, calibration accuracy may be affected by the calibration samples per new item. In Study 2, the number of examinees to each new item is set as 300, 400, 500, 600, 700.

In study 1, random design and adaptive design are considered. There are some researches adopted random design to assign the new items to the examinees during CAT due to its convenient implementation and acceptable calibration precision (e.g., Wainer and Mislevy, 1990; Ban et al., 2001; Chen et al., 2012; He et al., 2017). And match-b selection method (MATB) is selected for adaptive design in this study, which matches the mean of b-parameters with the provisional $\hat{\theta}$ of examinee (Zheng, 2016). Every time an examinee reaches a seeding location, the distance between his or her current $\hat{\theta}$ and the mean of provisional b-parameters was computed for each new item, and the item with the shortest absolute distance was selected. In order to obtain the initial parameter of new items, this study uses a data-based method, that is, the new items are first randomly assigned to a sub-group of examinees and are pre-estimated item parameters, then for the remaining examinees, these new items are selected adaptively according to their initial parameters to fit the examinees' current ability. The item parameters of each new item are updated each time they receive a fixed number of new responses (van der Linden and Ren, 2015; Zheng, 2016; He et al., 2019), in this study, the fixed number of new responses was set 20. The proportion of the sample size used in two different phases was specified as 1:1 in this study.

Evaluation Criteria

The calibration accuracy of the new items was evaluated by root mean square error (RMSE) and bias. They quantify the recovery between the estimated and true parameter values, and the calculation formulas based on vector are as follows (He and Chen, 2019; He et al., 2019):

$$RMSE_x = \sqrt{(\sum_{r=1}^{R}\sum_{j=1}^{M}(\hat{x}_j^{(r)} - x_j^{(r)})^2)/(R \times M)} \quad (16)$$

$$bias_x = (\sum_{r=1}^{R}\sum_{j=1}^{M}(\hat{x}_j^{(r)} - x_j^{(r)}))/(R \times M) \quad (17)$$

Where x denotes the specific element in the item parameter vector, such as a-parameter, b_{fj}-parameters, R and M denotes replications and the number of new items respectively.

In order to evaluate the overall recovery of b-parameters under different categories, the average RMSE and bias of b-parameters, named mean(b), are defined as follows:

$$RMSE_{mean(b)} = \sqrt{(\sum_{r=1}^{R}\sum_{j=1}^{M}\sum_{t}^{f_j}(\hat{b}_{jt}^{(r)} - b_{jt}^{(r)})^2)/(R \times M \times f_j)} \quad (18)$$

$$bias_{mean(b)} = (\sum_{r=1}^{R}\sum_{j=1}^{M}\sum_{t}^{f_j}(\hat{b}_{jt}^{(r)} - b_{jt}^{(r)}))/(R \times M \times f_j) \quad (19)$$

Smaller RMSE indicates higher calibration precision. If bias is close to 0, the calibration could be regarded as unbiased.

Results and Conclusion

Study 1

The results of Study 1 are shown in **Tables 1**, **2** and **Figure 1**, using two separate criteria (RMSE and bias) to evaluate the calibration results under different combinations. As can be seen from **Tables 1**, **2** and **Figure 1**, (1) the RMSE values obtained by the combination of random design, Poly-Sq-Ini method and MEM (the combination denoted by C2) were the smallest, and the bias obtained by C2 also had better performance, although not always the best. Which provided the basis for the simulation design of Study 2. (2) The calculation of initial item parameters had a great influence on the calibration results, Poly-Sq-Ini method had better performance under most experimental combinations, the bias had the same trend as RMSE, which showed that the Poly-Sq-Ini method is a feasible method. (3) Comparing OEM and MEM, when adaptive design was adopted, OEM and MEM generated quite comparable RMSE and bias values, when random design was adopted, there are two aspects, MEM was more accurate than OEM if Poly-Sq-Ini method was adopted to compute initial item parameters, otherwise OEM was more accurate than MEM. (4) Comparing random design and adaptive design, the RMSE of b-parameters generated by random design were smaller than those by adaptive design, although the a-parameters generated by random design were not absolutely superior, the most accurate a-parameters still came from random design. The result seems counter-intuitive, one possible explanation for this result is that the simulated examinee's ability distribution is normal, random design leads to an approximately normal distribution of ability for each new item. For adaptive design, the distributions of ability received by each new item may be skewed (Zheng, 2016). The other possible explanation is that the proportion of the sample size used in random phase and adaptive phase would affect the calibration results (Chen et al., 2012).

TABLE 1 | RMSE under different combinations.

Calibration design	Method of calculating initial item parameters	Calibration method	RMSE			
			a	*b1*	*b2*	*b3*
Random	Poly-Sq-Ini	OEM	0.2047	0.2696	0.1567	0.2377
		MEM	0.2022	0.1705	0.1522	0.2009
	Poly-Ini	OEM	0.2892	0.1789	0.1705	0.2306
		MEM	0.2632	0.2142	0.1847	0.2595
Adaptive	Poly-Sq-Ini	OEM	0.2266	0.2651	0.2108	0.2501
		MEM	0.2259	0.2700	0.2101	0.2433
	Poly-Ini	OEM	0.2324	0.3106	0.2005	0.3179
		MEM	0.2324	0.3116	0.2070	0.3231

TABLE 2 | Bias under different combinations.

Calibration design	Method of calculating initial item parameters	Calibration method	bias			
			a	*b1*	*b2*	*b3*
Random	Poly-Sq-Ini	OEM	0.1258	−0.1310	−0.0549	0.0261
		MEM	0.0483	−0.0423	−0.0422	−0.0398
	Poly-Ini	OEM	0.2380	0.0727	−0.0367	−0.1391
		MEM	0.2163	0.1065	−0.0292	−0.1589
Adaptive	Poly-Sq-Ini	OEM	0.0286	0.0777	0.0099	−0.0482
		MEM	0.0296	0.0783	0.0126	−0.0472
	Poly-Ini	OEM	0.1744	0.2182	0.0120	−0.1887
		MEM	0.1751	0.2159	0.0056	−0.1875

Study 2

The results of Study 2 are shown in **Tables 3**, **4** and **Figures 2**–**5**. As can be seen from **Table 3** and **Figure 2**, under various categories, with the increase of calibration sample size, the RMSE of b-parameters were decreasing, but the decline extent was decreasing also. While the calibration sample size had little effect on the RMSE of a-parameters, even under 2-categories and 5-categories, the RMSE increases with the increase of sample size. In addition, it was an interesting observation, the RMSE of b-parameters under different category of the same item were different. In general, the RMSE of the middle category were smaller, while the RMSE of the beginning and ending category

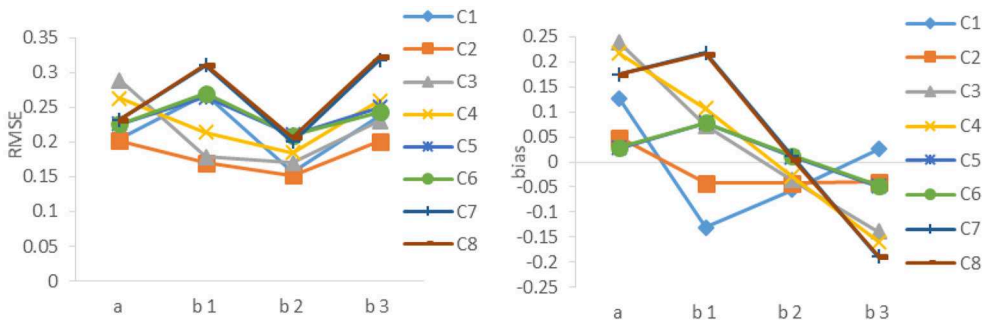


FIGURE 1 | RMSE and bias of *a*- parameter and *b*-parameters under different combinations. C1 denotes the combination of Random, Poly-Sq-Ini and OEM; C2 denotes the combination of Random, Poly-Sq-Ini and MEM; C3 denotes the combination of Random, Poly-Ini and OEM; C4 denotes the combination of Random, Poly-Ini and MEM; C5 denotes the combination of Adaptive, Poly-Sq-Ini and OEM; C6 denotes the combination of Adaptive, Poly-Sq-Ini and MEM; C7 denotes the combination of Adaptive, Poly-Ini and OEM; C8 denotes the combination of Adaptive, Poly-Ini and MEM.

TABLE 3 | RMSE of different calibration sample size under different categories.

Categories	RMSE	Calibration sample size				
		300	400	500	600	700
$f = 2$	a	0.2730	0.2716	0.2683	0.2656	0.2722
	b_1	0.2495	0.2259	0.2216	0.2078	0.2060
	b_2	0.2876	0.2660	0.2602	0.2554	0.2470
	Mean(b)	0.2706	0.2481	0.2427	0.2338	0.2286
$f = 3$	a	0.2189	0.2141	0.2119	0.2074	0.2033
	b_1	0.2413	0.2237	0.1954	0.1919	0.1865
	b_2	0.2127	0.1827	0.1723	0.1673	0.1568
	b_3	0.2674	0.2395	0.2270	0.2249	0.2156
	Mean(b)	0.2439	0.2187	0.2014	0.1993	0.1899
$f = 4$	a	0.2166	0.2150	0.2138	0.2149	0.2081
	b_1	0.2989	0.2866	0.2599	0.2458	0.2262
	b_2	0.2232	0.1968	0.1760	0.1634	0.1577
	b_3	0.2357	0.2016	0.1908	0.1610	0.1659
	b_4	0.2996	0.2611	0.2564	0.2337	0.2294
	Mean(b)	0.2722	0.2432	0.2345	0.2098	0.2007
$f = 5$	a	0.2340	0.2407	0.2353	0.2301	0.2208
	b_1	0.2837	0.2616	0.2604	0.2503	0.2491
	b_2	0.1929	0.1706	0.1662	0.1583	0.1511
	b_3	0.1693	0.1451	0.1419	0.1346	0.1210
	b_4	0.1950	0.1743	0.1633	0.1600	0.1462
	b_5	0.2672	0.2565	0.2368	0.2356	0.2257
	Mean(b)	0.2284	0.2095	0.2044	0.1976	0.1873

TABLE 4 | Bias of different calibration sample size under different categories.

Categories	Bias	Calibration sample size				
		300	400	500	600	700
$f = 2$	a	0.1517	0.1561	0.1488	0.1564	0.1611
	b_1	−0.0231	−0.0193	−0.0289	−0.0253	−0.0336
	b_2	−0.0976	−0.0912	−0.0979	−0.0945	−0.1047
	Mean(b)	−0.0603	−0.0553	−0.0634	−0.0599	−0.0692
$f = 3$	a	0.0479	0.0415	0.0546	0.0500	0.0398
	b_1	−0.0451	−0.0479	−0.0424	−0.0457	−0.0589
	b_2	−0.046	−0.0365	−0.0395	−0.0403	−0.0502
	b_3	−0.0398	−0.0385	−0.0502	−0.0435	−0.0478
	Mean(b)	−0.0365	−0.0409	−0.0440	−0.0432	−0.0523
$f = 4$	a	−0.0491	−0.0445	−0.0602	−0.0477	−0.0449
	b_1	−0.086	−0.0829	−0.089	−0.1059	−0.0957
	b_2	−0.0491	−0.0354	−0.0444	−0.0544	−0.0451
	b_3	−0.0298	−0.0186	−0.0227	−0.0238	−0.0115
	b_4	0.0032	0.0132	0.0239	0.0204	0.0347
	Mean(b)	−0.0404	−0.0309	−0.0330	−0.0409	−0.0294
$f = 5$	a	−0.1217	−0.1305	−0.1289	−0.1232	−0.1199
	b_1	−0.1567	−0.1473	−0.1699	−0.1519	−0.1568
	b_2	−0.0752	−0.0662	−0.0789	−0.0665	−0.0737
	b_3	−0.0238	−0.0155	−0.0239	−0.0126	−0.0211
	b_4	0.0192	0.0273	0.0294	0.0390	0.0315
	b_5	0.0817	0.0990	0.1018	0.1168	0.1031
	Mean(b)	−0.0309	−0.0205	−0.0283	−0.0150	−0.0233

were larger. The possible explanation for this result is that the *b*-parameters in GRM were monotonically increasing, and most of the examinees' scores were concentrated on the middle category. Thus there were relatively few examinees with the lowest score and the highest score, and the sample size would affect the estimation accuracy of new items.

As can be seen from **Table 3** and **Figure 3**, the RMSEs of *a*-parameter under 3-categories and 4-categories did not show noticeable difference under the same calibration sample size, and they were noticeably smaller than those under 2-categories and 5-categories, while the mean(b) of *b*-parameters under 3-categories and 5-categories had similar RMSE values under the same calibration sample size, and they were smaller than those under 2-categories and 4-categories.

It can be seen from **Table 4** and **Figures 4**, **5**, the bias of new items had the same trend as the RMSE, The

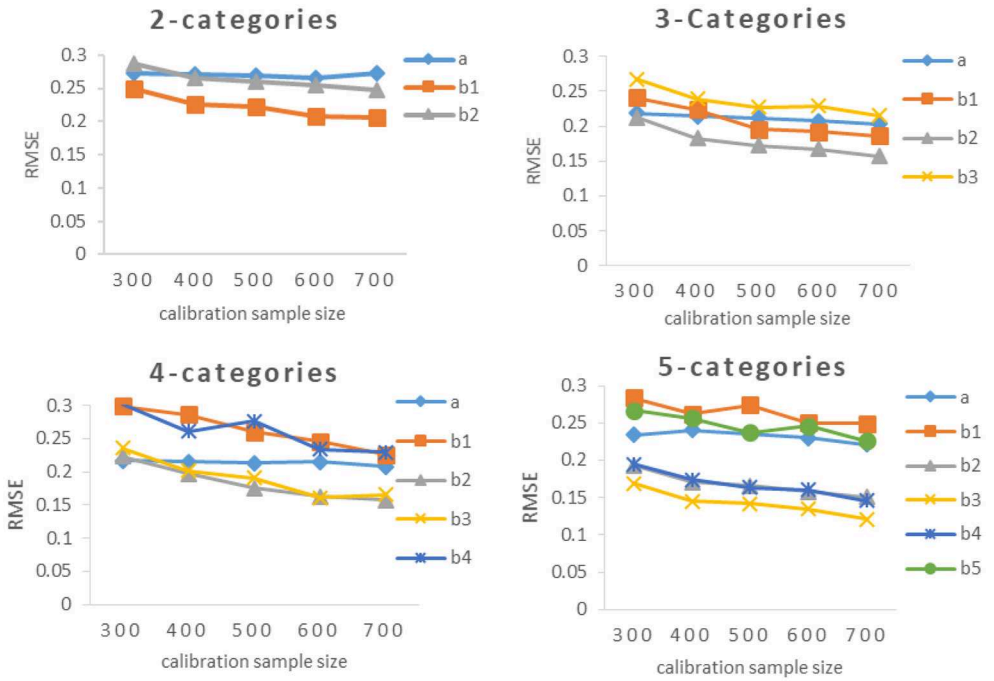


FIGURE 2 | RMSE of *a*- parameter and *b*-parameters under different categories.

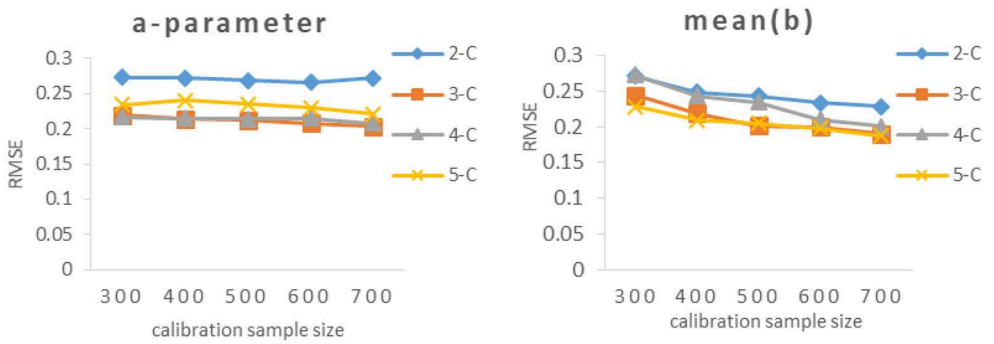


FIGURE 3 | RMSE of different calibration sample size under different categories. 2-C, 2-categories; 3-C, 3-categories; 4-C, 4-categories; 5-C, 5-categories. **Figure 5** also has the same definition.

smaller the value of RMSE, the closer the value of bias was to 0.

EMPIRICAL STUDY

In this paper, an online calibration method based on GRM is proposed, which has a good performance in simulation study. What is the performance on real data? Because the construction of the real CAT item pool is expensive, it is difficult to organize and arrange large-scale CAT tests also. This study used the response data of 500 examinees on 10 polytomous items (3-categories) in HSK4 (Chinese proficiency test) to conduct an empirical study. Detailed steps are as follows.

Step 1: 500 examinees were randomly divided into two parts. One was the training set, including the response data of 300 examinees. The other was the testing set, including the response data of 200 examinees.

Step 2: The ability parameters of examinees and item parameters are estimated through the training set, then the estimated item parameters are taken as the true parameters.

Step 3: For the testing set, the K-fold cross validation method (Tan et al., 2014) is used to simulate and generate the operational items and new items in CAT. In this study, leave-one-out approach was used, that is, each test chose one as new item, and the remaining nine items were as operation items.

Step 4: According to the responses of 200 examinees on 9 operational items and the true values of the corresponding

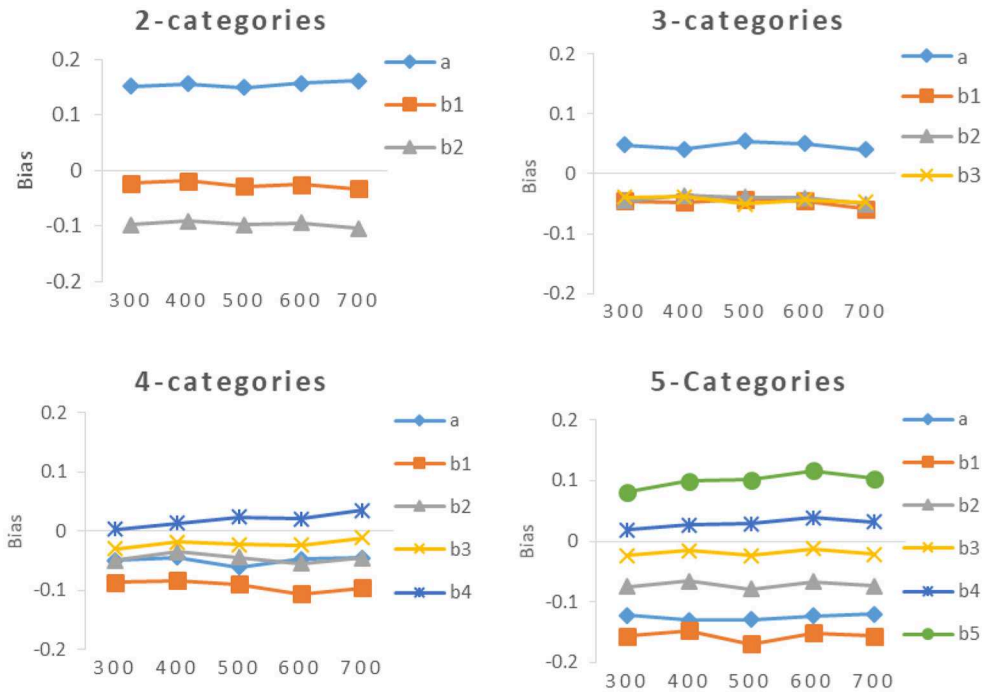


FIGURE 4 | Bias of *a*- parameter and *b*-parameters under different categories.

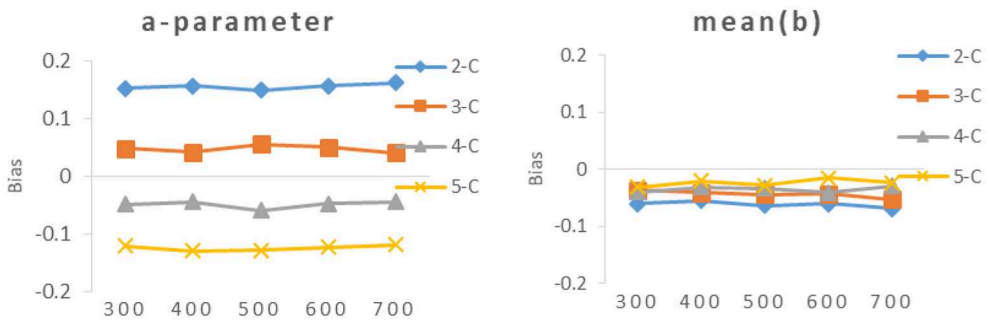


FIGURE 5 | Bias of different calibration sample size under different categories.

item parameters, the ability values of 200 examinees were estimated.

Step 5: According to the examinee's ability values obtained in step 4 and their responses to the new item, the parameters of the new item were estimated by the new method proposed in this study.

Step 6: Each time a different item was selected as the new item, and then the work in step 3∼5 was repeated so that the estimated parameters of each new item could be obtained. Then the RMSE between the estimated parameters and the true parameters were calculated.

Because of the limited real data, this study only analyzed the calibrated sample of 200. The results of the analysis were as follows:$RMSE_a = 0.4067$, $RMSE_{b1} = 0.4778$, $RMSE_{b2} = 0.3218$, $RMSE_{b3} = 0.3029$.

DISCUSSION AND FUTURE DIRECTIONS

This research extended OEM and MEM to GRM for online calibration, detailed description of algorithms were given in the article. While online calibration is a complex process, there are many factors affecting the calibration accuracy. In order to make online calibration efficient and practicable under GRM, various factors should be explored clearly. Two simulation studies were conducted to investigate the calibration results under various conditions. The results showed: (1) both OEM and MEM were able to generate reasonably new item parameters with 700 examinees per item, and each has its own merits. (2) The Poly-Sq-Ini method had better performance than Poly-Ini method under most experimental conditions. (3) Compared to the random calibration design,

the adaptive calibration design do not improve the calibration accuracy in most conditions. (4) The calibration sample size had an effect on the calibration accuracy. In most conditions, the calibration accuracy increases with the increase of sample size. (5) The number of categories of new items also affected the calibration results, the calibration accuracy of 3-categories items was higher than that of 2-categories, and so on.

In addition, a supplementary study was conducted to investigate the calibration accuracy of GRM online calibration under different CAT scenarios. Eight CAT scenarios, which were fully crossed by sample sizes (2,000 and 3,000) and test lengths (variable-length, fixed-length with 10, 20, and 30 respectively), were investigated. The ability estimation results of CAT and the calibration results of new items under various CAT scenarios were listed in **Tables A1**–**A3**. As can be seen from **Table A1**, for the fixed-length CAT, the estimation accuracy of ability increased with the increase of test length under the same sample size. The RMSE value of variable-length CAT was close to that of test length 10 in fixed-length CAT, which indicated that the test length was about 10 under specified cumulative information. All ability bias values in all CAT scenarios were very close to 0. It showed that the simulated CAT can provide accurate ability estimates for the examinees. As can be seen from **Tables A2**, **A3**, (1) the calibration accuracy was acceptable in various CAT scenarios, which showed the robustness of online calibration method under GRM. (2) The estimation accuracy of ability had an effect on the calibration accuracy, but the effect was not monotonous, and there was fluctuation. (3) Under two different sample sizes, the calibration accuracy is higher when the test length is 20.

Several future directions for research can be identified. First, in this paper, the *b*-parameters are randomly selected from the normal distribution and then sort in ascending. The true values of *b*-parameters of new items are random, the following scenarios are possible, such as the *b*-parameters under all categories of an item are less than 0, or are greater than 0, and the difference between adjacent categories is very large or so small. Different scenarios may lead to different calibration results, online calibration based on deliberately designed true parameters of new items is the next research content.

Second, in this paper, only the match-*b* method is considered in the adaptive design, other adaptive design methods are not discussed. There are some adaptive calibration design that practicable and perform well under dichotomously scored models (He and Chen, 2019; He et al., 2019). How to extend these adaptive designs to GRM, and whether it will get the same conclusion as dichotomously scored models are the directions of future research.

Third, the number of categories discussed in this paper was up to 5, which means that the new items can be 2, 3, 4, and 5 categories. If there are more than 5-categories items, whether the new online calibration method is still valid is worthy of further study.

Fourth, there is an interesting phenomenon in the bias of the 5-categories condition. The lower *b*-parameters (*b*1, *b*2) have negative bias, and the higher *b*-parameters (*b*4, *b*5) have positive bias. Does it have anything to do with the calibration methods. Other calibration methods will be extended to GRM in further studies, and observe whether similar phenomenon will also occur. So as to investigate whether the phenomenon It is related to the calibration method, whether it is related to the number of categories of new items, or other factors.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SD, ZL, and JX designed experiments. JX and FL carried out experiments. JX analyzed experimental results and wrote the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61967009, 31360237, 61877031, and 31660279), Jiangxi Education Science Foundation (GJJ160282), Postgraduate Innovation Fund Project of Jiangxi Normal University (YC2019-B055), Chinese Test International research fund project (CTI2017B06), Collaborative Innovation Center for Teacher Quality Monitoring, Evaluation and Service in Jiangxi Province (JXJSZLC05).

REFERENCES

Baker, F. B., and Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques, 2nd Edn*. New York, NY: Marcel Dekker.

Ban, J. C., Hanson, B. A., Wang, T. Y., Yi, Q., and Harris, D. J. (2001). A comparative study of on-line pretest item—calibration/scaling methods in computerized adaptive testing. *J. Educ. Meas.* 38, 191–212. doi: 10.1111/j.1745-3984.2001.tb01123.x

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

Chang, H.-H., and Ying, Z. L. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Ann. Stat.* 37, 1466–1488. doi: 10.1214/08-AOS614

Chang, H.-H., and Zhang, J. M. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika* 67, 387–398. doi: 10.1007/BF02294991

Chang, Y. C. I., and Lu, H. Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika* 75, 140–157. doi: 10.1007/s11336-009-9133-0

Chen, P., and Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika* 81, 674–701. doi: 10.1007/s11336-015-9482-9

Chen, P., and Xin, T. (2014). "Online calibration with cognitive diagnostic assessment," in *Advancing Methodologies to Support Both Summative and Formative Assessments*, eds Y. Cheng and H.-H. Chang (Charlotte, NC: Information Age), 287–313.

Chen, P., Xin, T., Wang, C., and Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika* 77, 201–222. doi: 10.1007/s11336-012-9255-7

Cheng, Y., and Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br. J. Math. Stat. Psychol.* 62, 369–383. doi: 10.1348/000711008X304376

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Earlbaum.

He, Y., and Chen, P. (2019). Optimal online calibration designs for item replenishment in adaptive testing. *Psychometrika*. doi: 10.1007/s11336-019-09687-0. [Epub ahead of print].

He, Y., Chen, P., and Li, Y. (2019). New efficient and practicable adaptive designs for calibrating items online. *Appl. Psychol. Meas.* doi: 10.1177/0146621618824854

He, Y. H., Chen, P., Li, Y., and Zhang, S. (2017). A new online calibration method based on lord's bias-correction. *Appl. Psychol. Meas.* 41, 456–471. doi: 10.1177/0146621617697958

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Meijer, R. R., and Nering, M. L. (1999). Computerized adaptive testing: overview and introduction. *Appl. Psychol. Meas.* 23, 187–194. doi: 10.1177/01466219922031310

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Appl. Psychol. Meas.* 14, 59–71. doi: 10.1177/014662169001400106

Olsson, U., Drasgow, F., and Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika* 47, 337–347. doi: 10.1007/BF02294164

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika* 34, 1–97. doi: 10.1007/BF03372160

Samejima, F. (1996). "The graded response model," in *Handbook of Modern Item Response Theory*, eds W. J. van der Linden and R. K. Hambleton (New York, NY: Springer), 85–100. doi: 10.1007/978-1-4757-2691-6_5

Sands, W. A., Waters, B. K., and McBride, J. R. (eds.). (1997). *Computerized Adaptive Testing: From Inquiry to Operation*. Washington DC: American Psychological Association.

Tan, P.-N., Steinbach, M., and Kumar, V. (2014). *Introduction to Data Mining*. New York, NY: Pearson New International Edition.

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

van der Linden, W. J., and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer.

van der Linden, W. J., and Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika* 80, 263–288. doi: 10.1007/s11336-013-9391-8

Wainer, H., and Mislevy, R. J. (1990). "Item response theory, item calibration, and proficiency estimation," in *Computerized Adaptive Testing: A Primer*, ed H. Wainer (Hillsdale, NJ: Lawrence Erlbaum), 65–102.

Wang, C., and Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika* 76, 363–384. doi: 10.1007/s11336-011-9215-7

Wang, C., Chang, H.-H., and Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Appl. Psychol. Meas.* 37, 99–122. doi: 10.1177/0146621612463422

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

Xiong, J. H., Luo, H., and Wang, X. Q., and Ding, S. L. (2018). The online calibration based on graded response model. *J. Jiangxi Normal Univ.* 42, 62–66.

You, X. F., Ding, S. L., and Liu, H. Y. (2010). Parameter estimation of the raw item in computerized adaptive testing. *Acta Psychol. Sin.* 42, 813–820. doi: 10.3724/SP.J.1041.2010.00813

Zheng, Y. (2014). *New methods of online calibration for item bank replenishment* (Unpublished doctoral thesis). University of Illinois at Urbana, Champaign.

Zheng, Y. (2016). Online calibration of polytomous items under the generalized partial credit model. *Appl. Psychol. Meas.* 40, 434–450. doi: 10.1177/0146621616650406

Zheng, Y., and Chang, H.-H. (2017). A comparison of five methods for pretest item selection in online calibration. *Int. J. Quant. Res. Educ.* 4, 133–158. doi: 10.1504/IJQRE.2017.086500

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.

APPENDIX

Table A1 | Estimation accuracy of ability under different CAT scenarios.

Sample size	Test length	RMSE	Bias
2,000	Variable-length	0.1904	0.0007
	10	0.1924	−0.0004
	20	0.1340	−0.0008
	30	0.1105	−0.0012
3,000	Variable-length	0.1882	0.0033
	10	0.2012	−0.0001
	20	0.1286	−0.0024
	30	0.1057	0.0050

For variable-length CAT, the cumulative information was set to 25.

Table A2 | RMSE of new item parameters under different CAT scenarios.

Sample size	Test length	RMSE			
		a	*b1*	*b2*	*b3*
2,000	Variable-length	0.2483	0.2109	0.1802	0.2294
	10	0.2345	0.2224	0.1557	0.2182
	20	0.2169	0.1954	0.1545	0.2242
	30	0.2232	0.2060	0.1685	0.2357
3,000	Variable-length	0.2337	0.1921	0.1620	0.2203
	10	0.2302	0.2571	0.1668	0.2143
	20	0.2121	0.2102	0.1640	0.2078
	30	0.2069	0.2012	0.1664	0.2235

Table A3 | Bias of new item parameters under different CAT scenarios.

Sample size	Test length	Bias			
		a	*b1*	*b2*	*b3*
2,000	Variable-length	0.0998	−0.0005	−0.0330	−0.0685
	10	−0.0719	−0.0889	−0.0088	0.0615
	20	−0.0001	−0.0464	−0.0011	0.0272
	30	0.0589	−0.0168	−0.0211	−0.0269
3,000	Variable-length	0.0939	−0.0117	−0.0239	−0.0458
	10	−0.0650	−0.1364	−0.0486	0.0287
	20	−0.0228	−0.0187	−0.0106	0.0011
	30	0.0613	−0.0152	−0.0232	−0.0486

Taking 3-categories items as example and the calibration sample size is set to 500.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 February 2020
doi: 10.3389/fpsyg.2020.00305

The Development of a Multidimensional Diagnostic Assessment With Learning Tools to Improve 3-D Mental Rotation Skills

Shiyu Wang[1]*, *Yiling Hu*[2], *Qi Wang*[3], *Bian Wu*[2], *Yawei Shen*[1] and *Martha Carr*[1]

[1] Quantitative Methodology Program, Department of Educational Psychology, University of Georgia, Athens, GA, United States, [2] Department of Educational Information Technology, East China Normal University, Shanghai, China, [3] Measurement and Statistics Program, Department of Educational Psychology and Learning System, Tallahassee, FL, United States

This study reported on development and evaluation of a learning program that integrated a multidimensional diagnostic assessment with two different learning interventions with the aim to diagnose and improve three-dimensional mental rotation skills. The multidimensional assessment was built upon the Diagnostic Classification Model (DCM) framework that can report the binary mastery on each specific rotation skill. The two learning interventions were designed to train students to use a holistic rotation strategy and a combined analytic and holistic strategy, respectively. The program was evaluated through an experiment paired with multiple exploratory and confirmatory statistical analysis. Particularly, the recently proposed joint models for response times and response accuracy within dynamic DCM framework is applied to assess the effectiveness of the learning interventions. Compared with the traditional assessment on spatial skills, where the tests are timed and number correct is reported as a measure for test-takers' performances, the developed dynamic diagnostic assessment can provide an informative estimate of the learning trajectory for each participant in terms of the strengths and weaknesses in four fine-grained spatial rotation skills over time. Compared with an earlier study that provided initial evidence of the effectiveness of building a multidimensional diagnostic assessment with training tools, the present study improved the assessment and learning intervention design. Using both response times and response accuracy, thus current study additionally evaluated the newly developed program by investigating the effectiveness of two interventions across gender, country and rotation strategy.

Keywords: mental rotation skills, learning program, diagnostic assessment, rotation strategy, longitudinal diagnostic model

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Qiwei He,
Educational Testing Service, United States
Xin Qiao,
University of Maryland, College Park, United States

***Correspondence:**
Shiyu Wang
swang44@uga.edu

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

Received: *12 November 2019*
Accepted: *10 February 2020*
Published: *26 February 2020*

Citation:
Wang S, Hu Y, Wang Q, Wu B, Shen Y and Carr M (2020) The Development of a Multidimensional Diagnostic Assessment With Learning Tools to Improve 3-D Mental Rotation Skills. Front. Psychol. 11:305. doi: 10.3389/fpsyg.2020.00305

1. INTRODUCTION

Spatial ability has long been considered as an important dimension of human intelligence through the studies in various populations and settings (e.g., Carroll, 1993; Eliot, 2012). It is an emerging area of interest to educators as spatial ability has been linked to better performance in mathematics and science achievement (Brownlow and Miderski, 2001; Thompson et al., 2013). The notion of spatial ability varies across studies. Different types of spatial skills have been measured including

spatial perception, visualization and mental rotation (e.g., Perry, 2013; Weckbacher and Okamoto, 2014). Among these various spatial factors, mental rotation ability involves a cognitive visualization process to mentally rotate two-dimensional (2-D) or three-dimensional (3-D) objects. These two forms of mental rotation, particularly 3-D mental rotation, have been commonly associated with mathematics and science achievement (Voyer et al., 1995). Virtually, all 2-D and 3-D mental rotation tests involve presenting a target item and several solutions and the test taker has to mentally rotate the target to select the correct solution. One problem in this area is that little is known about the psychometric qualities of spatial skills tests or how or why students' performance differs as a function of test items. There are several possible causes of these problems. It may be the degree of rotation or manipulation needed, the complexity of the items, or the strategies used to solve items. Two strategies have been identified in literature: analytic/verbal and holistic (Glück et al., 2002). Holistic strategies involve rotating the entire object whereas analytic strategies involve matching parts of rotated objects to determine the correct answer. Both these two strategies can produce good outcomes but the holistic strategies are typically considered better examples of spatial processing and they seem to be more efficient and effective for more cognitive demanding spatial items; specifically items that require multiple, simultaneous rotations or that are complex (Wang and Carr, 2014). Some research studies also found that the combined analytic and holistic strategy might be more efficient than the sole holistic or analytic strategy, and it can decrease the gender difference (Stieff et al., 2014). Existing literature about mental rotation strategy also concluded that male and female students, Chinese Speakers and English Speakers may use different rotation strategies when solving spatial rotation questions (Weiss et al., 2003; Geiser et al., 2008; Li and O'Boyle, 2013; Li et al., 2014; Stieff et al., 2014).

While most studies in the literature focused on measuring the spatial ability or on investigating how the spatial ability is related to test-takers' characteristics, there has been a lack of research on investigating the factors that are related to the improvement of spatial skills. There are emerging evidence indicating that spatial ability can be improved (Uttal et al., 2013) and evidence that improving spatial skills results in improved mathematics (e.g., Cheng and Mix, 2014). Efforts to improve spatial skills have involved having participants practice on existing spatial skills tests or have involved extensive training in several aspects of spatial skills (e.g., isometric drawing). However, these instruction are time consuming and are not responsive to individual students' strengths and weaknesses.

This present study reported the development of a learning program that aims to improve mental rotation skills from a new perspective. This computer-based learning program integrates multiple multidimensional assessments with different learning interventions. Particularly, the embedded multidimensional assessments were built upon the Diagnostic Classification Model (DCM) framework. This is a family of restricted latent class models that can provide information concerning whether or not students have mastered each of a group of specific skills. These psychometric models have been used to design assessments that measure fine-grained skills or latent attributes across various domains, such as math skills (Bradshaw et al., 2014) and depression (Wang et al., 2019a). In addition to these applications of cross-sectional cognitive diagnostic assessment, the recently development of dynamic DCMs (e.g., Kaya and Leite, 2016; Li et al., 2016; Wang et al., 2017, 2018; Chen et al., 2018b; Zhan et al., 2019) enable the possibility of developing longitudinal cognitive diagnostic assessments to track skill learning and skill acquisition over time. This current study serve as the first attempt to develop the learning program within the longitudinal cognitive diagnostic assessment framework. Another important objective of this study is to evaluate the effectiveness of the developed learning program. Multiple exploratory and confirmatory analysis were conducted to evaluate the cognitive diagnostic assessment and learning interventions. Particularly, students' demographic information, such as gender, country and the rotation strategy, were collected and integrated with one of the recently developed dynamic DCMs, the joint model of response times and response accuracy (Wang et al., 2018, 2019b), to evaluate the learning interventions.

The rest of the paper is organized as follows. We first provide background on the test questions for measuring mental rotation skills, the Purdue Spatial Visualization Test: Visualization of Rotations (PSVT: R) and the revised PSVT:R. Second, we introduce the joint model of response times and response accuracy within dynamic DCM framework. This is followed by the description of the development of a new spatial rotation learning program. An experiment study is then presented to evaluate the learning program and understand students' learning behavior. We report the results from this experiment in the following section. Finally, the discussion section addresses implications for psychometrics and training mental rotation skills, limitations of the current study and future research study.

2. PSVT: R AND REVISED PSVT:R

The Purdue Spatial Visualization Test: Visualization of Rotations (PSVT: R), developed by Guay (1976), is one of the most popular tests that targets on measuring spatial visualization ability in 3-D mental rotation of individuals aged 13 years or older. This test has been frequently used in STEM education (Maeda and Yoon, 2013), and has shown in general good internal consistency reliability through several studies (Guay, 1976; Branoff, 2000; Alkhateeb, 2004). The PSVT: R consists of 30 items including 13 symmetrical and 17 non-symmetrical 3-D objects that are drawn in 2-D isometric format. Each item featured a reference object that had undergone a rotation. Test-takers then considered a new object and attempted to determine which of five options corresponded to the same rotation as the reference object. This test was revised by Yoon (2011) to correct the 10 figural errors identified by Yue (2006) and the format of the instrument was modified to avoid possible measurement errors. The revised test is named as revised PSVT:R. Since then, the revised PSVT:R has been used in several studies to investigate the psychometric properties of the test questions through Item Response Theory (IRT) Models (Maeda et al., 2013). They were also used to explore the association of the spatial ability of undergraduate students

with gender, STEM majors and gifted program membership (Yoon and Mann, 2017).

3. DCMS AND DYNAMIC DCMS FOR RESPONSE TIMES AND RESPONSE ACCURACY

Diagnostic Classification Model (DCM), or Cognitive Diagnosis Models (CDM), has emerged as an important statistical tool to help with diagnosing students' learning outcomes, such as skills and abilities that students have at the completion of a course or a learning program. These models assume that there are a number of pre-specified attributes measured by the assessment. A student's latent attribute profile is denoted by a multidimensional binary random vector with element 1 to indicate one possess a specific attribute and 0 to denote the lack of that particular attribute. In this way, DCMs can provide feedback regarding the measured skills. This allows for changes to be made in instruction, which can hopefully enhance students' learning. Research continues to document the benefits of DCMs as a framework for classifying students into educationally relevant skill profiles, and they have been used to study English-language proficiency (Templin and Hoffman, 2013; Chiu and Köhn, 2015), fraction subtraction (de la Torre and Douglas, 2004), pathological gambling (Templin and Henson, 2006), skills found in large-scale testing programs (Bradshaw et al., 2014; Li et al., 2015; Ravand, 2016), and Mental Rotation Skills (Culpepper, 2015).

The traditional DCMs are useful to classify attribute profiles at a given point in time. Recently research has begun to consider the role of DCMs to track learning and skill acquisition in a longitudinal fashion (Kaya and Leite, 2016; Li et al., 2016; Wang et al., 2017, 2018; Chen et al., 2018b; Zhan et al., 2019). In this type of research, the multidimensional binary latent skills for each student are assumed to be time-dependent and the purpose is to track the change of these binary skills overtime. Furthermore, in addition to the traditional product data, that is the response accuracy, the process data, such as the response times, are utilized to assess students' skill change over time. The joint model of response times and response accuracy (Wang et al., 2019b) used in this study is such an example. This joint model consists of a dynamic response model and a dynamic response time model. The dynamic response model includes a DCM as the measurement model to describe how test-takers respond to the assessment items with their attribute profiles at a given point of time, and a higher-order hidden Markov model that describes how the latent attribute profile changes from one time point to another, depending on the individual covariates (Wang et al., 2017). Like the traditional DCM, the dynamic DCM produces the output of the parameter estimation that quantify the psychometric properties for each item. It in addition can provide an estimate of students' learning trajectories in terms of the change of fine-grained skills over time. The estimated coefficients of the transition model from which can be used to identify the factors that are related to the transition probability and to evaluate the intervention. The dynamic response time model assumes students' latent speed on answering an item changes with the change of the latent attribute profile. It is thus directly connected with the dynamic response model through the latent attribute profile to provide additional information. The original work by Wang et al. (2019b) only considers the latent individual covariate in the dynamic response time model. In our study we will include students' demographic variables and problem-solving strategies to further investigate the between and within latent classes transitions. The details of this model are described in the Method section.

4. THE DEVELOPMENT OF A NEW SPATIAL ROTATION LEARNING PROGRAM

The new spatial rotation learning program reported in this study was developed on the basis of the findings from a previous research study (Wang et al., 2017). That old learning program was developed with the revised PSVT:R (Yoon, 2011) and consisted of five testing modules and four learning modules. Each of these modules contained 10 test questions. Four fine-grained mental rotation skills measuring the degree and direction of rotation were measured by test questions. That is (1) $x90$: 90° x-axis, (2) $y90$: 90° y-axis, (3)$x180$: 180° x-axis, and (4) $y180$: 180° y-axis. These four distinct yet related skills were identified to be measured by the revised PSVT:R through several previous studies (e.g., Maeda et al., 2013; Culpepper, 2015; Wang et al., 2017). To use this program, students first answered 10 questions in a testing module without any feedback to their answers then proceeded to a learning module in which they received feedback about their answers to the previous 10 questions and used a learning intervention to practice rotations. With such a design, test-takers need to finish 50 testing questions without feedback and to practice 40 additional questions with feedback and intervention. Positive findings of benefits of practice, an enhanced intervention, and the value of knowing some of the attributes, on the probability of making a transition to a master of a spatial skill, were demonstrated through a previous analysis (Wang et al., 2017). However, it was also found that a number of items had low psychometric qualities. This means that the students with low ability on spatial skills can easily guess the correct answer or the students with high spatial ability might easily miss the correct answer. These items provided less diagnostic information on measuring the spatial skills. Another finding was that students' performance in the 5th testing module is relatively lower than the 4th testing module, indicating there might be a fatigue factor due to the long testing and learning (it took about roughly 1 h and 15 min on average for students to finish this learning program). The following subsections provide details on the development of a new learning program based upon this old version of learning software.

4.1. The Learning Program Structure

Compared with the old version, the whole structure of the learning program was redesigned to have two testing modules and two learning modules. The structure of the learning program is summarized by the flow chart in **Figure 1**. Specifically, this program starts with a testing module, followed by two

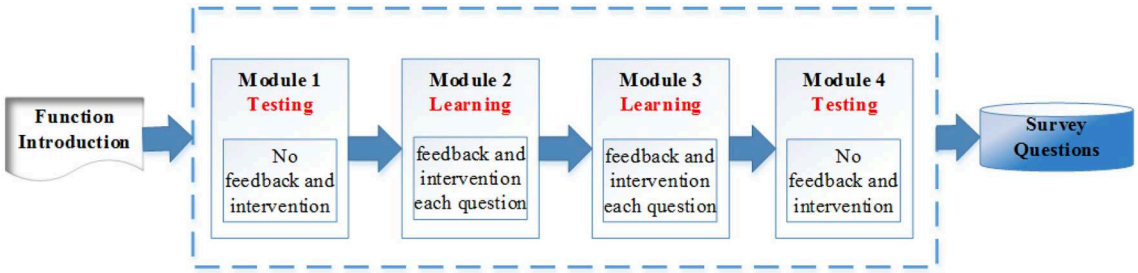


FIGURE 1 | Spatial rotation learning program structure.

consecutive learning modules, and finally ends with a testing module. The main purpose of module 1 and 4 is to accurately measure the four binary spatial skills at a given point in time. The two learning modules, model 2 and 3, aim to improve test-takers' mental rotation skills. The orders of these four modules are carefully designed thus are not exchangeable. The rationale of the design of these modules are summarized in section 4.2. Interventions are only provided in the learning modules. Each module contains 10 different questions, and they are selected based on various of item characteristics to reflect their functioning of assessing or improving the skills. A survey is provided at the end of the program to collect the test-takers' demographic information, the rotation strategy used by them during the test and their opinions about this learning program.

4.2. The Design of Module Blueprint

As described in the introduction, the learning program used the revised PSVT:R questions to measure four rotation skills. In fact, the original revised PSVT: R has 30 questions, and Wang et al. (2017) developed another 20 new items following the same item format so that a total of 50 questions are available to use in our study. Based on the learning program structure, 40 questions were selected from the existing 50 questions to assemble the four modules. These questions were selected based on different item characteristics, which can be measured from both a qualitative and quantitative point of view. The qualitative properties include the skill(s) measured by each item and the shape of the item. A very important component in the DCM based assessment, is a Q matrix (Tatsuoka, 1985), that specifies the rotation skill(s) measured by each item. The Q matrix is usually pre-determined by panels of subject-matter experts or estimated and validated based on the response data (e.g., Xu and Shang, 2017). In this study, we used the Q matrix in Wang et al. (2017), which was built based on the findings from Guay (1980) and Culpepper (2015). According to this Q matrix, each of the available 50 questions measures 1 or 2 skills. The shape of an item reveals the complexity in visualizing the 3-D object. The current 50 questions include symmetrical and non-symmetrical 3-D objects that are drawn in 2-D isometric format. The quantitative properties of the questions can be described by the difficulty and discrimination of the item. The difficulty of the PSVT: R items has been analyzed based on classical test theory and item response theory (Yoon, 2011; Maeda et al., 2013). The discrimination of the items describes how one item can discriminate/differentiate the students with low spatial ability from those with high spatial ability. Previous studies used the two parameter and three parameter logistic models (Maeda et al., 2013) and the deterministic input, noisy, "and" gate model (DINA; Junker and Sijtsma, 2001) to get item discrimination parameter estimation (Culpepper, 2015). In order to accurately measure students' mental rotation skills and to detect the possible learning effect, we design the two testing modules to have balanced and similar item quality. For the two learning modules, the main purpose is to help students improve their spatial skills and keep their motivation of using the learning intervention. Thus, the first learning module contains the relative easy items with simple shapes, with the purpose to minimize the side effect of lack of interest in learning due to frustration of providing too many wrong answers (as they are informed their answer is right or wrong in the learning module). The second learning module contains relatively harder and moderate to complex shape of items. In addition, the analysis on the learning data (Wang et al., 2017) revealed that the four attributes might have a hierarchical structure that implies that students who have mastery of 180 rotations should also be skilled at 90° rotation. In other words, the 90° rotation is the prerequisite for the 180° rotation. Thus, it's reasonable to guide students to learn the prerequisite skill first. Based on all above analysis, the finalized targeted properties of the items in the four modules are presented in **Tables 1**, **2**. The next section summarizes the details of the selection of 40 questions based on both quantitative and qualitative analysis.

4.3. Item Pre-analysis and Validation

4.3.1. Quantitative and Qualitative Analysis

We conducted both qualitative and quantitative analysis to the 50 available questions from Wang et al. (2017) in order to select 40 from them to assemble the four modules based on the blueprint. For the quantitative aspect, using the data from a previous research study (Culpepper, 2015), a Rasch model was fitted to the 50 questions to produce the item difficulty parameters. Six raters with high spatial abilities were invited to rate the difficulty of each item, and their scores were highly positively correlated with the estimated difficulty parameters from the Rasch model, ranging from 0.89 to 0.94. The item discrimination, $1 - s_j - g_j$, is defined based on the Deterministic Input, Noisy "And" gate (DINA; Junker and Sijtsma, 2001) model, which describes how

TABLE 1 | The targeted properties of the items in four modules.

Index	Module 1	Module 2	Module 3	Module 4
Difficulty	Balanced	Easy	Moderate-high	Balanced
Discrimination	Balanced	Low-moderate	Moderate-high	Balanced
Shape	Balanced	Simple-moderate	Complex-moderate	Balanced

TABLE 2 | The skill(s) measured by the number of questions across four modules.

Attribute	Module 1	Module 2	Module 3	Module 4	Total
*x*90	1	2	1	1	5
*x*180	1	2	1	1	5
*y*90	1	2	1	1	5
*y*180	1	3	2	1	7
*x*90, *y*90	2	1	2	2	7
*x*90, *y*180	2	0	1	2	5
*x*180, *y*90	2	0	2	2	6

well an item can discriminate subjects who master all the required attributes for the item from subjects who do not master any of the required attributes. The larger the discrimination index, the more diagnostic information the item can provide. For the qualitative aspect, a spatial skill domain expert examined the shape of object in each question, and rated the complexity of the shape to the scale of 1–5. The higher the score, the more difficult for this object to be visualized as a 3-D object. Based on the characteristic of the 50 available questions, a heuristic automatic test assembly algorithm was developed to select 40 questions to assemble the four modules. This test assembly algorithm was developed by authors based on Armstrong's et al. (1992) Phase II algorithm to guarantee the four modules match the program blueprint (**Tables 1**, **2**). The item positions in each module are in ascending order of item difficulty (from easy to hard).

4.3.2. 3-D Model Building

The original PSVT: R presented the 3-D object in a 2-D isometric format. In the current study, in order to accurately measure the four fine-grained mental rotation skills that target on degree and direction of rotation only, all the objects in the 50 questions were reconstructed based on 3-D model building in computer and an example is presented in **Figure 2**. The 3-D models were constructed using 3ds Max 2016 developed by Autodesk. The questions in the testing modules and the learning modules are all like the one presented in **Figure 2**, which include a reference item that is rotated. Test-takers are presented a new object and they must select one answer from the five options that corresponds to the ending position of the new object, rotated the same way as the reference item. In the testing module, test-takers are not informed about whether their questions are correct or wrong. And in the learning module, they are informed immediately about the correct answer correct or not after each question. In addition, in the learning module, test-takers have the chance to interact with a learning intervention to practice rotation. The next subsection describes the intervention design.

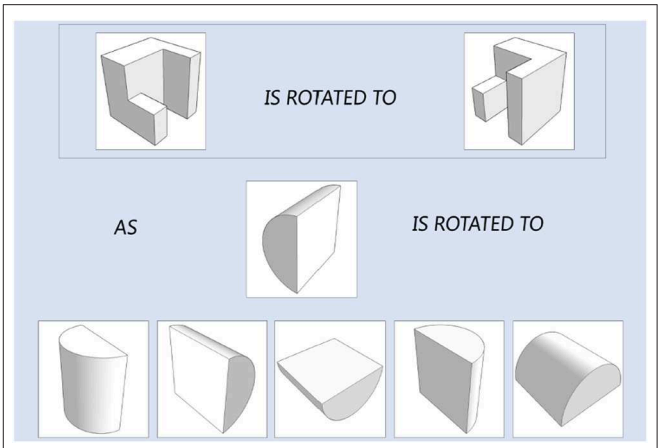


FIGURE 2 | The 3D model for the objects/figures in an item.

4.4. Learning Intervention Design

4.4.1. Two Learning Interventions

We developed two types of learning interventions by using C++ with Visual Studio 2012. One version is animation plus interaction as shown in **Figure 3**. The left panel of **Figure 3** shows the testing items, the top panel on the right shows animation of rotating the reference object from the initial position to the final position and the bottom panel on the right allows users to rotate the testing object from the initial position to the final correct position by following the rotation path from the reference one. This type of intervention intends to train test-takers with the *holistic strategy*. The other intervention has the same functions as the first one and with an additional coloring feature (**Figure 4**). One of the facets of both reference and testing objects in three panels was draw with pink color. This is designed to help test-takers figure out the final position of the testing object by mapping the pink facet in the initial position to its final position. This coloring is more like training the test-takers using an analytic strategy. Combined with the rotation functions from the right panels, the second intervention intends to train test-takers with a *combined analytic* and *holistic strategy*.

4.4.2. Learning Routine

In both versions of intervention, test-takers follow the same learning routine of three steps: (a) solve the displayed testing question on the left panel with the top and the bottom right panels invisible, and hit the check answer button to receive the feedback; (b) the two panels on the right are then displayed and test-takers can press the rotate button on the top right panel to watch the rotation animation of the reference object; (c) test-takers need to further rotate the testing object in the bottom right panel to the correct position. During this process, test-takers are allowed to repeat step (b) and (c).

5. METHOD

5.1. Experiment Study

5.1.1. Sample

The participants in this experiment were undergraduate students, 18 years or older, enrolled in three Universities, one in

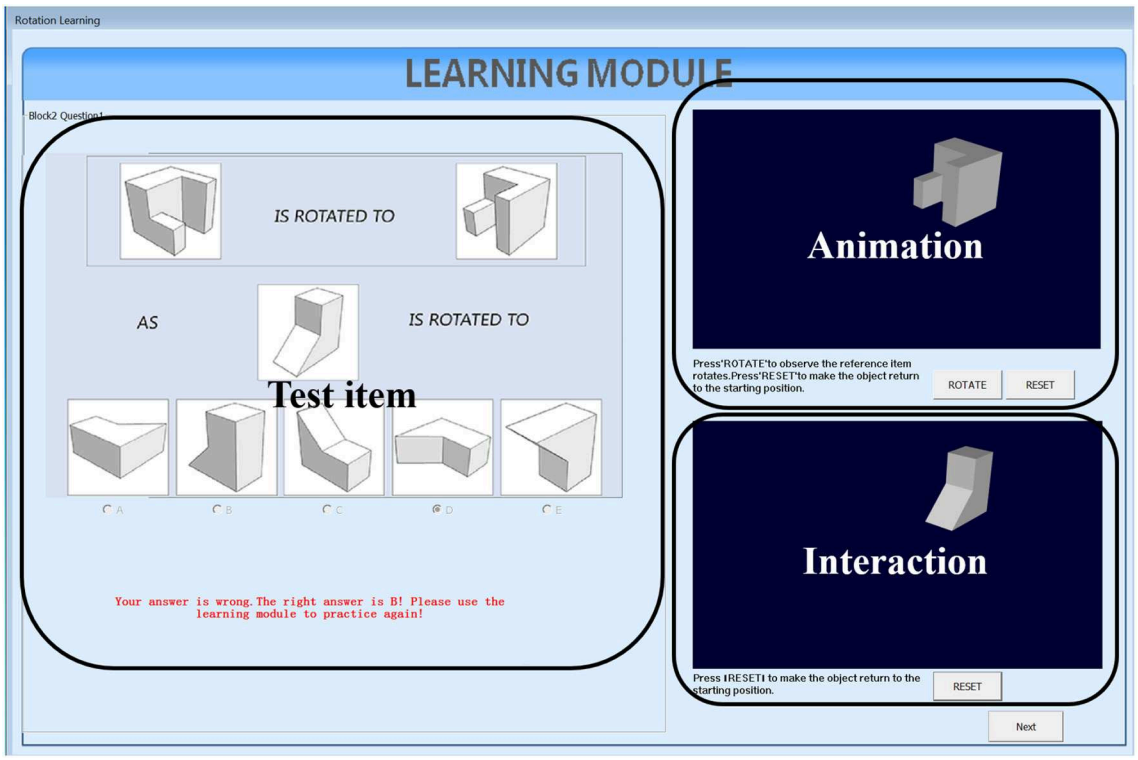


FIGURE 3 | Non-colored intervention.

United States and two in China. Recruiting participants in two countries can help us investigate whether there is cultural difference in terms of learning spatial skills. In order to get enough sample size, participants in both countries were recruited by two ways. The first was to recruit participants through the Educational Psychology or Psychology Research Participant Pool. Participants from this source were rewarded 1 course credit after completing this study. The second was to recruit participants through flyers and email announcement. For those participants, they were paid with a base amount of money and can earn additional amount of payment for each question answered correctly. From Spring 2017 to Summer 2017, recruitment through the above two approaches yielded 585 participants. Because of the various sources of recruiting participants, we fitted a mixture learning model (Zhang and Wang, 2018) to exclude some participants who were identified to be not engaged in the experiment. These participants' response data did not reflect the measured latent attributes and cannot be used to evaluate the learning program. Through this procedure, a total of 548 students were included for final data analysis.

5.1.2. Procedures and Variables

The experiment was conducted in the computer lab in each University. The two types of learning interventions (colored and non-colored) that corresponding to the combined and holistic rotation training strategy were randomly assigned among the participants. Before starting the learning program, the participants first watched the instruction about how to use the learning program. Researchers in the computer lab also gave directions on how to use the program and they were available to answer questions during the experiment. Participants were informed that they had as much time as they wanted to complete this assessment. They were told that this study was conducted to understand how people solve and learn spatial rotation tasks. The participants who received the payment instead of the course credit were informed that the payment were based on the number of questions answered correctly. On average, it took 30 min for the participants to finish the experiment.

The participants' binary responses and their response time to each of the 40 test questions were recorded by the software directly. In addition to these response data, a survey after each participant completed the experiment collected participants' demographic information, such as gender (female and male), the country (China and US), and the strategy participants used to solve the questions (Analytic, Holistic and Hybrid). The information about the rotation strategy used by each participant was collected based on a self-report question in the survey. These covariates will help us further evaluate the developed diagnostic assessment and learning interventions across different populations. In the survey, participants also provided their opinions about whether the learning module can help them learn rotation skills on the Likert scale [1 (not helpful)–5 (very helpful)].

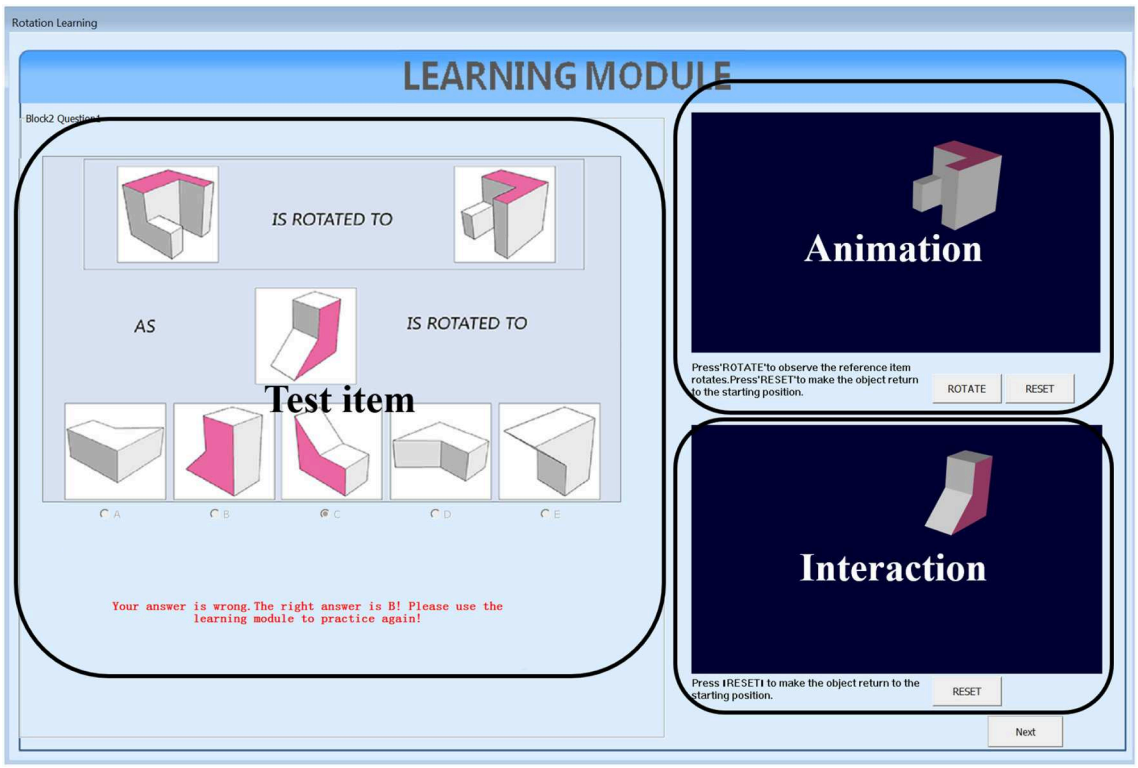


FIGURE 4 | Colored intervention.

5.2. Exploratory Statistical Analysis

5.2.1. Descriptive Statistic

Descriptive statistics, such as the number of participants (N), the means and standard deviations of the module scores in the learning program and the module completion time, are presented in **Tables 3**, **4** by participants' characteristics, such as gender, country, strategy used to solve the problem and the intervention. A slightly increase of the mean score in module 4 compared with module 1 can be observed. The module 3 contains the items that are most difficult while module 2 consists of the items that are easiest among the four modules. Thus, the average score in module 3 is the lowest and the average score in module 2 is the highest among the four modules across different groups. Note that, we can hardly eyeball the "growth" based on the descriptive statistics in different modules, as each module has different item difficulty. In addition, the evaluation of the learning program should target on the population who have relatively low spatial rotation skills. However, in order to recruit participants as many as possible in a short time, we did not conduct a separate pretest to exclude the participants who already had a high spatial rotation ability. Thus, the final 548 sample may mix a proportion of participants who do not need to improve their spatial skills. Fortunately, the joint learning models presented in the later section can consider the item difficulty and help us identify the participants who already mastered the four skills in the very beginning. In terms of the completion time, participants spent least time on completing module 2, which is consistent with that module 2 is the easiest one. Though module 3 contains the most difficult items, participants on average spent less time on it compared with the module 1, which is relatively easier. This might be due to the warm-up effect for module 1, in which participants were still not very familiar with the questions or due to the improvement of their spatial rotation skills so that they can apply those skills more quickly in module 3. The distribution of participants over country and intervention are roughly balanced, while for gender and rotation strategy, the distributions are unbalanced. The large proportion of the female participants and combined rotation strategy used by participants are mainly due to our convenience sampling procedure and self-report of the strategy in the survey.

5.2.2. Clustering Analysis on Items

A very important component used in the joint model of response times and response accuracy is the Q matrix, which gives the information on which attributes are measured by each item. The previous research studies on Q matrix estimation or validation are in general conducted in through a confirmatory way that assumes students' responses follow a specific DCM (e.g., Xu and Shang, 2017). In this study, we conduct exploratory clustering analysis on items, using not only responses but also response times. The item group results from the cluster analysis can be used to compare with the existing Q and further valid it in the future. One clustering algorithm that accounts for both continuous and categorical data is K-prototype (Huang, 1997).

TABLE 3 | Descriptive statistics for 548 participants (response scores).

Variable		Module score				
		N	1	2	3	4
Gender	Female	401	7.00 (1.87)	8.76 (1.30)	6.34 (2.02)	7.15 (1.85)
	Male	147	7.76 (1.80)	9.16 (1.05)	6.71 (1.95)	7.70 (1.85)
Country	US	223	6.94 (1.85)	8.63 (1.37)	5.74 (1.92)	7.27 (1.82)
	China	325	7.38 (1.89)	9.03 (1.13)	6.91 (1.92)	7.31 (1.89)
Strategy	Analytic	62	7.37 (2.03)	8.84 (1.16)	6.56 (2.09)	7.52 (1.80)
	Holistic	63	6.71 (2.02)	8.79 (1.05)	6.11 (1.89)	7.14 (1.88)
	Combined	423	7.25 (1.83)	8.88 (1.29)	6.46 (2.01)	7.29 (1.87)
Intervention	Color	264	7.07 (2.00)	9.06 (1.14)	6.58 (1.87)	7.13 (1.91)
	Non-color	284	7.32 (2.00)	8.68 (1.32)	6.30 (2.11)	7.44 (1.81)

The numbers in the brackets are the standard deviation. The total score for each module is 10.

TABLE 4 | Descriptive statistics for 548 participants (response time).

		Module completion time (minute)				
		N	1	2	3	4
Gender	Female	401	8.76 (4.30)	4.71 (2.12)	7.23 (3.49)	6.84 (3.30)
	Male	147	8.13 (3.84)	4.06 (1.94)	6.52 (3.21)	5.98 (2.53)
Country	US	223	7.17 (3.88)	3.84 (1.76)	6.15 (3.28)	6.15 (3.27)
	China	325	9.56 (4.12)	5.01 (2.17)	7.64 (3.40)	6.92 (2.99)
Strategy	Analytic	62	9.55 (4.80)	4.96 (2.45)	7.99 (4.33)	6.67 (3.24)
	Holistic	63	7.27 (3.10)	3.74 (1.28)	5.78 (2.62)	5.48 (2.45)
	Combined	423	8.64 (4.19)	4.59 (2.11)	7.08 (3.33)	6.77 (3.18)
Intervention	Color	264	8.37 (4.07)	3.99 (1.75)	6.21 (3.15)	6.68 (3.29)
	Non-color	284	8.80 (4.29)	5.05 (2.25)	7.81 (3.49)	6.55 (2.98)

The numbers in the brackets are the standard deviation.

We apply this method to group items in each module based on the categorical responses and continuous response times in which. The number of clusters, M, is determined by the Silhouette index (Rousseeuw, 1987), which is commonly used in clustering analysis (e.g., Rendón et al., 2011; Hämäläinen et al., 2017). This index measures the similarity of an item to its cluster compared to other clusters and its value ranges from -1 to 1. A value of 1 is ideal as it suggests that data point is far away from other clusters. On the contrary, value of -1 is not preferred because it indicates that the data point is closer to other clusters than to its own. In our study, we use the Global Silhouette value, which is the average of the total silhouette values for all items of each cluster, to determine the number of clusters (Bolshakova and Azuaje, 2003). For all four modules, the average Silhouette values were highest when $M = 2$. Based on this, we group items into two clusters for each module. Note that the items can be in general classified as two types based on the Q matrix. One are simple items which measure only one attribute, the other are complex items which measure more than one attributes. The clustering results from K-prototype indicated that for each module, all simple items were grouped together and most complex items were grouped into another cluster. We note that four complex items, item 6 and 7 in module 1, item 23 in module 3, and item 35 in module 4, were grouped with simple items instead. Based on the current Q matrix, these four items all measure attributes $x90$ and $y90$. To explore the reason of mismatching of these four items, we compared them with item 20, 29, and 36, which also measure attributes $x90$ and $y90$. It was found that the 3D objects in item 6, 7, 23, and 35 are in relative simple shapes compared with those for item 20, 29, and 36, as shown in **Figure 5**. Moreover, the response accuracy and response times on item 6, 7, 23, and 35 were closer to simple items than the complex items measuring the same attributes. For example, the mean response time for item 35, simple and complex items in module 4 are 36.96, 27.04, and 50.33 s, respectively, and the mean correct response proportion for these three groups are 0.7, 0.76, and 0.54, respectively.

5.3. Confirmatory Statistical Analysis

5.3.1. The Joint Model of Response Time and Response Accuracy

In a longitudinal set up, such as the one in our study, the multidimensional binary latent skills for an individual i at time t are denoted as $\alpha_i(t) = (\alpha_{i1}(t), ..., \alpha_{iK}(t))'$, with t indexes time and $k = 1, ..., K$ indexes attributes and $\alpha_{ik}(t) = 0$ indicating non-mastery and 1 meaning mastery. Test-takers' responses are also time dependent, and the i_{th} test-taker's responses to J questions at time t can be denoted as $\mathbf{Y}_i(t) = (Y_{i1}(t), ..., Y_{iJ}(t))$, with $Y_{ij}(t) = 1$ if the test-taker responded correctly to item j at time t, and 0 otherwise. In addition, the computer records the response time on completing each test question for each test taker, denoted by $L_i(t) = (L_{i1}(t), ..., L_{iJ}(t))$. Both $\mathbf{Y}_i$ and L_i are used to provide an estimate of each test-taker's learning trajectory in terms of the change of fine-grained skills over time based on responses and also an estimate of their initial latent speed and the change of the speed due to the latent attribute profile and other covariates.

Specifically, the joint model proposed by Wang et al. (2019b) consists of a dynamic response model and a dynamic response time model. The dynamic response model includes two components. For each time t, a measurement model is used to model $P(Y_{ij}(t)|\alpha_i(t))$. An example is

$$P\left(Y_{ij}(t) = 1|\alpha_i(t), s_j(t), g_j(t)\right) = \begin{cases} 1 - s_j(t) & \text{if } \alpha_i(t) \succeq q_j, \\ g_j(t) & \text{otherwise,} \end{cases} \quad (1)$$

where q_j denotes the skills measured by item j. The notation $\succeq$ indicates the test-taker i with latent attribute profile $\alpha_i(t)$ has mastered all the required skills for item j at time t. The model describes by Equation (1) is the DINA model, which uses two parameters to describe the correct response probability to each item given the test-takers' latent profile and the required skills for that item. For example, if the test-taker's latent profile at this time is $(1, 1, 0, 0)'$, meaning he mastered the 90° rotations along the x and y axes. If item j only requires 90° rotation along

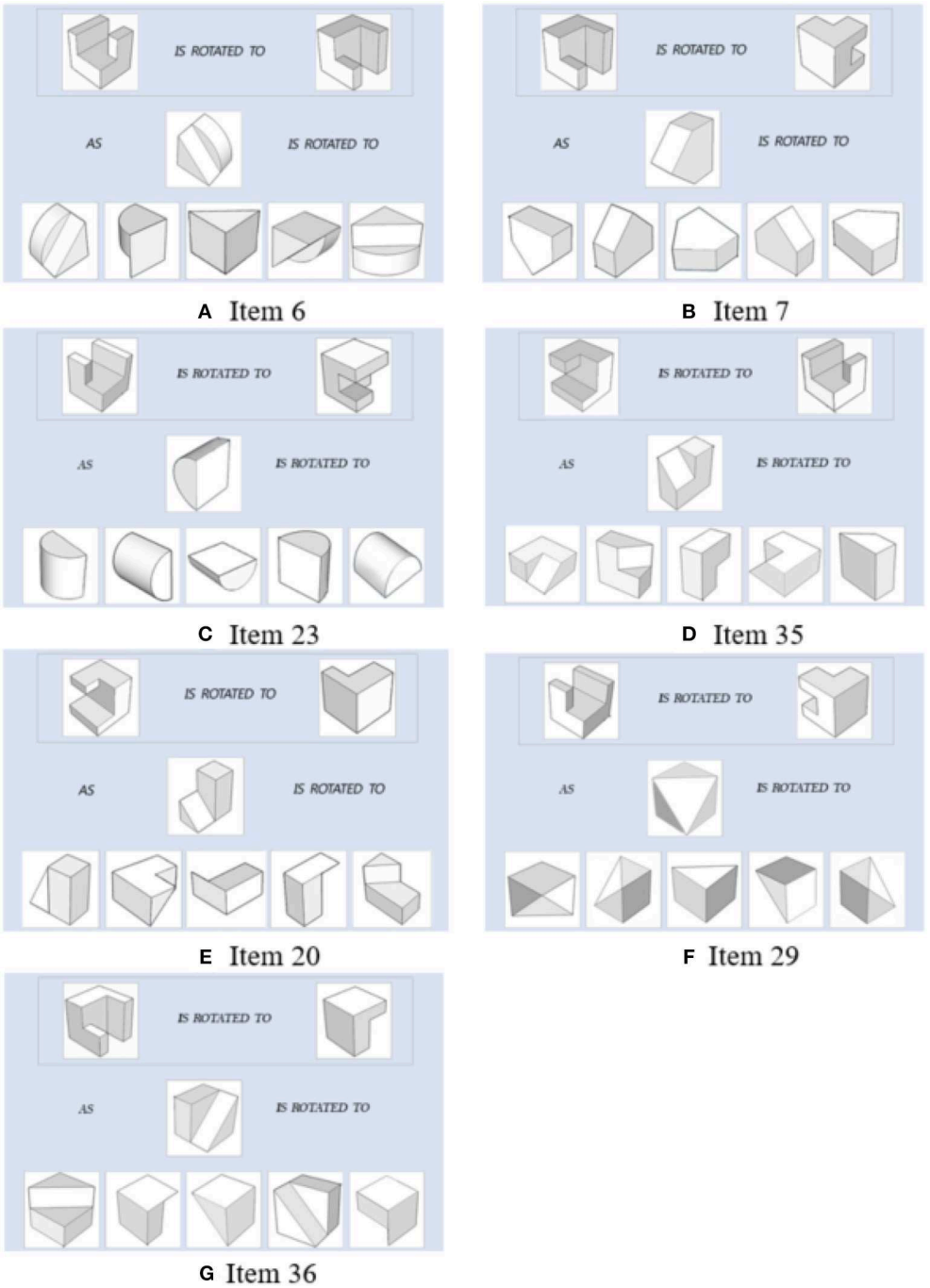


FIGURE 5 | Items measuring *x*90 and *y*90. Items 6,7,23, and 35 are clustered with simple items. Items 20, 29, and 36 are clustered with other complex items.

x axis, then this test-taker has probability $1 - s_j(t)$ to answer this item correctly. The term $s_j(t)$ is the slipping probability that refers to the probability that the test-taker misses item j at time t that his level of mastery suggests he would be expected to answer correctly to it. In the other case, if this item j requires the 180° rotation along x axis, and this test-taker does not master this required skill, then he has the probability $g_j(t)$ to answer this item correctly. This probability, $g_j(t)$, is called the guessing probability that describes the chance that the test-taker correctly answers a question that his level of mastery would suggest he should not. The DINA model is a very simple DCM model with a conjunctive structure. It assumes only two correct response probabilities for each item. Many popular CDMs, such as the models assumes a compensatory structure or in more general

forms can also be a candidate for this measurement portion. In the subsequent section, we will conduct a measurement model selection procedure to determine the most appropriate measurement model for our data.

The second component in the dynamic response model is a transition model, which describes how the latent attribute profile changes from one time point to another. This transition model assumes non-decreasing skill trajectories and conditional independence of attribute-wise transitions given the previous attribute pattern, and hence, it focuses on modeling the transition of each skill from non-mastery (0) to mastery (1), depending on several latent and observed covariates. To model the transition probability, we first assume the transition of an unlearned skill from 0 to 1, depends on a general learning ability. This general ability is denoted as a latent continuous variable for each test-taker i as θ_i and the number of learned skills. In addition, as one of the primary objectives of this study is to compare the two types of interventions on improvement of the rotation skills across gender, country and problem solving strategy, the variables reflect this information are also included in this model. In summary, the covariates we considered in the transition model are the main effects of general learning ability θ, the mastered skill(s), the gender, country, intervention, rotation strategy, as well as the two-way interactions between intervention and gender, intervention and country and intervention and rotation strategy. It can be written as,

$$\begin{aligned} & logit(P(\alpha_{ik}(t+1)=1|\alpha_{ik}(t)=0)) \\ =\ & \lambda_0+\lambda_\theta\theta_i+\lambda_\alpha\sum_{l\neq k}\alpha_{il}(t)+\lambda_g*\text{gender}_i \\ +\ & \lambda_c*\text{country}_i+\lambda_I*IV_i+\lambda_{st1}*\text{Strategy}_{1i}+\lambda_{st2}*\text{Strategy}_{2i} \\ +\ & \lambda_{gI}*\text{gender}_i*IV_i+\lambda_{cI}*\text{country}_i*IV_i \\ +\ & \lambda_{Ist1}*IV_i*\text{Strategy}_{1i}+\lambda_{Ist2}*IV_i*\text{Strategy}_{2i}. \end{aligned} \tag{2}$$

Here $\sum_{l\neq k}\alpha_{il(t)}$ quantifies the number of mastered skills at time t. IV_i, gender_i and country_i are dummy variables representing the two levels of each categorical variable. The Strategy_1 and Strategy_2 are the two dummy variables denoting the three levels of the rotation strategies used by the test-takers. Each of the component in the coefficient vector $\boldsymbol{\lambda} = (\lambda_\theta, \lambda_\alpha, \lambda_g, \lambda_c, \lambda_I, \lambda_{st1}, \lambda_{st2}, \lambda_{gI}, \lambda_{cI}, \lambda_{Ist1}, \lambda_{Ist2})'$ describes how the corresponding covariate influences the odds of skill transition from 0 to 1. These estimated values can help us evaluate the designed learning program.

Finally, the dynamic response time model is built based on a log-normal distribution. That is, the model assume the log of response time on each question follows a normal distribution, where the mean depends on a time intensity parameter(γ_j), the test taker's initial latent speed (τ_i), and the covariates that may influence the speed during the learning process. The variance of the distribution is characterized by a time discrimination parameter (a_j). The log-normal response time model is chosen based on the analysis from a previous research study that used the same experiment data set (Zhang and Wang, 2018). The key part of the dynamic response time model is on defining a latent covariate that connects the latent attribute profile and identifying several observed covariates that may impact the speed. In our case, we use a fixed effect model as the following specific form.

$$log(L_ij(t)) \sim N(\gamma_j-(\tau_i+\sum_{h=1}\phi_h Cov_h), \frac{1}{a_j}). \tag{3}$$

The quantity $\sum_{h=1}\phi_h Cov_h)$ in Equation (3) describes the different covariates that may impact the speed. Specifically,

$$\begin{aligned} \sum_{h=1}\phi_h Cov_h\ =\ & \phi_\alpha G(\alpha_i,\text{q}_\text{j})+\phi_\text{g}*\text{gender}_\text{i} \\ & +\ \phi_c*\text{country}_i+\phi_I*IV_i+\phi_{st1}*\text{Strategy}_{1i} \\ & +\ \phi_{st2}*\text{Strategy}_{2i} \\ & +\ \phi_{gI}*\text{gender}_i*IV_i+\phi_{cI}*\text{country}_i*IV_i \\ & +\ \phi_{Ist1}*IV_i*\text{Strategy}_{1i}+\phi_{Ist2}*IV_i*\text{Strategy}_{2i}. \end{aligned} \tag{4}$$

The $G(\alpha_i, \text{q}_\text{j})$ is the latent covariate that connects the learning trajectory α_i with the response time model. We define $G(\alpha_i, \text{q}_\text{j}) = 1$ is $\alpha_i(t) \succeq \text{q}_\text{j}$ and 0 otherwise. In this way, this covariate classify the change of speed into 2 classes on each item. The other observed covariates in (4) are the same as those in the transition model (2), and we are interested in investigating whether those covariates can give us additional information on the respond speed after controlling the latent learning trajectory. Such information are useful to evaluate the developed learning interventions.

In summary, the confirmatory joint model of response times and response accuracy can produce a learning trajectory for each test taker. In our case, if the latent profile is described based on the order of $x90$, $y90$, $x180$, and $y180$, and for a participant with the initial latent profile as $(0, 1, 0, 0)'$, indicating one masters only the 90° rotation along y axis, then joint model can provide an estimate of the latent profile after each stage of the learning program. The improvement of a specific rotation skill can be observed as the change from non-mastery (0) to mastery (1). In addition, the estimated coefficients in the transition model (λs) and dynamic response model (ϕs) can be used to evaluate the effectiveness of the learning program cross different populations defined by various latent and observed covariates.

5.3.1.1. Selection of the response measurement model

Before fitting the joint model, we first need to select the appropriate measurement model for responses. The models we consider are DINA, the deterministic-input, nopisy-or-gate model (DINO; Templin and Henson, 2006), the reduced reparameterized unified model (RRUM; Hartz, 2002), linear logistic model (LLM; Maris, 1999), the additive CDM (ACDM; de la Torre, 2011), and generalized DINA (G-DINA; de la Torre, 2011). These models are the representatives of DCMs that either have conjunctive/compensatory assumptions or belong to a family of models that have more general assumptions. To select the most appropriate model, we performed both test-level and item-level model selection procedures, treating each module as a mini test. These procedures were conducted using package *GDINA* (Ma et al., 2019) with Expectation-Maximization algorithm in R version 3.5.1 (R Core Team, 2018). For the test-level model selection, the Akaike information criterion (AIC)

TABLE 5 | Model-data fit indices.

	Module 1		Module 2		Module 3		Module 4	
Model	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**	**AIC**	**BIC**
DINA	5996.65	6147.37	3629.94	3780.66	6452.84	6603.56	5485.19	5635.91
DINO	5993.90	6144.62	3632.55	3783.27	6469.02	6619.74	5498.81	5649.53
GDINA	5979.02	6181.41	3633.79	3793.12	6464.39	6658.17	5498.36	5700.76
RRUM	5985.95	6162.51	3631.89	3786.91	6458.71	6630.96	5496.97	5673.53
LLM	5982.26	6158.81	3631.75	3786.78	6459.18	6631.43	5498.22	5674.77
ACDM	5982.03	6158.59	3631.84	3786.87	6458.79	6631.04	5499.01	5675.57

and Bayesian information criterion (BIC) were used. **Table 5** represents the values of AIC and BIC for multiple models at module level. For module 2–4, both AIC and BIC suggest that the best measurement model is DINA. However, for module 1, AIC suggests the GDINA and BIC suggests the DINO. The BIC value from the DINA model is very close to the DINO. For the item-level model selection, we apply the Wald test (de la Torre and Ma, 2016; Ma and de la Torre, 2016) to determine the most appropriate model for each item. The reduced models with p-values less than the pre-specified α level were rejected. If all reduced models were rejected for an item, the GDINA model was used as the best model; if more than one reduced models were retained, the reduced model with the largest p-values is selected as the most appropriate model with prioritizing DINA and DINO. Before doing that, we note that there are in fact 21 items that measure only one attribute. For these items, all types of DCMs are equivalent to the DINA model. The For the rest 19 items, the Wald test suggests that DINA model fits best for 12 of them. Other reduced models, such as RRUM, ACDM, and DINO, fit best for the rest 7 items. The details of the Wald test rest are summarized in **Table 1A** in Appendix. Both the test-level and item-level results suggest the DINA model fits most of the test questions, and also given its simple format, we choose to use the DINA model as the measurement model in the joint model.

5.3.1.2. Model convergence result

The joint model was calibrated through a Metropolis-Hastings within Gibbs Sampler (Wang et al., 2019b) through R (R Core Team, 2018). The MCMC chain convergence was evaluated by the Gelman-Rubin proportional scale reduction factor (PSRF) (Gelman and Rubin, 1992), commonly known as $\hat{R}$. Based on this criterion, this fitted model converged quickly as that shown in **Figure 6**. We can observe that after about 15,000 iterations, the maximum Gelman-Rubin proportional scale reduction factor among all parameters fell below 1.2, indicating that parameter estimates have stabilized.

5.3.2. Item Analysis for Testing and Learning Modules

5.3.2.1. Item parameters

The joint model of response accuracy and response times is able to estimate two types of item parameters for each item: the slipping and guessing parameters from the DINA model and the item discrimination and item intensity parameters from the log-normal response time model. The distribution of the estimated guessing and slipping parameters for the 40 rotation questions are summarized in terms of boxplots in **Figure 7**. The items in the testing modules and learning modules are presented separately to better compare their characteristics. In each of the boxplot, the x axis denotes the item type in terms of the attributes measured by that item and the y axis denotes the estimated parameter value for an item with certain measured skills. The distribution of the slipping and guessing parameters had the similar pattern for the items in the testing modules (module 1 and 4) and learning modules (module 2 and 3). Specifically, the items require only one simple skill, such as $x90$ or $y90$, tend to have large guessing parameters and small slipping parameters. The items require one complex skill, such as $y180$, or two skills, have small guessing parameters and large slipping parameters. The variation of the same type of item parameters is larger for the items in the learning module than that in the testing modules. Similarly, the distribution of the estimated time intensity and time discrimination parameters of the 40 items are documented in **Figure 8**. Again, the distribution of these parameters had the similar pattern in the testing and learning modules. That is, the items require one simple skill tend to have small time intensity parameters and large time discrimination parameters. The items require one complex skill or two skills tend to have large time intensity parameters and small time discrimination parameters. The average values of DINA model parameters and response time model parameters for each module are presented in **Table 6**. The distribution of these parameters are relatively consistent with the test assembly requirement presented in **Table 1**. That is, the two testing modules (module 1 and 4) were assembled with items that had balanced item quality, while the two learning modules (module 2 and 3) were designed based on their corresponding learning functions.

TABLE 6 | The mean item parameters for each module.

Function	**Modules**	s	g	$1-s-g$	a	γ
Testing	Module 1	0.182	0.560	0.259	1.657	3.288
	Module 4	0.193	0.514	0.293	1.560	2.970
Learning	Module 2	0.065	0.757	0.178	1.978	2.680
	Module 3	0.272	0.418	0.310	1.757	3.093

Next, we focus on the analysis with some items that were identified to have extreme item parameters. The item with the largest guessing parameter, which is a new item created based on

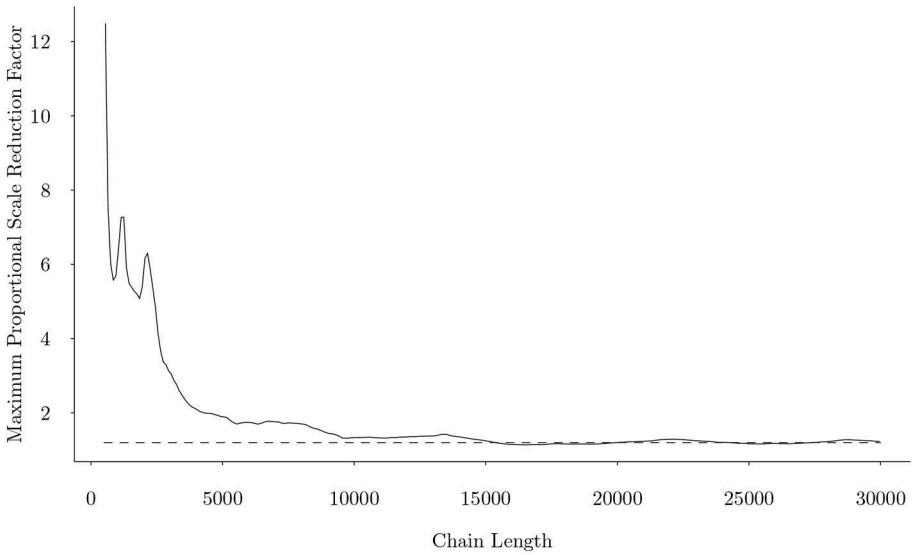


FIGURE 6 | The maximum univariate Gelman-Rubin proportional scale reduction factor from the joint model as a function of number of iterations, when uniform initial attribute patterns were used. Dotted line represents the cutoff of 1.2.

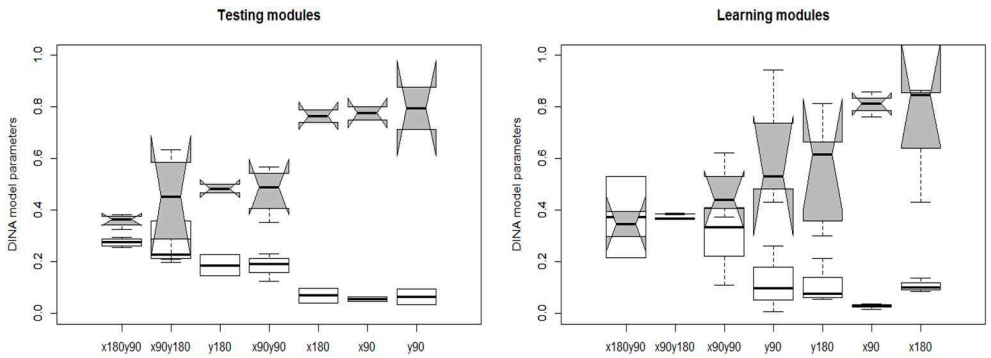


FIGURE 7 | The estimated DINA model item parameters. The white whisker diagram represents the slipping parameters *s* and the gray ones represent the guessing parameters *g*.

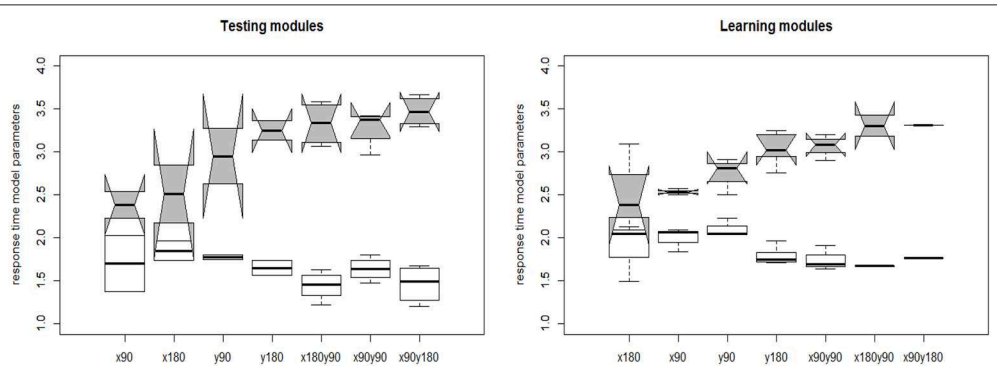


FIGURE 8 | The estimated response time model item parameters. The white whisker diagram represents the time discrimination parameter *a* and the gray ones represent the time intensity parameters γ.

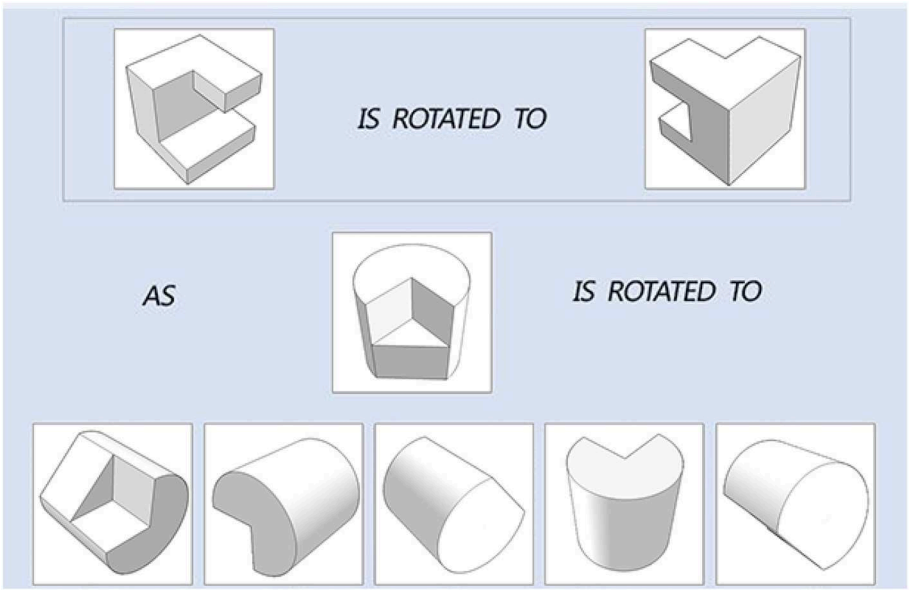


FIGURE 9 | The item (ID: N10) with largest guessing parameter *g*.

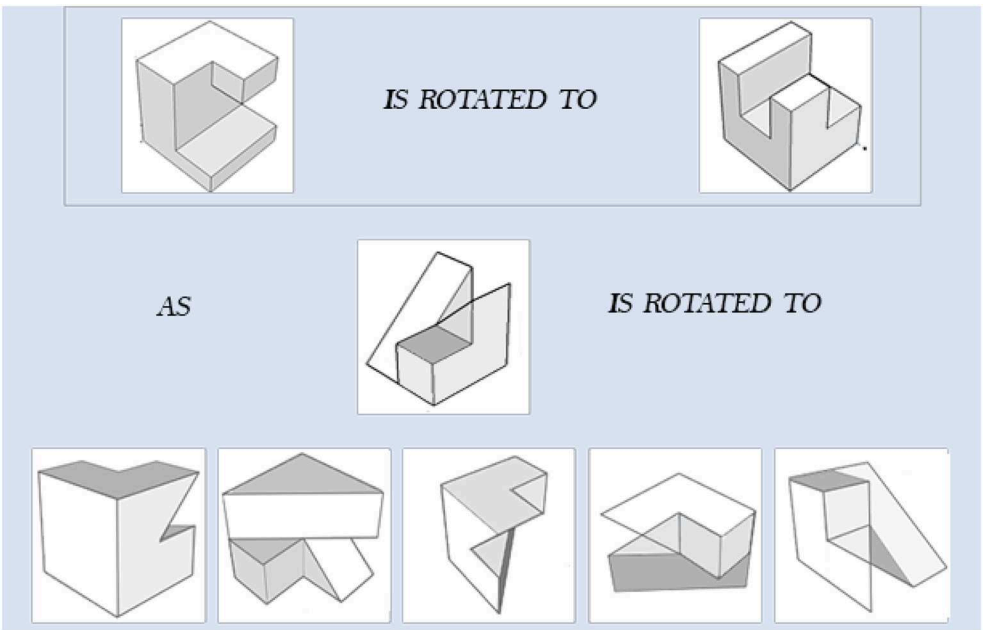


FIGURE 10 | The item (ID:30) with the largest slipping parameter *s*.

the revised PSVT: R, is presented in **Figure 9**. This item is also the one that has the largest item discrimination parameter. The reference object in this item measures the 180° rotation along *x* axis. If the participants can recognize the rotation is along the *x* axis, they can easily exclude the four distractors and select the correct option (the 4*th* one). It may be due to this reason, this item has the largest guessing probability. The distractors of this new item need to be further refined in the future to better diagnose the test-takers' rotation skills. For the current learning program, this item is the second question of the first learning module, thus the main function is to help test-takers learn the rotation. The item with the largest slipping parameter is presented in **Figure 10**. It has a relatively large time intensity parameter as well. It measures 180° rotation along the *x* axis, and 90° rotation along the *y* axis, and the object in this item has the most complex shape. Again, this item may not have a good

TABLE 7 | The estimated coefficients from the transition model.

Variable	Notation	Mean	*SD*	95% credible interval	
				Lower bound	Upper bound
θ	λ_{θ}	2.821*	0.703	1.443	4.199
Learned skills	λ_{α}	0.471*	0.217	0.046	0.896
Gender	λ_{g}	−0.350	0.346	−1.028	0.328
Country	λ_{c}	−0.225	0.296	−0.805	0.355
IV	λ_{IV}	0.165	0.355	−0.531	0.861
Strategy 1	λ_{st1}	−0.101	0.255	−0.601	0.399
Strategy 2	λ_{st2}	0.101	0.443	−0.767	0.969
Gender*IV	λ_{gI}	0.226	0.312	−0.386	0.838
Country*IV	λ_{cI}	−0.004	0.285	−0.563	0.555
IV*Strategy 1	λ_{Ist1}	0.216	0.233	−0.241	0.673
IV*Strategy 2	λ_{Ist2}	0.280	0.468	−0.637	1.197

*IV, Intervention; Dummy coding of the categorical variables: gender (female 1, male −1), Country (US 1, China −1), IV: colored (1), non-colored(-1), Strategy 1: (Compare Analytic Strategy and Holistic Strategy with Hybrid Strategy), Strategy 2: (Compare Analytic with Holistic Strategy); *$p < 0.05$.*

diagnostic function. However, in the current learning program, it is the last question in the second learning module, and the main purpose is to improve test-taker's rotation skills.

5.3.2.2. Reliability analysis for two testing modules

A reliability analysis was conducted to evaluate the two testing modules (module 1 and 4). In our study, classification consistency index (CCI; Cui et al., 2012) was chosen to estimate the test reliability. The CCI is the probability of classifying a randomly selected examinee consistently according to two administrations of a test. The range of CCI is between 0 and 1, and a higher values indicate a larger reliability. The CCI for module 1 and module 4 are 0.729 and 0.931.

5.4. Evaluation the Effectiveness of the Learning Program

The learning program is evaluated using the joint model results. The rest of this section reports the results from the dynamic response model and dynamic response time model portion of the joint model.

5.4.1. Dynamic Response Model Result

The estimated coefficients from the transition model are documented in **Table 7**. Based on the 95% creditable interval, only the general learning ability θ and the learned skills were statistically related to odds of the transition probability. This indicates that after controlling the latent variables and based on the response accuracy across different time points, the two learning interventions (colored and non-colored) have the same effectiveness in improving the spatial rotation skills across gender, country, and the rotation strategy.

Next, we evaluate the learning program by investigating the overall growth of spatial skills. The output from the dynamic response model indicates that at the initial time point, that is when the participants finished the first testing module and before they received the first learning module, 59.5% participants were estimated as mastery of four rotation skills. Because those participants had already mastered the four skills before receiving the learning modules, we excluded them from the following analysis to better evaluate the learning program. We refer the rest 222 participants who at least missed one rotation skill in the beginning as the non-masters. The overall effectiveness of the learning program is evaluated on summarizing the growth of the non-masters.

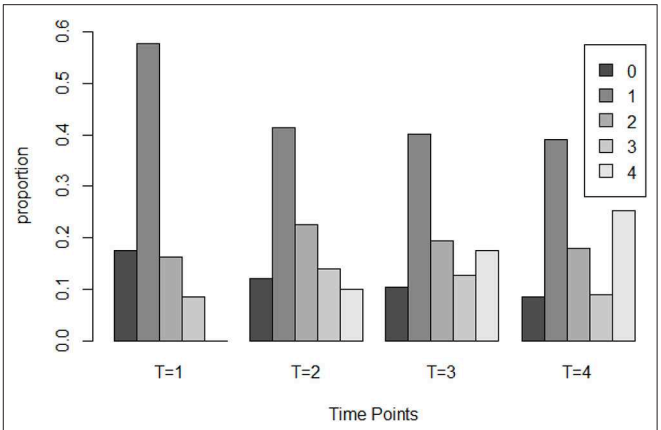


FIGURE 11 | The distribution of the number of mastered skills for non-master group at four time point.

5.4.1.1. The overall growth of non-masters

We first report a paired t test result that compares the test score from module 1 and module 4, as the items in these two modules have similar psychometric properties and can be treated as a pretest and a post-test. On average, for the non-masters, the module 4 test score ($M = 6.032, SD = 1.780$) is significantly higher than the module 1 test score ($M = 5.716, SD = 1.638$) and with a small to median effect size, $t(221) = 2.060, p = 0.04, r = 0.137$. Then the results from the dynamic response model using the item score in the four modules are explored. The overall learning trajectory, denoted as the distribution of the number of mastered skills at each time point, is documented in **Figure 11**. From there we can observe a "growth" of the rotation skill as the number of non-masters who mastered none of the skills reduced from 17.6% in the beginning of the experiment to 8.5% at the end of the experiment. There are also about 25.2% non-masters mastered four skills in the end. **Table 8** further documents the proportion of people who mastered each skill after test module 1 and 4. The results from a χ^2 test that compares the paired proportion indicates a significant increase of mastery for each skill with medium effect size (Cohen's h). This demonstrates the newly developed learning program can significantly improve the non-masters' four spatial rotation skills.

Next, we further investigate how the learning trajectory is influenced by the general learning ability θ. Based on the results from the transition model (**Table 7**), we can conclude that for a specific rotation skill, the odds of transition from non-mastery to mastery is significantly positively related to the general learning ability θ, ($\hat{\lambda}_{\theta} = 2.821, p < 0.05$) and the

TABLE 8 | The Skill Mastery Rate (proportion of participants that master each skill).

Skill	Time 1	Time 4	Difference	*p*-value	Cohen's *h*
*x*90	0.812	0.876	0.064	< 0.01	0.290
*x*180	0.648	0.761	0.113	< 0.01	0.374
*y*90	0.761	0.836	0.075	< 0.01	0.301
*y*180	0.628	0.730	0.102	< 0.01	0.322

Module 1 and Module 4 represent time 1 and time 2.

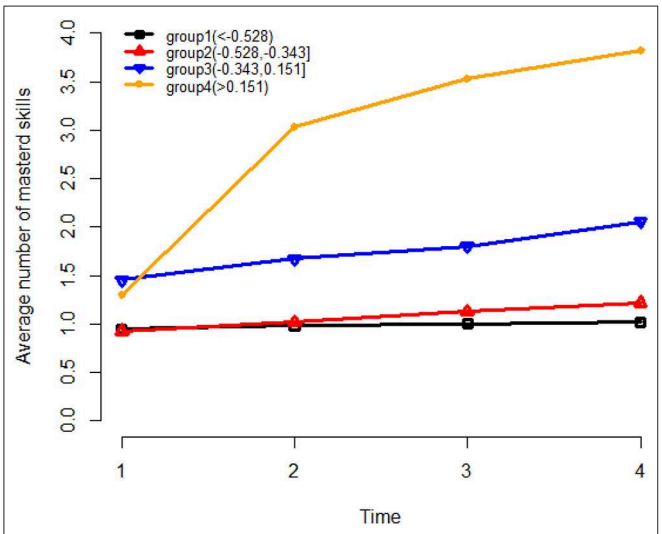


FIGURE 12 | The average number of mastered skills at each time point for four learning ability groups. −0.528, −0.343, and 0.151 are the 1st, 2nd, and 3rd quartile of θ values.

number of mastered skills ($\hat{\lambda}_\alpha = 0.471, p < 0.05$). To further explore these two variables, the non-masters were divided into four groups based on their estimated general learning ability θ. For each group, the number of mastered skills at each stage of the experiment was investigated. The three cut off points were selected as the 1st (−0.528), 2nd (−0.343), and 3rd quantile (0.151) of the estimated general learning ability so that group 1 consists of participants with the lowest learning ability and group 4 consists of participants with the highest learning ability. **Figure 12** presents the average number of mastered skills at each time point for each of the four groups. From there we can see that, for the participants with low learning ability (group 1), their learning rate was the lowest. While for the high learning ability participants (group 4), the learning rate is the highest (starts with around 1.5 skills and can master more 3–4 skills). This figure also illustrates how the learned skills can help learn the un-mastered skills. For the participants starting with more than one skills (group 3 and group 4), they learned much faster than the participants starting with 1 or <1 skill (group 1 and group 2).

5.4.2. Dynamic Response Time Model Result

The estimated coefficients for covariates (ϕ) in the dynamic response time model are presented in **Table 9**. First, on average, the participants who mastered the required skills for an item spent 1.38 s more on completing this question, compared with

TABLE 9 | The estimated ϕs from the response time model.

Variable	Notation	Mean	*SD*	95% credible interval	
				Lower bound	Upper bound
$G(\alpha_i(t), q_j)$	ϕ_α	−0.327*	0.028	−0.382	−0.272
Gender	ϕ_g	−0.081*	0.017	−0.114	−0.048
Country	ϕ_c	0.084*	0.016	0.053	0.115
IV	ϕ_{IV}	0.072*	0.023	0.027	0.117
Strategy1	ϕ_{st1}	0.034*	0.012	0.010	0.058
Strategy2	ϕ_{st2}	−0.093*	0.033	−0.158	−0.028
Gender*IV	ϕ_{gI}	0.017	0.018	−0.018	0.052
Country*IV	ϕ_{cI}	0.017	0.016	−0.014	0.048
IV*Strategy1	ϕ_{Ist1}	0.011	0.013	−0.014	0.036
IV*Strategy2	ϕ_{Ist2}	0.038	0.032	−0.025	0.101

*IV, Intervention; Dummy coding of the categorical variables: gender (female 1, male −1), Country (US 1, China −1), IV: colored (1), non-colored(-1), Strategy 1: (Compare Analytic Strategy and Holistic Strategy with Hybrid Strategy), Strategy 2: (Compare Analytic with Holistic Strategy); *$p < 0.05$.*

those who did not master all the required skills ($\hat{\phi}_\alpha = 0.327, p < 0.05$). Given the participants who had the same learning trajectory, the male participants completed a question faster than female participants ($\hat{\phi}_g = -0.081, p < 0.05$); the participants in US completed a question faster than participants from China ($\hat{\phi}_c = 0.084, p < 0.05$); the participants using colored intervention completed a question faster than participants using non-colored intervention ($\hat{\phi}_{IV} = 0.072, p < 0.05$); and finally, the average response time of participants who used analytic strategy and who used holistic strategy were shorter than the one who used a combined strategy ($\hat{\phi}_{st1} = 0.034, p < 0.05$), and the participants using a holistic strategy completed a question faster than participants using an analytic strategy ($\hat{\phi}_{st2} = -0.093$, $p < 0.05$).

5.4.3. Survey Questions for Validation

According to the survey collected at the end of experiment, 68% participants rated greater or equal to 3 regarding the questions, "Do you think the learning program is helpful or not." This question used the 5 points Likert scale with 1 indicates "not very helpful" and 5 denotes "very helpful."

6. DISCUSSION

This study investigated the possibility of developing a learning program that integrates a multidimensional diagnostic assessment with two different learning interventions with the purpose to diagnose and improve the 3-D mental rotation skills. The program was evaluated through an experiment paired with the statistical analysis from a joint model of response accuracy and response times. Compared with the traditional assessment on spatial skills, where the tests are timed and number correct is reported as a measure for test-takers' performances, the proposed diagnostic assessment through the analysis from the joint model can provide an informative estimate of the learning trajectory for each participant in terms of the strengths and weaknesses in four fine-grained mental rotation skills over

time. The response times are also utilized to discover additional information about learning across different covariates. While the earlier study (Wang et al., 2017) provided initial evidence of the effectiveness of building a multidimensional diagnostic assessment with training tools, the present study improved the assessment and learning intervention design and evaluated the newly developed program by investigating the effectiveness of two interventions across gender, country and rotation strategy.

The results from the joint learning model demonstrated that learning of a specific rotation skill is significantly related to a general learning ability and the mastered skills. **Figure 7** illustrates that it is difficult for test-takers who mastered none of four rotation skills to improve over a short time training. **Table 8** indicates the learning of the four rotation skills may follow a hierarchical structure, as the $x90^{\circ}$ rotation might be the easiest one to learn and $y180^{\circ}$ is the most difficult to learn. Thus, to train the test-takers with extremely low spatial ability, it's better to start with a relatively simple and single rotation then transfer to more complex task. This in fact supports the current learning program that first provides an easy learning module then a more challenging one.

However, the current learning program is not adaptive, meaning all the participants received the same learning modules. The results from this study can guide a future design of the adaptive intervention that targets at the weakness of the specific spatial skill and provide the appropriate learning materials. In addition, the output from the dynamic response model portion of the joint model indicates the learning programs with the two designed interventions had the same effectiveness to improve the response accuracy across gender, country and rotation strategy. However, the dynamic response time model reveals the speed difference between the female and male participants, participants using colored and non-colored intervention and participants using three different rotation strategies. Such additional information from the dynamic response time are also helpful in designing an adaptive learning system in the future.

The output of the item parameter estimations from the joint learning model provides new insights into the revised PSVT: R test questions as well. As reviewed in the beginning of this paper, the PSVT:R and revised PSVT: R test questions have been used in many research studies and in generally were reported to have high reliability. The item parameters estimation from the joint learning model indicates that some test questions, especially the ones measure a simple rotation skill can have large guessing parameter, and the ones with complex object and combination of multiple difficult rotation skills may not have good diagnostic information to differentiate the participates with low spatial ability from those with high spatial ability. Carefully examining the distractors may improve their diagnostic functioning. Lastly, another important component in the joint model, is the Q matrix which links the items and the measured attributes. The correct inference from the joint model and the diagnostic assessment relies on how accurate the Q matrix is. The current study used the Q matrix from a previous study, which was mainly specified based on subject experts' opinions. An exploratory clustering method was used to validate the Q matrix, using both response times and response accuracy. It was found that attributes defined in the Q matrix did not contain the information about the degree of complexity of the objects. In the future, we will further validate this Q matrix using many recent techniques in psychometrics (e.g., Chen et al., 2018a).

DATA AVAILABILITY STATEMENT

The real data set analyzed in this article is not publicly available, because it is part of an ongoing research project from the research team. Requests to access the dataset should be directed to SW, swang44@uga.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board, Office of the Vice President for Research at University of Georgia (IRB ID: STUDY00004215). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SW contributed to design the learning program, conducted the experiment, data analysis and draft, and revised the manuscript. YH and BW contributed to design the learning program, conducted the experiment, and drafted the part of the manuscript. QW contributed to help with the design of learning program and conducted the experiment, and also helped to draft the introduction section. YS contributed to the part of data analysis and drafted the part of the manuscript. MC contributed to provide suggestions on designing the learning program.

FUNDING

The development of the spatial learning program was supported by the Junior Faculty Seed Grant from the Owens Institute for Behavioral Research, The University of Georgia. The statistical analysis of this research was supported by 2019 NAEd/Spencer Post-doctoral Research Fellowship Program.

REFERENCES

Alkhateeb, H. M. (2004). Spatial visualization of undergraduate education majors classified by thinking styles. *Percept. Motor Skills* 98, 865–868. doi: 10.2466/pms.98.3.865-868

Armstrong, R. D., Jones, D. H., and Wu, I. -L. (1992). An automated test development of parallel tests from a seed test. *Psychometrika* 57, 271–288. doi: 10.1007/BF02294509

Bolshakova, N., and Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Process.* 83, 825–833. doi: 10.1016/S0165-1684(02)00475-9

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

Branoff, T. J. (2000). Spatial visualization measurement: a modification of the purdue spatial visualization test-visualization of rotations. *Eng. Des. Graph. J.* 64, 14–22.

Brownlow, S., and Miderski, C. A. (2001). *How Gender and College Chemistry Experience Influence Mental Rotation Ability*. Atlanta, GA: ERIC.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York, NY: Cambridge University Press.

Chen, Y., Culpepper, S. A., Chen, Y., and Douglas, J. (2018a). Bayesian estimation of the dina q matrix. *Psychometrika* 83, 89–108. doi: 10.1007/s11336-017-9579-4

Chen, Y., Culpepper, S. A., Wang, S., and Douglas, J. (2018b). 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

Cheng, Y.-L., and Mix, K. S. (2014). Spatial training improves children's mathematics ability. *J. Cogn. Dev.* 15, 2–11. doi: 10.1080/15248372.2012.725186

Chiu, C.-Y., and Köhn, H.-F. (2015). The reduced RUM as a logit model: parameterization and constraints. *Psychometrika* 81, 350–370. doi: 10.1007/s11336-015-9460-2

Cui, Y., Gierl, M. J., and Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Educ. Meas.* 49, 19–38. doi: 10.1111/j.1745-3984.2011.00158.x

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

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

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

de la Torre, J., and Ma, W. (2016). "Cognitive diagnosis modeling: a general framework approach and its implementation in R," in *A Short Course at the Fourth Conference on Statistical Methods in Psychometrics* (New York, NY: Columbia University).

Eliot, J. (2012). *Models of Psychological Space: Psychometric, Developmental, and Experimental Approaches*. Springer Science & Business Media.

Geiser, C., Lehmann, W., Corth, M., and Eid, M. (2008). Quantitative and qualitative change in children's mental rotation performance. *Learn. Individ. Differ.* 18, 419–429. doi: 10.1016/j.lindif.2007.09.001

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

Glück, J., Machat, R., Jirasko, M., and Rollett, B. (2002). Training-related changes in solution strategy in a spatial test: an application of item response models. *Learn. Individ. Differ.* 13, 1–22. doi: 10.1016/S1041-6080(01)00042-5

Guay, R. (1976). *Purdue Spatial Visualization Test*. Lafayette, IN: Purdue University.

Guay, R. (1980). "Spatial ability measurement: a critique and an alternative," in *the Annual Meeting of the American Educational Research Association* (Boston, MA).

Hämäläinen, J., Jauhiainen, S., and Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms* 10:105. doi: 10.3390/a10030105

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality* (Ph.D. thesis), University of Illinois at Urbana-Champaign, Champaign, IL, United States.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. *DMKD* 3, 34–39.

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kaya, Y., and Leite, W. L. (2016). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling an evaluation of model performance. *Educ. Psychol. Meas.* 77, 369–388. doi: 10.1177/0013164416659314

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

Li, H., Hunter, C. V., and Lei, P.-W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Lang. Test.* 33, 391–409. doi: 10.1177/0265532215590848

Li, Y., and O'Boyle, M. (2013). How sex and college major relate to mental rotation accuracy and preferred strategy: an electroencephalographic (EEG) investigation. *Psychol. Rec.* 63, 27–42. doi: 10.11133/j.tpr.2013.63.1.003

Li, Y., O'Boyle, M. W., Liu, L., Zeng, X., Zhang, J., Zhu, J., et al. (2014). The influence of native acquisition of chinese on mental rotation strategy preference: an EEG investigation. *Psychol. Rec.* 64, 321–328. doi: 10.1007/s40732-014-0028-9

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

Ma, W., de la Torre, J., Sorrel, M., and Jiang, Z. (2019). *GDINA: The Generalized DINA Model Framework*. R package version 2.7. Retrieved from: http://CRAN.R-project.org/package=GDINA

Maeda, Y., and Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: visualization of rotations (PSVT:R). *Educ. Psychol. Rev.* 25, 69–94. doi: 10.1007/s10648-012-9215-x

Maeda, Y., Yoon, S. Y., Kim-Kang, G., and Imbrie, P. (2013). Psychometric properties of the revised PSVT:R for measuring first year engineering students' spatial ability. *Int. J. Eng. Educ.* 29, 763–776.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Perry, P. C. (2013). *Influences on visual spatial rotation: science, technology, engineering, and mathematics (STEM) experiences, age, and gender* (Ph.D. thesis), Notre Dame of Maryland University, Baltimore, MD, United States.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assess.* 34, 782–799. doi: 10.1177/0734282915623053

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *Int. J. Comput. Commun.* 5, 27–34.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65.

Stieff, M., Dixon, B. L., Ryu, M., Kumi, B. C., and Hegarty, M. (2014). Strategy training eliminates sex differences in spatial problem solving in a stem domain. *J. Educ. Psychol.* 106:390. doi: 10.1037/a0034823

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *J. Educ. Behav. Stat.* 10, 55–73.

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

Templin, J. L., and Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educ. Meas. Issues Pract.* 32, 37–50. doi: 10.1111/emip.12010

Thompson, J. M., Nuerk, H.-C., Moeller, K., and Kadosh, R. C. (2013). The link between mental rotation ability and basic numerical representations. *Acta Psychol.* 144, 324–331. doi: 10.1016/j.actpsy.2013.05.009

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., et al. (2013). The malleability of spatial skills: a meta-analysis of training studies. *Psychol. Bull.* 139:352. doi: 10.1037/a0028446

Voyer, D., Voyer, S., and Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychol. Bull.* 117:250.

Wang, D., Gao, X., Cai, Y., and Tu, D. (2019a). Development of a new instrument for depression with cognitive diagnosis models. *Front. Psychol.* 10:1306. doi: 10.3389/fpsyg.2019.01306

Wang, L., and Carr, M. (2014). Working memory and strategy use contribute to gender differences in spatial ability. *Educ. Psychol.* 49, 261–282. doi: 10.1080/00461520.2014.960568

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2017). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

Wang, S., Zhang, S., Douglas, J., and Culpepper, S. (2018). Using response times to assess learning progress: a joint model for responses and response times. *Meas. Interdiscipl. Res. Perspect.* 16, 45–58. doi: 10.1080/15366367.2018.1435105

Wang, S., Zhang, S., and Shen, Y. (2019b). A joint modeling framework of responses and response times to assess learning outcomes. *Multivar. Behav. Res.* 55, 49–68. doi: 10.1080/00273171.2019.1607238

Weckbacher, L. M., and Okamoto, Y. (2014). Mental rotation ability in relation to self-perceptions of high school geometry. *Learn. Individ. Differ.* 30, 58–63. doi: 10.1016/j.lindif.2013.10.007

Weiss, E., Siedentopf, C., Hofer, A., Deisenhammer, E., Hoptman, M., Kremser, C., et al. (2003). Sex differences in brain activation pattern during a visuospatial cognitive task: a functional magnetic resonance imaging study in healthy volunteers. *Neurosci. Lett.* 344, 169–172. doi: 10.1016/S0304-3940(03)00406-3

Xu, G., and Shang, Z. (2017). Identifiability of latent structures in restricted latent class models. *J. Am. Stat. Assoc.* 113, 1284–1295. doi: 10.1080/01621459.2017.1340889

Yoon, S. Y. (2011). *Psychometric properties of the revised Purdue spatial visualization tests: visualization of rotations (the revised PSVT-R)* (Ph.D. thesis), Purdue University, West Lafayette, IN, United States.

Yoon, S. Y., and Mann, E. L. (2017). Exploring the spatial ability of undergraduate students: association with gender, stem majors, and gifted program membership. *Gift. Child Q.* 61, 313–327. doi: 10.1177/0016986217722614

Yue, J. (2006). "Spatial visualization by isometric drawing," in *Proceedings of the 2006 IJMEINTERTECH Conference* (Union, NJ).

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

Zhang, S., and Wang, S. (2018). Modelling learner heterogeneity: a mixture learning model with responses and response times. *Front. Psychol.* 9:2339. doi: 10.3389/fpsyg.2018.02339

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.

APPENDIX

TABLE 1A | The Wald test results for selecting response measurement model.

Item	Model	*p*-value	Adjusted *p*-value
Item 1	GDINA	NA	NA
Item 2	GDINA	NA	NA
Item 3	GDINA	NA	NA
Item 4	DINA	0.3407	1
Item 5	GDINA	NA	NA
Item 6	DINO	0.1682	1
Item 7	DINO	0.3288	1
Item 8	RRUM	0.5466	1
Item 9	RRUM	0.3861	1
Item 10	DINA	0.3447	1
Item 11	GDINA	NA	NA
Item 12	GDINA	NA	NA
Item 13	GDINA	NA	NA
Item 14	GDINA	NA	NA
Item 15	GDINA	NA	NA
Item 16	GDINA	NA	NA
Item 17	GDINA	NA	NA
Item 18	GDINA	NA	NA
Item 19	GDINA	NA	NA
Item 20	DINA	0.3018	1
Item 21	GDINA	NA	NA
Item 22	GDINA	NA	NA
Item 23	DINA	0.2031	1
Item 24	GDINA	NA	NA
Item 25	GDINA	NA	NA
Item 26	GDINA	NA	NA
Item 27	ACDM	0.9331	1
Item 28	DINA	0.6384	1
Item 29	RRUM	0.088	0.792
Item 30	RRUM	0.0579	0.6374
Item 31	GDINA	NA	NA
Item 32	GDINA	NA	NA
Item 33	GDINA	NA	NA
Item 34	GDINA	NA	NA
Item 35	DINO	0.7287	1
Item 36	DINA	0.8787	1
Item 37	DINA	0.0974	1
Item 38	DINA	0.7211	1
Item 39	DINA	0.0881	1
Item 40	DINA	0.1309	1

The p-value and adjusted p-value are from the Wald test between the selected reduced DCM model and GDINA model. Thus, those values are NA for the items are best fitted with GDINA model.

frontiers in Psychology

METHODS
published: 28 February 2020
doi: 10.3389/fpsyg.2020.00224

Attribute Discrimination Index-Based Method to Balance Attribute Coverage for Short-Length Cognitive Diagnostic Computerized Adaptive Testing

Yutong Wang[1], *Xiaojian Sun*[2], *Weifeng Chong*[1] *and Tao Xin*[1]*

[1] *Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China,*
[2] *School of Mathematics and Statistics, Southwest University, Chongqing, China*

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, United States

Reviewed by:
Yong Luo,
Educational Testing Service, United States
Wenchao Ma,
University of Alabama, United States
Zhehan Jiang,
University of Alabama, United States

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

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

Received: *04 August 2019*
Accepted: *31 January 2020*
Published: *28 February 2020*

Citation:
Wang Y, Sun X, Chong W and Xin T (2020) Attribute Discrimination Index-Based Method to Balance Attribute Coverage for Short-Length Cognitive Diagnostic Computerized Adaptive Testing. Front. Psychol. 11:224. doi: 10.3389/fpsyg.2020.00224

We propose a new method that balances attribute coverage for short-length cognitive diagnostic computerized adaptive testing (CD-CAT). The new method uses the attribute discrimination index (ADI-based method) instead of the number of items that measure each attribute [modified global discrimination index (MGDI)-based method] to balance the attribute coverage. Therefore, the information that each attribute provides can be captured. The purpose of the simulation study was to evaluate the performance of the new method, and the results showed the following: (a) Compared with uncontrolled attribute-balance coverage method, the new method produced a higher mastery pattern correct classification rate (PCCR) and attribute correct classification rate (ACCR) with both the posterior-weighted Kullback–Leibler (PWKL) and the modified PWKL (MPWKL) item selection method. (b) Equalization of ACCR (E-ACCR) based on the ADI-based method leads to better results, followed by the MGDI-based method. The uncontrolled method leads to the worst results regardless of item selection methods. (c) Both the ADI-based and MGDI-based methods produced acceptable examinee qualification rates, regardless of item selection methods, although they were relatively low for the uncontrolled condition.

Keywords: balance attribute coverage, cognitive diagnostic computerized adaptive testing, attribute discrimination index, equalization of attribute correct classification rate, examinee qualification rate

INTRODUCTION

Cognitive diagnostic assessment (CDA) has become popular in test theory research in recent years, which is developed to measure the cognitive skills of examinees (Leighton and Gierl, 2007; Gierl et al., 2008). Compared with classical test theory (CTT) and the most commonly used unidimensional item response theory (UIRT), which only provide overall scores to examinees, and multidimensional item response theory (MIRT), which provides both overall score and subscale scores, CDA can provide more detailed information about strengths and weaknesses of examinees for a specific content domain, so that administrators can identify whether or not examinees possess the attributes (Yao and Boughton, 2007; Lee et al., 2012). Evidence should be obtained of model

fit when IRT models are used in real test data, and it is the same with CDA models (Sinharay and Haberman, 2014). Otherwise, the misfit of models may lead to a misleading conclusion.

Computerized adaptive testing (CAT) combines test theory with computer technology to improve testing efficiency, which has become a promising method in psychological and educational measurement. CAT can provide equivalent or even higher measurement accuracy of examinees' latent skills, with reductions in test length of up to 50%, compared with traditional paper-and-pencil tests (Weiss, 1982). Further, items administered in the test are matched with examinees' estimated latent trait level (Mao and Xin, 2013; Chang, 2015). Recently, to maximize the benefits of both CDA and CAT, researchers have attempted to combine CDA with CAT and named it cognitive diagnostic CAT (Xu et al., 2003; McGlohen and Chang, 2008; Cheng, 2009a; CD-CAT).

In CD-CAT, many factors can affect the reliability and validity of the test, one of which is the balance of attribute-level coverage (Cheng, 2010; Mao and Xin, 2013). Cheng (2010) pointed out that it is very important to make sure that each attribute in the test has been measured adequately or the reliability of the test will not be reduced. Furthermore, test validity will be at risk because of inadequate attribute coverage (Cheng, 2010). To balance attribute coverage in CD-CAT, Cheng (2010) developed the modified maximum global discrimination index (MMGDI) to build the item selection method. The MMGDI method is based on the global discrimination index (GDI) developed by Xu et al. (2003). The mechanism of MMGDI is to accumulate the Kullback–Leibler (KL) information between conditional distribution given estimated pattern profile and conditional distribution given each of all possible candidate pattern profiles. However, there is a problem that the GDI method eliminates the coverage at the attribute level. To overcome that shortcoming, the MMGDI method uses the maximum priority index (MPI) method to balance attribute coverage (Cheng and Chang, 2009). In the simulation study, Cheng (2010) showed that the new item selection method not only improved the attribute correct classification rate (ACCR) and the rate of attribute master pattern (AMP) but also improved the validity of the test.

The findings from Cheng (2010) indicated that the correct classification rate had increased when the number of items measuring each attribute is adequate, which implied that there is a positive correlation between the numbers of items measuring each attribute and the correct classification rate. However, Finkelman et al. (2009) pointed out that, in some situations, even if the test contained adequate numbers of items to measure each attribute, different measurement accuracy could occur across the attributes. In other words, the number of items measuring each attributes maybe not the essential factor that affects the measurement accuracy of latent skills.

Note that based on the information that each item provided, CAT can produce accurate estimates of latent skills with lesser items. We can infer that the information each item provided may be the essential factor that affects the accuracy of latent skills and affects the attribute measurement precision. Consequently, we investigated the argument whether the information that each attribute provided can be utilized as the index to balance attribute coverage.

The purpose of the current study is to explore a new method based on the information provided by each attribute, instead of the number of items used in the test to measure each attribute in CD-CAT. The major benefit of this approach is to balance the attribute coverage in a short-length test. There are several reasons for choosing a short-length test: First, CDAs can be used to design as low-stake testing, and they help teachers or administrators to understand the performance of students and thus determine what should be done to improve the students' performance (Roussos et al., 2007; Hartz and Roussos, 2008; Mao and Xin, 2013; Kaplan et al., 2015). As a consequence, cognitive diagnostic tests would be conducted more frequently than traditional tests in some areas such as interim assessment (Roussos et al., 2007; Hartz and Roussos, 2008; Mao and Xin, 2013; Kaplan et al., 2015). When CD-CAT is applied to interim assessment, the AMPs of students should be obtained with short-length tests (Zheng and Chang, 2016). Second, to the best of our knowledge, among the studies focused on short-length test, there are only two applied that CD-CAT. The first one is practiced by Wang (2013), who introduced the mutual information (MI) item selection method in CD-CAT. And the second one is practiced by Zheng and Chang (2016), who developed two high-efficiency algorithms to select items in CD-CAT. But no study appears to have considered the situation that balances attribute coverage in the test.

The remainder of the present paper is organized as follows. The *Reduced Reparameterized Unified Model* section introduces the cognitive diagnostic model (CDM) that we have used in this study. The *Item Selection Methods* section presents two chosen methods, PWKL and MPWKL information for CD-CAT. After that, we introduce two methods to balance attribute coverage: one is to balance the number of items that measures each attribute and the other one is to balance the information that each attribute provides. In a further section, we report the results of a simulation study to evaluate the performance of the novel balanced attribute coverage method.

REDUCED REPARAMETERIZED UNIFIED MODEL

We used the reduced reparameterized unified model (RRUM) in the current study (Hartz, 2002), because previous studies have demonstrated that its prototype, the RUM, is very useful for formative assessment in practice (Jang, 2005; Wang et al., 2011). RRUM has gained more attention for educational assessment by researchers in recent years (Kim, 2011; Feng et al., 2013; Chiu et al., 2016). Chiu et al. (2016) also pointed out that RRUM has more flexibility than the "deterministic inputs, noisy 'and' gate" (DINA) model proposed by Junker and Sijtsma (2001). The item response function of the RRUM can be written as

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

where $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$ is the AMP of examinee i; η_i is the residual ability parameter of examinee i, which represents the latent trait account for attributes that are not included in the Q-matrix (McGlohen and Chang, 2008); K is the number of attributes. π_j^* represents the probability that examinee i possesses all of the required attributes for item j and correctly applies them, which is formulated as $\pi_j^* = \prod_{k=1}^{K} \pi_{jk}^{q_{jk}}$. And r_{jk}^* represents the ratio that examinee i lacks attribute k but correctly applies it to item j, which can be written as $P\left(Y_{ijk} = 1|\alpha_{ik} = 0\right)$, and examinee i possesses attribute k and correctly applies it to item j, which can be written as $P\left(Y_{ijk} = 1|\alpha_{ik} = 1\right)$, so r_{jk}^* can be described as

$$r_{jk}^* = \frac{P\left(Y_{ijk} = 1|\alpha_{ik} = 0\right)}{P\left(Y_{ijk} = 1|\alpha_{ik} = 1\right)}, \tag{2}$$

where q_{jk} is the attribute that item j measured, and $q_{jk} = 1$ presents if item j measures attribute k, otherwise $q_{jk} = 0$.

ITEM SELECTION METHODS

Posterior-Weighted Kullback–Leibler Information Method

KL information assumes that all candidate AMPs, α_c, share $\frac{1}{2^K}$ probabilities equally that belong to the true AMP for each examinee at each step of item selection. Cheng (2009a,b) commented that this assumption was unnecessary and may lead to low test efficiency. Cheng also pointed out that different candidate AMPs should have different probabilities to be the true AMP, and then he proposed a new item selection method that considered the posterior probability of examinees' responses. That modified approach was termed PWKL information:

$$PWKL_j\left(\hat{\alpha}\right) = \sum_{c=1}^{2^K}\left\{\left[\sum_{x=0}^{1}\log\left(\frac{P\left(X_j = x|\hat{\alpha}\right)}{P\left(X_j = x|\alpha_c\right)}\right)P\left(X_j = x|\hat{\alpha}\right)\right]L\left(\alpha_c|X_{t-1}\right)\right\}, \tag{3}$$

and

$$L\left(\alpha_c|X_{t-1}\right) \propto \left(\prod_{j=1}^{t-1} P\left(x_j = 1|\alpha_c\right)^{x_j}\left[1 - P\left(x_j = 1|\alpha_c\right)\right]^{1-x_j}\right)p\left(\alpha_c\right),$$

where $L\left(\alpha_c|X_{t-1}\right)$ is the likelihood function, X_{t-1} is response vector of $t-1$ items, and $p\left(\alpha_c\right)$ is the prior distribution of α_c. The item t will be selected for a specific examinee with maximum PWKL information. Simulation studies have shown that PWKL information outperformed KL information and Shannon entropy (SHE) algorithms in most aspects (Cheng, 2009a,b; Wang, 2013).

Modified Posterior-Weighted Kullback–Leibler Information Method

The MPWKL method modifies the PWKL method to lead to a more reasonable result, especially in short-length test (Kaplan et al., 2015). The PWKL method uses point estimate, whereas the MPWKL method uses the entire posterior distribution. Thus, more information can be gained from the MPWKL than the PWKL method. The MPWKL information method is shown as follows:

$$MPWKL_{ij} = \sum_{d=1}^{2^k}\left\{\sum_{c=1}^{2^k}\left[\sum_{x=0}^{1}\log\left(\frac{P\left(X_{ij} = x|\alpha_d\right)}{P\left(X_{ij} = x|\alpha_c\right)}\right)P\left(X_{ij} = x|\alpha_d\right)\pi\left(\alpha_c|X_{n-1}\right)\right]\pi\left(\alpha_c\middle|X_{n-1}\right)\right\}. \tag{4}$$

METHODS FOR BALANCING ATTRIBUTE COVERAGE

Balance Attribute Coverage Based on Number of Items That Measure Each Attribute

Cheng and Chang (2009) introduced the MPI method to select items to meet the constraints in IRT-based CAT. Later, Cheng (2010) extended the MPI method to CD-CAT for balancing attribute coverage. The definition of the attribute-balance index (ABI) is

$$ABI_j = \prod_{k=1}^{K}\left(\frac{B_k - b_k}{B_k}\right)^{q_{jk}}, \tag{5}$$

where B_k is the lower bound of the number of items required to measure attribute k, b_k is the number of items measuring attribute k that has already been selected, and q_{jk} is the element of Q-matrix. The value of ABI is non-negative. By combining ABI and PWKL information methods, the modified global discrimination index (MGDI) is formulated as

$$MGDI_j = ABI_j * PWKL\left(\hat{\alpha}\right) = \prod_{k=1}^{K}\left(\frac{B_k - b_k}{B_k}\right) * PWKL_j\left(\hat{\alpha}\right) \tag{6}$$

An item with maximum MGDI will be administered as the next item for a specific examinee. Cheng (2010) named it maximum MGDI (MMGDI) item selection method. It is worth noting that the MMGDI method will be used to select the next item if ABI is larger than 0; otherwise, the PWKL information method will be used. When $q_{jk} = 0$, which means item j does not measure attribute k, then $\left[\frac{(B_k - b_k)}{B_k}\right]^{q_{jk}} = 1$, which does not affect MGDI_j.

Balance Attribute Coverage Based on Attribute Discrimination Index

As mentioned in the *Introduction*, in some situations, even though adequate items are used to measure each attribute, the estimated accuracy may differ across attributes (Finkelman et al.,

2009). The number of items measuring each attribute may be the necessary condition to improve the AMPs' accuracy. However, the information that each attribute provides may also be an essential factor to increase the test accuracy. Therefore, not only measuring each attribute with the number of items but also information that each attribute provides can be used to balance attribute coverage.

Henson et al. (2008) developed the attribute discrimination index (ADI) to compute the information each attribute provided. Then Finkelman et al. (2010) developed a binary programming method based on ADI to assemble tests automatically for CDM. ADI aims to compute the expected KL information between any two AMPs, with all the attributes holding constant except the target attribute, within the ideal response pattern (IRP; Tatsuoka, 1995). Considering that the test that measures K attributes will produce $2^K(2^K-1)$ possible comparisons regardless of hierarchy among attributes, a $(2^K \times 2^K)$ matrix $\mathbf{D}_j$ will be used to contain all these values. $\mathbf{D}_j$ can be written as follows:

$$D_{juv} = E_{\alpha_u}\left[\log\left(\frac{P_{\alpha_u}\left(x_j\right)}{P_{\alpha_v}\left(x_j\right)}\right)\right] = P_{\alpha_u}(1)\log\left(\frac{P_{\alpha_u}(1)}{P_{\alpha_v}(1)}\right) + P_{\alpha_u}(0)\log\left(\frac{P_{\alpha_u}(0)}{P_{\alpha_v}(0)}\right), \quad (7)$$

where $P_{\alpha_u}\left(x_j\right)$ and $P_{\alpha_v}\left(x_j\right)$ are response probabilities of item j given AMPs α_u and α_v, respectively. D_{juv} represents the degree to which a master (non-master) differed from non-master (master) for the target attribute (Henson et al., 2008).

There are $2^{(K-1)}$ comparisons of AMPs that differ only for the target attribute k. Note that the KL information between two AMPs is not symmetric. Therefore, two ADIs can be calculated for item j: one is the power that discriminates the master from non-master for the target attribute and the other one discriminates the non-master from master. The formulations of these two ADIs are

$$ADI_{jk1} = \sum_{\alpha_u,\alpha_v \in\, \Omega_1} \omega_{k1} D_{juv}, \quad (8)$$

$$ADI_{jk0} = \sum_{\alpha_u,\alpha_v \in\, \Omega_0} \omega_{k0} D_{juv}, \quad (9)$$

where $\omega_{k1} = p\left(\alpha_u|\alpha_k = 1\right)$, $\Omega_{k1} \equiv \left\{\alpha_{uk} = 1 \; and \; \alpha_{vk} = 0 \; and \; \alpha_{um} = \alpha_{vn} \; \forall \; m \neq n\right\}$, and $\omega_{k0} = p\left(\alpha_u|\alpha_k = 0\right)$, $\Omega_{k0} \equiv \left\{\alpha_{uk} = 0 \; and \; \alpha_{vk} = 1 \; and \; \alpha_{um} = \alpha_{vn} \; \forall \; m \neq n\right\}$. In general, ω_{kg} is the weight of D_{juv}. Two situations need to be considered: First, there is no idea about the prior information of examinees population; then all AMPs are equally likely, which means $\omega_{kg} = \frac{1}{2^{(K-1)}}$; second, the situation in which each AMP has different prior information and the estimates of the joint probabilities of the AMPs will be used as the weight of D_{juv} (Henson et al., 2008). Henson et al. (2008) defined the ADIs under the first situation as $\mathrm{ADI}_{(\mathrm{A})}$ and the second as $\mathrm{ADI}_{(\mathrm{B})}$. Noting that $\mathrm{ADI}_{(\mathrm{A})}$ is related to items and unrelated to the knowledge states of examinees, therefore, this index can be used to represent the degree that the attribute is being measured by items. As a consequence, the $\mathrm{ADI}_{(\mathrm{A})}$-based ABI ($\mathrm{ADI}_{\mathrm{A}}$-ABI) can be defined as

$$ADI_{(A)} - ABI_j = \prod_{k=1}^{K}\left(\frac{ADI_{(A)k} - adi_{(A)k}}{ADI_{(A)k}}\right)^{q_{jk}}, \quad (10)$$

where $ADI_{(A)k}$ is the lower bound ADI of attribute k and the value of $ADI_{(A)k}$ is the average of $ADI_{(A)k1}$ and $ADI_{(A)k0}$ (Finkelman et al., 2010); $adi_{(A)k}$ represents ADI of attribute k that has already been selected.

The difference between the number of items measuring each attribute-based (MGDI-based) ABI and $\mathrm{ADI}_{(\mathrm{A})}$-based ABI is that B_k and b_k are both positive integers and ABIs are nonnegative, whereas $ADI_{(A)k}$ and $adi_{(A)k}$ include any values that larger than 0. $\mathrm{ADI}_{(\mathrm{A})}$-ABI outcomes can produce negative values in some situations, which are undesirable. Hence, we constrain negative values to 0 when $ADI_{(A)} - ABI_j < 0$. By combining $\mathrm{ADI}_{(\mathrm{A})}$-$\mathrm{ABI}_j$ with PWKL or MPWKL information, the ADI-based item selection method can be written as

$$I_j\left(\hat{\alpha}\right) * \left[ADI_{(A)} - ABI\right] = I_j\left(\hat{\alpha}\right) * \prod_{k=1}^{K}\left(\frac{ADI_{(A)k} - adi_{(A)k}}{ADI_{(A)k}}\right)^{q_{jk}} \quad (11)$$

where $I\left(\hat{\alpha}\right)$ represents PWKL information or MPWKL information. If $\mathrm{ADI}_{(\mathrm{A})} - \mathrm{ABI} > 0$, the next item will be selected by Equation (10); otherwise, PWKL or MPWKL information method will be used to select the next item.

SIMULATION STUDY

Manipulated Factors

We conducted a simulation study to investigate the performance of the ADI-based method under different conditions. We manipulated four independent factors in the study.

Item Pool

In this study, we had designed three item pools, which all contained 775 items and measured five attributes in total. Item pools were constructed based on the study of Huebner et al. (2018) and Wang et al. (2011). In item pool 1, item parameters π_j^* and r_{jk}^* were generated from uniform distributions $U(0.75, 0.95)$ and $U(0.15, 0.50)$, respectively. Considering that r_{jk}^* was relatively large, hence, we labeled item pool 1 as the low discrimination (LD) item pool. In item pool 2, high discrimination (HD) item pool, item parameters π_j^* and r_{jk}^* were generated from uniform distributions $U(0.75, 0.95)$ and $U(0.05, 0.40)$, respectively. In item pool 3, hybrid discrimination (HyD) item pool, item parameter π_j^* was also generated from uniform distributions $U(0.75, 0.95)$, but r_{jk}^*s were generated from uniform distributions $U(0.05, 0.50)$ contained in both low and high discriminations. **Tables 1**, **2** present the descriptive statistics of LD, HD, and HyD item pools.

Examinee Populations

We generated three examinee populations, each one containing 3,200 examinees. The first population (denote as *Unif*)

TABLE 1 | Descriptive statistics of item parameters of LD item pool, HD item pool, and HyD item pool.

		π^*	r_1^*	r_2^*	r_3^*	r_4^*	r_5^*
LD item pool	Min	0.750	0.151	0.153	0.151	0.152	0.152
	Max	0.950	0.499	0.496	0.500	0.500	0.499
	Mean	0.848	0.327	0.326	0.328	0.335	0.329
	SD	0.058	0.100	0.101	0.100	0.099	0.107
HD item pool	Min	0.750	0.053	0.051	0.050	0.051	0.050
	Max	0.949	0.400	0.399	0.400	0.400	0.400
	Mean	0.850	0.217	0.230	0.233	0.227	0.225
	SD	0.056	0.100	0.103	0.102	0.097	0.104
HyD item pool	Min	0.750	0.052	0.051	0.051	0.052	0.052
	Max	0.950	0.495	0.500	0.499	0.498	0.498
	Mean	0.854	0.266	0.270	0.269	0.278	0.282
	SD	0.059	0.125	0.125	0.131	0.124	0.129

LD item pool, low discrimination item pool; HD item pool, high discrimination item pool; HyD item pool, hybrid discrimination item pool.

TABLE 2 | Descriptive statistics of attribute discrimination index for each attribute of LD item pool, HD item pool, and HyD item pool.

		A_1	A_2	A_3	A_4	A_5
LD item pool	Number of items	341	341	341	341	341
	Sum of ADI_k	136.959	139.476	139.460	136.731	139.873
	Mean of ADI_k	0.402	0.409	0.409	0.401	0.410
HD item pool	Number of items	400	400	400	400	400
	Sum of ADI_k	179.699	166.479	169.338	171.138	174.643
	Mean of ADI_k	0.449	0.416	0.423	0.428	0.437
HyD item pool	Number of items	377	355	363	382	355
	Sum of ADI_k	173.688	165.489	164.094	167.221	151.328
	Mean of ADI_k	0.461	0.466	0.452	0.438	0.426

LD item pool, low discrimination item pool; HD item pool, high discrimination item pool; HyD item pool, hybrid discrimination item pool; A_1-A_5, attribute 1 to attribute 5; ADI, attribute discrimination index.

assumed that the AMP of each examinee, $\boldsymbol{\alpha}$, was generated from a uniform distribution of 32 possible pattern profiles with probability 1/32. Thus, each AMP had 100 examinees; meanwhile, each examinee had a 0.5 chance to master each attribute. Considering that correlations among attributes are common in practice, we used a multivariate normal distribution to describe the relationship among attributes for the second and third populations (denote as *Norm*) (de la Torre and Douglas, 2004; Cheng, 2009b; Kunina-Habenicht et al., 2012; Liu et al., 2016). The mastery probabilities for the five attributes were defined as 0.45, 0.50, 0.55, 0.60, and 0.65, respectively, in both populations. The correlations among attributes were set at 0.5 (low correlation) for the second population and 0.8 (high correlation) for the third population. **Table 3** represents the frequencies of examinees who possess each possible number of attributes.

We obtained nine subgroups by crossing item pools and examinee populations. These combinations were as follows: LD item pool with the uniform distributed population (*LD-unif*); LD item pool with the normal distributed population and 0.5 attribute correlation (*LD-norm-0.5*); LD item pool with the normal distributed population and 0.8 attribute correlation (*LD-norm-0.8*); HD item pool with the uniform distributed population (*HD-unif*); HD item pool with the normal distributed population and 0.5 attribute correlation (*HD-norm-0.5*); HD item pool with the normal distributed population and 0.8 attribute correlation (*HD-norm-0.8*); HyD item pool with the uniform distributed population (*HyD-unif*); and HyD item pool with the normal distributed population and 0.5 attribute correlation (*HyD-norm-0.5*); and HyD item pool with the normal distributed population and 0.8 attribute correlation (*HyD-norm-0.8*).

Constraints of Attribute-Balance Coverage

We considered three levels of constraint: Level 1 did not constrain the coverage of attribute balance, whereas level 2 and level 3 added a constraint to it. Level 2 used the method developed by Cheng (2010), who balanced attribute coverage via the number of items measuring each attribute. In Cheng's simulation study, he set the lower bound of item number that measures each attribute at 4 ($B_k = 4$) for a 30-item test; in the current study, we set the lower bound at 2 ($B_k = 2$) for a 10-item test. Level 3 used the method proposed in the current study that balance attribute coverage

TABLE 3 | Frequencies of examinees exhibiting each possible number of attributes in each population.

Number of attributes		0	1	2	3	4	5
Number of examinees	Unif	100	500	1,000	1,000	500	100
	Norm-0.5	166	250	378	489	650	1,267
	Norm-0.8	382	225	222	279	418	1,674

TABLE 4 | Results of mastery pattern correct classification rate (PCCR).

	Uncontrolled		MGDI based		ADI based	
	PWKL	MPWKL	PWKL	MPWKL	PWKL	MPWKL
LD-unif	0.398	0.391	0.582	0.590	0.580	0.582
LD-norm-0.5	0.470	0.458	0.579	0.579	0.591	0.598
LD-norm-0.8	0.507	0.515	0.551	0.557	0.570	0.575
HD-unif	0.378	0.410	0.705	0.706	0.675	0.675
HD-norm-0.5	0.486	0.481	0.686	0.693	0.678	0.685
HD-norm-0.8	0.579	0.578	0.678	0.682	0.692	0.702
HyD-unif	0.390	0.395	0.686	0.678	0.665	0.661
HyD-norm-0.5	0.465	0.443	0.632	0.635	0.647	0.646
HyD-norm-0.8	0.530	0.530	0.633	0.638	0.659	0.642

Uncontrolled, attribute-balance coverage not considered; MGDI-based, balance attribute coverage via MMGDI method; ADI-based, balance attribute coverage via ADI method; PWKL, posterior-weighted Kullback–Leibler information method; MPWKL, modified posterior-weighted Kullback–Leibler information method; LD-unif, low discrimination item pool, uniform distribution of examinees and ignorable correlation among attributes; LD-norm-0.5, low discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; LD-norm-0.8, low discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8; HD-unif, high discrimination item pool, uniform distribution of examinees and ignorable correlation among attributes; HD-norm-0.5, high discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; HD-norm-0.8, high discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8; HyD-unif, hybrid discriminating item pool, uniform distribution of examinees and ignorable correlation among attributes; HyD-norm-0.5, hybrid discriminating item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; HyD-norm-0.8, hybrid discriminating item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8.

via the information that each attribute provided (ADI), with 1 as the lower bound of information ($ADI_{(A)k} = 1$). The reason that setting $\mathrm{ADI}_{(\mathrm{A})k} = 1$ was that as can be seen from **Table 2**, 1 was the lower bound of information for each attribute that can provide approximately two items that measure each attribute, which means level 3 and level 2 had the same constraints.

Item Selection Methods

Cheng (2010) used KL information method to select items successively, whereas many studies have demonstrated that PWKL information method performed better than KL information method in terms of pattern and ACCR (Cheng, 2009a,b; Mao and Xin, 2013; Wang, 2013; Hsu and Wang, 2015; Zheng and Chang, 2016). And the MPWKL information method may perform even better than PWKL (Kaplan et al., 2015). Thus, we adopted both the PWKL and MPWKL information methods in the current study.

We generated a total of 54 conditions study (3 item pools × 3 examinee populations × 3 constraints of attribute-balance coverage × 2 item selection methods). We fixed the number of items in the test to 10 in all conditions. The first item was selected randomly from the item pool, with a maximum a posteriori (MAP) method used to estimate the examinee's AMP, and the prior information of AMP assumed to follow a uniform distribution. The study procedures were implemented by R software.

Evaluation Criteria

We evaluated results against five criteria: mastery pattern correct classification rate (PCCR), ACCR, equalization of ACCR (E-ACCR), item exposure index, and examinee qualification rate. E-ACCR is the ratio between the standard deviation of ACCR and the mean of ACCR, which represents the stability of ACCR. Examinee qualification rate means the proportion of examinees who satisfy the prescribed constraints (e.g., a minimum of two items that measure each attribute under the MGDI-based method in this study), which ranged from 0 to 1. The computation of PCCR and ACCR is as follows:

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

$$ACCR_k = \frac{\sum_{i=1}^{N} I\left(\alpha_{ik} = \hat{\alpha}_{ik}\right)}{N},$$

where N is the number of examinees and $I(\ldots)$ is an indicator function. And item exposure index can be expressed as

$$\chi^2 = \sum_{j=1}^{N} \frac{\left(\exp_j - \frac{J}{N}\right)^2}{\frac{J}{N}},$$

$$\exp_j = \frac{N_j^{administered}}{N},$$

where J is the number of items, $N_j^{administered}$ is the number of items administered to examinees.

RESULTS

Table 4 lists the estimates of PCCR for each condition. The data summarized in the table make several meaningful points. First, the MPWKL information method performs similarly or even better than the PWKL information method for both LD and HD item pools, regardless of the methods that constrain attribute coverage and distribution of the population. Second, compared with uncontrolled conditions, both the PWKL and MPWKL information methods lead to better PCCR outcomes when attribute coverage was controlled, and there are only minor differences between the MGDI-based and ADI-based methods. Third, the ADI-based attribute-balance method performs better than the MGDI-based method in normal distribution populations with 0.8 attribute correlation, regardless of the quality of the item pool. Fourth, the PCCRs in HyD item pools are quite complex. Both the ADI-based and MGDI-based attribute-balance methods perform better than uncontrolled conditions. However, the MPWKL information method does not always perform better than the PWKL information method in all conditions.

Figures 1–3 depict the ACCR for each condition, and **Table 5** represents the summary of ACCR and E-ACCR. They document the following results: First, the MPWKL information method has a similar performance or even outperforms the PWKL information method with ACCR for both LD and HD item pools with all populations under coverage controlled conditions and E-ACCR in most cases. Second, the coverage of ACCR and E-ACCR under uncontrolled conditions performs the worst, whereas they are comparable between the MGDI-based and ADI-based methods. And most of the E-ACCRs of the MGDI-based method perform slightly worse than the ADI-based method. Third, in the LD and HD item pools, when the PWKL information method was employed, the E-ACCR for uncontrolled conditions yields worse results than does the MGDI-based method; meanwhile, the ADI-based method leads to the best results. Fourth, in the HyD item pool, the ACCRs and E-ACCRs with both the ADI-based and MGDI-based attribute-balance methods outperform uncontrolled conditions; meanwhile, the ADI-based attribute-balance method performs the best under the condition of HyD-norm-0.8.

The results of the item exposure rate and examinee qualification rate for each condition are summarized in **Table 6**. The following results can be drawn from the table: First, both PWKL and MPWKL information methods lead to acceptable item exposure, regardless of attribute-balance constraints, quality of item pool, and population distribution. However, the MGDI-based attribute coverage constraint gains the worst outcomes. When the ADI-based attribute coverage constraint is used, it mitigates the worst result but better than the uncontrolled attribute coverage constraint for uniform distribution populations with HD and HyD item pools. Second, compared with uncontrolled attribute coverage constraint, the examinee qualification rates of both MGDI-based and ADI-based attribute coverage constraints produce perfect results, regardless of item selection methods. In addition, MGDI-based and ADI-based attribute coverage constraints lead to consistent examinee qualification rates with both PWKL and MPWKL information methods. Moreover, an unexpected result appears that examinee qualification rates for uniform distribution populations with HD and HyD item pools are extremely low.

DISCUSSION AND CONCLUSION

CD-CAT captures the advantages of both CDA and CAT, allowing the diagnosis of strengths and weaknesses of examinees with fewer items. CD-CAT can be used for low-stakes testing, so it can be adopted to provide detailed information on examinees for educators regularly (Hartz and Roussos, 2008; Mao and Xin, 2013; Kaplan et al., 2015). Thus, educators can provide remedial instruction for those examinees who need help. It is worth noting that the test length of CD-CAT should not be too long, in order to avoid increasing the burden on students. It should deviate from the original orientation by using a computer-based test to reduce students' burden and improve the efficiency of testing and learning if students do not take the test too long.

It is critical to consider the structure of short tests to assess the knowledge states of examinees comprehensively in CD-CAT. It is also important that each attribute should be measured adequately. Cheng (2010) used the number of items measuring each attribute to balance the coverage of attributes. The current study uses the information that each attribute provided to balance attribute coverage, as proposed by Henson et al. (2008). The simulation study was conducted to evaluate the performance of the new method, and the results showed that compared with the uncontrolled attribute coverage under the PWKL and MPWKL information methods, the ADI-based attribute-balance coverage method (the new method) improved both PCCR and ACCR. The reason is that when the attribute-balance coverage constraint is not controlled, some attributes may not be measured adequately; thus, the ADI is small for many examinees. Henson et al. (2008) demonstrated that the correlations are quite high between ADI and correct classification rates. Therefore, ADI can be used as the indicator of correct classification rates reasonably. Moreover, Cheng (2010) pointed out that the smallest ACCR dominated the PCCR, and he described this phenomenon as similar to Liebig's *law of the minimum*, which means the shortest stave is the most important factor that affects the capacity of a barrel with staves. In sum, considering that some ADIs are slightly smaller when attribute-balance coverage is not controlled, the ACCRs for some attributes are lower. As a consequence,

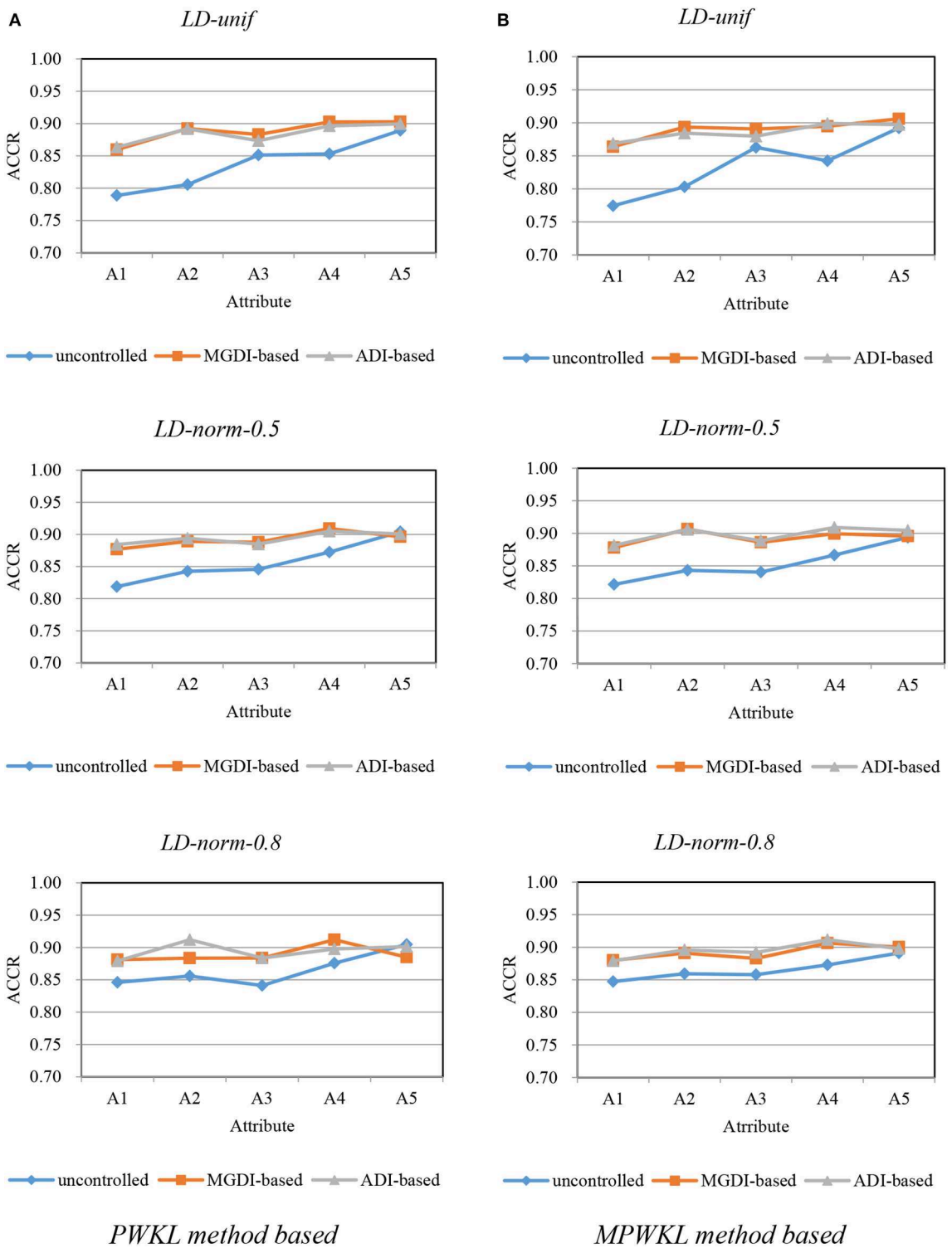


FIGURE 1 | Attribute correct classification rates (ACCRs) under posterior-weighted Kullback–Leibler (PWKL) and modified PWKL (MPWKL) information methods for low discrimination (LD) item pools. **(A)** PWKL method based. **(B)** MPWKL method based.

the PCCRs under uncontrolled conditions are lower than those of MGDI-based and ADI-based attribute-balance coverage methods.

The present results also show that, compared with the uncontrolled method, both the ADI-based and MGDI-based attribute-balance coverage methods produce noticeable better

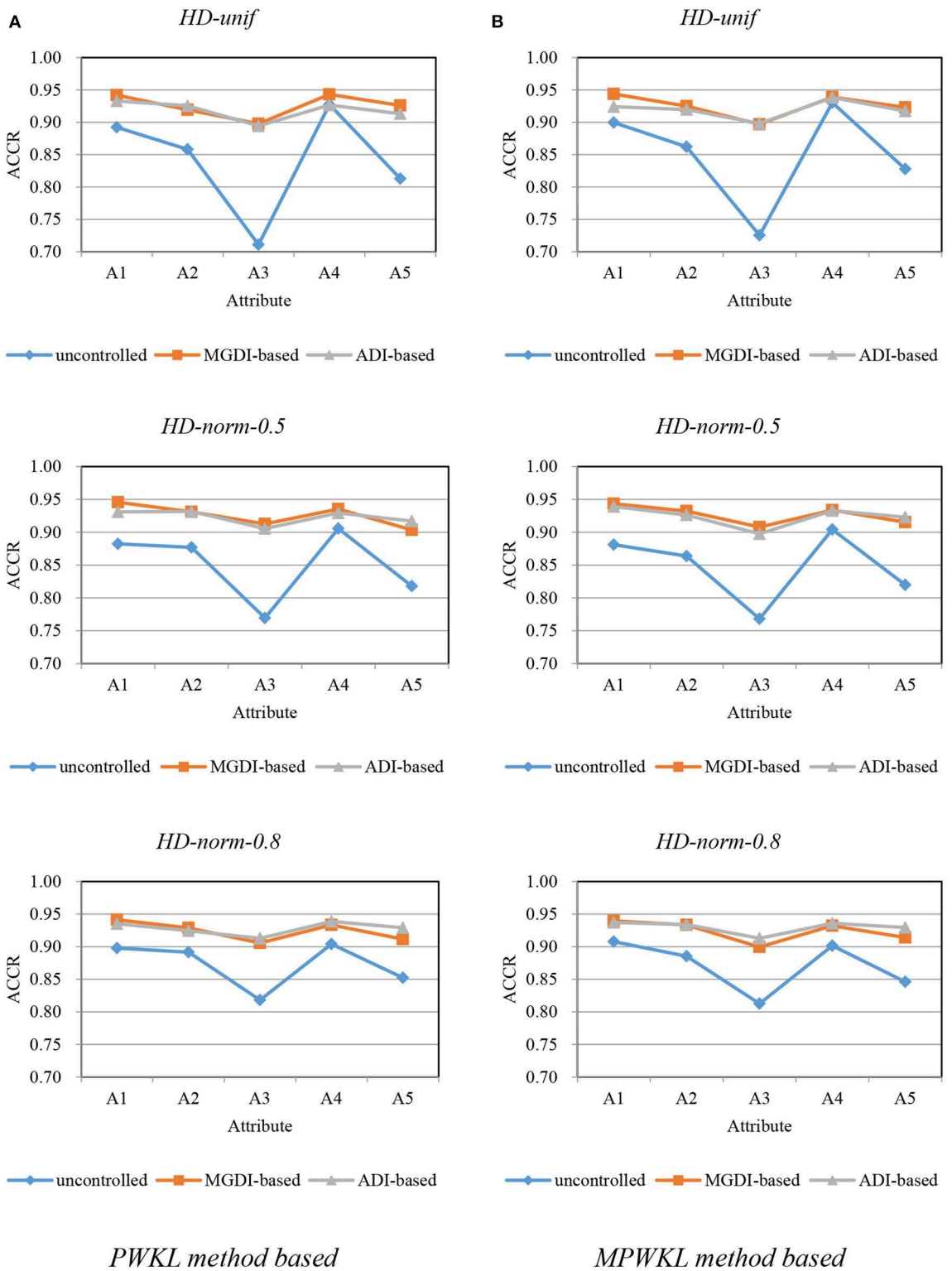


FIGURE 2 | Attribute correct classification rates (ACCRs) under posterior-weighted Kullback–Leibler (PWKL) and modified PWKL (MPWKL) information methods for high discrimination (HD) item pools. **(A)** PWKL method based. **(B)** MPWKL method based.

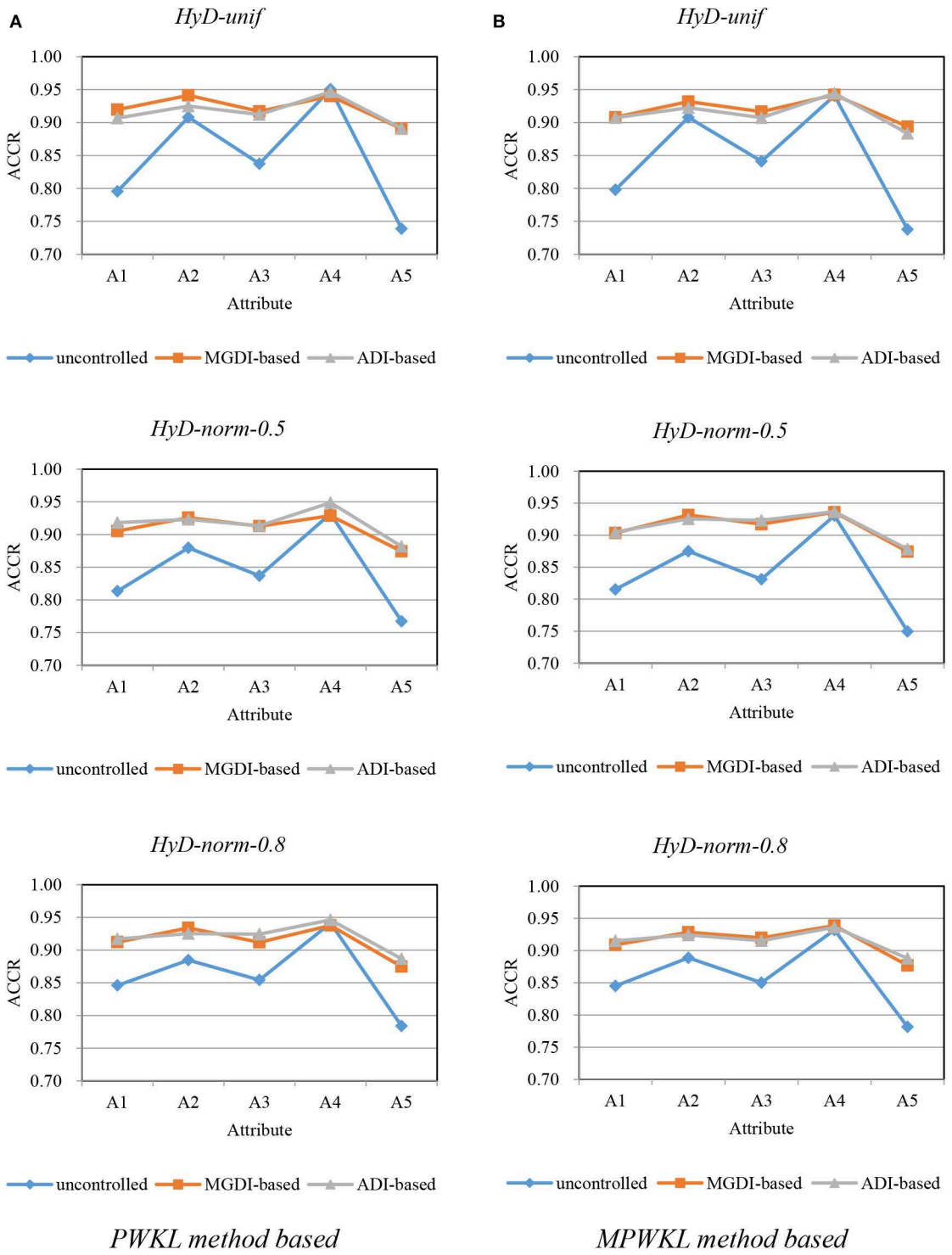


FIGURE 3 | Attribute correct classification rates (ACCRs) under posterior-weighted Kullback–Leibler (PWKL) and modified PWKL (MPWKL) information methods for hybrid discrimination (HyD) item pools. **(A)** PWKL method based. **(B)** MPWKL method based.

results of PCCR and E-ACCR and slightly better ones of ACCR. Although there are no noticeable differences of E-ACCR between the ADI-based method and the MGDI-based method, the ADI-based method performs slightly better for most conditions. We infer that the ADI-based attribute-balance coverage method produces more stable ACCR than the other

TABLE 5 | Summary of ACCR and E-ACCR.

		Uncontrolled		MGDI based		ADI based	
		PWKL	**MPWKL**	**PWKL**	**MPWKL**	**PWKL**	**MPWKL**
LD-unif	M	0.838	0.835	0.888	0.890	0.885	0.886
	SD	0.040	0.047	0.018	0.016	0.016	0.013
	E-ACCR	4.773	5.629	2.027	1.798	1.808	1.467
LD-norm-0.5	M	0.857	0.853	0.892	0.893	0.894	0.898
	SD	0.033	0.028	0.012	0.011	0.009	0.012
	E-ACCR	3.851	3.283	1.345	1.232	1.007	1.336
LD-norm-0.8	M	0.865	0.866	0.889	0.892	0.895	0.895
	SD	0.026	0.017	0.013	0.011	0.013	0.012
	E-ACCR	3.006	1.963	1.462	1.233	1.453	1.341
HD-unif	M	0.840	0.849	0.926	0.926	0.919	0.920
	SD	0.084	0.079	0.019	0.018	0.015	0.015
	E-ACCR	10.000	9.305	2.052	1.944	1.632	1.630
HD-norm-0.5	M	0.850	0.847	0.926	0.927	0.923	0.924
	SD	0.056	0.054	0.017	0.015	0.011	0.016
	E-ACCR	6.588	6.375	1.836	1.618	1.192	1.732
HD-norm-0.8	M	0.873	0.871	0.924	0.924	0.928	0.930
	SD	0.037	0.040	0.015	0.017	0.010	0.010
	E-ACCR	4.238	4.592	1.623	1.840	1.078	1.075
HyD-unif	M	0.846	0.845	0.922	0.919	0.916	0.913
	SD	0.085	0.082	0.021	0.019	0.021	0.022
	E-ACCR	10.047	9.704	2.278	2.067	2.293	2.410
HyD-norm-0.5	M	0.846	0.840	0.909	0.913	0.917	0.914
	SD	0.063	0.068	0.022	0.025	0.024	0.023
	E-ACCR	7.447	8.095	2.420	2.738	2.617	2.516
HyD-norm-0.8	M	0.862	0.859	0.914	0.915	0.920	0.916
	SD	0.058	0.056	0.025	0.024	0.021	0.018
	E-ACCR	6.729	6.519	2.735	2.623	2.283	1.965

Uncontrolled, attribute-balance coverage not considered; MGDI-based, balance attribute coverage via MMGDI method; ADI-based, balance attribute coverage via ADI method; PWKL, posterior-weighted Kullback–Leibler information method; MPWKL, modified posterior-weighted Kullback–Leibler information method; LD-unif, low discrimination item pool, uniform distribution of examinees and ignorable correlation among attributes; LD-norm-0.5, low discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; LD-norm-0.8, low discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8; HD-unif, high discrimination item pool, uniform distribution of examinees and ignorable correlation among attributes; HD-norm-0.5, high discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; HD-norm-0.8, high discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8; HyD-unif, hybrid discriminating item pool, uniform distribution of examinees and ignorable correlation among attributes; HyD-norm-0.5, hybrid discriminating item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; HyD-norm-0.8, hybrid discriminating item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8; E-ACCR, equalization of attribute correct classification rate.

two methods. Besides, regardless of item selection methods, all examinees satisfied the prescribed constraints when the ADI-based and MGDI-based methods have been used, whereas the uncontrolled method failed for some examinees.

It is worth noting that when attribute-balance coverage is uncontrolled, the examinee qualification rates for HD and HyD item pools with uniform distribution populations are extremely poor under both item selection methods, for still unknown reasons. Therefore, a further study of that effect is needed.

Some future studies can be conducted to improve and enhance the application of the ADI-based attribute-balance coverage method. First, a variable-length CD-CAT can be conducted to evaluate the performance of the ADI-based method. Under variable-length CD-CAT, the measurement precision or standard error is fixed, and the number of items administered to each examinee is different. Second, there is only one RRUM model that has been used in the current study, which is a non-compensatory model. More models can be considered to verify the generalization of the ADI-based attribute-balance coverage method, especially for compensatory models. Third, the importance of each attribute to the item is assumed to be equal, but it is common that some traits are more important than others when more than one attribute is to be measured in practice (Wang et al., 2014). Thus, researchers need to take the relative importance of each attribute into account in a future study. Lastly, how to choose the lower bound of the ADI is an additional important issue. The value that has been used

TABLE 6 | Results of item exposure rate and examinee qualification rate for each condition.

		Uncontrolled		MGDI based		ADI based	
		PWKL	MPWKL	PWKL	MPWKL	PWKL	MPWKL
Item exposure rate	LD-unif	86.147	85.108	132.439	135.525	113.967	112.526
	LD-norm-0.5	78.412	80.639	117.650	116.717	99.591	98.479
	LD-norm-0.8	89.486	89.497	118.588	118.641	101.490	102.526
	HD-unif	107.043	105.655	135.684	134.428	105.052	106.532
	HD-norm-0.5	82.523	80.432	122.876	123.406	91.325	92.422
	HD-norm-0.8	97.435	98.674	130.523	129.354	96.410	96.595
	HyD-unif	108.915	106.501	140.560	137.609	105.359	106.776
	HyD-norm-0.5	77.463	77.452	127.192	127.473	91.535	91.059
	HyD-norm-0.8	86.915	86.931	128.974	128.897	92.644	93.363
Examinee qualification rate	LD-unif	0.432	0.422	1.000	1.000	1.000	1.000
	LD-norm-0.5	0.504	0.510	1.000	1.000	1.000	1.000
	LD-norm-0.8	0.580	0.574	1.000	1.000	1.000	1.000
	HD-unif	0.258	0.264	1.000	1.000	1.000	1.000
	HD-norm-0.5	0.429	0.424	1.000	1.000	1.000	1.000
	HD-norm-0.8	0.516	0.508	1.000	1.000	1.000	1.000
	HyD-unif	0.287	0.290	1.000	1.000	1.000	1.000
	HyD-norm-0.5	0.431	0.418	1.000	1.000	1.000	1.000
	HyD-norm-0.8	0.501	0.501	1.000	1.000	1.000	1.000

Uncontrolled, attribute-balance coverage not considered; MGDI-based, balance attribute coverage via MMGDI method; ADI-based, balance attribute coverage via ADI method; PWKL, posterior-weighted Kullback–Leibler information method; MPWKL, modified posterior-weighted Kullback–Leibler information method; LD-unif, low discrimination item pool, uniform distribution of examinees and ignorable correlation among attributes; LD-norm-0.5, low discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; LD-norm-0.8, low discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8; HD-unif, high discrimination item pool, uniform distribution of examinees and ignorable correlation among attributes; HD-norm-0.5, high discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; HD-norm-0.8, high discrimination item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8; HyD-unif, hybrid discriminating item pool, uniform distribution of examinees and ignorable correlation among attributes; HyD-norm-0.5, hybrid discriminating item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.5; HyD-norm-0.8, hybrid discriminating item pool, normal distribution of examinees and moderate correlation among attributes, correlations among attributes set at 0.8.

in the current study is a variation of the number of items measuring each attribute in the study of Cheng (2010), but how large the ADI should be to measure each attribute adequately is still unknown. Thus, studies that address the adequacy of the ADI in CD-CAT will provide some guidelines for further test administrations.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

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

FUNDING

This research was supported by the Cultural Experts and Four Groups of Talented People Foundation of China.

REFERENCES

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

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

Cheng, Y. (2009b). "Computerized adaptive testing for cognitive diagnosis," in *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, ed D. J. Weiss. Available online at: www.psych.umn.edu/psylabs/CATCentral/

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method. *Educ. Psychol. Meas.* 70, 902–913. doi: 10.1177/0013164410366693

Cheng, Y., and Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *Br. J. Math. Stat. Psychol.* 62, 369–383. doi: 10.1348/000711008X304376

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

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

Feng, Y., Habing, B. T., and Huebner, A. (2013). Parameter estimation of the reduced RUM using the EM algorithm. *Appl. Psychol. Meas.* 38, 137–150. doi: 10.1177/0146621613502704

Finkelman, M., Kim, W., and Roussos, L. A. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *J. Educ. Meas.* 46, 273–292. doi: 10.1111/j.1745-3984.2009.00081.x

Finkelman, M. D., Kim, W., Roussos, L., and Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Appl. Psychol. Meas.* 34, 310–326. doi: 10.1177/0146621609344846

Gierl, M. J., Wang, C., and Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT©. *J. Technol. Learn. Assess.* 6:1–53. Available online at: https://eric.ed.gov/?id=EJ838616

Hartz, S. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: blending theory with practice (Unpublished doctoral dissertation)*. University of Illinois at Urbana-Champaign, Champaign, IL, United States.

Hartz, S., and Roussos, L. (2008). The fusion model for skills diagnosis: blending theory with practicality. *ETS Res. Rep. Ser.* 2008:i-57. doi: 10.1002/j.2333-8504.2008.tb02157.x

Henson, R., Roussos, L., Douglas, J., and He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Appl. Psychol. Meas.* 32, 275–288. doi: 10.1177/0146621607302478

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

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

Jang, E. (2005). *A validity narrative: effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL (Unpublished doctoral dissertation)*. University of Illinois at Urbana-Champaign, Champaign, IL, United States.

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

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

Kim, Y.-H. (2011). Diagnosing eap writing ability using the reduced reparameterized unified model. *Lang. Test.* 28, 509–541. doi: 10.1177/0265532211400860

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *J. Educ. Meas.* 49, 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Lee, Y. S., de la Torre, J., and Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. *Asia Pac. Educ. Rev.* 13, 333–345. doi: 10.1007/s12564-011-9196-3

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

Liu, Y., Tian, W., and Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *J. Educ. Behav. Stat.* 41, 3–26. doi: 10.3102/1076998615621293

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

McGlohen, M., and Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behav. Res. Methods* 40, 808–821. doi: 10.3758/BRM.40.3.808

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., and Templin, J. L. (2007). "The fusion model skills diagnosis system," in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. Leighton and M. Gierl (Cambridge: Cambridge University Press), 275–318.

Sinharay, S., and Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educ. Meas. Issues Pract.* 33, 23–35. doi: 10.1111/emip.12024

Tatsuoka, K. K. (1995). "Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach," in *Cognitively Diagnostic Assessment*, eds P. D. Nichols, S. F. Chipman, and R. L Brennan (New York, NY: Routledge), 327–359.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educ. Psychol. Meas.* 73, 1017–1035. doi: 10.1177/0013164413498256

Wang, C., Chang, H. H., and Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *J. Educ. Meas.* 48, 255–273. doi: 10.1111/j.1745-3984.2011.00145.x

Wang, C., Zheng, C., and Chang, H. H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *J. Educ. Meas.* 51, 358–380. doi: 10.1111/jedm.12057

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

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

Yao, L., and Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Appl. Psychol. Meas.* 31, 83–105. doi: 10.1177/0146621606291559

Zheng, C., and Chang, H. H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Appl. Psychol. Meas.* 40, 608–624. doi: 10.1177/0146621616665196

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 March 2020
doi: 10.3389/fpsyg.2020.00384

Bayesian Estimation of the DINA Model With Pólya-Gamma Gibbs Sampling

Zhaoyuan Zhang[1], *Jiwei Zhang*[2]*, *Jing Lu*[1] *and Jian Tao*[1]

[1] *Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China,* [2] *Key Lab of Statistical Modeling and Data Analysis of Yunnan Province, School of Mathematics and Statistics, Yunnan University, Kunming, China*

With the increasing demanding for precision of test feedback, cognitive diagnosis models have attracted more and more attention to fine classify students whether has mastered some skills. The purpose of this paper is to propose a highly effective Pólya-Gamma Gibbs sampling algorithm (Polson et al., 2013) based on auxiliary variables to estimate the deterministic inputs, noisy "and" gate model (DINA) model that have been widely used in cognitive diagnosis study. The new algorithm avoids the Metropolis-Hastings algorithm boring adjustment the turning parameters to achieve an appropriate acceptance probability. Four simulation studies are conducted and a detailed analysis of fraction subtraction data is carried out to further illustrate the proposed methodology.

Keywords: Bayesian estimation, cognitive diagnosis models, DINA model, Pólya-Gamma Gibbs sampling algorithm, Metropolis-Hastings algorithm, potential scale reduction factor

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Gongjun Xu,
University of Michigan, United States
Yinghan Chen,
University of Nevada, Reno, United States

***Correspondence:**
Jiwei Zhang
zhangjw713@nenu.edu.cn

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

Received: *08 January 2020*
Accepted: *19 February 2020*
Published: *10 March 2020*

Citation:
Zhang Z, Zhang J, Lu J and Tao J (2020) Bayesian Estimation of the DINA Model With Pólya-Gamma Gibbs Sampling. Front. Psychol. 11:384. doi: 10.3389/fpsyg.2020.00384

1. INTRODUCTION

Modeling the interaction between examinee's latent discrete skills (attributes) and items at the item level for binary response data, cognitive diagnosis models (CDMs) is an important methodology to evaluate whether the examinees have mastered multiple fine-grained skills, and these models have been widely used in a variety of the educational and psychological researches (Tatsuoka, 1984, 2002; Doignon and Falmagne, 1999; Maris, 1999; Junker and Sijtsma, 2001; de la Torre and Douglas, 2004; Templin and Henson, 2006; DiBello et al., 2007; Haberman and von Davier, 2007; de la Torre, 2009, 2011; Henson et al., 2009; von Davier, 2014; Chen et al., 2015). With the increasing complexity of the problems in cognitive psychology research, various specific and general formulations of CDMs have been proposed to deal with the practical problems. There are several specific CDMs, widely known among them, are the deterministic inputs, noisy "and" gate model (DINA; Junker and Sijtsma, 2001; de la Torre and Douglas, 2004; de la Torre, 2009), the noisy inputs, deterministic, "and" gate model (NIDA; Maris, 1999), the deterministic input, noisy "or" gate model (DINO; Templin and Henson, 2006) and the reduced reparameterized unified model (rRUM; Roussos et al., 2007). In parallel with the specific CDMs, the

general CDMs have also made great progress, including the general diagnostic model (GDM; von Davier, 2005, 2008), the log-linear CDM (LCDM; Henson et al., 2009), and the generalized DINA (G-DINA; de la Torre, 2011). Parameter estimation has been a major concern in the application of CDMs. In fact, simultaneous estimations of items and examinee's latent discrete skills result in statistical complexities in the estimation task.

Within a fully Bayesian framework, a novel and highly effective Pólya-Gamma Gibbs sampling algorithm (PGGSA; Polson et al., 2013) based on the auxiliary variables is proposed to estimate the commonly used DINA model in this paper. The PGGSA overcomes the disadvantages of Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970; Chib and Greenberg, 1995; Chen et al., 2000), which requires to repeatedly adjust the specification of tuning parameters to achieve a certain acceptance probability and thus increases the computational burden. More specifically, the Metropolis–Hasting algorithm depends on the variance (tuning parameter) of the proposal distribution and is sensitive to step size. If the step size is too small, the chain will take longer to traverse the target density. If the step size is too large, there will be inefficiencies due to a high rejection rate. In addition, the Metropolis-Hastings algorithm is relatively difficulty to sample parameters with monotonicity or truncated interval restrictions. Instead, it can improve the accuracy of parameter estimation by employing strong informative prior distributions to avoid violating the restriction conditions (Culpepper, 2016).

The rest of this paper is organized as follows. Section 2 contains a short introductions of DINA model, its reparameterized form, and model identifications. A detailed implementation of PGGSA is shown in section 3. In section 4, four simulations focus on the performance of parameter recovery for the PGGSA, the results of comparing with the Metropolis-Hastings algorithm, the analysis of sensitivity of prior distributions for the PGGSA, the results of comparing with Culpepper (2015)'s Gibbs algorithm on the attribute classification accuracy and the estimation accuracy of class membership probability parameters. In addition, the quality of PGGSA is investigated using a fraction subtraction test data in section 5. We conclude the article with a brief discussion in section 6.

2. MODELS AND MODEL IDENTIFICATIONS

The DINA model focuses on whether the examinee i has mastered the k attribute, where $i = 1, \ldots, N, k = 1, \ldots, K$. Let α_{ik} be a dichotomous latent attribute variable with values of 0 or 1 indicating absence or presence of a attribute, respectively. $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})^{'}$ is a vector of K dimensional latent attributes for the ith examinee. Given the categorical nature of the latent classes, $\boldsymbol{\alpha}_i$ belongs to one of $C = 2^K$ attribute latent classes. If the ith examinee belongs to the cth classification, the attribute vector can be expressed as $\boldsymbol{\alpha}_c = (\alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cK})^{'}$. Considering a test consisting of J items, each item j is associated with a vector of K dimensional item attributes, $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jK})^{'}$, where

$$q_{jk} = \begin{cases} 1, & \text{if attribute } k \text{ is required by item } j, \\ 0, & \text{if attribute } k \text{ is not required by item } j. \end{cases}$$

Therefore, a $\boldsymbol{Q}$ matrix, $\boldsymbol{Q} = \{q_{jk}\}_{J \times K}$, can be obtained by the J item attribute vectors. The DINA model is conjunctive. That is, the examinee i must possess all the required attributes to answer the item j correctly. The ideal response pattern η_{ij} can be defined as follows

$$\eta_{ij} = \begin{cases} 1, & \text{if the examinee } i \text{ possesses all the required attributes for the item } j, \\ 0, & \text{if the examinee } i \text{ does not master at least one attribute for the item } j. \end{cases}$$

$\eta_{ij} = \mathrm{I}\left(\boldsymbol{\alpha}_i^{'}\boldsymbol{q}_j = \boldsymbol{q}_j^{'}\boldsymbol{q}_j\right) = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, where $\mathrm{I}(\cdot)$ denotes the indicator function. The parameters for a correct response to item j when given η_{ij} are denoted by s_j and g_j. The slipping parameter s_j and the guessing parameter g_j refer to the probability of incorrectly answering the item when $\eta_{ij} = 1$ and the probability of correctly guessing the answer when $\eta_{ij} = 0$, respectively. Let Y_{ij} denote the observed item response for the ith examinee to response jth item, $Y_{ij} = 1$ if the ith examinee correct answer the jth item, 0 otherwise. The parameters s_j and g_j are formally defined by

$$s_j = p\left(Y_{ij} = 0 \,|\eta_{ij} = 1\right) \text{ and } g_j = p\left(Y_{ij} = 1 \,|\eta_{ij} = 0\right).$$

The probabilities of observing response given attributes $\boldsymbol{\alpha}$ are represented by

$$f_{ij} = p\left(Y_{ij} = 1 \,|\boldsymbol{\alpha}_i, s_j, g_j\right) = \left(1 - s_j\right)^{\eta_{ij}}$$
$$g_j^{1-\eta_{ij}} = \begin{cases} 1 - s_j, & \eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} = 1, \\ g_j, & \eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} = 0. \end{cases} \quad (1)$$

and.

$$h_{ij} = 1 - p\left(Y_{ij} = 1 \,|\boldsymbol{\alpha}_i, s_j, g_j\right) = \left[1 - \left(1 - s_j\right)^{\eta_{ij}} g_j^{1-\eta_{ij}}\right]$$
$$= \begin{cases} s_j, & \eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} = 1, \\ 1 - g_j, & \eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} = 0. \end{cases} \quad (2)$$

2.1. The Reparameterized DINA Model

To describe the relationship between the attribute vector and the observed response, we can reexpress the DINA model as follow:

$$p\left(Y_{ij} = 1 \,|\boldsymbol{\alpha}_i\right) = g_j + \left(1 - s_j - g_j\right) \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}, \quad (3)$$

where the model discrimination index can be defined as $1 - s_j - g_j = \text{IDI}_j$ (de la Torre, 2008). Based on the traditional DINA model, we reparameterize s_j and g_j from the probability scale to the logit scale (Henson et al., 2009; DeCarlo, 2011; von Davier, 2014; Zhan et al., 2017). That is,

$$\zeta_j = \text{logit}\left(g_j\right),$$

$$\beta_j = \text{logit}\left(1 - s_j\right) - \text{logit}\left(g_j\right),$$

where $\text{logit}(x) = \log\left(x/\left(1-x\right)\right)$. Therefore, the reparameterized DINA model (DeCarlo, 2011) can be written as

$$\text{logit}\left[p\left(Y_{ij}=1\left|\boldsymbol{\alpha}_i, \varsigma_j, \beta_j\right.\right)\right] = \varsigma_j + \beta_j \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}, \tag{4}$$

where ς_j and β_j are the item intercept and interaction parameters, respectively.

2.2. The Likelihood Function of the Reparameterized DINA Model Based on the Latent Class

Suppose that the vector of item responses for the ith examinee can be denoted as $\boldsymbol{Y}_i = \left(Y_{i1}, \ldots, Y_{iJ}\right)'$. Let the vector of intercept and interaction parameters for J items be $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\varsigma} = \left(\varsigma_1, \ldots, \varsigma_J\right)$ and $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_J\right)$. Given the categorical nature of the latent classes, $\boldsymbol{\alpha}_i$ belongs to one of $C = 2^K$ attribute latent classes. For the ith examinee belonging to the cth classification, the attribute vector is expressed as $\boldsymbol{\alpha}_c = \left(\alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cK}\right)'$. According the Equation (4), the probability of observing $\boldsymbol{Y}_i$ that the ith examinee belonging to the cth latent class answers J items can be written as

$$p\left(\boldsymbol{Y}_i \left|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\varsigma}, \boldsymbol{\beta}\right.\right) = \prod_{j=1}^{J} \left[p\left(Y_{ij}=1\left|\boldsymbol{\alpha}_c, \varsigma_j, \beta_j\right.\right)\right]^{Y_{ij}} \left[1 - p\left(Y_{ij}=1\left|\boldsymbol{\alpha}_c, \varsigma_j, \beta_j\right.\right)\right]^{1-Y_{ij}} \tag{5}$$

where $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c$ denotes the examinee i belongs to the cth latent class. $p\left(Y_{ij}=1\left|\boldsymbol{\alpha}_c, \varsigma_j, \beta_j\right.\right)$ is the probability that the examinee i in class c correctly answers the item j.

Let $\pi_c = p\left(\boldsymbol{\alpha}_c\right)$ be the probability of examinees for each class c, $c = 1, \ldots, C$, and $\boldsymbol{\pi} = \left(\pi_1, \ldots, \pi_C\right)'$ is C dimensional vector of class membership probabilities, where $\sum_{c=1}^{C} \pi_c = 1$. Therefore, the probability of observing $\boldsymbol{Y}_i$ given item parameters $\boldsymbol{\varsigma}$, $\boldsymbol{\beta}$ and class membership probabilities $\boldsymbol{\pi}$ can be written as

$$p\left(\boldsymbol{Y}_i \left|\boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\pi}\right.\right) = \sum_{c=1}^{C} \pi_c p\left(\boldsymbol{Y}_i \left|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\varsigma}, \boldsymbol{\beta}\right.\right). \tag{6}$$

The likelihood function based on the latent class can be written as

$$p\left(\boldsymbol{Y} \left|\boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\pi}\right.\right) = \prod_{i=1}^{N} \sum_{c=1}^{C} \pi_c p\left(\boldsymbol{Y}_i \left|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\varsigma}, \boldsymbol{\beta}\right.\right). \tag{7}$$

2.3. Model Identification

The model identification is an important cornerstone for estimating parameters and practical applications. Chen et al. (2015), Xu and Zhang (2016), and Xu (2017) discuss the DINA model identification conditions. Gu and Xu (2019) further provide a set of sufficient and necessary conditions for the identifiability of the DINA model. That is,

Condition 1: *(1) The Q-matrix is complete under the DINA model and without loss of generality, we assume the Q-matrix takes the following form:*

$$Q = \begin{pmatrix} \mathcal{I}_K \\ Q^* \end{pmatrix}_{J \times K}, \tag{8}$$

where $\mathcal{I}_K$ is the $K \times K$ identify matrix and Q^* is a $(J-K) \times K$ submarix of Q.

(2) Each of the K attributes is required by at least three items.

Condition 2: *Any two different columns of the submatrix Q^* in (8) are distinct.*

Under the above two conditions, Gu and Xu (2019) give the following identifiability result.

Theorem (Sufficient and Necessary Condition) *Conditions* 1 *and* 2 *are sufficient and necessary for the identifiability of all the DINA model parameters.*

3. PÓLYA-GAMMA GIBBS SAMPLING ALGORITHM

Polson et al. (2013) propose a new data augmentation strategy for fully Bayesian inference in logistic regression. The data augmentation approach appeals to a new class of Pólya-Gamma distribution rather than Albert and Chib (1993)'s data augmentation algorithm based on a truncated normal distribution. Next, we introduce the Pólya-Gamma distribution.

Definition: Let $\{T_k\}_{k=1}^{+\infty}$ is a iid random variable sequences from a Gamma distribution with parameters λ and 1. That is, $T_k \sim \text{Gamma}\left(\lambda, 1\right)$. A random variable W follows a Pólya-Gamma distribution with parameters $\lambda > 0$ and $\tau \in R$, denoted $W \sim \text{PG}\left(\lambda, \tau\right)$, if

$$W \overset{D}{=} \frac{1}{2\pi} \sum_{k=1}^{+\infty} \frac{T_k}{\left(k - \frac{1}{2}\right)^2 + \frac{\tau^2}{4\pi^2}}, \tag{9}$$

where $\overset{D}{=}$ denotes equality in distribution. In fact, the Pólya-Gamma distribution is an infinite mixture of gamma distributions which provide the plausibility to sample from Gamma distributions.

Based on Polson et al. (2013, p. 1341, Equation 7)'s Theorem 1, the likelihood contribution of the ith examinee to answer the jth item can be expressed as

$$L\left(\varsigma_j, \beta_j, \boldsymbol{\alpha}_i\right) = \frac{\left[\exp\left(\varsigma_j + \beta_j \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}\right)\right]^{Y_{ij}}}{1 + \exp\left(\varsigma_j + \beta_j \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}\right)}$$

$$\propto \exp\left[k_{ij}\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)\right]\times\int_0^{\infty}\exp\left[-\frac{W_{ij}\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2}{2}\right]p\left(W_{ij}\,|1,0\right)dW_{ij},\tag{10}$$

where $k_{ij}=Y_{ij}-\frac{1}{2}$. $p\left(W_{ij}\,|1,0\right)$ is the conditional density of W_{ij}. That is, $W_{ij}\sim \mathrm{PG}\,(1,0)$. The auxiliary variable W_{ij} follows a Pólya-Gamma distribution with parameters (1, 0). Biane et al. (2001) provide proofs of Equation (10). In addition, Polson et al. (2013) further discuss Equation (10). Therefore, the full conditional distribution of $\boldsymbol{\varsigma},\boldsymbol{\beta},\boldsymbol{\alpha}$ given the auxiliary variables W_{ij} can be written as

$$p\left(\boldsymbol{\varsigma},\boldsymbol{\beta}_j,\boldsymbol{\alpha}\,|\boldsymbol{W},\boldsymbol{Y}\right)\propto\left\{\prod_{i=1}^{N}\prod_{j=1}^{J}\left[\exp\left[k_{ij}\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)\right]\exp\left[-\frac{W_{ij}\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2}{2}\right]\right]\right\}\times\left\{\prod_{j=1}^{J}\left[p\left(\varsigma_j\right)p\left(\beta_j\right)\right]\right\}\left\{\prod_{i=1}^{N}p\left(\boldsymbol{\alpha}_i\right)\right\}.\tag{11}$$

where $p(\varsigma),p(\beta)$, and $p(\boldsymbol{\alpha})$ are the prior distributions, respectively. The joint posterior distribution based on the latent classes is given by

$$p\left(\boldsymbol{\varsigma},\boldsymbol{\beta}_j,\boldsymbol{\alpha},\boldsymbol{\pi},\boldsymbol{W}\,|\boldsymbol{Y}\right)\propto\left\{\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{c=1}^{C}\left[p\left(Y_{ij}=y_{ij}\,|\varsigma_j,\beta_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\right)f\left(W_{ij}\,|\varsigma_j,\beta_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\right)\right]\right\}\times\left\{\prod_{j=1}^{J}\left[p\left(\varsigma_j\right)p\left(\beta_j\right)\right]\right\}\left\{\prod_{c=1}^{C}p\left(\pi_c\right)\right\}.$$

where $p(\varsigma),p(\beta)$, and $p(\pi)$ are the prior distributions, respectively.

Step 1: Sampling the auxiliary variable W_{ij}, given the item intercept and interaction parameters ς_j,β_j and $\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c$. According to Equation (10), the full conditional posterior distribution of the random auxiliary variable W_{ij} is given by

$$f\left(W_{ij}\,|\varsigma_j,\beta_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\right)\propto\exp\left[-\frac{W_{ij}\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2}{2}\right]p\left(W_{ij}\,|1,0\right),\tag{12}$$

According to Biane et al. (2001) and Polson et al. (2013; p. 1341), the density function $p\left(W_{ij}\,|1,0\right)$ can be written as

$$p\left(W_{ij}\,|1,0\right)=\sum_{v=0}^{\infty}(-1)^{v}\frac{(2k+1)}{\sqrt{2\pi W_{ij}}}\exp\left[-\frac{(2k+1)^2}{8W_{ij}}\right].\tag{13}$$

Therefore, $f\left(W_{ij}\,|\varsigma_j,\beta_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\right)$ is proportional to

$$\sum_{v=0}^{\infty}(-1)^{v}\frac{(2k+1)}{\sqrt{2\pi W_{ij}}}\exp\left[-\frac{(2k+1)^2}{8W_{ij}}-\frac{W_{ij}\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2}{2}\right].\tag{14}$$

Finally, the specific form of the full conditional distribution of W_{ij} is as follows

$$W_{ij}\sim\mathrm{PG}\left(1,\left|\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right|\right).\tag{15}$$

Next, the Gibbs samplers are used to draw the item parameters.

Step 2: Sampling the intercept parameter ς_j for each item j. The prior distribution of ς_j is assumed to follow a normal distribution, that is, $\varsigma_j\sim N\left(\mu_\varsigma,\sigma_\varsigma^2\right)$. Given $\boldsymbol{Y}$, $\boldsymbol{W}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$, the fully condition posterior distribution of ς_j is given by

$$p\left(\varsigma_j\,|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\alpha},\beta_j\right)\propto\prod_{i=1}^{N}\left\{\frac{\left[\exp\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)\right]^{Y_{ij}}}{1+\exp\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)}f\left(W_{ij}\,|\varsigma_j,\beta_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\right)\right\}p\left(\varsigma_j\right),\tag{16}$$

where $f\left(W_{ij}\,|\varsigma_j,\beta_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\right)$ is equal to the following equation (the details see Polson et al., 2013; p. 1341)

$$f\left(W_{ij}\left|\varsigma_j,\beta_j,\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\right.\right)=\left\{\cosh\left(2^{-1}\left|\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right|\right)\right\}\frac{2^0}{\Gamma(1)}$$
$$\times\sum_{v=0}^{\infty}(-1)^v\frac{(2k+1)}{\sqrt{2\pi W_{ij}}}\exp\left[-\frac{(2k+1)^2}{8W_{ij}}-\frac{W_{ij}\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2}{2}\right]. \quad (17)$$

After rearrangement, the full conditional posterior distribution of ς_j can be written as follows

$$p\left(\varsigma_j\left|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\alpha},\beta_j\right.\right)\propto\prod_{i=1}^{N}\left\{\frac{\left[\exp\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)\right]^{Y_{ij}}}{1+\exp\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)}\right. \quad (18)$$
$$\left.\left[\cosh\left(2^{-1}\left|\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right|\right)\right]\times\exp\left[-\frac{\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2W_{ij}}{2}\right]\right\}p\left(\varsigma_j\right).$$

Therefore, the fully condition posterior distribution of ς_j follow normal distribution with mean

$$\mathrm{Var}_{\varsigma_j}\times\left(\mu_\varsigma\sigma_\varsigma^{-2}+\left(\sum_{i=1}^{N}W_{ij}\right)\left(\frac{\sum_{i=1}^{N}2Y_{ij}-1-2\beta_jW_{ij}\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}}{2\sum_{i=1}^{N}W_{ij}}\right)\right),$$

and variance

$$\mathrm{Var}_{\varsigma_j}=\left(\sigma_\varsigma^{-2}+\left(\sum_{i=1}^{N}W_{ij}\right)\right)^{-1}.$$

Step 3: Sampling the interaction parameter β_j for each item j. The prior distribution of β_j is assumed to follow a truncated normal distribution to satisfy the model identification restriction (Junker and Sijtsma, 2001; Henson et al., 2009; DeCarlo, 2012; Culpepper, 2015). That is, $\beta_j\sim N\left(\mu_\beta,\sigma_\beta^2\right)\mathrm{I}\left(\beta_j>0\right)$. Similarly, given $\boldsymbol{Y}$, $\boldsymbol{W}$, $\boldsymbol{\varsigma}$, and $\boldsymbol{\alpha}$, the full condition posterior distribution of β_j is given by

$$p\left(\beta_j\left|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\alpha},\boldsymbol{\varsigma}\right.\right)\propto\prod_{i=1}^{N}\left\{\frac{\left[\exp\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)\right]^{Y_{ij}}}{1+\exp\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)}\right. \quad (19)$$
$$\left.\left[\cosh\left(2^{-1}\left|\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right|\right)\right]\times\exp\left[-\frac{\left(\varsigma_j+\beta_j\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2W_{ij}}{2}\right]\right\}p\left(\beta_j\right).$$

Therefore, the fully condition posterior distribution of ς_j follow the truncated normal distribution with mean

$$\mathrm{Var}_{\beta_j}\times\left(\mu_\beta\sigma_\beta^{-2}+\sum_{i=1}^{N}\left[\left(\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2W_{ij}\right]\left(\frac{\sum_{i=1}^{N}\left(2Y_{ij}\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}-\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}-2\varsigma_jW_{ij}\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)}{2\sum_{i=1}^{N}\left[\left(\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2W_{ij}\right]}\right)\right)$$

and variance

$$\mathrm{Var}_{\beta_j}=\left\{\sigma_\beta^{-2}+\sum_{i=1}^{N}\left[\left(\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)^2W_{ij}\right]\right\}^{-1} \quad (20)$$

Step 4: Sampling the attribute vector $\boldsymbol{\alpha}_i$ for each examinee i. Given $\boldsymbol{Y}$, $\boldsymbol{W}$, $\boldsymbol{\varsigma}$, and $\boldsymbol{\beta}$, we can update the ith examinee's attribute vector $\boldsymbol{\alpha}_i$ from the following multinomial distribution

$$\boldsymbol{\alpha}_i\left|\boldsymbol{Y}_i,\boldsymbol{W}_i,\boldsymbol{\varsigma},\boldsymbol{\beta}\right.\sim\mathrm{Multinomial}\left(1,\left[\lambda_{i1},\ldots,\lambda_{iC}\right]\right). \quad (21)$$

where the probability that the attribute vector $\boldsymbol{\alpha}_i$ belongs to the cth$(c=1,\ldots,C)$ class can be written as

$$\lambda_{ic}=P\left(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c\left|\boldsymbol{Y}_i,\boldsymbol{W}_i,\boldsymbol{\varsigma},\boldsymbol{\beta},\boldsymbol{\pi}\right.\right)$$
$$=\frac{\pi_cp\left(\boldsymbol{Y}_i\left|\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c,\boldsymbol{\varsigma},\boldsymbol{\beta}\right.\right)f\left(\boldsymbol{W}_i\left|\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c,\boldsymbol{\varsigma},\boldsymbol{\beta}\right.\right)}{\sum_{c=1}^{C}\pi_cp\left(\boldsymbol{Y}_i\left|\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c,\boldsymbol{\varsigma},\boldsymbol{\beta}\right.\right)f\left(\boldsymbol{W}_i\left|\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c,\boldsymbol{\varsigma},\boldsymbol{\beta}\right.\right)}. \quad (22)$$

Step 5: Sampling the class membership probabilities $\boldsymbol{\pi}$. The prior of $\boldsymbol{\pi}$ is assumed to follow a Dirichlet distribution. I.e., $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C) \sim$Dirichlet$(\delta_0, \ldots, \delta_0)$. The full condition posterior distribution of the class membership probabilities $\boldsymbol{\pi}$ can be written as

$$\boldsymbol{\pi} \left| \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C \right. \sim \text{Dirichlet}\left(\delta_0 + \sum_{i=1}^{N} \text{I}\left(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_1\right), \ldots, \delta_0 + \sum_{i=1}^{N} \text{I}\left(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_C\right) \right). \quad (23)$$

4. SIMULATION STUDY

4.1. Simulation 1

4.1.1. Simulation Design

In this simulation study, the purpose is to assess the performance of the Pólya-Gamma Gibbs sampling algorithm. Considering the test length is $J = 30$, and the number of the attribute is set equal to $K = 5$. The Q-matrix is shown in **Table 1**, where the design of Q-matrix satisfies Gu and Xu (2019)'s DINA model identification conditions. For the true values of the class membership probabilities, we only consider the most general case that the class membership probabilities are flat though all class, i.e., $\pi_c = \frac{1}{2^K}$, $c = 1, \ldots, C$, where $C = 2^K$. Next, two factors and their varied test conditions are simulated. (a) two sample sizes ($N = 1000, 2000$) are considered; (b) Following Huebner and Wang (2011) and Culpepper (2015), four noise levels are considered to explore the relationship between noise level and recovery by constraining the true values of the item parameters. For each item, (b1) low noise level (LNL) case: $s_j = g_j = 0.1$; the corresponding true values of reparameterized parameters are $\zeta_j = -2.1972$, $\beta_j = 4.3945$; (b2) high noise level (HNL) case : $s_j = g_j = 0.2$; the corresponding true values of reparameterized parameters are $\zeta_j = -1.3863$, $\beta_j = 2.7726$; (b3) slipping higher than guessing (SHG) case: $s_j = 0.2$, $g_j = 0.1$; the corresponding true values of reparameterized parameters are $\zeta_j = -2.1972$, $\beta_j = 3.5835$; (b4) guessing higher than slipping (GHS) case: $s_j = 0.1$, $g_j = 0.2$; the corresponding true values of reparameterized parameters are $\zeta_j = -1.3863$, $\beta_j = 3.5835$. Fully crossing the different levels of these two factors yield 8 conditions.

TABLE 1 | The Q matrix design in the simulation study 1.

	Attribute Q(matrix)						Attribute Q(matrix)				
Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	0	1	1

4.1.2. Priors

Based on the four noise levels, the corresponding four kinds of non-informative prior are used. I.e.,

$$\begin{aligned}
&(b1)\ \zeta_j \sim N\left(-2.1972, 10^5\right), \beta_j \sim N\left(4.3945, 10^5\right) \text{I}\left(\beta_j > 0\right);\\
&(b2)\ \zeta_j \sim N\left(-1.3863, 10^5\right), \beta_j \sim N\left(2.7726, 10^5\right) \text{I}\left(\beta_j > 0\right);\\
&(b3)\ \zeta_j \sim N\left(-2.1972, 10^5\right), \beta_j \sim N\left(3.5835, 10^5\right) \text{I}\left(\beta_j > 0\right);\\
&(b4)\ \zeta_j \sim N\left(-1.3863, 10^5\right), \beta_j \sim N\left(3.5835, 10^5\right) \text{I}\left(\beta_j > 0\right),
\end{aligned}$$

where the purpose of using non-informative priors is to eliminate the influence of prior uncertainty on posterior inferences. Similarly, the non-informative Dirichlet prior distribution is employed for the class membership probabilities $\boldsymbol{\pi}$. I.e., $(\pi_1, \ldots, \pi_C) \sim$ Dirichlet $(1, \ldots, 1)$.

4.1.3. Convergence Diagnostics

As an illustration of the convergence of parameter estimates, we only consider the low noise level (LNL) case and the number of examinees is 1,000. Two methods are used to check the convergence of parameter estimates. One is the "eyeball" method to monitor the convergence by visually inspecting the history plots of the generated sequences (Hung and Wang, 2012; Zhan et al., 2017), and another method is to use the Gelman-Rubin method (Gelman and Rubin, 1992; Brooks and Gelman, 1998) to check the convergence of parameter estimates.

To implement the MCMC sampling algorithm, chains of length 20,000 with an initial burn-in period 10,000 are chosen. Four chains started at overdispersed starting values are run for each replication. The trace plots of Markov Chains for three randomly selected items and class membership probabilities are shown in **Figure 1**. In addition, the potential scale reduction factor (PSRF; Brooks and Gelman, 1998) values of all parameters are <1.1, which ensures that all chains converge as expected. The trace plots of PSRF values are shown in the simulation 2.

4.1.4. Evaluation Criteria for Convergence and Accuracy of Parameter Estimations

The accuracy of the parameter estimates is measured by two evaluation criteria, i.e., Bias and Mean Squared Error (MSE). Let η be the interested parameter. Assume that $M = 25$ data sets are generated. Also, let $\widehat{\eta}^{(m)}$ be the posterior mean obtained from the mth simulated data set for $m = 1, \ldots, M$.

The Bias for parameter is defined as

$$\text{Bias}\left(\eta\right) = \frac{1}{M} \sum_{m=1}^{M} \left(\widehat{\eta}^{(m)} - \eta\right), \quad (24)$$

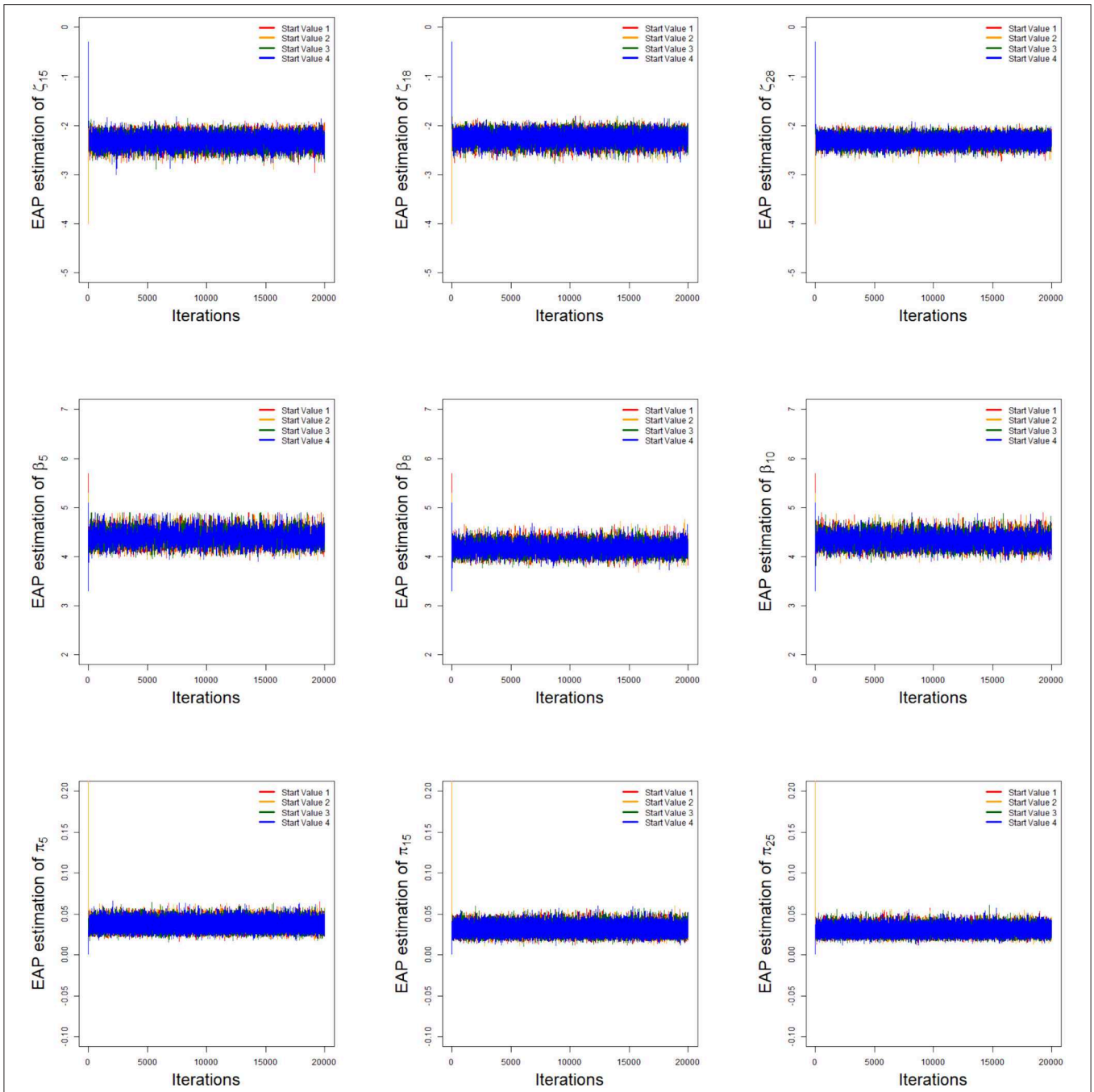


FIGURE 1 | The trace plots of the arbitrarily selected item and class membership probability parameters.

and the MSE for parameter is defined as

$$\mathrm{MSE}\,(\eta) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\eta}^{(m)} - \eta\right)^2. \tag{25}$$

For illustration purposes, we only show the Bias and MSE of $\boldsymbol{\varsigma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\pi}$ for the four noise levels based on 1,000 sample sizes in **Figures 2**, **3**. In the four noise levels, the Bias of $\boldsymbol{\varsigma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\pi}$ are near the zero values. However, the MSE of $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$ increase as the number of attributes required by the item increases. In the low noise level, the performances of the recovery for $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$ are well-based on the results of MSE, and the MSE of $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$ are <0.0250. The performances for the high noise level are worst in the four diagnosticity cases. Moreover, we find that when the item tests a attribute, the MSE of $\boldsymbol{\varsigma}$ is not much different from that of $\boldsymbol{\beta}$. However, the MSE of $\boldsymbol{\beta}$ is greater than that of $\boldsymbol{\varsigma}$ when the item

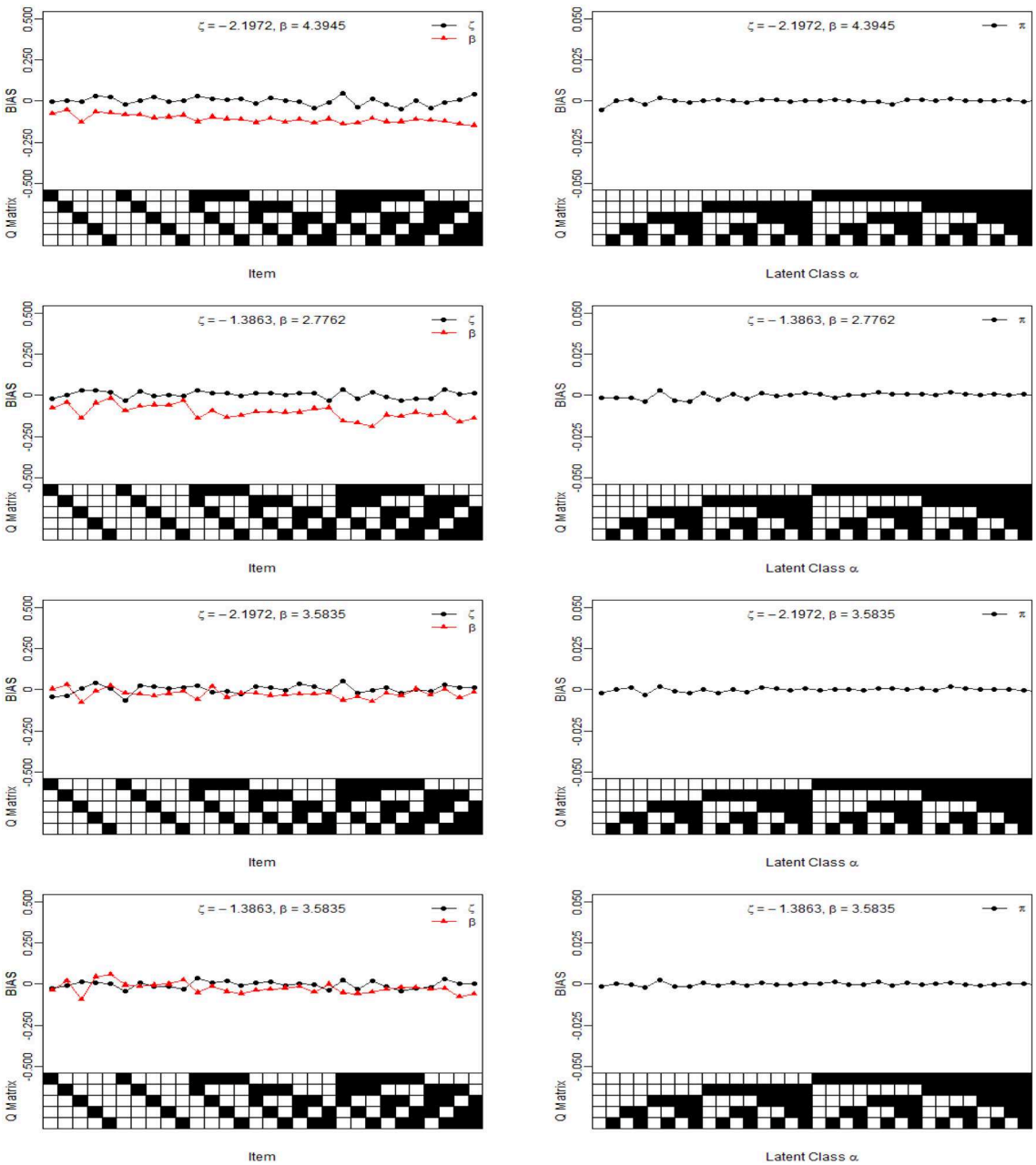


FIGURE 2 | The Bias of intercept, interaction and the latent class parameters under four different noise levels. The *Q* Matrix denotes the skills required for each item along the *x* axis, where the black square = "1" and white square = "0." The α_{ck} denotes the examinee who belongs to the *c*th latent class whether has mastered *k*th skill, where the black square = "1" for the presence of a skill and white square = "0" for the absence of a skill, $\alpha_c = (\alpha_{c1}, \ldots, \alpha_{cK})'$. Note the Bias values are estimated from 25 replications.

requires multiple attributes. The reason is due to a fact that the number of examinees for $\eta_{ij} = 1$ is almost equal to that of $\eta_{ij} = 0$ when the item tests a attribute, which is accurate for estimating the $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$. Along with the increase in the attributes required by the item, the number of examinees for $\eta_{ij} = 1$ reduces and the number of examinees for $\eta_{ij} = 0$ increases, thus resulting in the MSE of $\boldsymbol{\beta}$ higher than that of $\boldsymbol{\varsigma}$. Note that the MSE of $\boldsymbol{\beta}$ is dependent on the number of examinees for $\eta_{ij} = 1$.

The average Bias and MSE for $\boldsymbol{\varsigma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\pi}$ based on eight different simulation conditions are shown in **Table 2**. The following conclusions can be obtained. (1) Given a noise level, when the number of examinees increases from 1,000 to 2,000, the average MSE for $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$ show a decreasing trend. More specifically, when the number of examinees increases from 1,000 to 2,000, in the case of low noise level (LNL), the average MSE of $\boldsymbol{\varsigma}$ decreases from 0.048 to 0.034, the average MSE of $\boldsymbol{\beta}$ decreases

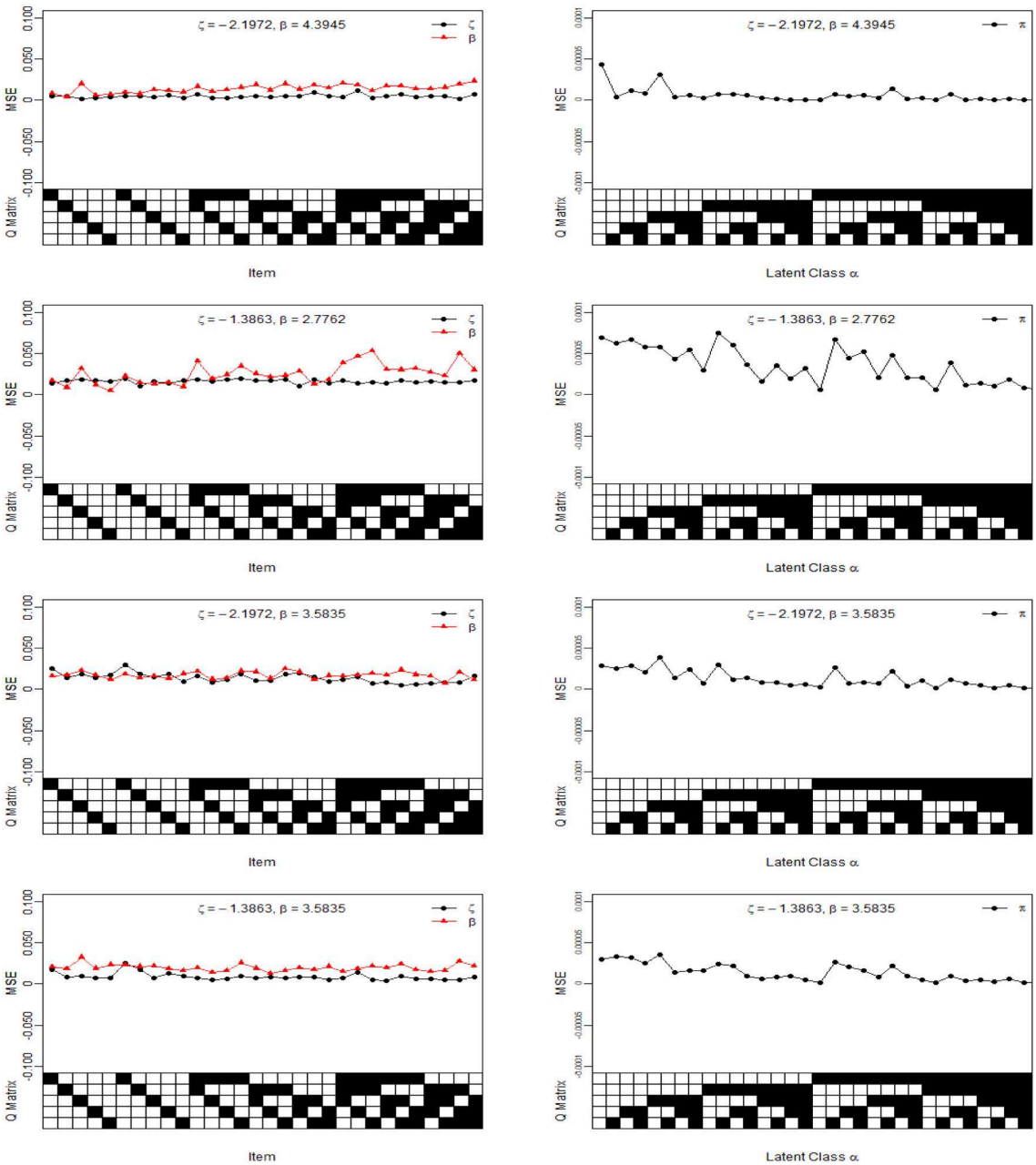


FIGURE 3 | The MSE of intercept, interaction and class membership probability parameters under four different diagnosticity cases. The Q Matrix denotes the skills required for each item along the *x* axis, where the black square = "1" and white square = "0." The α_{ck} denotes the examinee who belongs to the *c*th latent class whether has mastered *k*th skill, where the black square = "1" for the presence of a skill and white square = "0" for the absence of a skill, $\alpha_c = (\alpha_{c1}, \ldots, \alpha_{cK})'$. Note the MSE values are estimated from 25 replications.

from 0.0141 to 0.0107. In the case of high noise level (HNL), the average MSE of $\boldsymbol{\varsigma}$ decreases from 0.0163 to 0.0117, the average MSE of $\boldsymbol{\beta}$ decreases from 0.0254 to 0.0239. In the case of the slipping higher than the guessing (SHG), the average MSE of $\boldsymbol{\varsigma}$ decreases from 0.0139 to 0.0078, the average MSE of $\boldsymbol{\beta}$ decreases from 0.0172 to 0.0159. In the case of the guessing higher than the slipping (GHS), the average MSE of $\boldsymbol{\varsigma}$ decreases from 0.0088 to 0.0041, the average MSE of $\boldsymbol{\beta}$ decreases from 0.0198 to 0.0181. (2) Given a noise level, when the number of examinees increases from 1,000 to 2,000, In the case of four kinds of noises, the average MSE of $\boldsymbol{\pi}$ are basically the same and close to 0 under the conditions of four noise levels. (3) Compared with the other three noise level, the average MSE of $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$ are largest at high noise level. In summary, the Bayesian algorithm provides accurate estimates for $\boldsymbol{\varsigma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\pi}$ in term of various numbers of examinees.

4.2. Simulation 2

In this simulation study, we compare MH algorithm and PGGSA from two aspects: the accuracy and convergence. We consider 1,000 examinees to answer 30 items, and the number of the attribute is set equal to $K = 5$. The true values of ζ_j and β_j are set equal to -2.1972 and 4.3945 for each item. The corresponding true values of s_j and g_j are equal to 0.1 for each item. The class membership probabilities are flat though all classes, i.e., $\pi_c = \frac{1}{2^K}$, $c = 1, \ldots, C$, where $C = 2^K$. We specify the following non-informative priors to the PGGSA and MH algorithm: $\zeta_j \sim N\left(-2.1972, 10^5\right), \beta_j \sim N\left(4.3945, 10^5\right) \mathrm{I}\left(\beta_j > 0\right)$ and $(\pi_1, \ldots, \pi_C) \sim \mathrm{Dirichlet}\,(1, \ldots, 1)$.

It is known that an improper proposal distribution for MH algorithm can seriously reduce the acceptance probability of sampling. Most of the posterior samples are rejected. Therefore, the low sampling efficiency is usually unavoidable, and the reduction in the number of valid samples may lead to incorrect inference results. In contrast, our PGGSA takes the acceptance probability as 1 to draw the samples from fully condition posterior distributions. The following proposal distributions for the intercept and interaction parameters are considered in the process of implementing MH algorithm. The sampling details of MH algorithm, see **Appendix**. Note that the class membership probabilities are updated through the same way for the PGGSA and MH algorithms.

- Case 1: $\varsigma_j \sim N(\varsigma_j^{(r)}, 0.1), \beta_j \sim N(\beta_j^{(r)}, 0.1)\mathrm{I}\left(\beta_j > 0\right)$.
- Case 2: $\varsigma_j \sim N(\varsigma_j^{(r)}, 1), \beta_j \sim N(\beta_j^{(r)}, 1)\mathrm{I}\left(\beta_j > 0\right)$.

To compare the convergence of all parameters for the PGGSA and MH algorithm with different proposal distributions, the convergence of item and class membership probability parameters are evaluated by judging whether the values of PSRF are <1.1. From **Figure 4**, we find that the intercept, interaction and class membership probability parameters have already converged at the 5,000 step iterations for the PGGSA. The fastest convergence is the class membership probability parameters followed by intercept parameters. For the MH algorithm, some parameters do not converge after 5,000 step iterations for the proposal distributions with the variances of 0.1. The convergence of the proposal distributions with the variances of 1 is worse than the convergence of the proposal distributions with the variances of 0.1, even some parameters do not reach convergence at the end of the 10,000 step iterations. Moreover, the Bias and MSE are used to evaluate the performances of the two algorithms in **Table 3**. It has been proved that the selection of the proposal distribution has an important influence on the accuracy of parameter estimation. The process of finding the proper turning parameter is time consuming. In addition,

TABLE 2 | The average Bias and MSE for ς, $\boldsymbol{\beta}$, and $\boldsymbol{\pi}$.

	Number of examinees 1,000			
	LNL (b1)	**HNL (b2)**	**SHG (b3)**	**GHS (b4)**
BIAS				
ς	0.0023	0.0046	0.0042	−0.0039
$\boldsymbol{\beta}$	−0.1077	−0.1016	−0.0235	−0.0248
$\boldsymbol{\pi}$	−0.0000	−0.0000	−0.0000	−0.0000
MSE				
ς	0.0048	0.0163	0.0139	0.0088
$\boldsymbol{\beta}$	0.0141	0.0254	0.0172	0.0198
$\boldsymbol{\pi}$	0.0000	0.0000	0.0000	0.0000
	Number of examinees 2,000			
BIAS				
ς	0.0089	0.0089	0.0023	−0.0020
$\boldsymbol{\beta}$	−0.0890	0.0588	−0.0003	−0.0041
$\boldsymbol{\pi}$	−0.0000	0.0000	−0.0000	0.0000
MSE				
ς	0.004	0.0117	0.0078	0.0041
$\boldsymbol{\beta}$	0.0107	0.0239	0.0159	0.0181
$\boldsymbol{\pi}$	0.0000	0.0000	0.0000	0.0000

Note that the Bias and MSE denote the average Bias and MSE for the parameters. ς represents all intercept parameters, $\boldsymbol{\beta}$ represents all interaction parameters, $\boldsymbol{\pi}$ represents all class membership probabilities parameters.

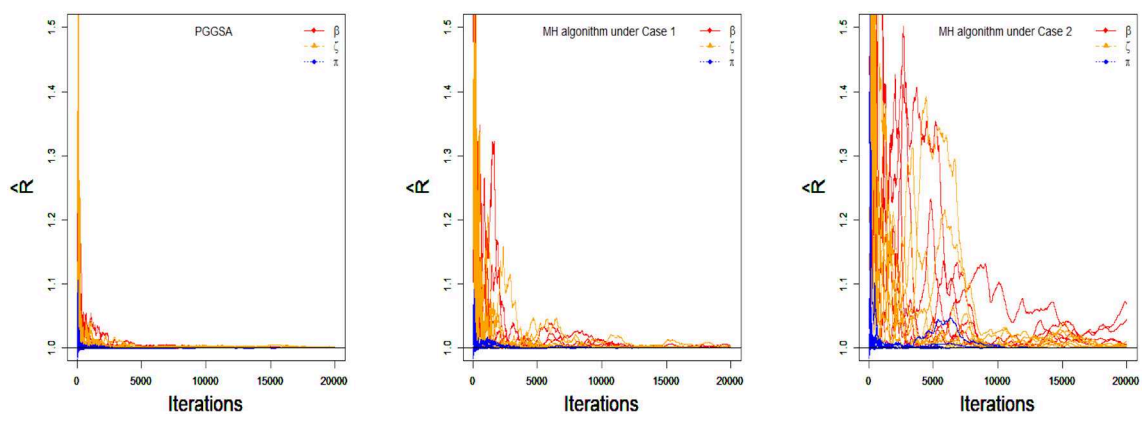


FIGURE 4 | The trace plots of PSRF values for the simulation study 2.

we investigate the efficiency of the two algorithms from the perspective of the time consumed by implementing them. On a desktop computer [Intel(R) Xeon(R) E5-2695 V2 CPU] with 2.4 GHz dual core processor and 192 GB of RAM memory, PGGSA and MH algorithm, respectively consume 3.6497 and 4.7456 h when Markov chain are run for 20,000 iterations for a replication experiment, where MH algorithm is used to implement the Case 1. In summary, PGGSA is more effective than MH algorithm in estimating model parameters.

4.3. Simulation 3

This simulation study is to show that PGGSA is sufficiently flexible to recover various prior distributions for the item and class membership probability parameters. The simulation design is as follows:

The number of the examinees is $N = 1,000$, and the test length is $J = 30$, and the number of the attributes is set equal to $K = 5$. The true values of item intercept and interaction parameters are -2.1972 and 4.3945 for each item at low noise level. The class membership probabilities are flat though all classes, i.e., $\pi_c = \frac{1}{2^K}, c = 1, \ldots, C$, where $C = 2^K$.

The non-informative Dirichlet prior distribution is employed for the class membership probabilities $\boldsymbol{\pi}$. I.e., $(\pi_1, \ldots, \pi_C) \sim$ Dirichlet $(1, \ldots, 1)$, and two kinds of prior distributions are considered for the intercept and interaction parameters:

(1) Informative prior: **Type I**: $\zeta_j \sim N(-2.1972, 0.5)$, $\beta_j \sim N(4.3945, 0.5)\, \mathrm{I}(\beta_j > 0)$; **Type II**: $\zeta_j \sim N(-2.1972, 1)$, $\beta_j \sim N(4.3945, 1)\, \mathrm{I}(\beta_j > 0)$;
(2) Non-informative prior: **Type III**: $\zeta_j \sim N(-2.1972, 10^3)$, $\beta_j \sim N(4.3945, 10^3)\, \mathrm{I}(\beta_j > 0)$; **Type IV**: $\zeta_j \sim N(-2.1972, 10^5)$, $\beta_j \sim N(4.3945, 10^5)\, \mathrm{I}(\beta_j > 0)$.

PGGSA is iterated 20,000 times. The first 10,000 iterations are discarded as burn-in period. 25 replications are considered in this simulation study. The PSRF values of all parameters for each simulation condition are <1.1. The Bias and MSE of the $\boldsymbol{\zeta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\pi}$ based on two kinds of prior distributions are shown in **Table 4**.

TABLE 3 | Evaluating accuracy of parameter estimation using the two algorithms in the simulation study 2.

	PGGSA		MH algorithm under Case 1		MH algorithm under Case 2	
	Bias	MSE	Bias	MSE	Bias	MSE
$\boldsymbol{\zeta}$	0.0023	0.0048	0.0016	0.0069	0.0021	0.0081
$\boldsymbol{\beta}$	−0.1077	0.0141	−0.1042	0.0152	−0.1087	0.0174
$\boldsymbol{\pi}$	−0.0000	0.0000	−0.0007	0.0005	−0.0004	0.0011

Note that the Bias and denote the average Bias and MSE for the parameters. $\boldsymbol{\zeta}$ represents all intercept parameters, $\boldsymbol{\beta}$ represents all interaction parameters, $\boldsymbol{\pi}$ represents all latent class probabilities parameters.

TABLE 4 | Evaluating the accuracy of parameters based on different prior distributions in the simulation study 3.

Type of prior	Evaluation index	$\boldsymbol{\zeta}$	$\boldsymbol{\beta}$	$\boldsymbol{\pi}$
Type I	Bias	0.0024	−0.1044	−0.0000
	MSE	0.0047	0.0134	0.0000
Type II	Bias	0.0026	−0.1059	−0.0000
	MSE	0.0047	0.0138	0.0000
Type III	Bias	0.0022	−0.1068	−0.0000
	MSE	0.0048	0.0140	0.0000
Type IV	Bias	0.0023	−0.1077	−0.0000
	MSE	0.0048	0.0141	0.0000

Note that the Bias and denote the average Bias and MSE for the parameters. $\boldsymbol{\zeta}$ represents all intercept parameters, $\boldsymbol{\beta}$ represents all interaction parameters, $\boldsymbol{\pi}$ represents all latent class probabilities parameters.

TABLE 5 | Evaluating accuracy of attribute and class membership probability parameter estimations using PGGSA and Gibbs algorithm in the simulation study 4.

		Attribute(α)		CMP(π)	
Noise level	Algorithm	CPCR	AAMA	Bias	MSE
LNL	PGGSA	0.8740	0.9693	−0.0000	0.0000
	Gibbs	0.8722	0.9688	−0.0000	0.0000
HNL	PGGSA	0.5643	0.8696	−0.0000	0.0000
	Gibbs	0.5697	0.8718	−0.0000	0.0000
SHG	PGGSA	0.7480	0.9336	−0.0000	0.0000
	Gibbs	0.7429	0.9308	−0.0000	0.0000
GHS	PGGSA	0.8436	0.9310	−0.0000	0.0000
	Gibbs	0.8484	0.9338	−0.0000	0.0000

Note that the CMP denotes the class membership probability. Bias and MSE denote the average Bias and MSE for the class membership probability parameters.

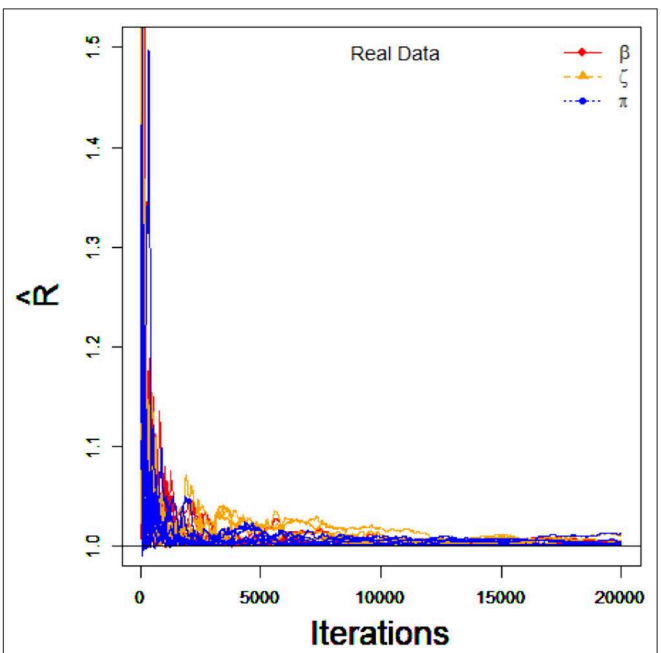


FIGURE 5 | The trace plots of PSRF values for the real data.

4.3.1. Result Analysis

From **Table 4**, we find that the Bias and MSE of $\boldsymbol{\varsigma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ are almost the same under different prior distributions. More specifically, the Bias of $\boldsymbol{\varsigma}$ ranges from 0.0022 to 0.026, $\boldsymbol{\beta}$ ranges from −0.1077 to −0.1044, and the Bias of $\boldsymbol{\pi}$ under the two kinds of prior distributions is equal to −0.0000. In addition, the MSE of $\boldsymbol{\varsigma}$ ranges from 0.0047 to 0.0048, $\boldsymbol{\beta}$ ranges from 0.0134 to 0.0141, and the MSE of $\boldsymbol{\pi}$ under the two kinds of prior distributions is equal to −0.0000. This shows that the accuracy of parameter estimation can be guaranteed by PGGSA, no matter what the informative prior or non-informative distributions are chosen.

4.4. Simulation 4

The main purpose of this simulation study is to compare PGGSA and Culpepper (2015)'s Gibbs sampling algorithm (Geman and Geman, 1984; Tanner and Wong, 1987; Gelfand and Smith, 1990; Albert, 1992; Damien et al., 1999; Béguin and Glas, 2001; Sahu, 2002; Bishop, 2006; Fox, 2010; Chen et al., 2018; Lu et al., 2018) on the attribute classification accuracy and the estimation accuracy of class membership probability parameter ($\boldsymbol{\pi}$).

The number of the examinees is $N = 1,000$. Considering the test length is $J = 30$, and the number of the attribute is set equal to $K = 5$. The Q-matrix is shown in **Table 1**. Four noise levels are considered in this simulation, i.e., LNL, HNL, SHG, and GHS. The true values of item parameters under the four noise levels, see the simulation study 1. For the true values of the class membership probabilities, we only consider the most general case that the class membership probabilities are flat though all classes, i.e., $\pi_c = \frac{1}{2^K}, c = 1, \ldots, C$, where $C = 2^K$.

For the prior distributions of the two algorithms, we use the non-informative prior distributions to eliminate the influence of the prior distributions on the posterior inference. The non-informative Dirichlet prior distribution is employed for the class membership probabilities $\boldsymbol{\pi}$. I.e., $(\pi_1, \ldots, \pi_C) \sim \text{Dirichlet}(1, \ldots, 1)$, and the non-informative prior distributions of item parameters under the two algorithms based on the four noise levels are set as follows

- (**LNL case**): PGGSA: $\zeta_j \sim N\left(-2.1972, 10^5\right), \beta_j \sim N\left(4.3945, 10^5\right) \mathrm{I}\left(\beta_j > 0\right)$. v.s. Gibbs algorithm: $s_j \sim Beta\left(1,1\right), g_j \sim Beta\left(1,1\right) \mathrm{I}\left(g_j < 1 - s_j\right)$;
- (**HNL case**): PGGSA: $\zeta_j \sim N\left(-1.3863, 10^5\right), \beta_j \sim N\left(2.7726, 10^5\right) \mathrm{I}\left(\beta_j > 0\right)$. v.s. Gibbs algorithm: $s_j \sim Beta\left(1,1\right), g_j \sim Beta\left(1,1\right) \mathrm{I}\left(g_j < 1 - s_j\right)$;
- (**SHG case**): PGGSA: $\zeta_j \sim N\left(-2.1972, 10^5\right), \beta_j \sim N\left(3.5835, 10^5\right) \mathrm{I}\left(\beta_j > 0\right)$. v.s. Gibbs algorithm: $s_j \sim Beta\left(1,1\right), g_j \sim Beta\left(1,1\right) \mathrm{I}\left(g_j < 1 - s_j\right)$;
- (**GHS case**): PGGSA: $\zeta_j \sim N\left(-1.3863, 10^5\right), \beta_j \sim N\left(3.5835, 10^5\right) \mathrm{I}\left(\beta_j > 0\right)$. v.s. Gibbs algorithm: $s_j \sim Beta\left(1,1\right), g_j \sim Beta\left(1,1\right) \mathrm{I}\left(g_j < 1 - s_j\right)$.

PGGSA and Gibbs algorithm are iterated 20,000 times. The first 10,000 iterations are discarded as burn-in period for the two algorithms. Twenty-five replications are considered for the two algorithms in this simulation study. The PSRF values of all parameters for each simulation condition are <1.1. Culpepper's the R "dina" package is used to implement the Gibbs sampling.

The correct pattern classification rate (CPCR), the average attribute match rate (AAMR) are used as the evaluation criteria to evaluate the attributes. These statistics are defined as

$$\text{CPCR} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(\boldsymbol{\alpha}_i = \widehat{\boldsymbol{\alpha}_i}\right), \ \text{AAMA} = \frac{1}{N \times K}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathrm{I}\left(\boldsymbol{\alpha}_{ik} = \widehat{\boldsymbol{\alpha}_{ik}}\right). \tag{26}$$

TABLE 6 | The Q matrix design and MCMC estimations of $\boldsymbol{\varsigma}$ and $\boldsymbol{\beta}$.

	Attribute(Q Matrix)					$\widehat{\varsigma}$			$\widehat{\beta}$		
Item	α_1	α_2	α_3	α_4	α_5	EAP	*SD*	HPDI	EAP	*SD*	HPDI
1	1	0	0	0	0	−2.3274	0.0277	[−2.4998, −1.9766]	3.3884	0.0662	[2.8484, 3.8721]
2	1	1	1	1	0	−1.2990	0.0225	[−1.5639, −1.0087]	3.4200	0.0947	[2.8714, 4.0615]
3	1	0	0	0	0	−1.2247	0.0276	[−1.5357, −1.0000]	4.2999	0.0294	[3.9575, 4.4999]
4	1	1	1	1	1	−1.8944	0.0358	[−2.2841, −1.5472]	3.8815	0.1217	[3.2857, 4.4977]
5	0	0	1	0	0	−1.7971	0.1042	[−2.4667, −1.2948]	2.9899	0.1145	[2.5007, 3.6131]
6	1	1	1	1	0	−2.3961	0.0113	[−2.4999, −2.1653]	3.7058	0.0817	[3.1377, 4.2461]
7	1	1	1	1	0	−2.1109	0.0322	[−2.4999, −1.8117]	4.3549	0.0223	[4.0401, 4.4998]
8	1	1	0	0	0	−1.3433	0.0409	[−1.7158, −1.0005]	4.1817	0.0558	[3.7427, 4.4999]
9	1	0	1	0	0	−1.6266	0.0566	[−2.0725, −1.1512]	4.2735	0.0384	[3.8794, 4.4998]
10	1	0	1	1	1	−1.5226	0.0246	[−1.8180, −1.2110]	4.1072	0.0796	[3.5678, 4.4999]
11	1	0	1	0	0	−1.7813	0.0681	[−2.3048, −1.2903]	4.0454	0.0884	[3.5121, 4.4999]
12	1	0	1	1	0	−2.3802	0.0119	[−2.4998, −2.1534]	4.2212	0.0481	[3.7945, 4.4994]
13	1	1	1	1	0	−1.8221	0.0399	[−2.2142, −1.4328]	3.5878	0.1009	[2.9818, 4.1937]
14	1	1	1	1	1	−2.4279	0.0058	[−2.4999, −2.2647]	3.8646	0.0982	[3.3310, 4.4741]
15	1	1	1	1	0	−2.4298	0.0060	[−2.4999, −2.2551]	4.0033	0.0765	[3.5339, 4.4946]

Note that α_1 denotes the skill of subtract basic fractions, α_2 denotes the skill of reduce and simplify, α_3 denotes the skill of separate whole from fraction, α_4 denotes the skill of borrow from whole, α_5 denotes the skill of convert whole to fraction. EAP denotes expected a posteriori *estimator. SD denotes standard deviation. HPDI denotes 95% highest posterior density intervals (HPDI).*

TABLE 7 | The posterior probability distribution of the latent class parameters for the Fraction Subtraction Test.

Latent classes					$\widehat{\pi}$		
α_1	α_2	α_3	α_4	α_5	EAP	*SD*	HPDI
0	0	0	0	0	1.909%	0.0003	[0.0000, 0.0542]
1	0	0	0	0	0.766%	0.0000	[0.0000, 0.0208]
0	1	0	0	0	1.743%	0.0002	[0.0000, 0.0504]
0	0	1	0	0	1.299%	0.0001	[0.0000, 0.0367]
0	0	0	1	0	2.001%	0.0002	[0.0000, 0.0533]
0	0	0	0	1	1.790%	0.0002	[0.0000, 0.0540]
1	1	0	0	0	0.677%	0.0000	[0.0000, 0.0190]
1	0	1	0	0	1.898%	0.0001	[0.0000, 0.0443]
1	0	0	1	0	0.756%	0.0000	[0.0000, 0.0203]
1	0	0	0	1	0.822%	0.0000	[0.0000, 0.0222]
0	1	1	0	0	1.162%	0.0001	[0.0000, 0.0339]
0	1	0	1	0	1.808%	0.0002	[0.0000, 0.0507]
0	1	0	0	1	1.943%	0.0003	[0.0000, 0.0567]
0	0	1	1	0	1.242%	0.0001	[0.0000, 0.0330]
0	0	1	0	1	1.165%	0.0001	[0.0000, 0.0328]
0	0	0	1	1	1.778%	0.0002	[0.0000, 0.0486]
1	1	1	0	0	10.146%	0.0039	[0.0002, 0.2029]
1	1	0	1	0	0.709%	0.0000	[0.0000, 0.0198]
1	1	0	0	1	0.764%	0.0000	[0.0000, 0.0205]
1	0	1	1	0	0.546%	0.0000	[0.0000, 0.0140]
1	0	1	0	1	1.782%	0.0001	[0.0000, 0.0419]
1	0	0	1	1	0.751%	0.0000	[0.0000, 0.0201]
0	1	1	1	0	1.326%	0.0001	[0.0000, 0.0370]
0	1	1	0	1	1.181%	0.0001	[0.0000, 0.0357]
0	1	0	1	1	1.675%	0.0002	[0.0000, 0.0473]
0	0	1	1	1	1.167%	0.0001	[0.0000, 0.0335]
1	1	1	1	0	9.680%	0.0002	[0.0667, 0.1264]
1	1	1	0	1	11.119%	0.0038	[0.0001, 0.2078]
1	1	0	1	1	0.688%	0.0000	[0.0000, 0.0195]
1	0	1	1	1	0.429%	0.0000	[0.0000, 0.0119]
0	1	1	1	1	1.119%	0.0001	[0.0000, 0.0320]
1	1	1	1	1	34.142%	0.0004	[0.2998, 0.3844]

Note that α_1 denotes the skill of subtract basic fractions, α_2 denotes the skill of reduce and simplify, α_3 denotes the skill of separate whole from fraction, α_4 denotes the skill of borrow from whole, α_5 denotes the skill of convert whole to fraction.

where $\widehat{\boldsymbol{\alpha}}_i = (\widehat{\alpha_{i1}}, \alpha_{i2}, \ldots, \alpha_{iK})'$ represents examinee i's estimated attribute patterns. Next, the evaluation results of the accuracy of the two algorithms for attribute patterns and class membership probability parameters are shown in **Table 5**.

In **Table 5**, we find that the results of the attributes classification accuracy (CPCR and AAMA criteria) are basically the same for PGGSA and Gibbs algorithm under four kinds of noise levels. More specifically, the values of CPCR and AAMA for two algorithms under the HNL case are lowest. At the LNL case, the values of CPCR and AAMA for two algorithms are the highest. In addition, the CPCR value for the SHG case is lower than the CPCR value for the GHS, while the corresponding AAMA values are basically the same for the SHG case and GHS case. This indicates that slipping parameters ($\boldsymbol{s}$) have important influence on the CPCR. In term of the two algorithms, the Bias and MSE of the classification membership parameters ($\boldsymbol{\pi}$) are basically the same and close to zero under the four noise levels.

5. EMPIRICAL EXAMPLE

In this example, a fraction subtraction test data is analyzed based on Tatsuoka (1990), Tatsuoka (2002), and de la Torre and Douglas (2004). The middle school students of 2,144 take part in this test to response 15 fraction subtraction items, where five attributes are measured, including subtract basic fractions, reduce and simplify, separate whole from fraction, borrow from whole, and convert whole to fraction. We choose 536 of 2,144 students in this study. These students are divided into 2^5 latent classes based on the five attributes. The reparameterized DINA model is used to analyze the cognitive response data.

The priors of parameters are also the same as the simulation 1. I.e., the non-informative priors are used in this empirical example analysis. To implement PGGSA, chains of length 20,000 with an initial burn-in period 10,000 are chosen. The PSRF is used to evaluate the convergence of each parameters. The trace plots of PSRF values for all parameters is shown in **Figure 5**. We find that the values of PSRF are <1.1.

The Q matrix, the expected *a posteriori* (EAP) estimators of the item parameters, the corresponding standard deviation (SD), and 95% highest posterior density intervals (HPDIs) of these item parameters are shown in **Table 6**. Based on the **Table 6**, we transform intercept and interaction parameters into traditional slipping and guessing parameters to analyze item characteristics. We find that the expected *a posteriori* (EAP) estimations of the five items with the lowest slipping are item 3, item 8, item 9, item 10, and item 11 in turn. The EAP estimations of slipping parameters for the five items are 0.0461, 0.0585, 0.0708, 0.0754, and 0.1039. This shows that these items are not easy to slipping compared with the other ten items. In addition, the EAP estimations of five items with the highest guessing are item 3, item 2, item 8, item 10, and item 11 in turn. The EAP estimations of guessing parameters for the five items are 0.2271, 0.2143, 0.2069, 0.1790, and 0.1441. Furthermore, we find that items 3, 8, 10, and 11 have low slipping parameters and high guessing parameters, which indicates that these items are more likely to be guessed correctly.

The EAP estimations of the class membership probabilities, $\widehat{\pi}_c$, $c = 1, \ldots, 32$, and the corresponding SD and 95% HPDI are reported in **Table 7**. The top five classes that the majority of examinees are classified into these classes are respectively "11111,""11101,""11100,""11110," and "00010."The estimation results show that $\widehat{\pi}_{32} = 34.142\%$ of the examinees have mastered all the five skills, and $\widehat{\pi}_{28} = 11.119\%$ of the examinees have mastered the four skills except the skill of borrow from whole, and the examinees who only have mastered the three skills of subtract basic fractions, reduce and simplify, separate whole from fraction account for $\widehat{\pi}_{17} = 10.146\%$, and $\widehat{\pi}_{27} = 9.680\%$ of the examinees have mastered the four skills except the skill of convert whole to fraction, and the examinees who only have mastered

a skill of skill of borrow from whole account for $\widehat{\pi}_3 = 2.001\%$. In addition, among the thirty-two classes, the class with the lowest number of the examinees is $\widehat{\pi}_{30} = 0.429\%$. I.e., when the examinees have mastered the skills of subtract basic fractions, separate whole from fraction, borrow from whole, and convert whole to fraction, the proportion of examinees who do not master the skill of reduce and simplify is very low. According to the $\widehat{\pi}_3 = 1.743\%$ and $\widehat{\pi}_{30} = 0.429\%$, we find that the skill of reduce and simplify is easier to master than the other four skills.

6. CONCLUSION

In this paper, a novel and effective PGGSA based on auxiliary variables is proposed to estimate the widely applied DINA model. PGGSA overcomes the disadvantages of MH algorithm, which requires to repeatedly adjust the specification of tuning parameters to achieve a certain acceptance probability and thus increases the computational burden. However, the computational burden of the PGGSA becomes intensive especially as the CDMs become more complex, when a large number of examinees or the items is considered, or a large number of the MCMC sample size is used. Therefore, it is desirable to develop a standing-alone R package associated with C++ or Fortran software for more extensive CDMs and large-scale cognitive assessment tests.

In addition, Pólya-Gamma Gibbs sampling algorithm can be used to estimate many cognitive diagnosis models, which is not limited to the DINA model. These cognitive diagnostic models include DINO (Templin and Henson, 2006), Compensatory RUM (Hartz, 2002; Henson et al., 2009), and log-linear CDM (LCDM; von Davier, 2005; Henson et al., 2009) and so on. More specifically, first of all, the parameters of these cognitive diagnosis models are reparameterized, and then the logit link function is used to link these parameters with the response. Further, we can use Pólya-Gamma Gibbs sampling algorithm to estimate these reparameterized cognitive diagnosis models. Discussions of the reparameterized cognitive diagnosis models based on logit link function, see Henson et al. (2009).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://cran.r-project.org/web/packages/CDM/index.html.

AUTHOR CONTRIBUTIONS

JZ completed the writing of the article. ZZ and JZ provided article revisions. JZ, JL, and ZZ provided the key technical support. JT provided the original thoughts.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00384/full#supplementary-material

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibb sampling. *J. Educ. Stat.* 17, 251–269. doi: 10.3102/10769986017003251

Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J am Stat Assoc*. 88, 669–679. doi: 10.2307/2290350

Béguin, A. A., and Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* 66, 541–561. doi: 10.1007/BF02296195

Biane, P., Pitman, J., and Yor, M. (2001). Probability laws related to the Jacobi theta and Riemann zeta Functions, and Brownian Excursions. *B Am Math Soc.* 38, 435–465.

Bishop, C. M. (2006). "Slice sampling," in *Pattern Recognition and Machine Learning*. New York, NY: Springer, 546.

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

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York, NY: Springer.

Chen, Y., Culpepper, S. A., Chen, Y., and Douglas, J. A. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika* 83, 89–108. doi: 10.1007/s11336-017-9579-4

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

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Am. Stat.* 49, 327–335. doi: 10.1080/00031305.1995.10476177

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

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika* 81, 1142–1163. doi: 10.1007/s11336-015-9477-6

Damien, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by auxiliary variables. *J. R. Stat. Soc. B* 61, 331–344. doi: 10.1111/1467-9868.00179

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

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

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

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika.* 69, 333–353. doi: 10.1007/BF02295640

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: the DINA model, classification, latent class sizes, and the Q-matrix. *Appl Psychol Meas.* 35, 8–26. doi: 10.1177/0146621610377081

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Appl. Psychol. Meas.* 36, 447–468. doi: 10.1177/0146621612449069

DiBello, L. V., Roussos, L. A., and Stout, W. F. (2007). "Review of cognitively diagnostic assessment and a summary of psychometric models," in *Handbook of Statistics*, Psychometrics, Vol. 26, eds C. R. Rao and S. Sinharay (Amsterdam: Elsevier), 979–1030.

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

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85, 398–409. doi: 10.1080/01621459.1990.10476213

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

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741. doi: 10.1109/TPAMI.1984.4767596

Gu, Y., and Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika* 84, 468–483. doi: 10.1007/s11336-018-9619-8

Haberman, S. J., and von Davier, M. (2007). "Some notes on models for cognitively based skill diagnosis," in *Handbook of Statistics*, Psychometrics, Vol. 26, eds C. R. Rao and S. Sinharay (Amsterdam: Elsevier), 1031–1038.

Hartz, S. M. (2002). *A Bayesian framework for the unifified model for assessing cognitive abilities: blending theory with practicality* (Unpublished doctoral dissertation), University of Illinois at Urbana-Champaign, Champaign, IL, United States.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109. doi: 10.1093/biomet/57.1.97

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

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

Hung, L.-F., and Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *J. Educ. Behav. Stat.* 37, 231–255. doi: 10.3102/1076998611402503

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Lu, J., Zhang, J. W., and Tao, J. (2018). Slice-Gibbs sampling algorithm for estimating the parameters of a multilevel item response model. *J. Math. Psychol.* 82, 12–25. doi: 10.1016/j.jmp.2017.10.005

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Stat. Assoc.* 108, 1339–1349. doi: 10.1080/01621459.2013.829001

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., and Templin, J. L. (2007). "The fusion model skills diagnosis system," in *Cognitive Diagnostic Assessment for Education: Theory and Applications*, eds J. P. Leighton and M. J. Gierl (Cambridge: Cambridge University Press), 275–318.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *J. Stat. Comput. Simul.* 72, 217–232. doi: 10.1080/00949650212387

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–550. doi: 10.1080/01621459.1987.10478458

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *J. R. Stat. Soc. C Appl. Stat.* 51, 337–350. doi: 10.1111/1467-9876.00272

Tatsuoka, K. K. (1984). *Analysis of Errors in Fraction Addition and Subtraction Problems*. Champaign, IL: Computer-Based Education Research Laboratory, University of Illinois at Urbana-Champaign.

Tatsuoka, K. K. (1990). "Toward an integration of item-response theory and cognitive error diagnosis." In *Diagnostic monitoring of skill and knowledge acquisition*. eds N. Frederiksen, R. Glaser, A. Lesgold, and M. Shafto (Hillsdale, NJ: Erlbaum) 453–488.

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

von Davier, M. (2005). *A General Diagnostic Model Applied to Language Testing Data*. Research report no. RR-05-16. Princeton, NJ: Educational Testing Service.

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

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: two variants of a model equivalency. *Br. J. Math. Stat. Psychol.* 67, 49–71. doi: 10.1111/bmsp.12003

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

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

Zhan, P., Jiao, H. and Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *Brit J Math Stat Psy*. 71, 262–286. doi: 10.1111/bmsp.12114

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 24 March 2020
doi: 10.3389/fpsyg.2020.00525

Cognitive Diagnostic Models for Rater Effects

Xiaomin Li[1], Wen-Chung Wang[2] and Qin Xie[3]*

[1] Centre for Child and Family Science, The Education University of Hong Kong, Tai Po, Hong Kong, [2] Assessment Research Centre, The Education University of Hong Kong, Tai Po, Hong Kong, [3] Department of Linguistics and Modern Language Studies, The Education University of Hong Kong, Tai Po, Hong Kong

In recent decades, cognitive diagnostic models (CDMs) have been intensively researched and applied to various educational and psychological tests. However, because existing CDMs fail to consider rater effects, the application of CDMs to constructed-response (CR) items that involve human raters is seriously limited. Given the popularity of CR items, it is desirable to develop new CDMs that are capable of describing and estimating rater effects on CR items. In this study, we developed such new CDMs within the frameworks of facets models and hierarchical rater models, using the log-linear cognitive diagnosis model as a template. The parameters of the new models were estimated with the Markov chain Monte Carlo methods implemented in the freeware JAGS. Simulations were conducted to evaluate the parameter recovery of the new models. Results showed that the parameters were recovered fairly well and the more data there were, the better the recovery. Implications and applications of the new models were illustrated with an empirical study that adopted a fine-grained checklist to assess English academic essays.

Keywords: cognitive diagnostic models, facets models, hierarchical rater models, rater effect, item response theory

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Wenchao Ma,
The University of Alabama, United States
Jung Yeon Park,
KU Leuven, Belgium

****Correspondence:***
Xiaomin Li
xmli@eduhk.hk

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

Received: *09 October 2019*
Accepted: *05 March 2020*
Published: *24 March 2020*

Citation:
Li X, Wang W-C and Xie Q (2020) Cognitive Diagnostic Models for Rater Effects. Front. Psychol. 11:525. doi: 10.3389/fpsyg.2020.00525

In the past few decades, extensive research has been conducted in the area of cognitive diagnosis, and a wide range of cognitive diagnostic models (CDMs) (also called diagnostic classification models; DCMs) has been developed to provide fine-grained information about students' learning strengths and weaknesses (Tatsuoka, 1985; Templin, 2004; de la Torre, 2011; Chiu and Douglas, 2013; Hansen and Cai, 2013). Popular CDMs include the deterministic inputs, noisy and gate (DINA) model (Haertel, 1989; Junker and Sijtsma, 2001; de la Torre and Douglas, 2004), the deterministic input, noisy or gate model (Templin and Henson, 2006), and the reduced reparameterized unified model (Hartz, 2002). Unlike unidimensional item response theory (IRT), which provides a single score for a student's proficiency on a latent continuum, CDMs offer a profile of multiple binary (mastery or non-mastery) statuses of certain knowledge or skills.

In applications of CDMs, item responses to multiple-choice items, for example, are assumed to be objectively scored. In many situations, such as educational assessment, performance appraisal, psychological diagnosis, medical examination, sports competition, and singing contests, responses to constructed-response (CR) or performance-based items are evaluated by human raters. Different raters often exhibit different degrees of severity. There are two major approaches to rater effects in the IRT framework. One is to treat raters as a third facet, in addition to the item and person facets, to highlight the impact of rater effects on the item scores. Examples are the Rasch facets models (Linacre, 1989) and the random-effect facets model (Wang and Wilson, 2005). The other approach is to employ signal detection theory to describe raters' judgment. Examples include the hierarchical

rater model (HRM; Patz et al., 2002) and the latent class extension of signal detection theory (DeCarlo et al., 2011). The facets approach and the HRM approach have very different assumptions regarding rater behaviors, as discussed below. The resulting measures of a person (ratee) can only be considered fair and valid for individual comparison if rater effects are directly accounted for in the IRT models.

Rater effects can happen in the CDM framework when raters are recruited to mark item responses. In this study, we adapt these two approaches (facets and HRM) to the CDM framework to account for rater effects. Based on the same logic above, the resulting profiles (a set of binary latent attributes) of persons (ratees) are fair and valid for individual comparison only when rater effects are directly accounted for in the CDMs. The remainder of this paper is organized as follows. First, the facets and HRM approaches within the IRT framework are briefly introduced. Second, these two approaches are adapted to the CDM framework to create new CDMs to account for rater effects. Third, a series of simulations are conducted to evaluate the parameter recovery of the new CDMs, and their results are summarized. Fourth, an empirical example about essay writing is provided to demonstrate applications of the new models. Finally, conclusions are drawn and suggestions for future studies are provided.

INTRODUCTION TO THE FACETS AND HRM APPROACHES

The Facets Approach

In the facets approach, raters are treated as instruments to measure ratees, just like items are. Raters are recruited to provide their own expertise to make judgments of ratees' performance; therefore, the more raters there are, the more reliable the measurement of the ratees. In the facets model (Linacre, 1989), the log-odds (logit) of scoring k over $k-1$ on item j for ratee i judged by rater r is defined as:

$$\log(P_{ijkr}/P_{ij(k-1)r}) = \theta_i - \beta_{jk} - \eta_r \quad (1)$$

where P_{ijkr} and $P_{ij(k-1)r}$ are the probabilities of receiving a score of k and k - 1, respectively, for ratee i on item j from rater r; θ_i is the latent (continuous) trait of ratee i and is often assumed to follow a normal distribution; β_{jk} is the kth threshold of item j; η_r is the severity of rater r. A positive (negative) η_r decreases (increases) the probability of receiving a high score. Equation 1 can be easily generalized to more than three facets.

In Equation 1, a rater has a single parameter η_r to account for the rater's degree of severity, meaning that the rater holds a constant degree of severity throughout all ratings. In reality, it is likely that a rater exhibits some fluctuations in severity when giving ratings. If so, Equation 1 is too stringent, and the assumption of constant severity needs to be relaxed. To account for the intra-rater fluctuations in severity, Wang and Wilson (2005) proposed adding a random-effect parameter to the facets model, which can be expressed as:

$$\log(P_{ijkr}/P_{ij(k-1)r}) = \theta_i - \beta_{jk} - (\eta_r + \zeta_{ir}) \quad (2)$$

where ζ_{ir} is assumed to follow a normal distribution, with mean 0 and variance σ_r^2; others have been defined in Equation 1; θ and ζ are assumed to be mutually independent. Where appropriate, slope parameters can be added and covariates (e.g., gender) can be incorporated to account for variations in θ and η (Wang and Liu, 2007). The facets models have been widely used to account for rater effects in practice (Engelhard, 1994, 1996; Myford and Wolfe, 2003, 2004).

The HRM Approach

In the HRM approach, it is argued that thorough scoring rubrics can (in theory) be programmed into computers so human raters are no longer needed (Patz et al., 2002). However, until computer scoring is made possible (e.g., it is not cost-effective to develop e-raters), human raters are still in demand but they are expected to function like scoring machines (clones) as closely as possible. Unfortunately, human judgment may deviate remarkably from machine scoring, which brings random noise to the ratings. Only when raters act exactly like scoring machines will a CR item provide as much information as an objective (machine-scorable) item does. Following this logic, increasing the number of raters will not increase the precision of ratee measurements.

The HRM involves two steps. In the first step, the scores provided by raters are treated as indicators of the latent (true, or ideal) category for ratee i's response to item j. Let ξ_{ij} be the latent category for ratee i on item j. The probability that rater r will assign a rating k given ξ_{ij} is assumed to be proportional to a normal density with a mean $\xi_{ij} - \phi_r$ and a standard deviation ψ_r:

$$P_{ijkr} \propto \exp\left[-\frac{1}{2\psi_r^2}[k - (\xi_{ij} - \phi_r)]^2\right] \quad (3)$$

where ϕ_r represents the severity for rater r: a value of 0 indicates the rater is most likely to provide the same rating as the latent (true) category, a negative value indicates that the rater tends to be lenient, a positive value implies that the rater tends to be severe, and ψ_r represents the rater's variability: the larger the value, the less reliable (consistent) the ratings.

In the second step, the latent category ξ_{ij} is used as the indicator of a ratee's ability via an IRT model such as the partial credit model (Masters, 1982):

$$P_{ijl} \equiv P(\xi_{ij} = l\,|\theta_i) = \frac{\exp\sum_{k=0}^{l}(\theta_i - \delta_{jk})}{\sum_{m=0}^{M_j}\exp\sum_{k=0}^{m}(\theta_i - \delta_{jk})} \quad (4)$$

$$\text{logit}(P_{ijl}) \equiv \log(P_{ijl}/P_{ij(l-1)}) = \theta_i - \delta_{jk} \quad (5)$$

where M_j is the maximum score of item j, δ_{jk} is the kth step parameter of item j, θ_i is the latent trait for person i. By defining $\sum_{k=0}^{0}(\theta_i - \delta_{jk}) \equiv 0$ and $\sum_{k=0}^{m}(\theta_i - \delta_{jk}) \equiv \sum_{k=1}^{m}(\theta_i - \delta_{jk})$, the probability of scoring 0 is $P_{ij0} = \frac{1}{\sum_{m=0}^{M_j}\exp\sum_{k=0}^{m}(\theta_i - \delta_{jk})}$. Note that ξ_{ij} in Equation 4 is latent rather than observed in the standard partial credit model.

A problem in the HRM, also noted by Patz et al. (2002), is that a relatively small value for ψ_r would lead to difficulties in determining a unique value for ϕ_r because the posterior

distribution of ϕ_r is almost uniform (DeCarlo et al., 2011). Another limitation of the HRM is that it can account for a rater's severity and inconsistency, but not for other rater effects, such as centrality. To resolve these problems, DeCarlo et al. (2011) extended the HRM by incorporating a latent class extension of the signal detection theory as:

$$P_{ijk^*r} = F\left[a_{jr}(\xi_{ij} - c_{jkr})\right], \quad (6)$$

where P_{ijk^*r} denotes the probability of assigning a rating less than or equal to k (denoted as k^*) given ξ_{ij}; F can be a cumulative normal or logistic distribution; a_{jr} is a slope (sensitivity) parameter for rater r on item j; c_{jkr} is the kth ordered location parameter of item j for rater r. Like ψ_r in Equation 3, a_{jr} depicts how sensitive or reliable the ratings are for rater r on item j. A close investigation of c_{jkr} can reveal rater severity and centrality. Further, by including an autoregressive time series process and a parameter for overall growth, the HRM approach is also feasible for longitudinal data (Casabianca et al., 2017).

THE LOG-LINEAR COGNITIVE DIAGNOSIS MODEL

Cognitive diagnostic models have been applied to large-scale educational assessments such as the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), the National Assessment of Educational Progress (NEAP), and the Test of English as a Foreign Language (TOEFL) to obtain information about students' cognitive abilities (Tatsuoka et al., 2004; Xu and von Davier, 2008; Chiu and Seo, 2009; Templin and Bradshaw, 2013). In these datasets, both multiple-choice items and CR items are used. For example, in the PIRLS reading comprehension test, approximately half of the items require examinees to write down their responses, which are then marked by human raters. In these studies of fitting CDMs to large-scale educational assessments, rater effects were not considered simply because existing CDMs could not account for rater effects. To resolve this problem, we developed new CDMs for rater effects within both the facets and HRM frameworks. We adopted the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) as a template because it includes many CDMs as special cases. Nevertheless, the new models developed in this study can also apply easily to other general CDMs, such as the general diagnostic model (von Davier, 2008) or the generalized DINA model (de la Torre, 2011).

Under the LCDM, the probability of success (scoring 1) on item j for person i is defined as:

$$P_{ij1} \equiv P(X_{ij} = 1|\boldsymbol{\alpha}_i) = \frac{\exp\left(\lambda_{j,0} + \boldsymbol{\lambda}_j^T h\left(\boldsymbol{\alpha}_i, \mathbf{q}_j\right)\right)}{1 + \exp\left(\lambda_{j,0} + \boldsymbol{\lambda}_j^T h\left(\boldsymbol{\alpha}_i, \mathbf{q}_j\right)\right)} \quad (7)$$

$$\text{logit}(P_{ij1}) \equiv \log[P_{ij1}/(1 - P_{ij1})] = \lambda_{j,0} + \boldsymbol{\lambda}_j^T h(\boldsymbol{\alpha}_i, \mathbf{q}_j) \quad (8)$$

where α_i is the latent profile of person i, $\lambda_{j,0}$ defines the probability of success for those persons who have not mastered any of the attributes required by item j; $\boldsymbol{\lambda}_j^T$ is a $(2^K - 1)$ by 1 vector of weights for item j; q_{jk} is the entry for item j in the Q-matrix; $h(\boldsymbol{\alpha}_i, \mathbf{q}_j)$ is a set of linear combinations of $\boldsymbol{\alpha}_i$ and $\mathbf{q}_j$; $\boldsymbol{\lambda}_j^T h(\boldsymbol{\alpha}_i, \mathbf{q}_j)$ can be written as:

$$\lambda_j^T h\left(\alpha_i, q_j\right) = \sum_{k=1}^{K} \lambda_{jk}\left(\alpha_{ik} q_{jk}\right) + \sum_{k=1}^{K} \sum_{v>k} \lambda_{jkv}\left(\alpha_{ik}\alpha_{iv} q_{jk} q_{jv}\right) + \ldots \quad (9)$$

For item j, the exponent includes an intercept term, all main effects of attributes, and all possible interaction effects between attributes. By constraining some of the LCDM parameters, many existing CDMs can be formed (Henson et al., 2009). For example, for a three-attribute item, the DINA model can be defined as:

$$P_{ij1} = \frac{\exp\left(\lambda_{j,0} + \lambda_{j,123}\alpha_{i1}\alpha_{i2}\alpha_{i3}\right)}{1 + \exp\left(\lambda_{j,0} + \lambda_{j,123}\alpha_{i1}\alpha_{i2}\alpha_{i3}\right)} \quad (10)$$

Although we concentrate on dichotomous responses in this study for illustrative purpose, Equation 7 can be extended to accommodate polytomous items. Let P_{ijk} and $P_{ij(k-1)}$ be the probabilities of scoring k and $k - 1$ on item j for person i, respectively. Equation 8 can be extended as:

$$\text{logit}(P_{ijk}) \equiv \log\left(P_{ijk}/P_{ij(k-1)}\right) = \lambda_{j,0,k-1} + \boldsymbol{\lambda}_j^T h(\boldsymbol{\alpha}_i, \mathbf{q}_j), \quad (11)$$

where $\lambda_{j,0,k-1}$ is the $(k - 1)$th intercept for item j. Equation 11 is based on adjacent-category logit. Actually, cumulative logit (Hansen, 2013) and other approaches are also feasible (Ma and de la Torre, 2016).

For the ease of understanding and interpretation, item parameters in the LCDM can be expressed as follows, which is commonly called as the guessing parameters (g_j) and slip parameters (s_j):

$$g_j = \frac{\exp\left(\lambda_{j,0}\right)}{1 + \exp\left(\lambda_{j,0}\right)} \quad (12)$$

$$s_j = 1 - \frac{\exp\left(\lambda_{j,0} + \boldsymbol{\lambda}_j^T(\boldsymbol{\alpha}_i, \mathbf{q}_j)\right)}{1 + \exp\left(\lambda_{j,0} + \boldsymbol{\lambda}_j^T(\boldsymbol{\alpha}_i, \mathbf{q}_j)\right)} \quad (13)$$

representing the probability of success without mastering all the required attributes, and the probability of failure with mastering all the required attributes, respectively.

NEW CDMs WITH THE FACETS APPROACH

All existing CDMs involve two facets: person and item. When items are marked by human raters, a third facet is needed to account for rater effects. To accomplish this, Equation 8 can be extended as:

$$\text{logit}(P_{ijr1}) \equiv \log[P_{ijr1}/(1 - P_{ijr1})] = \lambda_{j,0} - \eta_r + \boldsymbol{\lambda}_i^T h(\boldsymbol{\alpha}_i, \mathbf{q}_j) \quad (14)$$

where P_{ijr1} is the probability of success (scoring 1) on item j for person i marked by rater r; η_r is the severity of rater r; other terms

have been defined. A positive (negative) η_r decreases (increases) the probability of success. If $\eta_r = 0$ for all raters, Equation 14 simplifies to Equation 8. That is, Equation 14 is a three-facet extension of the LCDM.

When there is a concern about intra-rater variations in severity, η_r in Equation 14 can be replaced with $\eta_r + \zeta_{ir}$. Moreover, Equation 14 can be easily generalized to include more than three facets. For example, in the Test of Spoken English (TSE) assessment system, examinees' speaking tasks are marked on multiple criteria by human raters, so four facets are involved: ratee, task, rater, and criterion. In such cases, Equation 14 can be extended to four facets as:

$$\begin{aligned}\text{logit}(P_{\text{ijrs1}}) &\equiv \log[P_{\text{ijrs1}}/(1-P_{\text{ijrs1}})] \\ &= \lambda_{\text{j},0} - \eta_{\text{r}} - \gamma_{\text{s}} + \boldsymbol{\lambda}_{\text{i}}^{\text{T}} h(\boldsymbol{\alpha}_{\text{i}}, \mathbf{q}_{\text{j}}) \end{aligned} \quad (15)$$

where P_{ijrs1} is the probability of success (scoring 1) on task j along criterion s for examinee i marked by rater r; γ_s is the threshold of criterion s; other terms have been defined. Generalization to more facets is straightforward. For polytomous items, Equation 14 can be extended as:

$$\text{logit}(P_{\text{ijrk}}) \equiv \log\left(P_{\text{ijrk}}/P_{\text{ijr(k-1)}}\right) = \lambda_{\text{j},0,\text{k}-1} - \eta_{\text{r}} + \boldsymbol{\lambda}_{\text{i}}^{\text{T}} h(\boldsymbol{\alpha}_{\text{i}}, \mathbf{q}_{\text{j}}) \quad (16)$$

where P_{ijrk} and $P_{\text{ijr(k-1)}}$ be the probabilities of scoring k and $k-1$ on item j for examinee i marked by rater r, respectively; other terms have been defined.

NEW CDMs WITH THE HRM APPROACH

The signal detection model in the first step in the HRM approach can be defined as in Equation 3 or 6, with the constraint of $k = 0$ or 1 because of dichotomous items. For dichotomous items, Equations 3 and 6 become equivalent, except there is a single ψ_r for each rater in Equation 3, but multiple a_{jr} (across items) for each rater in Equation 6. The IRT model in the second step (Equation 4 or 5) can be replaced with a CDM like the LCDM. Using the LCDM as template, the new model can be written as:

$$P_{\text{ij1}} \equiv P(\xi_{ij} = 1 \,|\boldsymbol{\alpha}_{\text{i}}) = \frac{\exp\left(\lambda_{\text{j},0} + \boldsymbol{\lambda}_{\text{j}}^{\text{T}} h\left(\boldsymbol{\alpha}_{\text{i}}, \mathbf{q}_{\text{j}}\right)\right)}{1 + \exp\left(\lambda_{\text{j},0} + \boldsymbol{\lambda}_{\text{j}}^{\text{T}} h\left(\boldsymbol{\alpha}_{\text{i}}, \mathbf{q}_{\text{j}}\right)\right)} \quad (17)$$

$$\text{logit}(P_{\text{ij1}}) \equiv \log[P_{\text{ij1}}/(1-P_{\text{ij1}})] = \lambda_{\text{j},0} + \boldsymbol{\lambda}_{\text{j}}^{\text{T}} h\left(\boldsymbol{\alpha}_{\text{i}}, \mathbf{q}_{\text{j}}\right) \quad (18)$$

where ξ_{ij} is the latent binary category of person i on item j; other terms have been defined. Comparing Equations 17 and 7, one finds that the category is latent in Equation 17, but observed in Equation 7. For polytomous items, Equation 18 can be extended as:

$$\text{logit}(P_{\text{ijk}}) \equiv \log\left(P_{\text{ijk}}/P_{\text{ij(k-1)}}\right) = \lambda_{\text{j},0,\text{k}-1} + \boldsymbol{\lambda}_{\text{j}}^{\text{T}} h\left(\boldsymbol{\alpha}_{\text{i}}, \mathbf{q}_{\text{j}}\right) \quad (19)$$

PARAMETER ESTIMATION

Parameters in the new facets-CDM and HRM-CDMs can be estimated by utilizing Markov chain Monte Carlo (MCMC) methods (de la Torre and Douglas, 2004; Ayers et al., 2013), which treat parameters as random variables and repeatedly draw from their full conditional posterior distributions over a large number of iterations. In this study, the freeware JAGS (Version 4.2.0; Plummer, 2015) and the R2jags package (Version 0.5-7; Su and Yajima, 2015) in *R* (Version 3.3.0 64-bit; R Core Team, 2016) were used to estimate model parameters. JAGS uses a default option of the Gibbs sampler and offers a user-friendly tool for constructing Markov chains for parameters, so the derivation of the joint posterior distribution of the model parameters becomes attainable. We used the Gelman–Rubin diagnostic statistic (Gelman and Rubin, 1992) to assess convergence, in which a value smaller than 1.1 is typically regarded as convergence as a rule of thumb. In the facets-CDMs, the rater severity was constrained at a zero mean for model identification. Our pilot simulation supported the use of 10,000 iterations, with the first 5,000 iterations as burn-in and the remaining 5,000 iterations for the point estimates (expected *a posteriori*) and their standard errors by sampling one in every 10 values. The resulting Gelman–Rubin diagnostic statistic indicated no convergence problem.

Two simulation studies were conducted to evaluate the recovery of item parameters and person profiles for the two newly proposed models with rater effects. Moreover, we evaluated the effects of ignoring rater effects by comparing the proposed models (with rater effect) and standard models (without rater effect) in the simulations. In particular, Study I evaluated the item and person recovery of the facets-CDM under different rating designs. Study II assessed the implementation of the HRM-CDM. One hundred replications were conducted under each condition. For comparison, all simulated data were also analyzed with the standard CDMs, which did not consider rater effects.

SIMULATION STUDY I: FACETS-CDM

Design

Rating design is a practical issue because it involves resource allocation. A good rating design can save a great deal of resource while holding acceptable precision of ratee measurement. According to the procedures of Chiu et al. (2009), latent ability θ of 500 ratees were drawn from a multivariate normal distribution $MVN(0, \Sigma)$, with the diagonal and off-diagonal elements of the covariance matrix taking a value of 1 and 0.5, respectively. A correlation of 0.5 between attributes was specified to mimic moderate to medium correlations between attributes in educational settings. Assuming that the underlying continuous ability for the ith ratee was $\boldsymbol{\theta}_{\text{i}}^{\text{T}} = (\theta_{\text{i1}}, \theta_{\text{i2}}, \cdots, \theta_{\text{iK}})$, the profile pattern $\boldsymbol{\alpha}_{\text{i}}^{\text{T}} = (\alpha_{\text{i1}}, \alpha_{\text{i2}}, \cdots, \alpha_{\text{iK}})$ was determined by

$$\alpha_{\text{ik}} = \begin{cases} 1, & \text{if } \theta_{\text{ik}} \geq \Phi^{-1}\left(\dfrac{k}{K+1}\right), \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

The test consisted of 10 dichotomous items measuring five attributes, as shown in **Table 1**, and 10 raters. Dichotomous responses were simulated according to the facets-CDM (Equation 11). The generating intercepts ($\lambda_{j,0}$), main effects ($\lambda_{j,1}$), two-way interactions ($\lambda_{j,2}$), and rater severities (η_r) are listed in **Table 3**, and the resulting range of the guessing parameters and slip parameters was [0.08, 0.20] and [0.07, 0.19], respectively.

Four kinds of rating design were used: (a) completely crossed design, where every ratee was judged by every rater; (b) balanced incomplete design, where each ratee was judged by three raters and each rater judged 150 ratees; (c) unbalanced incomplete design, where each ratee was judged by three raters but different raters judged different numbers of ratees; (d) random design, where 20 ratees were judged by all raters and the remaining 480 ratees were judged by three raters randomly selected from the rater pool. The completely crossed design, although seldom used when there are a large number of ratees (e.g., several hundred), was adopted here to provide reference information about the parameter recovery of the facets-CDMs. In the three incomplete designs, raters were connected by a set of common ratees. Detailed specification of the incomplete designs is shown in **Table 2**.

Analysis

The generated data were analyzed with (a) the data-generating facets-CDM (saturated) model and (b) the standard CDM without considering rater effects, where the ratings given by the raters were treated as responses to virtual items with identical item parameters. Based on prior studies (e.g., Li and Wang, 2015; Zhan et al., 2019), a less informative normal prior was specified for all model parameters across the two models. Specifically, a normal prior with mean zero and standard deviation four was assumed for the intercepts ($\lambda_{j,0}$), main effects ($\lambda_{j,1}$), two-way interactions ($\lambda_{j,2}$), and rater severities (η_r). Moreover, a truncated normal distribution was specified to constraint the main effect parameters ($\lambda_{j,1}$) to be positive. In doing so, the probabilities of correct responses increased as a function of mastering each required attribute. To evaluate the recovery of item parameters, we computed the bias and root mean squared error (RMSE) of these estimates across replications. For person profiles, we computed the mean accurate recovery rate. In the completely crossed design, each item received 5,000 scores (500 ratees times 10 raters), each ratee received 100 scores (10 items times 10 raters), and each rater gave 5,000 scores (10 items times 500 ratees); in the three incomplete designs, each item received approximately 1,500 scores (500 ratees times 3 raters), each ratee received 30 scores (10 items times 3 raters) except 20 ratees received 100 scores (10 items times 10 raters) in the random design, and each rater gave approximately 1,500 scores (10 items times 150 ratees). In general, the more the data points, the better the parameter estimation and profile recovery. It was thus anticipated that when the facets-CDM was fit, the parameter estimation and recovery rates would be better in the completely crossed design than in the three incomplete designs. When the standard CDM was fit, the parameter estimation and recovery rates would be poor because the rater effects were not considered.

TABLE 1 | Q-matrix for the ten items in the simulations.

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	1	0	0	0
7	0	1	1	0	0
8	0	0	1	1	0
9	0	0	0	1	1
10	1	0	0	0	1

1s mean the attributes are required, and 0s mean the attributes are not required.

Results

Table 3 lists the generating values, the bias values, and the RMSE values for the two models under the four designs. When the facets-CDM was fit, the RMSE values were not large, ranging from 0.07 to 0.24 (M = 0.16) in the completely crossed design, from 0.10 to 0.52 (M = 0.23) in the balanced incomplete design, from 0.12 to 0.51 (M = 0.23) in the unbalanced incomplete

TABLE 2 | Number of ratees under the incomplete designs in simulation study I (Facets-CDM).

	Rater									
	1	2	3	4	5	6	7	8	9	10
Balanced										
	50	50	50							
		50	50	50						
			50	50	50					
				50	50	50				
					50	50	50			
						50	50	50		
							50	50	50	
								50	50	50
	50								50	50
	50	50								50
Total	150	150	150	150	150	150	150	150	150	150
Unbalanced										
	50	50	50							
		68	68	68						
			44	44	44					
				58	58	58				
					35	35	35			
						51	51	51		
							50	50	50	
								55	55	55
	40								40	40
	49	49								49
Total	139	167	162	170	137	144	136	156	145	144
Random										
Total	134	155	157	141	168	158	152	130	153	152

TABLE 3 | Generating values, bias, root mean square error (RMSE), and profile recovery rates (%) in simulation study I (Facets-CDM).

		Complete design				Balanced design				Unbalanced design				Random design			
		Facets-CDM		Standard CDM		Facets-CDM		Standard CDM		Facets-CDM		Standard CDM		Facets-CDM		Standard CDM	
Par.	Gen	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\lambda_{1,0}$	−2.00	−0.13	0.21	0.30	0.30	−0.21	0.28	0.08	0.11	−0.12	0.22	0.16	0.19	−0.12	0.22	0.17	0.19
$\lambda_{2,0}$	−1.40	−0.12	0.19	0.23	0.23	−0.19	0.25	0.04	0.07	−0.21	0.25	−0.02	0.06	−0.13	0.18	0.11	0.13
$\lambda_{3,0}$	−1.79	−0.13	0.20	0.27	0.27	−0.18	0.25	0.09	0.16	−0.18	0.24	0.09	0.13	−0.20	0.25	0.10	0.14
$\lambda_{4,0}$	−1.37	−0.15	0.21	0.20	0.20	−0.19	0.26	0.02	0.09	−0.17	0.23	0.08	0.09	−0.21	0.28	0.03	0.08
$\lambda_{5,0}$	−1.85	−0.12	0.21	0.28	0.28	−0.28	0.31	0.01	0.09	−0.16	0.21	0.10	0.15	−0.16	0.19	0.17	0.20
$\lambda_{6,0}$	−2.42	−0.14	0.23	0.33	0.33	−0.26	0.33	−0.06	0.16	−0.11	0.18	−0.05	0.18	−0.30	0.36	−0.05	0.13
$\lambda_{7,0}$	−1.57	−0.10	0.20	0.26	0.27	−0.12	0.22	0.01	0.16	−0.09	0.29	−0.03	0.16	−0.11	0.24	0.07	0.10
$\lambda_{8,0}$	−1.95	−0.10	0.21	0.31	0.32	−0.17	0.24	0.00	0.14	−0.18	0.23	−0.02	0.17	−0.19	0.24	−0.02	0.18
$\lambda_{9,0}$	−2.07	−0.11	0.20	0.32	0.33	−0.22	0.31	−0.06	0.23	−0.09	0.18	0.03	0.13	−0.12	0.27	0.07	0.21
$\lambda_{10,0}$	−1.69	−0.10	0.22	0.28	0.29	−0.18	0.28	−0.05	0.12	−0.08	0.29	0.02	0.16	−0.09	0.17	0.09	0.12
$\lambda_{1,1}$	3.72	−0.01	0.09	−0.62	0.62	−0.03	0.20	−0.63	0.65	−0.13	0.18	−0.84	0.84	−0.03	0.14	−0.60	0.62
$\lambda_{2,1}$	2.82	−0.06	0.10	−0.55	0.55	−0.15	0.21	−0.56	0.58	−0.19	0.24	−0.58	0.59	−0.18	0.20	−0.65	0.65
$\lambda_{3,1}$	3.41	0.03	0.08	−0.55	0.56	−0.01	0.18	−0.58	0.60	−0.05	0.17	−0.61	0.62	−0.05	0.11	−0.67	0.68
$\lambda_{4,1}$	2.91	0.01	0.08	−0.50	0.50	−0.06	0.10	−0.42	0.43	−0.09	0.14	−0.58	0.59	−0.04	0.18	−0.50	0.52
$\lambda_{5,1}$	3.38	0.00	0.10	−0.56	0.57	−0.05	0.10	−0.53	0.54	−0.03	0.12	−0.59	0.60	−0.08	0.14	−0.63	0.65
$\lambda_{6,1}$	1.26	0.04	0.10	−0.12	0.14	−0.02	0.18	0.02	0.16	−0.13	0.20	0.04	0.15	0.14	0.29	0.13	0.27
$\lambda_{7,1}$	1.04	−0.04	0.07	−0.20	0.20	−0.03	0.26	−0.01	0.21	−0.15	0.28	−0.04	0.19	−0.09	0.20	−0.10	0.17
$\lambda_{8,1}$	1.18	−0.05	0.09	−0.20	0.21	−0.03	0.15	−0.04	0.13	−0.04	0.22	0.00	0.25	−0.03	0.24	0.05	0.26
$\lambda_{9,1}$	1.00	−0.03	0.09	−0.16	0.17	−0.01	0.20	0.05	0.23	−0.09	0.23	0.08	0.26	−0.08	0.27	0.00	0.23
$\lambda_{10,1}$	0.96	0.00	0.07	−0.14	0.15	−0.01	0.15	0.06	0.14	−0.03	0.17	0.07	0.17	−0.04	0.10	−0.02	0.11
$\lambda_{6,2}$	2.16	−0.01	0.24	−0.42	0.48	0.07	0.32	−0.72	0.76	0.20	0.51	−0.79	0.82	0.00	0.48	−0.70	0.81
$\lambda_{7,2}$	2.14	0.07	0.19	−0.29	0.32	−0.13	0.39	−0.78	0.81	0.14	0.39	−0.58	0.66	0.24	0.40	−0.47	0.54
$\lambda_{8,2}$	1.98	0.09	0.20	−0.28	0.31	−0.03	0.22	−0.56	0.59	0.13	0.43	−0.51	0.61	0.26	0.39	−0.55	0.63
$\lambda_{9,2}$	2.05	0.08	0.15	−0.33	0.35	0.12	0.52	−0.44	0.63	0.20	0.34	−0.67	0.76	0.27	0.49	−0.38	0.51
$\lambda_{10,2}$	2.12	−0.02	0.21	−0.39	0.43	0.10	0.27	−0.52	0.59	0.19	0.28	−0.63	0.64	0.05	0.24	−0.55	0.59
η_1	0.57	−0.01	0.15			−0.01	0.18			0.09	0.17			−0.02	0.11		
η_2	0.59	0.01	0.13			0.05	0.17			0.11	0.18			−0.04	0.15		
η_3	0.70	0.04	0.16			0.04	0.17			0.12	0.18			−0.02	0.13		
η_4	1.83	0.00	0.17			0.05	0.15			0.10	0.18			−0.03	0.13		
η_5	−0.50	−0.04	0.16			−0.03	0.14			0.10	0.20			−0.07	0.18		
η_6	−0.56	0.03	0.16			−0.06	0.17			0.08	0.17			−0.09	0.15		
η_7	−0.10	0.01	0.18			0.03	0.23			0.09	0.18			−0.04	0.17		
η_8	−1.05	0.01	0.17			0.07	0.17			0.08	0.17			−0.03	0.17		
η_9	0.55	0.04	0.15			0.04	0.13			0.10	0.19			−0.04	0.14		
η_{10}	−2.03	0.02	0.19			0.10	0.24			0.16	0.21			0.04	0.14		
Profile recovery																	
Minimum		96.41		94.27		68.80		59.83		67.44		58.61		68.42		62.40	
Maximum		99.15		97.00		75.67		67.85		73.22		65.24		77.21		69.83	
Mean		97.58		95.78		71.02		63.50		70.14		62.16		72.62		66.20	
SD		0.66		0.64		2.14		2.77		1.45		1.83		2.89		2.29	

design, and from 0.10 to 0.49 (M = 0.22) in the random design. Such small RMSE values suggested good parameter recovery and they were similar to those found in common CDMs (e.g., De la Torre et al., 2010; Huang and Wang, 2014). With respect to the recovery of the latent profile, the mean recovery rate across profiles was 97.58% in the completely crossed design, 71.02% in the balanced incomplete design, 70.14% in the unbalanced incomplete design, and 72.62% in the random design. As expected, the parameter estimation and profile recovery were better in the completely crossed design than in the incomplete designs.

Focusing on results of the facets-CDM model, the profile recovery rates ranged from 67 to 69% in the three incomplete designs, where each item was rated by three raters. Such findings indicated that if one wishes to obtain a mean profile recovery rate of 70% from ten dichotomous items measuring five attributes, each item should be judged by three raters (i.e., each ratee received 30 scores). Moreover, as indicative by the results of the

completely crossed design, if each item is judged by ten raters (i.e., each ratee received 100 scores), the mean profile recovery rate could be as high as 98%.

When rater effects were ignored and the standard CDM was fit, the RMSE values became larger, ranging from 0.14 to 0.62 (M = 0.34) in the completely crossed design, from 0.07 to 0.81 (M = 0.34) in the balanced incomplete design, from 0.06 to 0.84 (M = 0.37) in the unbalanced incomplete design, and from 0.08 to 0.81 (M = 0.35) in the random design. The mean recovery rate across profiles was 95.78% in the completely crossed design, 63.50% in the balanced incomplete design, 62.16% in the unbalanced incomplete design, and 66.20% in the random design. Therefore, as expected, the parameter estimation in the standard CDM was worse than those in the facets-CDM. With respect to the recovery of the latent profile, both models yielded a higher recovery rate in the complete design than incomplete designs. This was because in the facets framework, when there are more raters, the measurements are more precise. In the complete design, these two models yielded almost identical recovery rates, which was because the mean rater effect was constrained at zero and thus canceled out. In the incomplete design, the mean rater effect was not canceled out, so the facets model consistently yielded a higher recovery rate (6–8% improvement) than the standard model.

TABLE 4 | Q-matrix of the 52 criteria in the empirical example.

	Attribute							Attribute					
Item	**1**	**2**	**3**	**4**	**5**	**6**	**Item**	**1**	**2**	**3**	**4**	**5**	**6**
1	1	1	1	1	1	0	27	0	0	1	0	0	0
2	1	0	0	0	0	0	28	0	0	1	0	0	0
3	1	0	0	0	0	0	29	0	0	1	0	0	0
4	0	1	1	1	0	0	30	0	0	1	0	0	0
5	0	1	1	1	0	0	31	0	0	1	0	0	1
6	1	0	1	1	1	0	32	0	0	1	0	0	1
7	1	1	1	1	0	0	33	0	0	1	0	0	0
8	1	1	0	0	0	0	34	0	0	1	0	0	0
9	1	1	0	0	0	0	35	0	0	1	0	0	0
10	1	1	0	0	0	0	36	0	0	1	1	0	0
11	1	0	0	1	0	0	37	0	0	1	0	0	0
12	1	1	0	1	0	0	38	0	0	1	0	0	0
13	1	0	0	0	0	0	39	0	0	1	1	0	0
14	0	1	0	0	1	0	40	0	0	1	1	0	0
15	0	1	0	0	0	0	41	0	0	0	1	0	0
16	0	1	0	0	0	0	42	0	0	0	1	0	0
17	0	1	0	0	0	0	43	0	0	0	1	0	0
18	0	1	0	0	0	0	44	0	0	0	1	0	0
19	0	1	0	0	0	0	45	0	0	1	1	0	0
20	0	1	0	0	0	0	46	0	0	0	1	0	1
21	0	1	1	1	0	0	47	0	0	0	1	0	1
22	0	1	1	1	0	0	48	0	0	1	0	0	1
23	0	1	1	1	0	0	49	0	0	0	0	0	1
24	0	1	1	1	0	0	50	0	0	0	0	0	1
25	0	1	0	0	0	0	51	0	0	1	1	1	0
26	0	0	1	0	0	0	52	0	0	1	1	1	0

1s mean the attributes are required, and 0s mean the attributes are not required.

TABLE 5 | Means and standard deviations for raters' scorings across all indicators in the empirical example.

Rater	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Mean	0.41	0.74	0.68	0.68	0.57	0.58	0.55	0.84	0.59
SD	0.28	0.26	0.29	0.26	0.28	0.31	0.27	0.18	0.33

TABLE 6 | Model fit statistics of the three models in the empirical example.

Model	***ppp***	**AIC**	**BIC**
DINA	0.36	17004	17625
Facets DINA	0.44	16690	17331
HRM DINA	0.56	10490	11167

ppp, posterior predictive p-value; AIC, Akaike's information criterion; BIC, Bayesian information criterion.

SIMULATION STUDY II: HRM-CDM

Design and Analysis

The settings were identical to those in simulation study I except only the completely crossed design was adopted and each ratee was judged by three or six raters. The (saturated) HRM-CDM (Equation 17), given the latent category, was used at the second step. At the first step, ϕ_r and ψ_r were fixed at 0 and 0.5, respectively, for all raters. Both the data-generating HRM-CDM and the standard CDM (without considering rater effects) were fit to the simulated data. In the standard CDM, multiple ratings given to the same item response were treated as independent responses. For example, the three sets of ratings given by three raters were analyzed as if the test was answered by three virtual examinees. Then, the posterior probability for each latent attribute (or latent profile) was averaged across the three virtual examinees to represent the examinee's final estimate. Like in simulation study I, it was expected that the more the raters (the more the data points), the better the parameter estimation and recovery rates. Further, when the standard CDM was fit, the parameter estimation and recovery rates would be poor because the rater effects were not considered.

Results

Detailed results for individual parameters are not presented due to space constraints but available on request. When the HRM-CDM was fit, the resulting RMSE values ranged from 0.11 to 0.67 (M = 0.35) and from 0.08 to 0.43 (M = 0.25) for three and six raters, respectively; the mean profile recovery rate was 61.12 and 79.12% for three and six raters, respectively. It appeared that the more the data points the better the parameter estimation and recovery rates when the HRM-CDM was fit. If one wishes to obtain a mean profile recovery rate of 80% from 10 dichotomous items measuring five attributes, it can be found from this simulation study that each item should be judged by six raters (i.e., each ratee received 60 scores). If each item is judged by only three raters (i.e., each ratee received 30 scores), the mean profile

TABLE 7 | Estimates for the guessing and slip parameters yielded by the three models in the empirical example.

Item	Guessing			Slip		
	DINA	Facets-DINA	HRM-DINA	DINA	Facets-DINA	HRM-DINA
1	0.49	0.49	0.48	0.03	0.02	0.00
2	0.47	0.47	0.42	0.02	0.03	0.00
3	0.49	0.49	0.49	0.09	0.08	0.12
4	0.46	0.46	0.46	0.05	0.06	0.00
5	0.24	0.30	0.12	0.34	0.36	0.43
6	0.50	0.49	0.49	0.05	0.00	0.00
7	0.16	0.17	0.00	0.52	0.50	0.70
8	0.09	0.11	0.00	0.18	0.20	0.23
9	0.38	0.41	0.28	0.01	0.04	0.00
10	0.31	0.31	0.14	0.23	0.21	0.30
11	0.18	0.17	0.01	0.26	0.27	0.31
12	0.46	0.46	0.38	0.07	0.07	0.01
13	0.48	0.48	0.43	0.08	0.08	0.06
14	0.49	0.49	0.49	0.05	0.00	0.00
15	0.15	0.14	0.00	0.45	0.44	0.56
16	0.48	0.48	0.48	0.08	0.08	0.06
17	0.47	0.47	0.46	0.15	0.15	0.17
18	0.26	0.41	0.23	0.23	0.25	0.31
19	0.46	0.45	0.44	0.15	0.13	0.16
20	0.38	0.24	0.31	0.10	0.07	0.07
21	0.25	0.16	0.09	0.31	0.22	0.13
22	0.49	0.48	0.48	0.12	0.06	0.01
23	0.50	0.49	0.49	0.01	0.01	0.00
24	0.48	0.48	0.48	0.11	0.13	0.06
25	0.45	0.46	0.45	0.07	0.08	0.05
26	0.07	0.12	0.00	0.79	0.81	1.00
27	0.31	0.33	0.02	0.66	0.67	0.96
28	0.40	0.47	0.40	0.15	0.18	0.19
29	0.47	0.47	0.46	0.04	0.05	0.03
30	0.44	0.37	0.29	0.31	0.26	0.39
31	0.39	0.36	0.19	0.45	0.42	0.59
32	0.30	0.42	0.25	0.24	0.26	0.31
33	0.18	0.10	0.02	0.37	0.32	0.41
34	0.08	0.15	0.01	0.54	0.56	0.77
35	0.42	0.47	0.33	0.21	0.24	0.19
36	0.23	0.34	0.06	0.25	0.26	0.27
37	0.45	0.47	0.42	0.11	0.12	0.09
38	0.13	0.17	0.01	0.56	0.59	0.70
39	0.48	0.48	0.48	0.05	0.04	0.00
40	0.00	0.01	0.00	0.93	0.95	1.00
41	0.01	0.07	0.00	0.54	0.58	0.88
42	0.37	0.35	0.39	0.26	0.24	0.41
43	0.23	0.33	0.12	0.28	0.30	0.34
44	0.21	0.33	0.05	0.33	0.38	0.49
45	0.19	0.21	0.02	0.07	0.18	0.02
46	0.28	0.27	0.09	0.02	0.11	0.00
47	0.49	0.49	0.48	0.04	0.07	0.01
48	0.49	0.49	0.49	0.01	0.02	0.00
49	0.49	0.49	0.49	0.00	0.02	0.00
50	0.04	0.24	0.00	0.43	0.69	0.95
51	0.33	0.47	0.24	0.13	0.31	0.37
52	0.03	0.30	0.00	0.40	0.66	0.92

recovery rate could be as low as 60%. When the standard CDM was fit, the RMSE values ranged from 0.25 to 0.91 ($M = 0.57$) and from 0.08 to 0.46 ($M = 0.30$) for three and six raters, respectively; the mean profile recovery rate was 56.34 and 70.84% for three and six raters, respectively. Taken together, as anticipated, ignoring rater effects by fitting the standard CDM would yield poor parameter estimation and profile recovery, and the fewer the raters, the worse the parameter and profile recovery. As for the recovery of latent profiles, the HRM-CDM outperformed the standard model, and its superiority (5–10% improvement) was more obvious when more raters were included.

A comparison between the facets-CDM and HRM-CDM revealed that the parameter estimation and profile recovery were better in the former than in the latter. This was mainly because each data point contributed to the parameter estimation directly in the facets-CDM, whereas the scores given by raters provided information about the latent category, which then provided information about the ratee and item parameters in the HRM-CDM. The corresponding JAGS codes for the facets-CDM and HRM-CDM are presented in **Appendix**.

REAL DATA APPLICATION

The empirical study involved a total of 287 university students, each producing one academic essay in English, which was judged by one or two teachers (out of nine) against a 52-item checklist. The checklist was developed on the basis of the Empirical Descriptor-based Diagnostic Checklist (Kim, 2011). Each item of the checklist was rated on a binary scale, where 1 = correct, 0 = incorrect. The 52 items aimed to measure six latent attributes of academic writing, namely, content, organization, grammar, vocabulary, conventions of the academic genre, and mechanics. The Q-matrix of the 52 items is shown in **Table 4**. The data matrix was three-dimensional: 287 examinees by 52 items by 9

TABLE 8 | Rater severity and variability yielded from the HRM DINA model in the empirical example.

Rater	1	2	3	4	5	6	7	8	9
Severity	0.40	0.02	0.01	0.02	0.01	0.02	0.24	0.04	0.02
SE	0.02	0.05	0.06	0.06	0.05	0.06	0.00	0.07	0.02
Variability	0.37	0.84	0.69	0.70	0.53	0.52	0.76	1.28	0.49
SE	0.01	0.05	0.04	0.04	0.03	0.03	0.04	0.10	0.02

TABLE 9 | Fair scores and observed scores for selected cases in the real data.

Student index	Estimated profile	Rater	Observed scores	Fair scores	Difference
21	1,1,1,1,1,0	1	23	40	−17
23	0,1,1,1,1,1	1	13	36	−23
30	0,0,0,1,0,0	1	15	22	−7
69	1,0,1,1,0,1	8	44	38	6
230	1,1,1,1,0,1	8	42	33	9

Estimated profiles were obtained by fitting facets-DINA model. Fair scores were calculated by DINA model with given person profile and item parameters.

raters. Because each item on the diagnostic checklist represented a concrete descriptor of the desirable quality of writing (e.g., item 4 "the essay contains a clear thesis statement"), the scoring rubrics were clear and simple for the raters to follow. Thus, the HRM framework appeared to be preferable to the facets approach. For completeness and illustrative simplicity, three models were fitted using JAGS, including (a) the standard DINA model, in which the ratings from raters were treated as responses to virtual items with identical item parameters; (b) the facet-DINA model; (c) the HRM-DINA model. A normal prior with mean zero and standard deviation four was specified for all item parameters across the three models, except that a log-normal distribution with mean zero and standard deviation four was specified for the variability parameter (ψ_r) in the HRM-DINA. For the facets-DINA model, a normal distribution with mean zero and standard deviation one was specified for rater parameters, and the mean severity across raters was fixed at zero for model identification. In the HRM-DINA model, the prior distributions for ϕ_r and ψ_r were set as $\phi_r \sim N(0, 1)$ and $\log(\psi_r) \sim N(0, 4)$, respectively.

Table 5 displays the means and standard deviations of the ratings on the 52 descriptors given by the nine raters. Rater 1 gave the lowest mean score ($M = 0.41$), whereas rater 8 gave the highest ($M = 0.84$). For model comparison, **Table 6** presents the posterior predictive *p*-values (Gelman et al., 1996) of the Bayesian chi-square statistic, Akaike's information criterion (AIC), and Bayesian information criterion (BIC) for the three models. The *p*-values suggested all models had a good fit. Both AIC and BIC indicated that the HRM-DINA model was the best-fitting model. **Table 7** lists the estimates for the guessing and slip parameters for the three models. The standard DINA model and the facets-DINA model produced very similar estimates. In comparison, the HRM-DINA model yielded smaller estimates for the guessing parameters and larger estimates for the slip parameters than the other two models.

Estimates for rater severity (ϕ) and variability (ψ) under the HRM-DINA model are presented in **Table 8**. Among the 9 raters, rater 1 was the most severe, followed by rater 7, while the others had severity measures around 0. Both rater 1 and rater 7 tended to assign ratings lower than what the ratees deserved (their severity parameters were positive). Furthermore, the estimates for rater variability ranged from 0.37 (rater 1) to 1.28 (rater 8), suggesting the raters exhibited moderate to high variability in their ratings.

Regarding the attribute estimates, the mastery probabilities of the six attributes were 50, 77, 76, 69, 63, and 73% for the standard DINA model, 53, 81, 77, 78, 83, and 79% for the facets-DINA model, and 50, 71, 68, 66, 75, and 74% for the HRM-DINA model. Among the 287 students, 77 students (27%) resulted in identical profile estimates with the three models, indicating moderate similarity on profile estimates across the three models.

To show the effects of ignoring rater effects, we picked up five students from the real data. For the selected cases, they were rated either by Rater 1, who tended to be the most severe, or by Rater 8, who tended to be most lenient. The differences between observed and fair scores (the expected score given the item and person parameters) are shown in **Table 9**. If one wants to admit students to some program according to their observed (raw) scores, then the ordering will be no. 69, 230, 21, 30, and 23, respectively. After taking into consideration of the rater effect by fitting the facets-DINA, we have fair score for each student. Now, if one wants to admit the five students according to the fair scores, then the ordering will be student no. 21, 69, 23, 230, and 30, respectively. Obviously, the two rank orderings were very different, which was because the former did not consider rater effect.

CONCLUSION AND DISCUSSION

Rater effects on CR items have been investigated extensively within the frameworks of IRT-facets and IRT-HRM, but not within those of CDMs. In this study, we adopted the facets and HRM frameworks and used the LCDM as a template to create new facets-CDM and HRM-CDM to accommodate rater effects. We also conducted simulations to evaluate parameter recovery of the new models under various conditions. Results indicate that model parameters could be estimated fairly well with JAGS package in R. Implications and applications of the new models were demonstrated with an empirical study that assessed English academic essays by university students. In the empirical study, the scales of the guessing and slip parameters for standard DINA and facets-DINA models were very similar, but they were very different from those for the HRM-DINA model, which was mainly because the HRM-DINA model was formed in a very different way from the other two models. Under the HRM-DINA model, among the 9 raters, raters 1 and 7 were the most severe. In addition, the rater variability ranged from 0.37 to 1.28, suggesting a moderate to high variability in their ratings.

Several limitations of the current study should be acknowledged. First, despite our efforts in testing the new models under different rating designs, the simulated conditions of the present study is not comprehensive. Future studies should be conducted to evaluate the performance of the new models under more comprehensive conditions, such as different test lengths, sample sizes, rater sizes, and rater designs. Second, a good CDM test depends on the quality of the Q-matrix (Lim and Drasgow, 2017). In this study, only one Q-matrix was used. How the facets- and HRM-CDMs perform with different Q-matrices needs further investigation. Third, like other simulation studies of CDMs, the data were analyzed with the data-generating models without looking to other potential sources of model-data misfit, such as mis-specification of the model or Q-matrix. Sensitivity analysis of the new models is warranted. Finally, the long computing time for MCMC methods may be a concern for potential users, especially for large scale data sets with long test length and large sample size. Future attempts are needed to develop more efficient and effective estimation programs.

Future studies can also be conducted to extend the new facets- and HRM-CDMs. For instance, the linear combination of parameters in the facets- or HRM-CDMs can be extended to account for interactions among facets (Jin and Wang, 2017). It is feasible to develop explanatory facets- or HRM-CDMs by incorporating covariates (e.g., gender or language background) to account for the variations in rater effects (Ayers et al., 2013). Large-scale educational testing services often recruit a large number of

raters (e.g., hundreds of raters), where it would be more efficient to treat rater severity as a random effect following some distributions (e.g., normal distributions). Finally, this study focuses on dichotomous items because the majority of existing CDMs focus on binary data. New facets- or HRM-CDMs can be developed to accommodate polytomous CR items, just as CDMs has been extended to accommodate polytomous items, as shown in Equations 11, 16, and 19, or those in the literature (Hansen, 2013; Ma and de la Torre, 2016).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Human Research Ethics Committee of The Education University of Hong Kong. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XL and W-CW conceived and designed the study, performed the simulations and analyses, interpreted the results, and wrote the manuscript. QX contributed the empirical data, provided critical comments on the study, and edited the whole manuscript. All authors provided the final approval of the version to publish.

REFERENCES

Ayers, E., Rabe-Hesketh, S., and Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *J. Classif.* 30, 195–224. doi: 10.1007/s00357-013-9130-y

Casabianca, J. M., Junker, B. W., Nieto, R., and Bond, M. A. (2017). A hierarchical rater model for longitudinal data. *Multivar. Behav. Res.* 52, 576–592. doi: 10.1080/00273171.2017.1342202

Chiu, C., and Seo, M. (2009). Cluster analysis for cognitive diagnosis: an application to the 2001 PIRLS reading assessment. *IERI Monogr. Ser. Issues Methodol. Large Scale Assess.* 2, 137–159.

Chiu, C.-Y., and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classif.* 30, 225–250. doi: 10.1007/s00357-013-9132-9

Chiu, C.-Y., Douglas, J. A., and Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika* 74, 633–665. doi: 10.1007/s11336-009-9125-0

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

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

De la Torre, J., Hong, Y., and Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *J. Educ. Meas.* 47, 227–249. doi: 10.1111/j.1745-3984.2010.00110.x

DeCarlo, L. T., Kim, Y. K., and Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *J. Educ. Meas.* 48, 333–356. doi: 10.1111/j.1745-3984.2011.00143.x

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *J. Educ. Meas.* 31, 93–112. doi: 10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *J. Educ. Meas.* 33, 56–70. doi: 10.1111/j.1745-3984.1996.tb00479.x

Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* 6, 733–807.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–511. doi: 10.1214/ss/1177011136

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26, 301–323. doi: 10.1111/j.1745-3984.1989.tb00336.x

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

Hansen, M., and Cai, L. (2013). Abstract: a hierarchical item response model for cognitive diagnosis. *Multivar. Behav. Res.* 48:158. doi: 10.1080/00273171.2012.748372

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality*. Doctoral dissertation, University of Illinois, Urbana-Champaign, IL.

Henson, R., Templin, J., and Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Huang, H.-Y., and Wang, W.-C. (2014). The random-effect DINA model. *J. Educ. Meas.* 51, 75–97. doi: 10.1111/jedm.12035

Jin, K. Y., and Wang, W. C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivar. Behav. Res.* 52, 391–402. doi: 10.1080/00273171.2017.1299615

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Lang. Test.* 28, 509–541. doi: 10.1177/0265532211400860

Li, X., and Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: the DINA model example. *J. Educ. Meas.* 52, 28–54. doi: 10.1111/jedm.12061

Lim, Y. S., and Drasgow, F. (2017). Nonparametric calibration of item-by-attribute matrix in cognitive diagnosis. *Multivar. Behav. Res.* 52, 562–575. doi: 10.1080/00273171.2017.1341829

Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.

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

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/BF02296272

Myford, C. M., and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *J. Appl. Meas.* 4, 386–422.

Myford, C. M., and Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *J. Appl. Meas.* 5, 189–227.

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large scale educational assessment data. *J. Educ. Behav. Stat.* 27, 341–384. doi: 10.3102/10769986027004341

Plummer, M. (2015). *coda (R Package Version 0.18-1, pp. 1–45)*. Vienna: The Comprehensive R Archive Network.

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

Su, Y., and Yajima, M. (2015). *R2jags (R Package Version 0.5-7, pp. 1–12)*. Vienna: The Comprehensive R Archive Network.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *J. Educ. Stat.* 12, 55–73. doi: 10.3102/10769986010001055

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

Templin, J. (2004). *Generalized Linear Mixed Proficiency Models for Cognitive Diagnosis*. Ph.D. dissertation, University of Illinois, Urbana-Champaign, IL.

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

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

von Davier, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* 61, 287–307. doi: 10.1002/j.2333-8504.2005.tb01993.x

Wang, W.-C., and Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educ. Psychol. Meas.* 67, 583–605. doi: 10.1177/0013164406296974

Wang, W.-C., and Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Appl. Psychol. Meas.* 29, 296–318. doi: 10.1177/0146621605276281

Xu, X., and von Davier, M. (2008). *Fitting the Structured General Diagnostic Model to NAEP Data (Research Report No. 08-27)*. Princeton, NJ: Educational Testing Service.

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

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.

APPENDIX

(1) JAGS code for the facets-CDM in Simulation Study I.

```
model<-function(){
   for (i in 1:n.p){
      for (k in 1:n.a){
         pi[i,k]~dunif(0,1)
         alpha[i,k]~dbern(pi[i,k])}
   eta2[i,1]<-0
   eta2[i,2]<-0
   eta2[i,3]<-0
   eta2[i,4]<-0
   eta2[i,5]<-0
   eta2[i,6]<-alpha[i,1]*alpha[i,2]
   eta2[i,7]<-alpha[i,2]*alpha[i,3]
   eta2[i,8]<-alpha[i,3]*alpha[i,4]
   eta2[i,9]<-alpha[i,4]*alpha[i,5]
   eta2[i,10]<-alpha[i,1]*alpha[i,5]

      for (j in 1:n.i){
         for (k in 1:n.a) {w[i,j,k]<- alpha[i,k]*q[j,k]}
      eta1[i,j]<-prod(w[i,j,k]
         for (r in 1:n.r){
           logit(prob[i,j,r])<-lamda0[j]+lamda1[j]*eta1[i,j]+lamda2[j]*eta12[i,j]+rater[r]
           resp[i,j,r]~dbern(prob[i,j,r])}}}
for (r in 1:n.r) {rater[r]~dnorm(mean.r, pr.r)}
for (j in 1:n.i) {
   lamda0[j]~dnorm(mean.lamda0, pr.lamda0)
   lamda1[j]~dnorm(mean.lamda1, pr.lamda1)
   lamda2[j]~dnorm(mean.lamda2, pr.lamda2)}}
```

(2) JAGS code for the HRM-CDM in Simulation Study II.

```
model<-function(){
   for (i in 1:n.p){
      for (k in 1:n.a){
         pi[i,k]~dunif(0,1)
         alpha[i,k]~dbern(pi[i,k])}

      for (j in 1:n.i){
         for (k in 1:n.a) {w[i,j,k]<- alpha[i,k]*q[j,k]}
      eta[i,j]<-prod(w[i,j,k]

         for (r in 1:n.r){
           logit(p[i,j,r])<-lamda0[j]+lamda1[j]*eta[i,j]
           resp[i,j,r]~dbern(p[i,j,r])
           rating.prob[i,j,r]<-exp((-0.5)*pow((1-resp[i,j,r]-mu.rater[r]),2)*pow(sigma.rater[r],-2))
           rating[i,j,r]~dbern(rating.prob[i,j,r])}}}

   for (j in 1:n.i){
     lamda0[j]~dnorm(mean.lamda, pr.lamda)
     lamda1[j]~dnorm(mean.lamda, pr.lamda)}
   for (r in 1:n.r) {
         mu.rater[r]~dnorm(mean.mu.rater, pr.mu.rater)
         sigma.rater[r]~dlnorm(mean.sigma.rater, pr.sigma.rater)}}
```

frontiers in Psychology

ORIGINAL RESEARCH
published: 15 May 2020
doi: 10.3389/fpsyg.2020.00944

Spectral Clustering Algorithm for Cognitive Diagnostic Assessment

Lei Guo[1,2], *Jing Yang*[3]* *and Naiqing Song*[2,4,5]*

[1] Faculty of Psychology, Southwest University, Chongqing, China, [2] Southwest University Branch, Collaborative Innovation Center of Assessment Toward Basic Education Quality, Chongqing, China, [3] School of Mathematics and Statistics, Northeast Normal University, Changchun, China, [4] Basic Education Research Center, Southwest University, Chongqing, China, [5] Urban and Rural Education Research Center, Southwest University, Chongqing, China

In cognitive diagnostic assessment (CDA), clustering analysis is an efficient approach to classify examinees into attribute-homogeneous groups. Many researchers have proposed different methods, such as the nonparametric method with Hamming distance, K-means method, and hierarchical agglomerative cluster analysis, to achieve the classification goal. In this paper, according to their responses, we introduce a spectral clustering algorithm (SCA) to cluster examinees. Simulation studies are used to compare the classification accuracy of the SCA, K-means algorithm, G-DINA model and its related reduced cognitive diagnostic models. A real data analysis is also conducted to evaluate the feasibility of the SCA. Some research directions are discussed in the final section.

Keywords: cognitive diagnostic assessment, spectral clustering, K-means, G-DINA model, classification accuracy

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Wenyi Wang,
Jiangxi Normal University, China
Feiming Li,
Zhejiang Normal University, China

***Correspondence:**
Jing Yang
yangj014@nenu.edu.cn
Naiqing Song
songnq@swu.edu.cn

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

Received: *03 January 2020*
Accepted: *16 April 2020*
Published: *15 May 2020*

Citation:
Guo L, Yang J and Song N (2020) Spectral Clustering Algorithm for Cognitive Diagnostic Assessment. Front. Psychol. 11:944. doi: 10.3389/fpsyg.2020.00944

INTRODUCTION

In the past decades, there has been a significant increasing interest in cognitive diagnostic assessment (CDA) that allows for the purpose of identifying the presence or absence of specific fine-grained attributes required for solving problems on a test in educational and psychological assessment. Researchers have proposed a variety of methods to classify examinees into several categories by matching their attribute profiles. To sum up, there have been two major kinds of approaches till now. One of them usually uses cognitive diagnosis models (CDMs) to estimate the attribute profile for each examinee, which can be called parametric technique. The differences between these CDMs are assumptions about how cognitive attributes affect examinees' responses in CDAs. The deterministic input; noisy "and" gate (DINA; Junker and Sijtsma, 2001), and noisy input; deterministic "and" gate model (NIDA; Junker and Sijtsma, 2001) are the typical conjunctive models, which require examinees must master all required attributes, thus even lacking one required attribute will lead to a totally wrong response. Disjunctive models, such as the deterministic input; noisy "or" gate model (DINO; Templin and Henson, 2006), suppose that if one has mastered a subset of required attributes, even merely one, the probability of a correct response will be sufficiently high. Other specific, interpretable CDMs include the linear logistic model (LLM; Maris, 1999) the additive CDM (A-CDM; de la Torre, 2011) and the reduced reparameterized

unified model (RRUM; Hartz, 2002). To subsume the above reduced models, some general CDM frameworks has been proposed, such as the log-linear CDM (LCDM; Henson et al., 2009) the generalized DINA (G-DINA; de la Torre, 2011) model and the general diagnostic model (GDM; von Davier, 2008). The major advantage of general CDMs is that they have the largest flexibility of fetting response data which is set under the CDM framework, and it always should be taken into account at first when doing parameter estimation.

The superiority of parametric models is conciseness. However, one big issue inherently exists in the parametric technique, i.e. sample size. Several researchers have investigated the influence of sample size on estimation accuracy of the model parameters and pattern/attribute correct classification rate (de la Torre et al., 2010; Chen and de la Torre, 2013; Minchen et al., 2017). Although the results represented that sample size had a negligible impact on correct classification rate, most previous studies obtained this conclusion by setting the number of examinees no less than 500. So, there is no evidence to draw the inference that no effect on correct classification rate when using small sample size (may be less than 50 or 100). Virtually, the number of examinees in one class is not large for the most part. It is doubtful whether the performance of the parametric models is good or not when teachers implement the cognitive diagnostic test in class with a smaller sample size.

To address this issue, nonparametric techniques can be treated as alternative approaches to classify examinees into attribute-homogeneous groups, which is less restrictive and often computationally more efficient. Better yet, many nonparametric classification algorithms can be easily implemented in most statistical software packages. Based on the advantages of nonparametric techniques, many different methods have been proposed in the CDA. For example, three different methods of computing sum-scores (simple sum-scores, complex sum-scores, and weighted complex sum-scores) combined with model-based mastery sum-score cutoffs were proposed (Henson et al., 2007). Their results indicated that the correct classification rates of examinees' attribute profiles from model-based sum-scores and mastery sum-score cutoffs were able to compare with those correct classification rates from CDM. Chiu et al. (2009) used hierarchical agglomerative clustering and K-means methods to group examinees into different clusters possessing the same attribute profiles. Simulation results demonstrated that K-means method had better performance at the classification consistency and homogeneity of a cluster than that of hierarchical agglomerative clustering in most experimental conditions. Subsequently, Chiu and Douglas (2013) proposed a nonparametric procedure that merely relied on a given Q-matrix (Tatsuoka, 1985), and evaluated the examinees' attribute profiles by minimizing the distance measures (hamming distance, weighted hamming distance, and penalized hamming distance) between observed responses and the expected responses of a given attribute profile. Specifically, this procedure based on expected response patterns makes no direct use of item parameters of any CDMs. So, it required no parameter estimation, and can be used on a sample size as small as 1 (recall that the sample size is no less than 500 in CDMs based on existing studies). In addition, the existing studies have provided plenty of evidence that the nonparametric classification algorithms have good performance in CDA.

The primary objective of this paper is to introduce the method for implementing CDA using spectral clustering algorithm (SCA), which has become one of the most prevalent modern clustering methods in recent years. The SCA creates a graph of objects that require classifying based on the similarity measurement of each pair of objects (i.e. examinees in this paper). The more similar the examinees' attribute profiles are, the greater probability they can interrelate with each other in the graph. Next, the examinees' attribute profiles can be clustered by anatomizing the spectral graph, where the attribute profiles within a cluster have a strong connection and different clusters have a weak connection. Naturally, such algorithms have been widely applied in the field of image segmentation (Shi and Malik, 2000) neural information processing (Ng et al., 2002) biology (Zare et al., 2010) and large-scale assessment in psychology (Chen et al., 2017). However, no study has been done to investigate the performance of the SCA in CDA yet to our knowledge. And it is interesting to inspect the efficiency of the SCA for clustering examinees' into attribute-homogeneous groups under variedly underlying processes, such as conjunctive, disjunctive, additive, and saturated model (de la Torre, 2011).

In the next section, the G-DINA model and its related reduced models will be briefly reviewed. Subsequently, the K-means and SCA algorithms are detailedly introduced in the third section. This is followed by the simulation studies comparing SCA to K-means algorithm and CDMs mentioned in the second section are conducted in section "Simulation Studies," and the section "Analysis of Mixed Number Subtraction Data" concerns a real data study to examine the performance of the SCA. Finally, Summary and discussions are given in the final section.

COGNITIVE DIAGNOSTIC MODELS

First, some basic concepts and terms used in CDA are introduced. Consider J binary item response variables for each of the I examinees. Let X_{ij} represent the response of examinee i to item j, where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Let $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iK})$ denote the attribute profile of examinee i, where K is the number of attributes measured by the test. A value of $\alpha_{ik} = 1$ indicates the ith examinee masters the kth attribute and $\alpha_{ik} = 0$ otherwise. Let $\mathbf{q}_j = (q_{j1}, q_{j2}, ..., q_{jK})$ represent the jth row of the $\mathbf{Q}$-matrix that describes the relationship between items and attributes (Tatsuoka, 1995). $\mathbf{Q}$ is a $J \times K$ matrix with the entry $q_{jk} = 1$ indicating that item j requires attribute k, and $q_{jk} = 0$ otherwise.

The G-DINA Model

The G-DINA model is able to distinguish $2^{K_j^*}$ latent classes, where K_j^* is the number of required attributes for jth item, and $K_j^* = \sum_{k=1}^{K} q_{jk}$. For simplicity, the first K_j^* attributes are treated as the required attributes for jth item, and $\boldsymbol{\alpha}_{lj}^*$ is the reduced attribute vector corresponding to the columns of the required attributes

with $l = 1, ..., 2^{K_j^*}$. The probability of a correct response to jth item by examinees with attribute profile $\boldsymbol{\alpha}_{lj}^*$ can be denoted by $P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = P(\boldsymbol{\alpha}_{lj}^*)$. Then, the item response function (IRF) of the G-DINA model is as follow:

$$f\left[P(\boldsymbol{\alpha}_{lj}^*)\right] = \gamma_{j0} + \sum_{k=1}^{K_j^*} \gamma_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \gamma_{jkk'}\alpha_{lk}\alpha_{lk'} + ... + \gamma_{j12...K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}0 \quad (1)$$

where $f\left[P(\boldsymbol{\alpha}_{lj}^*)\right]$ represents $P(\boldsymbol{\alpha}_{lj}^*)$, $\log[P(\boldsymbol{\alpha}_{lj}^*)]$ and $\log it[P(\boldsymbol{\alpha}_{lj}^*)]$ in the identity, log and logit links, respectively. Moreover, γ_{j0} is the intercept for jth item, γ_{jk} is the main effect due to α_k, $\gamma_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\gamma_{j12...K_j^*}$ is the interaction effect due to $\alpha_1, ..., \alpha_{K_j^*}$. For more details about the G-DINA model, please refer to de la Torre (2011).

Related Reduced Models

It's conspicuous that the G-DINA model is a saturated model which can easily change into several popular reduced CDMs, including the DINA model, the DINO model, the *additive* CDM (*A*-CDM), etc. Note the symbol γ is used as item parameters across all these models in this paper. So, if we set all terms in the G-DINA model in identity link except γ_{j0} and $\gamma_{j12...K_j^*}$ to zero, the DINA model will be obtained, that is,

$$P(\boldsymbol{\alpha}_{lj}^*) = \gamma_{j0} + \gamma_{j12...K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (2)$$

If the intercept and main effect terms are remained with the follwing constraints: $\gamma_{jk} = -\gamma_{jk'k''} = ... = (-1)^{K_j^*+1}\gamma_{j12...K_j^*}$, for $k = 1, ..., K_j^*$, $k' = 1, ..., K_j^* - 1$, and $k'' > k', ..., K_j^*$. The DINO model can be given by

$$P(\boldsymbol{\alpha}_{lj}^*) = \gamma_{j0} + \gamma_{jk}\alpha_{lk} \quad (3)$$

By setting all interactions to zero in the identity-link G-DINA model, the *A*-CDM can be formulated as

$$P(\boldsymbol{\alpha}_{lj}^*) = \gamma_{j0} + \sum_{k=1}^{K_j^*} \gamma_{jk}\alpha_{lk} \quad (4)$$

Clearly, quite a few parameters of items and examinees require estimating in the saturated model and its related reduced CDMs. More often than not, one can use either marginalized maximum likelihood estimation (MMLE) or Bayesian approach with the Markov Chain Monte Carlo (MCMC) method to achieve parameter estimation.

CLUSTERING METHODS FOR COGNITIVE DIAGNOSIS

K-Means Method for Cognitive Diagnosis

K-means cluster analysis is widely used as the process of grouping a set of subjects into clusters so that subjects within a cluster have similarity in comparison to one another, but are dissimilar to subjects in other clusters. This approach finds the k centroid, where the coordinate of each centroid is the means of the coordinate of the subjects in the cluster and assigns every subject to the nearest centroid. Chiu et al. (2009) have made the best of K-means method in CDA already, and showed its effectiveness empirically for placing examinees in homogeneous groups. The algorithm in CDA can be summarized as follows (Please refer to Chiu et al.'s paper for details).

> Step 1: Select M initial K-dimensional cluster centroids.
> Step 2: Assign data points to clusters that have the closest centroid.
> Step 3:When all data points have been assigned, update the positions of the M centroids.
> Step 4: Repeat Steps 2 and 3 until the centroids no longer change.

Although K-means is a more than effective method for clustering, the starting values exercises a large impact on the classified performance for this method. Having poor starting values can result in converging to local optima (Steinley, 2003). So, many methods of choosing starting values for the K-means method have been proposed. Chiu et al. (2009) have investigated the performance of K-means method in CDA with two different kinds of starting values, called best and Ward's cases, respectively, which provided decent clustering results, and they should be considered in this study. Additionally, the K-means with random starting values will be deemed as the baseline to compare the classification performance to other two starting values. The introduction of starting values presents in section "The Selection of Starting Values" subsequently.

Spectral Clustering for Cognitive Diagnosis

As mentioned above, the SCA method was used in many research fields. For psychological assessment study, Chen et al. (2017) applied SCA to the context of exploratory item classification. Through constructing a graph of items, the similar items could be classified together and the dissimilar ones can be extracted based on the graphical structure. Intuitively, it is straightforward to wonder how the SCA performs on person classification in CDA. The SCA can be available in CDA context for the following reasons: (a) SCA creates a graph of examinees based on the similarity measurement of each pair of examinees, where examinees who possess the same attribute profiles tend to be connected. (b) Cai et al. (2005) wrote that "*The spectral clustering usually clusters the data points using the top eigenvectors of graph Laplacian, which is defined on the affinity matrix of data points*". In order to construct the affinity matrix for binary response data in CDA, the *Gaussian kernel* function can be

applied according to Ng et al. (2002). Then, one can use SCA to classify examinees. (c) both SCA and K-means method belong to clustering approach, and K-means is a component of the SCA method (Chen et al., 2017) which means both methods have the same parts of processing data to get clustering results. Chiu et al. (2009) had proved the feasibility of K-means in the aspect of classifying examinees into groups with same attribute profiles. So, the SCA should have a good chance of success in characterizing the same structure (i.e. attribute profiles) among examinees. We focus on the specific illustration and detail the core procedures on how to implement the SCA in CDA [for more details about the SCA, please refer to Von Luxburg (2007) and Chen et al. (2017)], now that the key point of this paper is not to introduce the SCA itself. One can easily operate this algorithm in CDA with following steps:

Step 1: Using response data to construct similarity matrix $\mathbf{S}$, which is a $I \times I$ square matrix with element,

$$S(\mathbf{X}_i, \mathbf{X}_{i'}) = \exp(- \parallel \mathbf{X}_i - \mathbf{X}_{i'} \parallel^2 / 2\sigma^2), \quad i, i' \in \{1, 2, ..., I\}, \tag{5}$$

where $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ are ith and i'th examinee's response vectors. Generally speaking, one may take $\sigma^2 = 1$ as assumption under standard normal distribution, and Eq. 5 can be considered as *Gaussian Kernel*. The SCA divided examinees into diverse clusters so that examinees in the same cluster tend to be similar, which means $S(\mathbf{X}_i, \mathbf{X}_{i'})$ value tends to be large if examinees i and i' belong to the same cluster. Meanwhile, those who are classified into different clusters tend to be differ from each other so as to the values become small.

Step 2: Construct a diagonal matrix $\mathbf{D}_{I\times I}$ and compute the normalized Laplacian matrix $\mathbf{L}_{I\times I}$ as follows:

$$D_{ii} = \sum_{i'=1}^{I} S_{ii'} \tag{6}$$

and

$$\mathbf{L}_{I\times I} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}} \tag{7}$$

where $\mathbf{I}$ is a $I \times I$ unit matrix.

Step 3: Compute the first M eigenvectors $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M$ of $\mathbf{L}_{I\times I}$.

Step 4: Let $\mathbf{U}_{I\times M}$ be the matrix containing the vectors $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M$ as columns.

Step 5: Derived the matrix $\mathbf{T}_{I\times M}$ from $\mathbf{U}_{I\times M}$ by normalizing the rows to norm 1, which is set $t_{im} = u_{im}/(\sum_m u_{im}^2)^{\frac{1}{2}}$.

Step 6: For i = 1,..., I, let $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, ..., Z_{iM})$ be the vector corresponding to the ith row of $\mathbf{T}_{I\times M}$.

Step 7: Cluster the points $\{\mathbf{Z}_i, i = 1, 2, ..., I\}$ with the K-means algorithm into M clusters.

Step 8: Assign the original points $\mathbf{X}_i$ to cluster j if and only if the points $\mathbf{Z}_i$ was assigned to cluster j.

According to these eight steps, examinees can be grouped into different clusters representing different attribute profiles. Currently, the R package "Kernlab" (Karatzoglou et al., 2004) can implement SCA availably.

The Selection of Starting Values

K-Means With Best Starting Values

In order to group examinees into the correct attribute profiles, Chiu et al. (2009) introduced the sum-score statistic, which was also used in Henson et al. (2007). For the ith examinee, the sum-score on attribute k can be defined as:

$$W_{ik} = \sum_{j}^{J} X_{ij} q_{jk} \tag{8}$$

Thus, $\mathbf{W}_i = (W_{i1}, W_{i2}, ...W_{iK})$ is the corresponding vector of K sum-scores. The matrix $\mathbf{W}_{I\times K}$ is then taken as the input of cluster analysis, with a fixed M clusters in CDA. Based on $\mathbf{W}_{I\times K}$ matrix, the K-means method assigns data point $\mathbf{W}_i$ to the mth cluster using Euclidean distance if

$$m = \arg \min_{u\in\{1,...,M\}} \parallel \mathbf{W}_i - \hat{\mathbf{c}}_u \parallel^2 \tag{9}$$

Where $\hat{\mathbf{c}}_u$ is the provisional centroids of the uth cluster during the iterative steps, and is calculated by averaging the observations in the cluster.

A key point of using K-means method is the selection of initial values. Let $\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, ...\alpha_{mK})'$ be the *unique* attribute profile in the universal set of attribute profiles, where m = 1, 2, ..., M and $M = 2^K$. For example, only four attribute profiles exist when K = 2, and they are $\alpha_1 = (0, 0)$, $\alpha_2 = (0, 1)$, $\alpha_3 = (1, 0)$, and $\alpha_4 = (1, 1)$, respectively. Then, the initial value matrix (denoted as $\mathbf{W}_{M\times K}$) in the 'best' scenario can be calculated as follow:

$$\mathbf{W}_{M\times K} = \mathbf{P}_{M\times J}\mathbf{Q}_{J\times K} \tag{10}$$

where $\mathbf{P}_{M\times J}$ is the expected response matrix with entry p_{mj} indicating that the probability of mth attribute profile correctly answering jth item. For instance, p_{mj} should be calculated according to Eq. 2 if the DINA model is selected (Chiu et al., 2009). Note that p_{mj} is used only as an ideal state for comparison in simulation study. When implementing K-means in practice, we have no idea about p_{mj} actually, thus other starting values, i.e. random and Ward's, will be selected.

Clustering With Ward's Starting Values

Ward's method is a general agglomerative hierarchical clustering approach originally presented by Ward (1963). The criterion of this manner is to minimize the total within-cluster variance. To implement this method, at each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centroids, and can be represented as the sum of square errors (SSE) statistic. Suppose that cluster p and q are next to be merged. Then, the SSE for the pth cluster is computed as follow:

$$SSE_p = \sum_{i=1}^{I_p} (\mathbf{Y}_{pi} - \bar{\mathbf{Y}}_p)'(\mathbf{Y}_{pi} - \bar{\mathbf{Y}}_p) \tag{11}$$

where I_p and I_q represent the number of data in clusters p and q, respectively. $\mathbf{Y}_{pi}$ is ith data point in cluster p, and $\bar{\mathbf{Y}}_p$ is the centroid of cluster p. Using the Eq. 11, the SSE for the qth cluster can be got. So, the pth and the qth clusters are merged into a new cluster if

$$SSE_{pq} - (SSE_p + SSE_q) = \frac{I_p I_q}{I_p + I_q}(\bar{\mathbf{Y}}_p - \bar{\mathbf{Y}}_q)'(\bar{\mathbf{Y}}_p - \bar{\mathbf{Y}}_q) \quad (12)$$

is the minimum among all pairs, where SSE_{pq} is the combined SSE for cluster p and q.

Subsequently, the initial values are determined according to the result of Ward's method in the 'Ward's starting values' scenario.

Clustering With Random Starting Values

The simplest method of choosing initial values is to utilize the random procedure. That means M data points may be selected randomly from the data set, and be treated as the M cluster centroid. Now that there is no prior knowledge guiding the way to choose the starting values in 'random' scenario, the randomness exerts a significant influence on the performance of this method. Then, with random starting values, the K-means and SCA can be considered as the baseline for the study.

Note that the 'best' starting value is used in K-means method but excluded in the SCA because the dimensionality of the matrix $\mathbf{W}_{M\times K}$ is different from the matrix $\mathbf{Z}_{I\times M}$. However, other two starting values can be both applied for SCA and K-means. The SCA and K-means are comparable as the following reasons: On the one hand, Chen et al. (2017) indicated that the K-means method is a component of the SCA algorithm. Meanwhile, the original materials used by both SCA and K-means method are raw response data actually. Only difference between these two methods is the mean to tackle raw response data. For the K-means method, in order to get the consistency theory, raw data was reconstructed through $\mathbf{W}_{I\times K} = \mathbf{X}_{I\times J}\mathbf{Q}_{J\times K}$, and $\mathbf{W}_{I\times K}$ matrix was used as input. On the other hand, according to the SCA, raw response data was reconstructed as $\mathbf{Z}_{I\times M}$ matrix through Steps 1 to 6 described in section "Spectral Clustering for Cognitive Diagnosis." And then, $\mathbf{Z}_{I\times M}$ matrix was treated as input in K-means method. Based on these evidences, clustering results from SCA are comparable with those from K-means method in essence.

SIMULATION STUDIES

The first goal of simulation studies is to investigate the effectiveness of clustering using the SCA in CDA, and compare SCA with K-means method in the aspect of classification accuracy further. These two methods pertain to clustering approach, and the last step of SCA needs to call K-means to accomplish clustering, which means both methods have the same parts of processing data to get clustering results. However, hamming distance is excluded in this paper because this method requires prior knowledge of cognitive processes to obtain the ideal response patterns. Then, measures of distance between observed response patterns and ideal response patterns can be calculated. It indicates that hamming distance method need to know the mechanism between attributes in advance (Chiu and Douglas, 2013). The SCA and K-means methods are unstinted in this constraint, clustering examinees according to their responses only.

Besides, it is not clear that the performance of K-means method is under some particular underlying processes (e.g. additive and saturated scenarios) because there is no research to compare K-means with the A-CDM and G-DINA model. So, the second goal is to examine the performances of the SCA and K-means methods in processing various response data sets generated by different CDMs, including the G-DINA, DINA, DINO, and A-CDM.

Simulation Design

To evaluate the performance of the SCA in clustering examinees, five factors were manipulated: the number of examinees I was set to 100 or 500; The number of attributes K equaled 3, 4 or 5; The item quality was defined by two parameters, which were denoted as $1 - P(1)$ and $P(1)$. Items with $1 - P(1)$, $P(1) \in U(0.05, 0.15)$ were labeled high quality, and items with $1 - P(1)$, $P(1) \in U(0.25, 0.35)$ were low quality (Ma et al., 2016); Generating models were G-DINA, DINA, DINO, and A-CDM model, respectively; Test length J = 5, 10, or 20. The generating rules of Q-matrix were as follows: (a) ensure that there were items at least require one attribute in Q-matrix. (b) the remaining items were selected from all $2^K - 1$ items randomly to satisfy the predetermined test length. For each condition, 100 replications were used.

The true attribute profiles $\boldsymbol{\alpha}$ were linked to an underlying multivariate normal distribution (Chiu et al., 2009) $\boldsymbol{\theta}_i \sim MVN(\mathbf{0}_K, \Sigma)$, where the covariance matrix Σ is

$$\begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}$$

Where ρ was set to 0.5, representing medium correlation between attributes. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots \theta_{iK})'$ express the latent continuous ability for examinee i, the attribute profile $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK})'$ was calculated by

$$\alpha_{ik} = \begin{cases} 1 & if\ \theta_{ik} \geq \Phi^{-1}\left(\frac{k}{K+1}\right), \\ 0 & otherwise. \end{cases} \quad (13)$$

Evaluation Criteria

To evaluate the performance of classifications in CDA, attribute correct classification rate (ACCR) and pattern correct classification rate (PCCR) are commonly used as the indicators. Nevertheless, they become available when examinees are classified into labeled sets, which is not the case with cluster analysis, for the reason that they manifest the consistency between the true and estimated attribute profiles. Only when the estimates of examinees' attribute profiles cognized can these indices be calculated. Obviously, it is not an issue when researchers use CDMs to analyze response data. However, the

cluster analysis classifies examinees into attribute-homogeneous groups, but it cannot provide information about the estimates of examinees' attribute profiles (i.e. labeling problem). So, ACCR and PCCR indices cannot be calculated in this case. Therefore, two indices which were applied in Chiu et al. (2009) paper were also used in this study. One was an indicator of agreement between partitions, called the Adjusted Rand Index (ARI), and the other was denoted as ω assessing the within-cluster homogeneity.

The ARI was modified from Rand index, and was originally proposed by Hubert and Arabie (1985). Given a set of I examinees $S = \{O_1, ..., O_I\}$, suppose that $U = \{u_1, ..., u_R\}$ and $V = \{v_1, ..., v_G\}$ represent two different partitions of the examinees in S. Supposed that U is the external criterion, i.e. true attribute profile in CDA, and V is a clustering result. The ARI assumes the generalized hypergeometric distribution as the model of randomness, i.e. the U and V partitions are picked at random such that the number of examinees in the clusters are fixed. Let I_{rg} be the number of examinees that are in both classes u_r and v_g, where $r = 1, 2,..., R$, and $g = 1, 2,..., G$. Let $I_{r\bullet}$ and $I_{\bullet g}$ be the number of examinees in class u_r and v_g, respectively. Then, the ARI can be shown as follows:

$$ARI = \frac{\sum_{r,g} C_{I_{rg}}^2 - \sum_r C_{I_{r\bullet}}^2 \sum_g C_{I_{\bullet g}}^2 / C_I^2}{\frac{1}{2}\left[\sum_r C_{I_{r\bullet}}^2 + \sum_g C_{I_{\bullet g}}^2\right] - \sum_r C_{I_{r\bullet}}^2 \sum_g C_{I_{\bullet g}}^2 / C_I^2} \quad (14)$$

which is limited between 0 and 1. The larger the ARI is, the higher agreement between partitions is. In Eq. 14, a binomial coefficient $C_{(\bullet)}^2$ is defined as 0 when the number of classified objects is 0 or 1.

In CDA, the index ω which can be used to evaluate the within-cluster homogeneity with respect to the true attribute profiles measures how similar examinees from the same cluster are to one another, and sums this over the clusters (Chiu et al., 2009). The formula for ω is given by

$$\omega = 1 - \frac{\sum_{i=2}^{I} \sum_{i'=1}^{i} \sum_{k=1}^{K} |\alpha_{ik} - \alpha_{i'k}| \, \mathrm{I}_{[\hat{c}_i = \hat{c}_{i'}]}}{\sum_{i=2}^{I} \sum_{i'=1}^{i} K \times I[\hat{c}_i = \hat{c}_{i'}]} \quad (15)$$

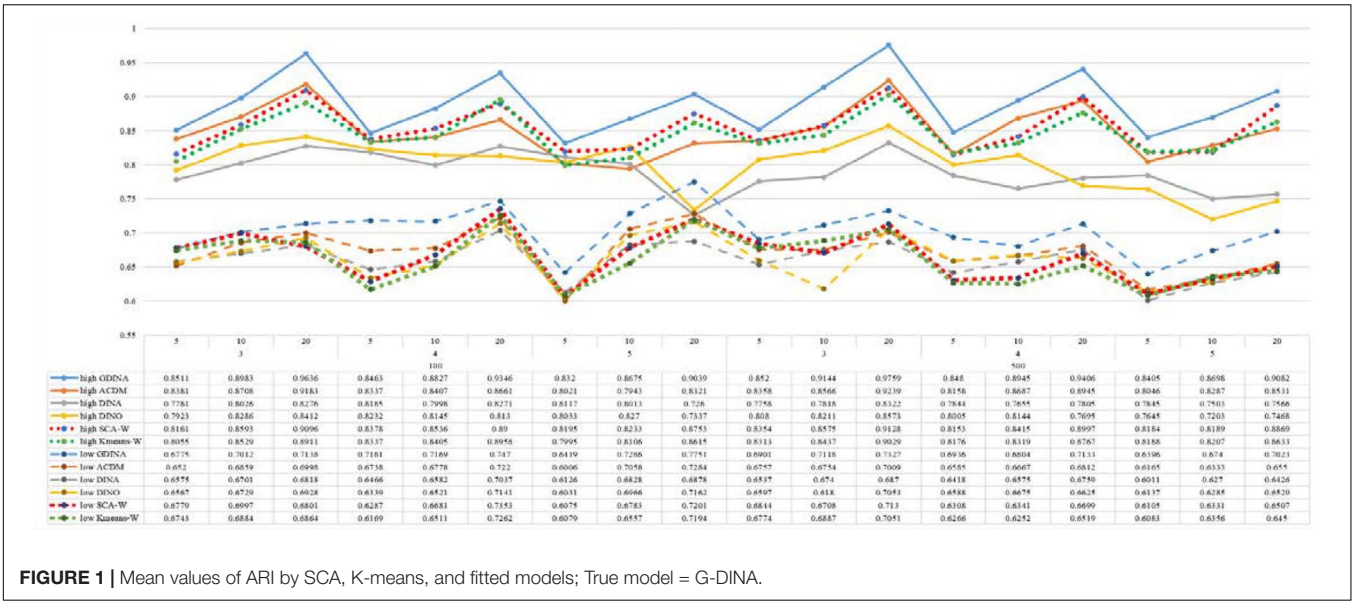
where $\hat{c}_i$ represents the classified result for the ith examinee, and $I_{[\hat{c}_i = \hat{c}_{i'}]}$ is the indicator function reflecting whether or not examinees i and i' are classified into same cluster. This index is also bounded between 0 and 1, and it equals 1 if true attribute profiles are the same for all pairs of examinees clustered together.

Results

Figures 1–8 totally demonstrate the means of ARI and ω for SCA, K-means, G-DINA model and its related reduced CDMs over 100 replications for each condition. Classification results of the true model are definitely the best, which provides the upper limit of comparison across all conditions. Oppositely, the random case just provides the lower limit of comparison to other settings, and it has indicated the worst performance among all methods based on simulation results. Although the "best" scenarios are treated as the best possible case for K-means to cluster response data, it has to use CDMs to get the expected response p_{mj} in advance, then $\mathbf{W}$ can be calculated. In this sense it is not indeed a nonparametric method. So, we mainly compare the performances of Ward's linkage for two clustering methods against the ones of other fitted CDMs in the following. The results of SCA with random, K-means with random and K-means with best do not present here.

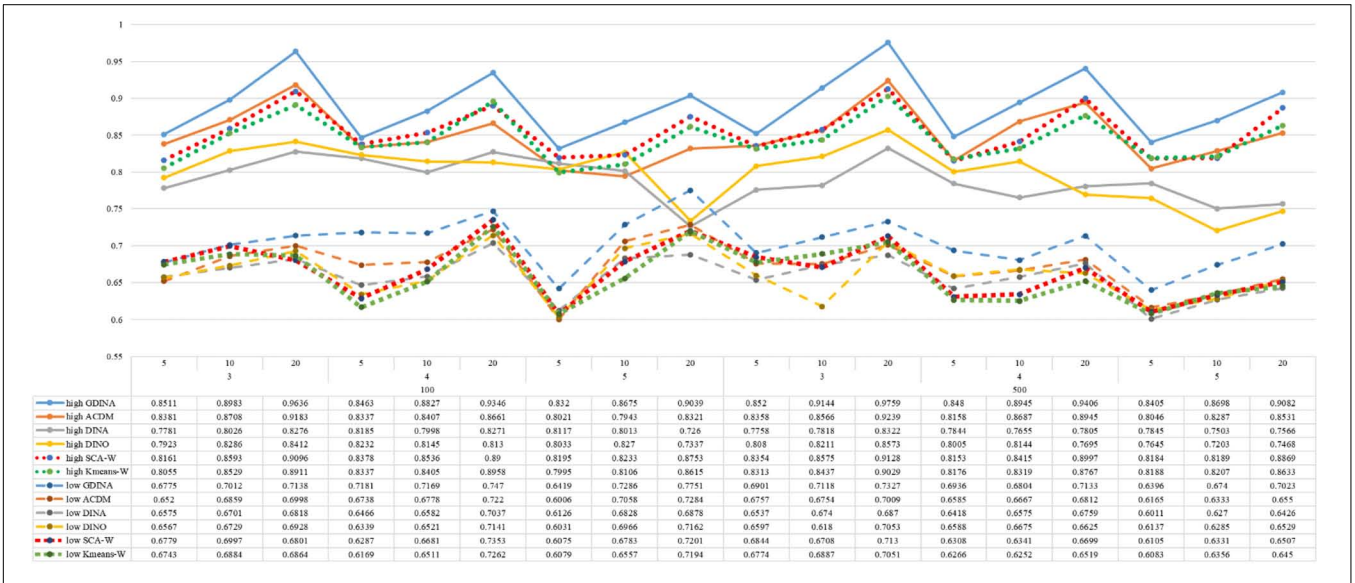
According to all results, the ARI and ω values are comparable between SCA and other methods (K-means and fitted CDMs) on the whole. In each Figure, the lines are clearly divided into two parts on account of item quality. The top half part presents high quality while the bottom half part presents low quality conditions. These results fully reflect the item quality, with a significant influence on accuracy of classification. Take **Figures 1**, **2** as an example, ARI values are all above 0.3, and ω values are all larger than 0.81 under the SCA with high quality. However, the lowest values of ARI and ω are 0.0284 and 0.6075, respectively, with low quality. **Figures 3–8** show the same results under different generating CDMs. It is noted that this deterioration is not unique for the SCA, moreover, the K-means and CDMs also have the same tendency. It demonstrates that item quality not only has a prominent influence on the performance of CDMs, but also has a dramatical effect on clustering methods. So, some important attentions should be paid to item quality in order to promote the classification accuracy in CDA regardless of the particular classification methods. As for two clustering methods, SCA can obtain higher ARI and ω values, representing more accurate clustering in most conditions, which can be concluded from that the red dot line (the legend denoted as SCA-W) is mostly above the green dot line (the legend denoted as Kmeans-W) in each parts.

For sample size, the impact of this factor on classification accuracy of these approaches is almost the same when other factors (e.g. attribute number, test length, item quality, and true models) are fixed, which means the clustering performance of SCA is comparable to K-means and other fitted CDMs. As can be seen those from eight figures, the ARI and ω values, soaring as the sample size, become large (from 100 to 500) on the whole. Since the relative advantage of cluster analysis applicated in small sample size, the main outcomes had been described under 100 sample size conditions (the left half part in each figure). Note that the similar results are presented in 500 sample size condition. When the G-DINA is the true model, the ARI and ω values of SCA are higher than those from K-means, DINA and DINO models (the red dot line is above) except that the ARIs in the conditions $K = 3$ and item quality is high, and $K = 4$ and item quality is low, respectively. This indicates SCA can be applied to most tests where there are a saturated underlying processes between attributes. As for A-CDM is the true model, we can see that SCA performs better than K-means, DINA and DINO models when item quality is high (except $K = 5$). Futhermore, SCA performs similarly as others in terms of ARIs (except $K = 3$ and $J = 5$ or 10), but ω values are consistently higher than other methods when item quality is low, which demonstrates stronger within-cluster homogeneity. This suggests SCA can also obtain decent classification accuracy when the cognitive mechanism is additive between attributes. Considering the true model is DINA model, the ARIs from SCA are almost higher than those from K-means and DINO model. Meanwhile, the ω values from SCA are also the highest among these three methods when item quality



	5	10 3	20	5	10 4 100	20	5	10 5	20	5	10 3	20	5	10 4 500	20	5	10 5	20
high GDINA	0.8511	0.8983	0.9636	0.8463	0.8827	0.9346	0.832	0.8675	0.9039	0.852	0.9144	0.9759	0.848	0.8945	0.9406	0.8405	0.8698	0.9082
high ACDM	0.8381	0.8708	0.9183	0.8337	0.8407	0.8661	0.8021	0.7943	0.8321	0.8358	0.8566	0.9239	0.8158	0.8687	0.8945	0.8046	0.8287	0.8531
high DINA	0.7781	0.8026	0.8276	0.8185	0.7998	0.8271	0.8117	0.8013	0.726	0.7758	0.7818	0.8322	0.7844	0.7655	0.7805	0.7845	0.7503	0.7566
high DINO	0.7923	0.8286	0.8412	0.8232	0.8145	0.813	0.8033	0.827	0.7337	0.808	0.8211	0.8573	0.8005	0.8144	0.7695	0.7645	0.7203	0.7468
high SCA-W	0.8161	0.8593	0.9096	0.8378	0.8536	0.89	0.8195	0.8233	0.8753	0.8354	0.8575	0.9128	0.8153	0.8415	0.8997	0.8184	0.8189	0.8869
high Kmeans-W	0.8055	0.8529	0.8911	0.8337	0.8405	0.8958	0.7995	0.8106	0.8615	0.8313	0.8437	0.9029	0.8176	0.8319	0.8767	0.8188	0.8207	0.8633
low GDINA	0.6775	0.7012	0.7138	0.7181	0.7169	0.747	0.6419	0.7286	0.7751	0.6901	0.7118	0.7327	0.6936	0.6804	0.7133	0.6396	0.674	0.7023
low ACDM	0.652	0.6859	0.6998	0.6778	0.6778	0.722	0.6006	0.7058	0.7284	0.6757	0.6754	0.7009	0.6585	0.6667	0.6812	0.6165	0.6333	0.655
low DINA	0.6575	0.6701	0.6818	0.6466	0.6582	0.7037	0.6126	0.6828	0.6878	0.6537	0.674	0.687	0.6418	0.6575	0.6759	0.6011	0.627	0.6426
low DINO	0.6567	0.6729	0.6928	0.6339	0.6521	0.7141	0.6031	0.6966	0.7162	0.6597	0.618	0.7053	0.6588	0.6675	0.6625	0.6137	0.6285	0.6529
low SCA-W	0.6779	0.6997	0.6801	0.6287	0.6681	0.7353	0.6075	0.6783	0.7201	0.6844	0.6708	0.713	0.6308	0.6341	0.6699	0.6105	0.6331	0.6507
low Kmeans-W	0.6743	0.6884	0.6864	0.6169	0.6511	0.7262	0.6079	0.6557	0.7194	0.6774	0.6887	0.7051	0.6266	0.6252	0.6519	0.6083	0.6356	0.645

FIGURE 1 | Mean values of ARI by SCA, K-means, and fitted models; True model = G-DINA.



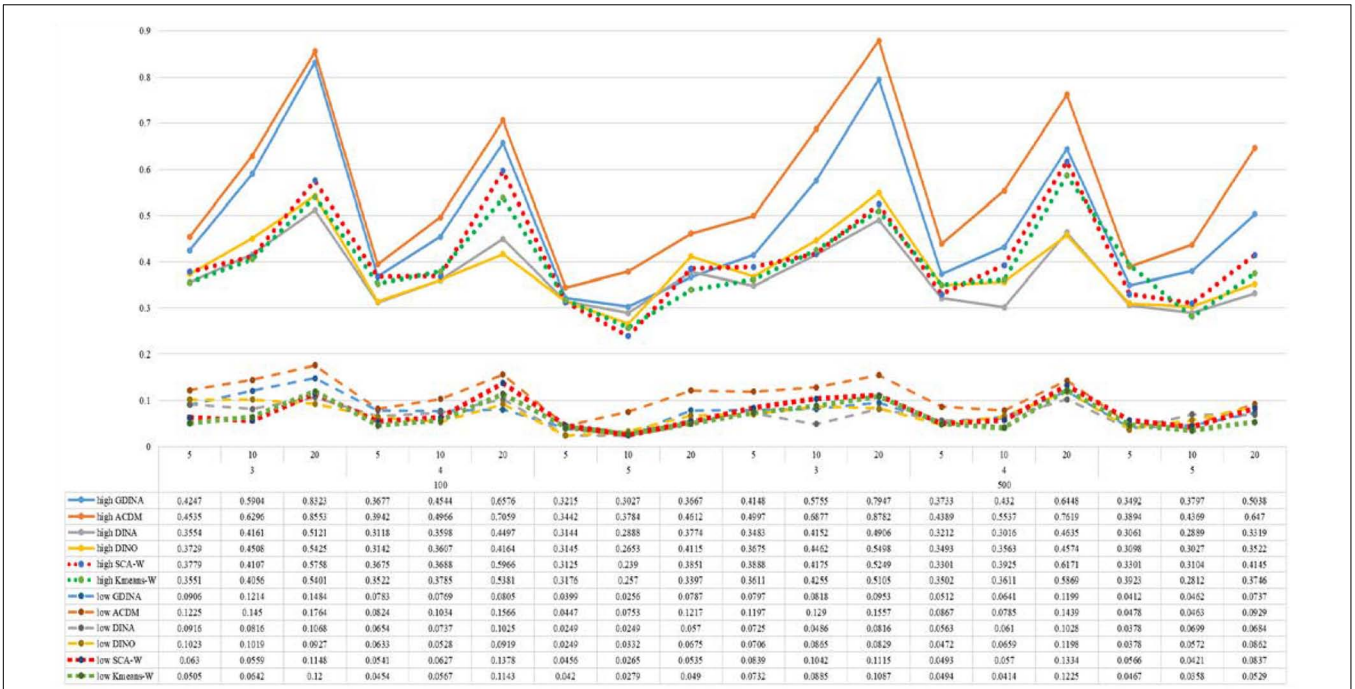
	5	10 3	20	5	10 4 100	20	5	10 5	20	5	10 3	20	5	10 4 500	20	5	10 5	20
high GDINA	0.8511	0.8983	0.9636	0.8463	0.8827	0.9346	0.832	0.8675	0.9039	0.852	0.9144	0.9759	0.848	0.8945	0.9406	0.8405	0.8698	0.9082
high ACDM	0.8381	0.8708	0.9183	0.8337	0.8407	0.8661	0.8021	0.7943	0.8321	0.8358	0.8566	0.9239	0.8158	0.8687	0.8945	0.8046	0.8287	0.8531
high DINA	0.7781	0.8026	0.8276	0.8185	0.7998	0.8271	0.8117	0.8013	0.726	0.7758	0.7818	0.8322	0.7844	0.7655	0.7805	0.7845	0.7503	0.7566
high DINO	0.7923	0.8286	0.8412	0.8232	0.8145	0.813	0.8033	0.827	0.7337	0.808	0.8211	0.8573	0.8005	0.8144	0.7695	0.7645	0.7203	0.7468
high SCA-W	0.8161	0.8593	0.9096	0.8378	0.8536	0.89	0.8195	0.8233	0.8753	0.8354	0.8575	0.9128	0.8153	0.8415	0.8997	0.8184	0.8189	0.8869
high Kmeans-W	0.8055	0.8529	0.8911	0.8337	0.8405	0.8958	0.7995	0.8106	0.8615	0.8313	0.8437	0.9029	0.8176	0.8319	0.8767	0.8188	0.8207	0.8633
low GDINA	0.6775	0.7012	0.7138	0.7181	0.7169	0.747	0.6419	0.7286	0.7751	0.6901	0.7118	0.7327	0.6936	0.6804	0.7133	0.6396	0.674	0.7023
low ACDM	0.652	0.6859	0.6998	0.6778	0.6778	0.722	0.6006	0.7058	0.7284	0.6757	0.6754	0.7009	0.6585	0.6667	0.6812	0.6165	0.6333	0.655
low DINA	0.6575	0.6701	0.6818	0.6466	0.6582	0.7037	0.6126	0.6828	0.6878	0.6537	0.674	0.687	0.6418	0.6575	0.6759	0.6011	0.627	0.6426
low DINO	0.6567	0.6729	0.6928	0.6339	0.6521	0.7141	0.6031	0.6966	0.7162	0.6597	0.618	0.7053	0.6588	0.6675	0.6625	0.6137	0.6285	0.6529
low SCA-W	0.6779	0.6997	0.6801	0.6287	0.6681	0.7353	0.6075	0.6783	0.7201	0.6844	0.6708	0.713	0.6308	0.6341	0.6699	0.6105	0.6331	0.6507
low Kmeans-W	0.6743	0.6884	0.6864	0.6169	0.6511	0.7262	0.6079	0.6557	0.7194	0.6774	0.6887	0.7051	0.6266	0.6252	0.6519	0.6083	0.6356	0.645

FIGURE 2 | Mean values of ω by SCA, K-means, and fitted models; True model = G-DINA.

is low, whereas the magnitudes of ωs are affected by test length when item quality is high. Specifically, ω values of SCA are higher than those from K-means, *A*-CDM and DINO when $J = 20$, and be inverse when $J = 5$ or 10. The results show that the performance of SCA is acceptable when item quality is low, or J > 20 if the underlying process is conjunctive among attributes. Providing that DINO is the true model, the ARIs from SCA are almost higher than those from K-means and DINA model. Similarly, the ω values from SCA are the highest among these three methods, especially higher than *A*-CDM when $K = 3$ and item quality is high. This implies SCA has a patchy performance when disjunctive process arose between attributes.

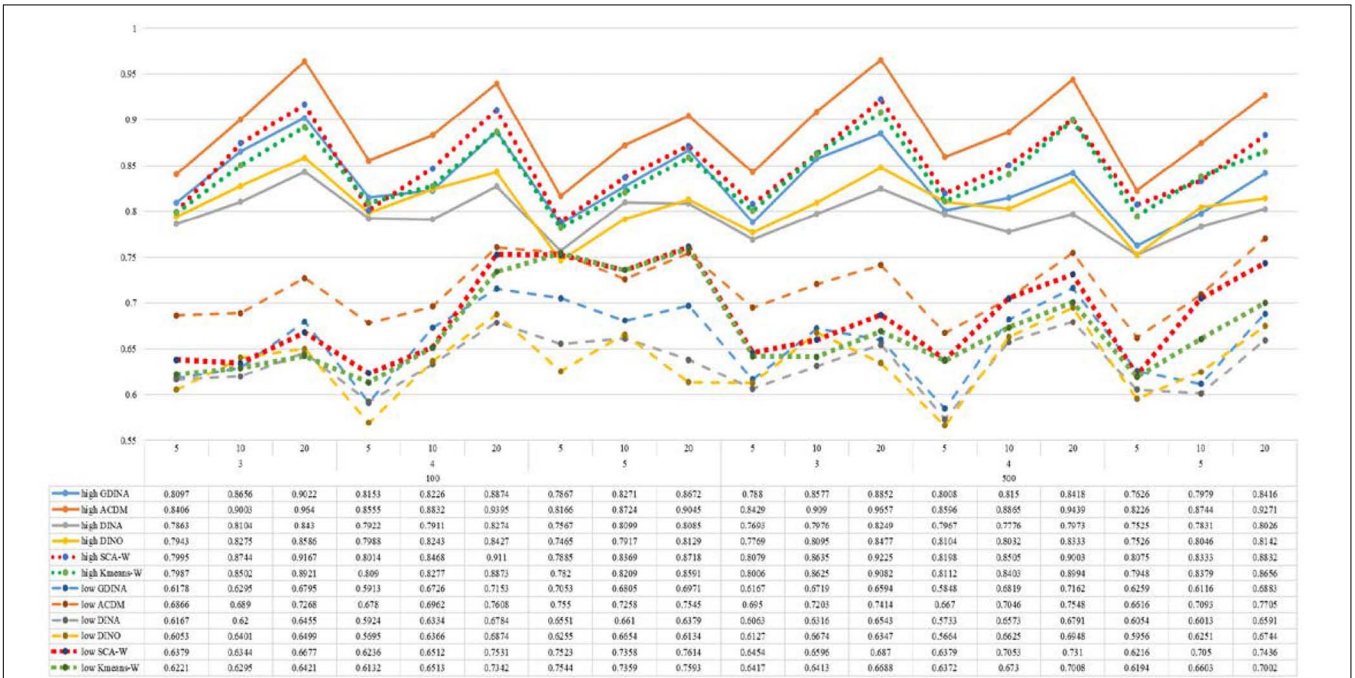
In addition, the number of attributes also affects the classification accuracy of SCA as same as CDMs. Generally speaking, with attribute number K increasing, the ARI and ω values decreases. Most results conform to this pattern as shown in **Figures 1**–**8**. However, this trend is not consistent across all conditions. For instance, in **Figure 1**, for condition $(I, J) = (100, 5)$, ARI values change from 0.3554 to 0.3628 under SCA, while ARI values change from 0.3446 to 0.3754 under K-means when K grows from 3 to 4. ω values change from 0.8161 (0.8055) to 0.8378 (0.8337) under SCA (K-means). Due to the randomness of generating Q-matrix in each replication, the K-means may arise some reversal results in some conditions. So, it may infer that the combination of q-vectors influences the effect of attribute number on classification accuracy.

Last, test length is a widely considered factor in CDA. Many studies have discussed the influence of this factor on classification



	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
		3			4			5			3			4			5		
					100									500				
high GDINA	0.4247	0.5904	0.8323	0.3677	0.4544	0.6576	0.3215	0.3027	0.3667	0.4148	0.5755	0.7947	0.3733	0.432	0.6448	0.3492	0.3797	0.5038
high ACDM	0.4535	0.6296	0.8553	0.3942	0.4966	0.7059	0.3442	0.3784	0.4612	0.4497	0.6877	0.8782	0.4389	0.5537	0.7619	0.3894	0.4369	0.647
high DINA	0.3554	0.4161	0.5121	0.3118	0.3598	0.4497	0.3144	0.2888	0.3774	0.3483	0.4152	0.4906	0.3212	0.3016	0.4635	0.3061	0.2889	0.3319
high DINO	0.3729	0.4508	0.5425	0.3142	0.3607	0.4164	0.3145	0.2653	0.4115	0.3675	0.4462	0.5498	0.3493	0.3563	0.4574	0.3098	0.3027	0.3522
high SCA-W	0.3779	0.4107	0.5758	0.3675	0.3688	0.5966	0.3125	0.239	0.3851	0.3888	0.4175	0.5249	0.3301	0.3925	0.6171	0.3301	0.3104	0.4145
high Kmeans-W	0.3551	0.4056	0.5401	0.3522	0.3785	0.5381	0.3176	0.257	0.3397	0.3611	0.4255	0.5105	0.3502	0.3611	0.5869	0.3923	0.2812	0.3746
low GDINA	0.0906	0.1214	0.1484	0.0783	0.0769	0.0805	0.0399	0.0256	0.0787	0.0797	0.0818	0.0953	0.0512	0.0641	0.1199	0.0412	0.0462	0.0737
low ACDM	0.1225	0.145	0.1764	0.0824	0.1034	0.1566	0.0447	0.0753	0.1217	0.1197	0.129	0.1557	0.0867	0.0785	0.1439	0.0478	0.0463	0.0929
low DINA	0.0916	0.0816	0.1068	0.0654	0.0737	0.1025	0.0249	0.0249	0.057	0.0725	0.0486	0.0816	0.0563	0.061	0.1028	0.0378	0.0699	0.0684
low DINO	0.1023	0.1019	0.0927	0.0633	0.0528	0.0919	0.0249	0.0332	0.0675	0.0706	0.0865	0.0829	0.0472	0.0659	0.1198	0.0378	0.0572	0.0862
low SCA-W	0.063	0.0559	0.1148	0.0541	0.0627	0.1378	0.0456	0.0265	0.0535	0.0839	0.1042	0.1115	0.0493	0.057	0.1334	0.0566	0.0421	0.0837
low Kmeans-W	0.0505	0.0642	0.12	0.0454	0.0567	0.1143	0.042	0.0279	0.049	0.0732	0.0885	0.1087	0.0494	0.0414	0.1225	0.0467	0.0358	0.0529

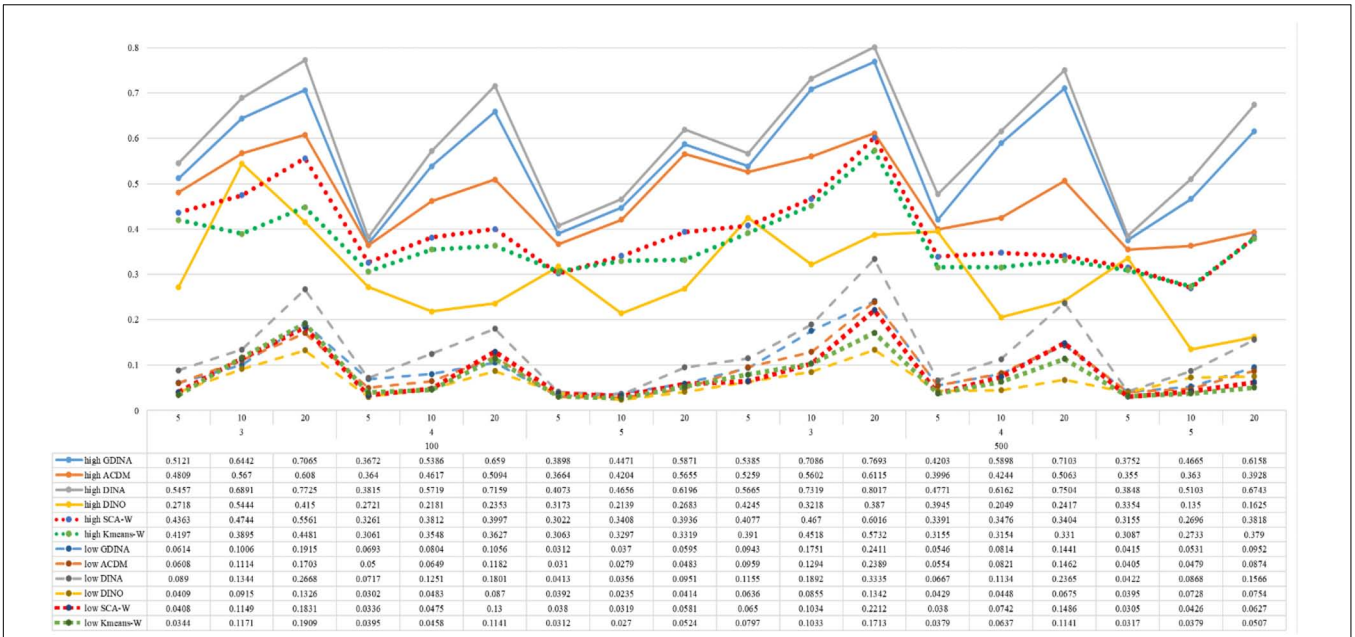
FIGURE 3 | Mean values of ARI by SCA, K-means, and fitted models; True model = *A*-CDM.



	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
		3			4			5			3			4			5		
					100									500				
high GDINA	0.8097	0.8656	0.9022	0.8153	0.8226	0.8874	0.7867	0.8271	0.8672	0.788	0.8577	0.8852	0.8008	0.815	0.8418	0.7626	0.7979	0.8416
high ACDM	0.8406	0.9003	0.954	0.8555	0.8832	0.9395	0.8166	0.8724	0.9045	0.8429	0.909	0.9557	0.8596	0.8865	0.9439	0.8226	0.8744	0.9271
high DINA	0.7863	0.8104	0.843	0.7922	0.7911	0.8274	0.7567	0.8099	0.8085	0.7693	0.7976	0.8249	0.7967	0.7776	0.7973	0.7525	0.7831	0.8026
high DINO	0.7943	0.8275	0.8586	0.7988	0.8243	0.8427	0.7465	0.7917	0.8129	0.7769	0.8095	0.8477	0.8104	0.8032	0.8333	0.7526	0.8046	0.8142
high SCA-W	0.7995	0.8744	0.9167	0.8014	0.8468	0.911	0.7885	0.8369	0.8718	0.8079	0.8635	0.9225	0.8198	0.8505	0.9003	0.8075	0.8333	0.8832
high Kmeans-W	0.7987	0.8502	0.8921	0.809	0.8277	0.8873	0.782	0.8209	0.8591	0.8006	0.8625	0.9082	0.8112	0.8403	0.8994	0.7948	0.8379	0.8656
low GDINA	0.6178	0.6295	0.6795	0.5913	0.6726	0.7153	0.7053	0.6805	0.6971	0.6167	0.6719	0.6594	0.5848	0.6819	0.7162	0.6259	0.6116	0.6883
low ACDM	0.6866	0.689	0.7268	0.678	0.6962	0.7608	0.755	0.7258	0.7545	0.695	0.7203	0.7414	0.667	0.7046	0.7548	0.6616	0.7093	0.7705
low DINA	0.6167	0.62	0.6455	0.5924	0.6334	0.6784	0.6551	0.661	0.6379	0.6063	0.6316	0.6543	0.5733	0.6573	0.6791	0.6054	0.6013	0.6591
low DINO	0.6053	0.6401	0.6499	0.5995	0.6366	0.6874	0.6255	0.6654	0.6134	0.6127	0.6674	0.6347	0.5664	0.6625	0.6948	0.5956	0.6251	0.6744
low SCA-W	0.6379	0.6344	0.6677	0.6236	0.6512	0.7531	0.7523	0.7358	0.7614	0.6454	0.6596	0.687	0.6379	0.7053	0.731	0.6216	0.705	0.7436
low Kmeans-W	0.6221	0.6295	0.6421	0.6132	0.6513	0.7342	0.7544	0.7359	0.7593	0.6417	0.6413	0.6688	0.6372	0.673	0.7008	0.6194	0.6603	0.7002

FIGURE 4 | Mean values of ω by SCA, K-means, and fitted models; True model = *A*-CDM.

accuracy (Chen et al., 2013; Chiu et al., 2009). From the results of these simulations, as J increases, the classification abilities of all methods tend to improve. Considering the shortest test length condition (J = 5), most ω values are no less than 0.8 when item quality is high, while most ω values are no less than 0.6 when item quality is low under the SCA procedure. Definitely, the longer the test length is, the more information about the examinees it provides, and more accurate classification will be obtained. This indicates the SCA can be affected by test length just like other methods.



	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
		3			4			5			3			4			5		
					100									500				
high GDINA	0.5121	0.6442	0.7065	0.3672	0.5386	0.659	0.3898	0.4471	0.5871	0.5385	0.7086	0.7693	0.4203	0.5898	0.7103	0.3752	0.4665	0.6158
high ACDM	0.4809	0.567	0.608	0.364	0.4617	0.5094	0.3664	0.4204	0.5655	0.5259	0.5602	0.6115	0.3996	0.4244	0.5063	0.355	0.363	0.3928
high DINA	0.5457	0.6891	0.7725	0.3815	0.5719	0.7159	0.4073	0.4656	0.6196	0.5665	0.7319	0.8017	0.4771	0.6162	0.7504	0.3848	0.5103	0.6743
high DINO	0.2718	0.5444	0.415	0.2721	0.2181	0.2353	0.3173	0.2139	0.2683	0.4245	0.3218	0.387	0.3945	0.2049	0.2417	0.3354	0.135	0.1625
high SCA-W	0.4363	0.4744	0.5561	0.3261	0.3812	0.3997	0.3022	0.3408	0.3936	0.4077	0.467	0.6016	0.3391	0.3476	0.3404	0.3155	0.2696	0.3818
high Kmeans-W	0.4197	0.3895	0.4481	0.3061	0.3548	0.3627	0.3063	0.3297	0.3319	0.391	0.4518	0.5732	0.3155	0.3154	0.331	0.3087	0.2733	0.379
low GDINA	0.0614	0.1006	0.1915	0.0693	0.0804	0.1056	0.0312	0.037	0.0595	0.0943	0.1751	0.2411	0.0546	0.0814	0.1441	0.0415	0.0531	0.0952
low ACDM	0.0608	0.1114	0.1703	0.05	0.0649	0.1182	0.031	0.0279	0.0483	0.0959	0.1294	0.2389	0.0554	0.0821	0.1462	0.0405	0.0479	0.0874
low DINA	0.089	0.1344	0.2668	0.0717	0.1251	0.1801	0.0413	0.0356	0.0951	0.1155	0.1892	0.3335	0.0667	0.1134	0.2365	0.0422	0.0868	0.1566
low DINO	0.0409	0.0915	0.1326	0.0302	0.0483	0.087	0.0392	0.0235	0.0414	0.0636	0.0855	0.1342	0.0429	0.0448	0.0675	0.0395	0.0728	0.0754
low SCA-W	0.0408	0.1149	0.1831	0.0336	0.0475	0.13	0.038	0.0319	0.0581	0.065	0.1034	0.2212	0.038	0.0742	0.1486	0.0305	0.0426	0.0627
low Kmeans-W	0.0344	0.1171	0.1909	0.0395	0.0458	0.1141	0.0312	0.027	0.0524	0.0797	0.1033	0.1713	0.0379	0.0637	0.1141	0.0317	0.0379	0.0507

FIGURE 5 | Mean values of ARI by SCA, K-means, and fitted models; True model = DINA.

ANALYSIS OF MIXED NUMBER SUBTRACTION DATA

Data Description

The data consist of 536 examinees' responses to 11 items taken from the mixed number fraction subtraction. The **Q**-matrix was modified from five attributes to three attributes, and they were previously used by Henson et al. (2009). The attributes defined for this study are (1) borrowing from a whole number, (2) separating a whole number from a fraction and (3) determining a common denominator. **Table 1** shows the 11 items and their required attributes. It should be pointed out that the data and the **Q**-matrix were got from R package 'CDM', and the item 12 was excluded from the original table as shown in Henson et al.'s paper. So, there were just 11 items in this study. Then, the SCA and K-means algorithm with Ward's linkage, and four CDMs were applied to classify examinees into different clusters.

Two major criteria evaluating the classified quality were used as those in Chiu et al. (2009) study, denoted as within-cluster mean of **W** (see Eq. 10 for the definition), and square root of mean squared residual (MSR) of **W**. Specifically, the mean of **W** reflects how well-separated cluster means are, which can provide good identification of examinees' overall patterns. And MSR of **W** shows that how homogeneous a cluster is. The MSR of **W** for cluster m is given by

$$MSR(m) = \frac{\sum_i^{I_m} \parallel \mathbf{W}_i^{(m)} - \bar{W}^{(m)} \parallel^2}{I_m} \tag{16}$$

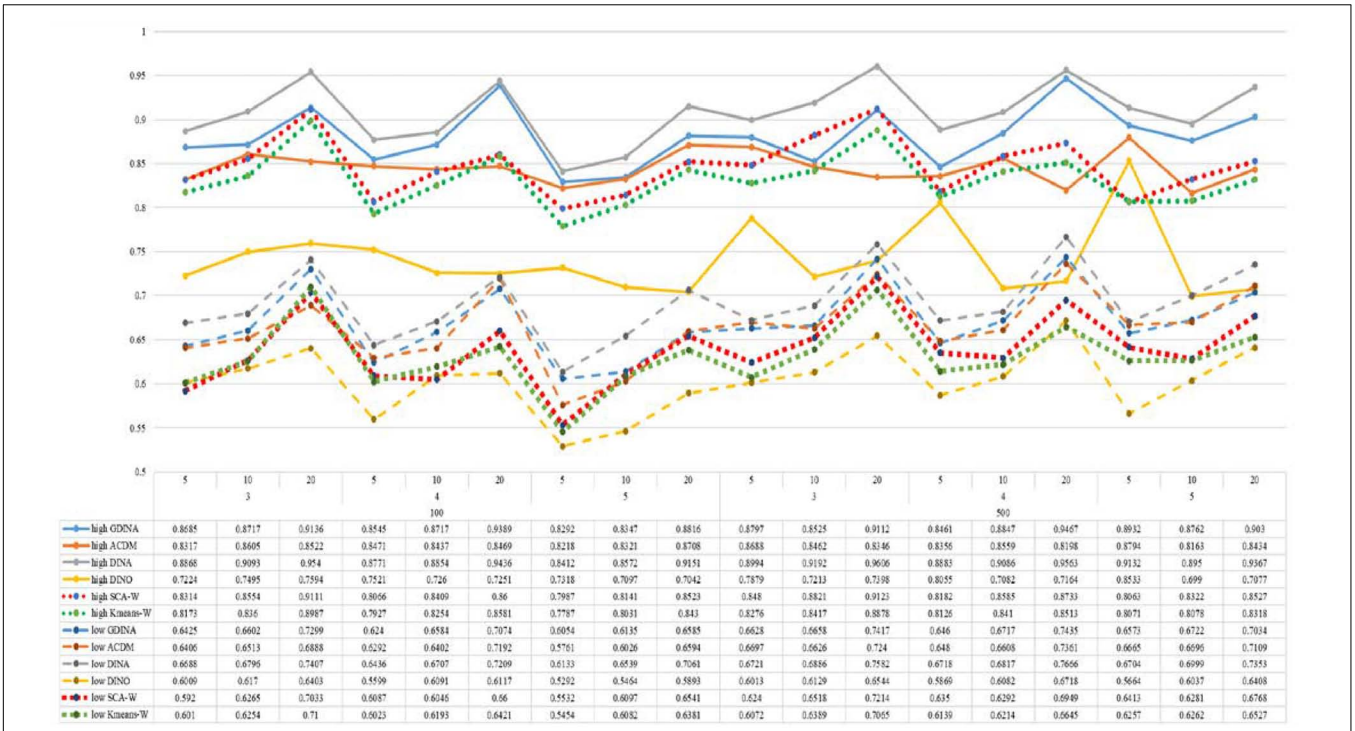
where I_m is the number of examinees grouped into cluster m. The smaller the MSR is, the more homogeneous a cluster is.

Meanwhile, we also report the cluster size and mean of sum-score as the auxiliary indicators. The classification results from SCA, K-means, and CDMs were sorted by means of sum-score, which can be used to infer attribute profiles in practice (Chiu et al., 2009). The rationale is that one may get higher sum-score if (s)he masters more attribute in a test usually.

Analysis and Results

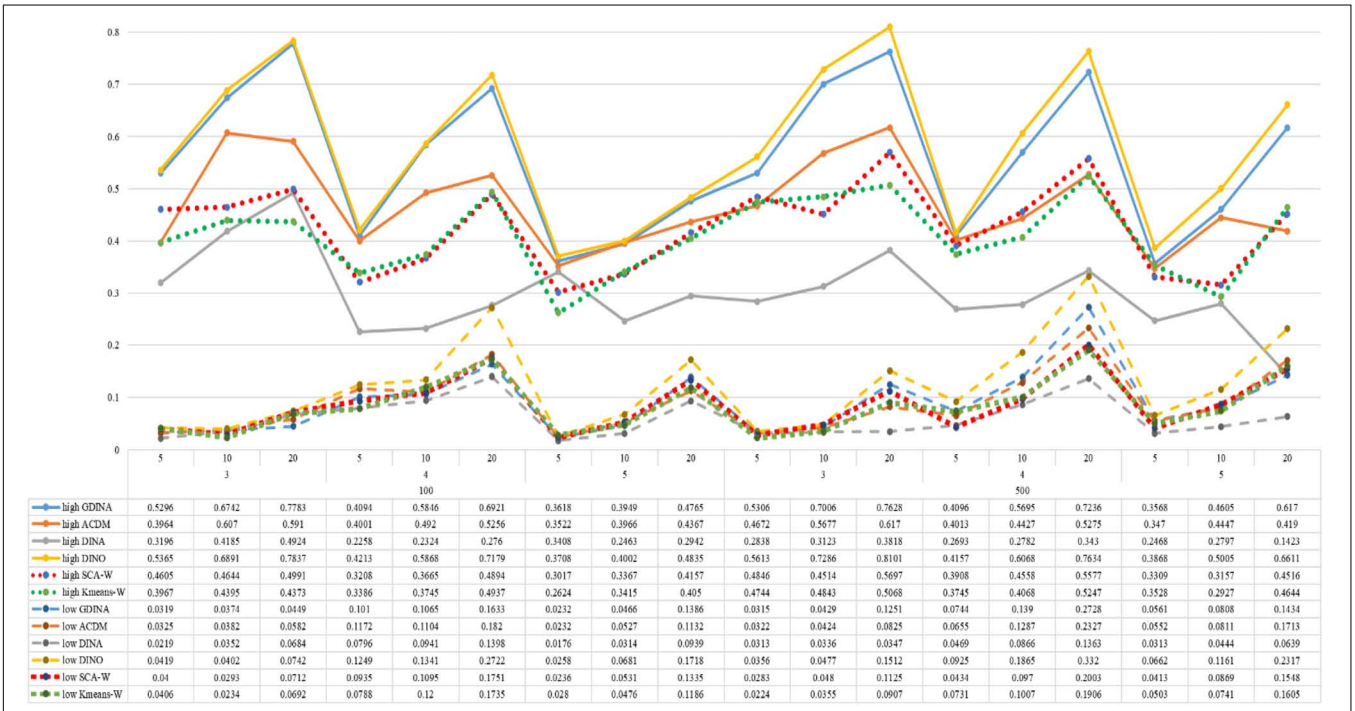
The data were analyzed by all methods through the statistic **W**. We only select the Ward's starting values due to their good performance in simulation studies. Note that the attribute profiles' labels were not available for clustering analysis, and the results from the SCA and K-means were sorted along with the means of sum-scores in the same cluster, illustrating how one can infer the examinees' attribute profiles. It means that the mean of sum-scores in certain cluster representing $\boldsymbol{\alpha} = (0, 0, 0)$ is definitely the smallest among eight attribute profiles, while the mean of sum-scores is the largest for profile $\boldsymbol{\alpha} = (1, 1, 1)$. Because of the acquirement of specific attribute profiles by using the CDMs, results are listed according to the size of attribute vectors.

When using multiple models to fit the same data, the Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarzer, 1976) were usually adopted to determine which model can provide a better fit result. For each of these two statistics, the fitted model with a the smaller value is selected among the set of competing models. **Table 2** shows the AIC and BIC for four CDMs fitting the fraction subtraction data. The AIC is the smallest under the G-DINA model, but the BIC is the smallest under A-CDM. According to previous study, if AIC and BIC contradict each other, the BIC may provide a better result for selecting model because BIC takes



	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
		3			4			5			3			4			5		
					100									500				
high GDINA	0.8685	0.8717	0.9136	0.8545	0.8717	0.9389	0.8292	0.8347	0.8816	0.8797	0.8525	0.9112	0.8461	0.8847	0.9467	0.8932	0.8762	0.903
high ACDM	0.8317	0.8605	0.8522	0.8471	0.8437	0.8469	0.8218	0.8321	0.8708	0.8688	0.8462	0.8346	0.8356	0.8559	0.8198	0.8794	0.8163	0.8434
high DINA	0.8868	0.9093	0.954	0.8771	0.8854	0.9436	0.8412	0.8572	0.9151	0.8994	0.9192	0.9606	0.8883	0.9086	0.9563	0.9132	0.895	0.9367
high DINO	0.7224	0.7495	0.7594	0.7521	0.726	0.7251	0.7318	0.7097	0.7042	0.7879	0.7213	0.7398	0.8055	0.7082	0.7164	0.8533	0.699	0.7077
high SCA-W	0.8314	0.8554	0.9111	0.8056	0.8409	0.86	0.7987	0.8141	0.8523	0.848	0.8821	0.9123	0.8182	0.8585	0.8733	0.8063	0.8322	0.8527
high Kmeans-W	0.8173	0.836	0.8987	0.7927	0.8254	0.8581	0.7787	0.8031	0.843	0.8276	0.8417	0.8878	0.8126	0.841	0.8513	0.8071	0.8078	0.8318
low GDINA	0.6425	0.6602	0.7299	0.624	0.6584	0.7074	0.6054	0.6135	0.6585	0.6628	0.6658	0.7417	0.646	0.6717	0.7435	0.6573	0.6722	0.7034
low ACDM	0.6406	0.6513	0.6888	0.6292	0.6402	0.7192	0.5761	0.6026	0.6594	0.6697	0.6626	0.724	0.648	0.6608	0.7361	0.6665	0.6696	0.7109
low DINA	0.6688	0.6796	0.7407	0.6436	0.6707	0.7209	0.6133	0.6539	0.7061	0.6721	0.6886	0.7582	0.6718	0.6817	0.7666	0.6704	0.6999	0.7353
low DINO	0.6009	0.617	0.6403	0.5599	0.6091	0.6117	0.5202	0.5464	0.5893	0.6013	0.6129	0.6544	0.5869	0.6082	0.6718	0.5664	0.6037	0.6408
low SCA-W	0.592	0.6265	0.7033	0.6087	0.6045	0.66	0.5532	0.6097	0.6541	0.624	0.6518	0.7214	0.635	0.6292	0.6949	0.6413	0.6281	0.6768
low Kmeans-W	0.601	0.6254	0.71	0.6023	0.6193	0.6421	0.5454	0.6082	0.6381	0.6072	0.6389	0.7065	0.6139	0.6214	0.6645	0.6257	0.6262	0.6527

FIGURE 6 | Mean values of ω by SCA, K-means, and fitted models; True model = DINA.



	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
		3			4			5			3			4			5		
					100									500				
high GDINA	0.5296	0.6742	0.7783	0.4094	0.5846	0.6921	0.3618	0.3949	0.4765	0.5306	0.7006	0.7628	0.4096	0.5695	0.7236	0.3568	0.4605	0.617
high ACDM	0.3964	0.607	0.591	0.4001	0.492	0.5256	0.3522	0.3966	0.4367	0.4672	0.5677	0.617	0.4013	0.4427	0.5275	0.347	0.4447	0.419
high DINA	0.3196	0.4185	0.4924	0.2258	0.2324	0.276	0.3408	0.2463	0.2942	0.2838	0.3123	0.3818	0.2693	0.2782	0.343	0.2468	0.2797	0.1423
high DINO	0.5365	0.6891	0.7837	0.4213	0.5868	0.7179	0.3708	0.4002	0.4835	0.5613	0.7286	0.8101	0.4157	0.6068	0.7634	0.3868	0.5005	0.6611
high SCA-W	0.4605	0.4644	0.4991	0.3208	0.3665	0.4894	0.3017	0.3367	0.4157	0.4846	0.4514	0.5697	0.3908	0.4558	0.5577	0.3309	0.3157	0.4516
high Kmeans-W	0.3967	0.4395	0.4373	0.3386	0.3745	0.4937	0.2624	0.3415	0.405	0.4744	0.4843	0.5068	0.3745	0.4068	0.5247	0.3528	0.2927	0.4644
low GDINA	0.0319	0.0374	0.0449	0.101	0.1065	0.1633	0.0232	0.0466	0.1386	0.0315	0.0429	0.1251	0.0744	0.139	0.2728	0.0561	0.0808	0.1434
low ACDM	0.0325	0.0382	0.0582	0.1172	0.1104	0.182	0.0232	0.0527	0.1132	0.0322	0.0424	0.0825	0.0655	0.1287	0.2327	0.0552	0.0811	0.1713
low DINA	0.0219	0.0352	0.0684	0.0796	0.0941	0.1398	0.0176	0.0314	0.0939	0.0313	0.0336	0.0347	0.0469	0.0866	0.1363	0.0313	0.0444	0.0639
low DINO	0.0419	0.0402	0.0742	0.1249	0.1341	0.2722	0.0258	0.0681	0.1718	0.0356	0.0477	0.1512	0.0925	0.1865	0.332	0.0662	0.1161	0.2317
low SCA-W	0.04	0.0293	0.0712	0.0935	0.1095	0.1751	0.0236	0.0531	0.1335	0.0283	0.048	0.1125	0.0434	0.097	0.2003	0.0413	0.0869	0.1548
low Kmeans-W	0.0406	0.0234	0.0692	0.0788	0.12	0.1735	0.028	0.0476	0.1186	0.0224	0.0355	0.0907	0.0731	0.1007	0.1906	0.0503	0.0741	0.1605

FIGURE 7 | Mean values of ARI by SCA, K-means, and fitted models; True model = DINO.

into account both the sample size and the number of parameters of the model (Chen et al., 2013). Based on this point, the A-CDM provides the best fit among these four CDMs.

Due to the space limitation, only the results obtained by the best fit model, *A*-CDM, are shown in the table. As can be seen in **Table 3**, the *A*-CDM intensively grouped most examinees

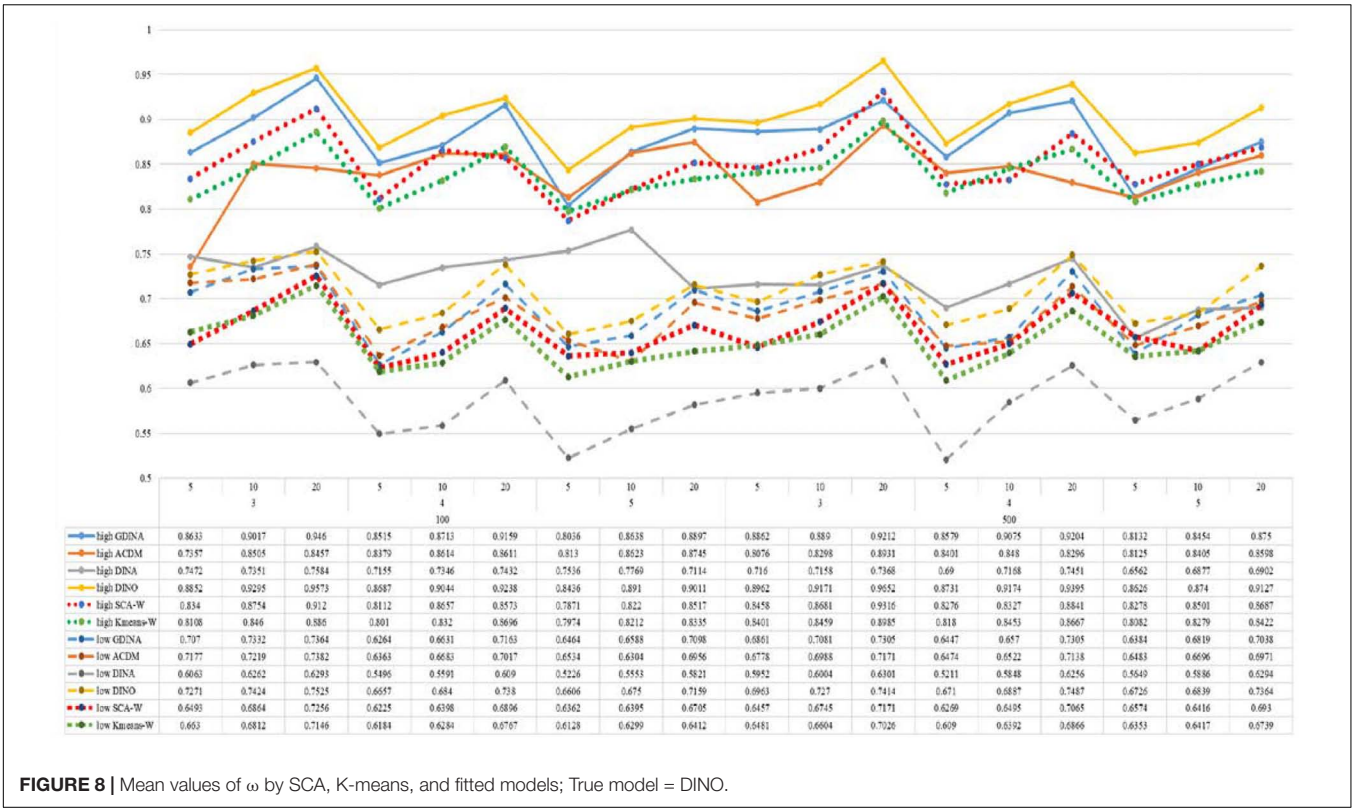


FIGURE 8 | Mean values of ω by SCA, K-means, and fitted models; True model = DINO.

into three main clusters, and the remaining clusters only have a few examinees. In addition, the differences among W_1 to W_3 are comparative large under the *A*-CDM, so it is benificial to identify attribute profiles of examinees. However, large MSRs are got by using this model, which means this empirical data are not clustered closely based on examinees' profiles, then apart cluster means may result in heterogeneous clustering.

In contrast, **Table 4** shows that the SCA classified most data to the profiles which stand for mastering only one attribute (denoted as $\boldsymbol{\alpha}^{(1)}$), the number of examinees is 137. The second largest cluster size is 75, and this cluster represents the profile $\boldsymbol{\alpha} = (1, 1, 1)$. Similarly, the K-means method also classified most data to the same profiles, with the clusters $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha} = (1, 1, 1)$ are both containing 100 examinees. The distances between the pairs of clusters in the SCA are larger than those in K-means method according to the values of **W**, which means that SCA can give well-separated clusters. In addition, the values of MSR under these two clustering methods are smaller than those under the *A*-CDM. Further, MSR under SCA are smaller than those under K-means, except one cluster (see the bold value on the second row). This is in accord with the results from simulation study that the SCA tends to form close and homogeneous clusters.

Finally, taking the *A*-CDM as the standard, **Table 5** presents the classification agreement of each two methods, including SCA, K-means, and *A*-CDM. The agreement between the *A*-CDM and SCA is slight higher than the other pairs with an ARI of 0.468 compared to an ARI of 0.443 for the agreement between the *A*-CDM and K-means. It indicates that SCA outperformed K-means for this data set.

SUMMARY AND DISCUSSION

The contribution of this study is to introduce the SCA into cognitive diagnosis and compare it with the K-means method and different CDMs in terms of classification accuracy. The clustering methods are computationally efficient and effctive for data with

TABLE 1 | Mixed number fraction subtraction and corresponding q-matrix.

Item number	Item	Q-matrix	Item number	Item	Q-matrix
1	$3\frac{1}{2} - 2\frac{3}{2}$	1 1 0	8	$2 - \frac{1}{3}$	1 0 1
2	$3 - 2\frac{1}{5}$	1 0 1	9	$4\frac{5}{7} - 1\frac{7}{4}$	1 1 1
3	$3\frac{7}{8} - 2$	1 0 1	10	$7\frac{3}{5} - \frac{4}{5}$	1 0 0
4	$4\frac{4}{12} - 2\frac{7}{12}$	1 0 0	11	$4\frac{1}{10} - 2\frac{8}{10}$	1 0 0
5	$4\frac{1}{3} - 2\frac{4}{3}$	1 1 0	13	$4\frac{1}{3} - 1\frac{5}{3}$	1 1 0
6	$\frac{11}{8} - \frac{1}{8}$	1 1 0			

TABLE 2 | AIC and BIC for four CDMs fitting fraction subtraction data.

Models	AIC	BIC
G-DINA	**5341.06**	5550.98
DINA	5534.39	5658.63
DINO	5517.80	5642.04
A-CDM	5363.15	**5525.94**

Bold values mean the best models that we need to select to fit the data.

TABLE 3 | Classification by A-CDM.

Profile	Size	Mean W			$\sqrt{MSR(m)}$	Mean Sum-score
		W_1	W_2	W_3		
(0 0 0)	127	0.93	0.51	0.66	1.35	1.07
(0 1 0)	109	1.55	1.94	1.37	1.34	2.83
(0 0 1)	6	3.41	1.33	3.00	0.75	4.00
(1 0 0)	9	4.70	3.56	1.22	1.19	6.11
(1 1 0)	40	3.62	4.55	1.45	1.35	7.33
(1 0 1)	6	2.59	2.83	3.50	1.03	8.00
(0 1 1)	44	4.37	2.14	2.93	1.60	4.80
(1 1 1)	195	7.09	4.63	3.62	1.37	9.88

TABLE 4 | Classification by SCA-Ward's and K-means-Ward's algorithm.

Size	Mean W			$\sqrt{MSR(m)}$	Mean sum-score
	W_1	W_2	W_3		
59 (79)	0.03 (0.04)	0.41 (0.70)	0.07 (0.09)	0.67 (0.79)	0.61 (0.73)
62 (40)	0.13 (0.00)	0.24 (0.00)	1.26 (1.00)	0.75 (**0.00**)	1.63 (1.00)
71 (100)	0.18 (0.05)	1.10 (1.04)	1.82 (1.61)	0.67 (0.74)	3.10 (2.70)
137 (71)	0.88 (0.87)	1.98 (1.24)	2.28 (2.82)	1.09 (1.18)	5.14 (4.93)
58 (52)	2.24 (2.06)	3.36 (3.35)	1.60 (1.35)	1.18 (1.30)	7.21 (6.75)
32 (52)	2.78 (2.62)	3.66 (2.87)	2.94 (3.60)	0.76 (0.77)	9.38 (9.08)
42 (42)	2.40 (2.57)	3.33 (4.00)	3.86 (2.81)	0.70 (0.93)	9.60 (9.38)
75 (100)	3.00 (2.67)	4.00 (3.97)	3.96 (4.00)	0.25 (0.55)	10.93 (10.67)

Results of K-means-Ward's present in parentheses.

TABLE 5 | ARI table for ACDM, SCA and *K*-means.

	***A*-CDM**	**SCA-Ward's**	**K-means-Ward's**
A-CDM	–	0.468	0.443
SCA-Ward's		–	0.427
K-means-Ward's			–

any sample size. It's easy and convenient to implement, and researchers only need to know the number of required attributes and their hierarchical structures. The previous study had shown that K-means has favorable performance in clustering examinees who possess the same attribute profiles (Chiu et al., 2009). In this study, we introduced the SCA for grouping examinees' attribute profiles into specific clusters in CDA. Then, the performance of SCA on classification accuracy was investigated under different factors, and some interesting findings were made based on simulation studies.

The most important factor affecting the classification accuracy of both clustering analysis and CDMs was item quality. Generally, the higher the item quality was, the higher the classification accuracy was. This is because the randomness (i.e. guessing and slipping behaviors) in the responses will decrease with high quality leading to a more aggregated cluster for the same attribute profile of examinees. Thus, it is not difficult to distinguish the differences between clusters.

With the number of attribute increasing, the ARI and ω values decrease for all methods. We know that the total number of attribute profiles in CDA is exponential in the number of attributes, i.e. 2^K which is also the magnitude of clusters to be identified. Obviously, the difficulty of accurately identifying attribute profiles from a large space is considerable. Besides, as test length increases, the classification abilities of all methods tend to improve. This results are consistent with previous studies. We chose short test length in simulation studies because, a) if giving students an "embedded assessment" at the end of an instruction period, we must prefer short tests to save lecture time (Wang, 2013). In addition, teachers also want to get the attribute profiles of students quickly with short test. b) some diagnostic tests that are commonly used in CDA do not have too many items, especially when the number of attributes is small. Based on our simulations, the SCA can yield considerable classification accuracy when test length is 20.

Simulation results presented here showed that the true CDM is always the best one to fit data. However, the underlying processes among attributes are various in real data actually, and it is hard to define the exact relationship between them. So, the simplicity of cluster analysis is an attractive selection without regard to specify the underlying processes in advance. As mentioned in section "Spectral Clustering for Cognitive Diagnosis," the SCA could simply implemented via the R package called 'Kernlab', which means it is very easy to master by teachers and practitioners. In this study, we investigated the performance of SCA under four specifical processes (saturated, additive, conjunctive and disjunctive) and compared it with other approaches. Overall, the SCA performed comparably to fitted CDMs, and it was basically superior to K-means method. Particularly, the ω values from SCA were highest when the true model was A-CDM (excluded the true model). The strength of cluster analysis was the application in small sample size, so we mainly focused on this point in this study. When the sample size was small, the effectiveness of SCA varied depending on the mechanism of attributes according to simulation results. So, integrating the role of generating CDMs and sample size, our usage recommendation is that the SCA is suitable for analyzing data in regard to saturated and additive underlying processes while it has slightly worse efficiency in conjunctive and disjunctive scenarios. With the sample size increased, it should be pointed out that the classification accuracy became better for all these approaches and the differences in classification accuracy between clustering analysis and CDMs were shrinked.

The ARIs are generally low for some conditions in this paper. These three setting factors (i.e. item quality, test length, and generating model) in this research are different from Chiu et al. (2009) study. As for reason, we can see that these three factors have significant effects on classification accuracy based on our simulation studies. So, it is not strange that the ARIs are lower than those of conclusions in Chiu et al.'s study. In point of fact, the ARIs are not very low when item quality is high in this study.

Just like K-means, the SCA also suffers from the *labeling problem*, and has difficulty in matching each cluster to a certain attribute profile. This is a major issue of clustering analysis for CDA. However, perhaps one can draw on the teachers' experience to help to determine the students' attribute profiles in the classroom. This issue will be one of our future directions.

Several directions for research can be identified. First, the hierarchy of attributes refers to structurally independent in this research, which means there is no prerequisite in every required attributes. So, the correlation exists between attributes is plausible in this case. However, there are other different structures among the attributes, such as linear, convergent, divergent, and unstructured hierarchical structures (Leighton et al., 2004). The hierarchy generally defines the educational and psychological ordering among the attributes required to solve a test problem, so it is reasonable to infer the attribute structures often exists in the test (Kim, 2001). Although the performance of SCA in one of the structures has been examined in this study, it can not directly generalize to other cases without investigation. So, the effect of different attributes structures need further studies.

Second, the fully connected graph, Gaussian Kernel (Eq. 5), was used to construct similarity matrix **S** in this study. However, there are different similarity graphs in the SCA, such as the *epsilon*-neighborhood graph and *k*-nearest neighbor graphs. Besides, two major methods, the unnormalized and the normalized spectral clustering, can be used to calculate Laplacian matrix. The current paper focused only on the normalized case. In the future, other similarity graphs and unnormalized spectral clustering method should be considered in the SCA to investigate the classification ability for the CDA.

Third, as an initial research to propose the SCA into CDA, the current study only investigated the SCA's performance for the dichotomous item responses. However, recent study proposed a general polytomous cognitive diagnosis model for a special type of graded responses to deal with non-dichotomous item responses (Ma and de la Torre, 2016). So, it is necessary to develop the clustering analysis to cope with the cognitive diagnostic test with both dichotomous and polytomous items. Thus, it may be reasonable to measure the similarity by methods based on rank correlation, such as in Chen et al. (2017). It is interesting to investigate how well the SCA performs for the graded responses.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the first corresponding author.

AUTHOR CONTRIBUTIONS

LG proposed the idea of the manuscript, wrote and revised the manuscript. JY wrote the simulation study code and revised the manuscript. NS organized and proofread the manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (31900793), the Research Program Funds of the Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University (2019-06-023-BZPK01 and 2018-06-002-BZPK01), and the Philosophy Society of the Ministry of Education (11jhq001).

REFERENCES

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Trans. Autom. Control* 19, 716–723.

Cai, D., He, X., and Han, J. (2005). Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.* 17, 1624–1637.

Chen, J., and de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Appl. Psychol. Meas.* 37, 419–437.

Chen, J., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Meas.* 50, 123–140. doi: 10.1177/0146621617707510

Chen, Y., Li, X., Liu, J., Xu, G., and Ying, Z. (2017). Exploratory item classification via spectral graph clustering. *Appl. Psychol. Meas.* 41, 579–599. doi: 10.1177/0146621617692977

Chiu, C., and Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classif.* 30, 225–250.

Chiu, C., Douglas, J., and Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika* 74, 633–665. doi: 10.1111/bmsp.12044

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199.

de la Torre, J., Hong, Y., and Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *J. Educ. Meas.* 47, 227–249.

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*. Unpublished doctoral dissertation, University of Illinois, Champaign, IL.

Henson, R., Templin, J., and Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *J. Educ. Meas.* 44, 361–376.

Henson, R., Templin, J., and Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210.

Hubert, L. J., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218.

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab - An S4 package for kernel methods in R. *J. Stat. Softw.* 11, 1–20.

Kim, S. (2001). *Towards a Statistical Foundation in Combining Structures of Decomposable Graphical Models*. Research Report No. 01-2. Yusong gu: Korea Advanced Institute of Science and Technology, Division of Applied Mathematics.

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

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

Ma, W., Iaconangelo, C., and de la Torre, J. (2016). Model similarity, model selection, and attribute. *Appl. Psychol. Meas.* 40, 200–217. doi: 10.1177/0146621615621717

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

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

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). "On spectral clustering: analysis and an algorithm," in *Proceedings of the 14th International Conference*

on Neural Information Processing Systems: Natural and Synthetic, eds T. Dietterich, S. Becker, and Z. Ghahramani (Cambridge, MA: MIT Press), 849–856.

Schwarzer, G. (1976). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.* 22, 888–905.

Steinley, D. (2003). Local optima in k-means clustering: what you don't know may hurt you. *Psychol. Methods* 8, 294–304. doi: 10.1037/1082-989X.8.3.294

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *J. Educ. Stat.* 12, 55–73.

Tatsuoka, K. K. (1995). "Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach," in *Cognitively Diagnostic Assessment*, eds P. Nichols, S. F. Chipman, and R. L. Brennan (Hillsdale, NJ: Erlbaum), 327–359.

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

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

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educ. Psychol. Meas.* 73, 1017–1035. doi: 10.1177/0146621616665196

Ward, J. H. (1963). Hierarchical Grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244.

Zare, H., Shooshtari, P., Gupta, A., and Brinkman, R. R. (2010). Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* 11:403. doi: 10.1186/1471-2105-11-403

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.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 03 June 2020
doi: 10.3389/fpsyg.2020.01051

The Impact of Sample Attrition on Longitudinal Learning Diagnosis: A Prolog

*Yanfang Pan and Peida Zhan**

Department of Psychology, Zhejiang Normal University, Jinhua, China

Missing data are hard to avoid, or even inevitable, in longitudinal learning diagnosis and other longitudinal studies. Sample attrition is one of the most common missing patterns in practice, which refers to students dropping out before the end of the study and not returning. This brief research aims to examine the impact of a common type of sample attrition, namely, individual-level random attrition, on longitudinal learning diagnosis through a simulation study. The results indicate that (1) the recovery of all model parameters decreases with the increase of attrition rate; (2) comparatively speaking, the attrition rate has the greatest influence on diagnostic accuracy, and the least influence on general ability; and (3) a sufficient number of items is one of the necessary conditions to counteract the negative impact of sample attrition.

Keywords: cognitive diagnosis, longitudinal learning diagnosis, missing data, sample attrition, Long-DINA model

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
James Soland,
University of Virginia, United States
Zhehan Jiang,
University of Alabama, United States
Yong Luo,
Educational Testing Service, United States

***Correspondence:**
Peida Zhan
pdzhan@gmail.com

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

Received: *25 November 2019*
Accepted: *27 April 2020*
Published: *03 June 2020*

Citation:
Pan Y and Zhan P (2020) The Impact of Sample Attrition on Longitudinal Learning Diagnosis: A Prolog. Front. Psychol. 11:1051. doi: 10.3389/fpsyg.2020.01051

INTRODUCTION

During the last few decades, to promote student learning, learning diagnosis (Zhan, 2020) or cognitive diagnosis (Leighton and Gierl, 2007) through objectively quantifying the learning status of fine-grained attributes (e.g., knowledge, skills, and cognitive processes) and providing diagnostic feedback has been increasingly valued. Longitudinal learning diagnosis identifies students' strengths and weaknesses of various attributes throughout a period of time, which also can be seen as an application of learning diagnosis through longitudinal assessments. Longitudinal learning diagnosis not only can be used to diagnose and track students' growth over time but also can be used to evaluate the effectiveness of diagnostic feedback and corresponding remedial teaching (Tang and Zhan, under review; Wang et al., 2020).

In recent years, to provide theoretical support for longitudinal learning diagnosis, several longitudinal learning diagnosis models (LDMs) have been proposed, which can be divided into two primary categories: the higher-order latent structure-based models (e.g., Huang, 2017; Lee, 2017; Zhan et al., 2019a) and the latent transition analysis-based models (e.g., Li et al., 2016; Kaya and Leite, 2017; Wang et al., 2018; Madison and Bradshaw, 2018). The former estimates the changes in higher-order latent ability over time, and from this, it infers the changes in the lower-order latent attributes. The latter estimates the transition probabilities from one latent class or attribute to another or to the same latent class or attribute. The diagnostic results of these two model types have a high consistency (Lee, 2017). Although the utility of these models has been evaluated by some simulation studies and a few applications, the harm of ubiquitous missing data in longitudinal designs has not yet been considered and studied.

In practice, missing data are hard to avoid, or even inevitable, in longitudinal learning diagnosis and other longitudinal studies. In this current study, we focused on a type of missing data that is common to longitudinal studies, namely, attrition (Little and Rubin, 2020, p. 10). Attrition refers to students dropping out prior to the end of the study and do not return. For instance, in school-level longitudinal learning diagnosis projects, some students may individually drop out before the end of the study because they move to other schools that are inaccessible to the researchers; all students in the class may even drop out altogether because of some unforeseen classroom instructional reasons (see the empirical example in Zhan et al., 2019a).

A higher percentage of attrition at each point in time means the remaining data at subsequent time points provide less diagnostic information, which may also challenge the robustness of measurement models. Some studies have previously employed a complete case analysis that deletes any students who dropped out (e.g., Zhan et al., 2019a). However, this is unfair to those students who were deleted in analysis, because they did not receive any diagnostic feedback. Secondly, it may produce biased results when students with complete data are systematically different from those with missing data. Longitudinal studies are particularly susceptible to such bias, as missing data accumulate over time due to attrition. Therefore, it is necessary to explore the impact of missing data caused by attrition on longitudinal learning diagnosis. This not only helps practitioners better understand the performance of existing longitudinal LDMs in specific test situations with missing data but also provides a reference to psychometricians for future research on the necessity of imputation methods for missing data in longitudinal learning diagnosis. However, as aforementioned, to our knowledge, the harm of ubiquitous missing data in longitudinal designs has not yet been considered and studied in the field of learning diagnosis.

As a prolog, this brief research report aims to explore the impact of various proportions of a common type of attrition (i.e., individual-level random attrition) on longitudinal learning diagnosis through a simulation study. For simplicity and without loss of generality, a simple version of the longitudinal higher-order deterministic-inputs, the noisy "and" gate (sLong-DINA) model (Zhan et al., 2019a) is used in this study. The rest of the paper starts with a brief review of the sLong-DINA model and different types of sample attrition. Subsequently, a simulation study was conducted to mimic the operational scenarios of attrition that may be considered by the sLong-DINA model. Finally, the authors summarize the findings and discuss potential directions for future research.

BACKGROUND

sLong-DINA Model

The sLong-DINA model is one of the representative models of the higher-order latent structural model-based longitudinal LDMs. Compared with the complete version, the special dimensions used to account for local item dependence among anchor items at different time points (see Paek et al., 2014) are ignored in the sLong-DINA model to reduce model complexity and computational burden.

Let y_{nit} be the response of person n (n = 1,..., N) to item i (i = 1,..., I) at time point t (t = 1,..., T). The sLong-DINA model can be expressed as follows:

First order:

$$\text{logit}(P(y_{nit} = 1|\alpha_{nt}, \gamma_{nm}, \lambda_{0it}, \lambda_{1it})) = \lambda_{0it} + \lambda_{1it} \prod_{k=1}^{K} \alpha_{nkt}^{q_{ikt}} \quad (1)$$

Second order:

$$\text{logit}(P(\alpha_{nkt} = 1|\theta_{nt}, \xi_k, \beta_k)) = \xi_k \theta_{nt} - \beta_k \quad (2)$$

Third order:

$$\theta_n = (\theta_{n1}, \ldots, \theta_{nT})' \sim MVN_T(\mu, \Sigma) \quad (3)$$

where α_{nt} = (α_{n1t},..., α_{nKt})′ denotes person n's attribute profile at time point t, $\alpha_{nkt} \in \{0, 1\}$, and α_{nkt} = 1 if person n masters attribute k (k = 1,..., K) at time point t and α_{nkt} = 0 if not; λ_{0it} and λ_{1it} are the intercept and interaction parameter for item i at time point t, respectively; $q_{ikt} \in \{0, 1\}$ is the element in an I-by-K Q_t-matrix at time point t, where q_{ikt} = 1 if item i requires attribute k at time point t and q_{ikt} = 0 if not; θ_{nt} is person n's general ability at time point t; ξ_k and β_k are the slope and difficulty parameters of attribute k at all time points, respectively, because the same latent structure is assumed to be measured at different time points; μ = (μ_1,..., μ_T)′ is the mean vector and Σ is a variance–covariance matrix:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ \vdots & \ddots & \\ \sigma_{1T} & \cdots & \sigma_T^2 \end{bmatrix}$$

where σ_{1T} is the covariance of the first and Tth general abilities. As a starting and reference point for subsequent time points, θ_{n1} is constrained to follow a standard normal distribution.

There are two reasons why we did not consider using a general or saturated model (e.g., Huang, 2017; Madison and Bradshaw, 2018). First, general models always need a large sample size to obtain a robust parameter estimate (Jiang and Ma, 2018; Ravand and Robitzsch, 2018). Thus, it is difficult for small-scale educational projects (e.g., school- and classroom-level assessments) to meet this requirement. Second, the parameters in general models are often hard to interpret in practice. Adequate parameter constraints are essential for obtaining interpretable and meaningful insights from the model, which are particularly important in educational and psychological applications to fulfill the need for accountability.

Sample Attrition

Sample attrition is one of the common sources of missing data in longitudinal studies (Little and Rubin, 2020) and refers to when students drop out prior to the end of the study and do not return. In practice, there are four typical types of sample attrition: individual-level random attrition, class-level random attrition, individual-level nonrandom attrition,

and class-level nonrandom attrition. More specifically, (a) the individual-level random attrition reflects the common scenario in which sample size decreased monotonically over time for individual reasons, such as illness, transferring to another school, and reluctance to participate; (b) the class-level random attrition can be seen as an extreme case of individual-level random attrition, where the whole class students drop out for some unpredictable reasons; for example, the testing time may conflict with other course time due to adjusting the curriculum schedule; (c) individual-level nonrandom attrition typically occurs when an individual has achieved a predetermined learning goal, such as mastering the target attributes; thus, some students may feel that there is no need to waste time on follow-up remediation and then quit the follow-up section(s); and (d) class-level nonrandom attrition may occur when the teacher finds that the vast majority of students (e.g., 80%) in the class have mastered the target attributes, then she/he may decide to quit the follow-up section(s) to ensure normal teaching progress. More discussions about sample attrition can be found in Goodman and Blum (1996) and Little and Rubin (2020).

This brief research aims to explore the impact of the individual-level random attrition, which is the simplest type of sample attrition, on longitudinal learning diagnosis. As this is a prolog or preliminary study, we hope that more researchers could continue to study the effects of different types of sample attrition and different types of missing data on longitudinal learning diagnosis (cf., Muthén et al., 2011; Zheng, 2017).

SIMULATION STUDY

Design and Data Generation

In the simulation study, three factors were manipulated. First, the sample size at the starting time point was varied to be either N = 200 or 400 students. According to the national situation in the authors' country, sample sizes of 200 and 400 translate to approximately 5 and 10 classes with 40 students in each. In real school-level longitudinal learning diagnosis projects, more classes and more students per class are rare. Second, the random attrition rate at each time point (from time point 2) equaled $M1$ = 0% (baseline), 5, 10, 20, 40, and 60% (all the decimal points that might occur in proportional sampling are deleted). The third manipulated variable was test length at each time point at two levels of relatively short (I_t = 15) and relatively long (I_t = 30).

According to the authors' practical experience in longitudinal learning diagnosis (e.g., Tang and Zhan, under review), two or three test times (i.e., one or two sessions of diagnostic feedback and/or remedial teaching) are sufficient for almost all students to master the target fine-grained attributes. Thus, three time points were considered (T = 3) in this brief study. In addition, four attributes (K = 4) were measured. The first four items for I_t = 15 and the first eight items for I_t = 30, respectively, were used as anchor items. The simulated Q-matrices were presented in **Figure 1**. In practice, it is common to use high-quality items as anchor items, and thus the anchor item parameters were fixed as $\lambda_{0it} = -2.197$ and $\lambda_{1it} = 4.394$. In such a case, the aberrant response (i.e., guessing and slipping) probabilities are approximately equal to 0.1. In addition, the results of Zhan et al. (2019b) indicate that assuming guessing and slipping parameters to follow a negative correlation is more realistic. Thus, non-anchor item parameters were generated from a bivariate normal distribution with a negative correlation coefficient as follows:

$$\begin{pmatrix}\lambda_{0it}\\ \lambda_{1it}\end{pmatrix} \sim MVN_2\left(\begin{pmatrix}-2.197\\ 4.394\end{pmatrix}, \begin{pmatrix}1.0 & -0.6\\ -0.6 & 1.0\end{pmatrix}\right)$$

This setting leads the guessing and slipping probabilities for all items to follow a positively skewed distribution (mean $\approx$ 0.1, minimum $\approx$ 0.01, and maximum $\approx$ 0.6). Attribute slope parameters were fixed at ξ_k = 1.5 for all attributes, and attribute difficulty parameters were fixed at $\beta = (-1, -0.5, 0.5, 1)$. For the general abilities on different time points, the correlations among them were set as 0.9. Between two consecutive time points, the overall mean growths were set at 1, and the overall scale changes were set at $\sqrt{1.25}$.

Furthermore, the response data without attrition (i.e., $M1$ = 0%) were generated from the sLong-DINA model based on the above-generated parameters. For the response data with attrition, a different proportion of students were randomly sampled as attrition from time point 2. Then, these selected students' responses were modified as missing (i.e., NA), and students who had been drawn out did not appear in the subsequent section(s). In other words, some students were dropped out from time point 2, while some others were dropped out until time point 3. The data were generated by using R software, and the data generation code is available from the authors.

Analysis

In this brief study, the parameters of the sLong-DINA model are estimated using the Bayesian Markov chain Monte Carlo method via Just Another Gibbs Sampler (JAGS) software. The prior distribution of the model parameters and the corresponding JAGS code are displayed in **Supplementary Table S1** in the online supporting materials. More details about how to use the JAGS code for Bayesian CDM estimation can be found in a tutorial by Zhan et al. (2019c).

Thirty replications were implemented in each condition. For each replication, two Markov chains with random starting points were used and 15,000 iterations were run for each chain. The first 10,000 iterations in each chain were discarded as burn-in. Finally, the remaining 10,000 iterations were used for the model parameter inferences. The potential scale reduction factor (PSRF; Brooks and Gelman, 1998) was computed to assess the convergence of each parameter. Values of PSRF less than 1.1 or 1.2 indicate convergence. The results indicated that PSRF was generally less than 1.1, suggesting acceptable convergence for the setting specified.

To evaluate parameter recovery, the bias and the root mean square error (RMSE) were computed as $\text{bias}(\hat{\nu}) = \sum_{r=1}^{R} \frac{\hat{\nu}_r - \nu}{R}$ and $\text{RMSE}(\hat{\nu}) = \sqrt{\sum_{r=1}^{R} \frac{(\hat{\nu}_r - \nu)^2}{R}}$, where $\hat{\nu}$ and ν are the estimated and true values of the model parameters, respectively; R is the total number of replications. In addition, the correlation

between the true values and estimated values (Cor) for some parameters (e.g., general abilities) were computed to evaluate the recovery. For attribute recovery, the attribute and pattern correct classification rate (i.e., ACCR and PCCR) were computed to evaluate the classification accuracy of individual attributes and profiles: $\text{ACCR} = \sum_{r=1}^{R}\sum_{n=1}^{N} {}^{\text{I}(\hat{\alpha}_{nkr}=\alpha_{nkr})}/_{NR}$ and $\text{PCCR} = \sum_{r=1}^{R}\sum_{n=1}^{N} {}^{\text{I}(\hat{\alpha}_{nr}=\alpha_{nr})}/_{NR}$, where $\text{I}(\cdot)$ is an indicator function. In reference to Zhan et al. (2019a), two kinds of PCCR were considered in this brief research, namely, the PCCR and the Longitudinal PCCR. The former focuses on whether K attributes can be correctly recovered at a given time point, while the latter focuses on whether all TK attributes can be correctly recovered (e.g., if $T = 3$, the pattern contains 12 attributes).

RESULTS

Figure 2 presents the recovery of item parameters. First, one of the most important results is that, with the increase of the attrition rate, the recovery of item parameters decreases, which manifests as larger bias, higher RMSE, and lower Cor. Second, increasing the number of classes

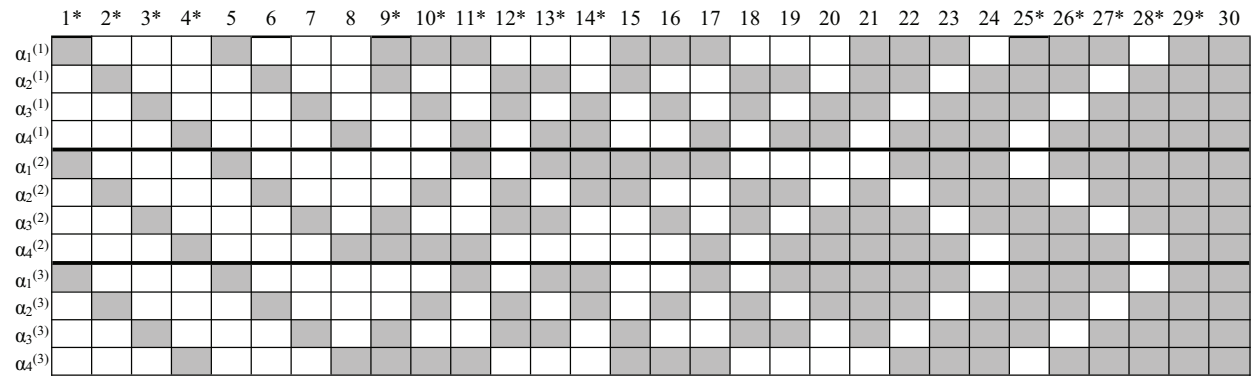


FIGURE 1 | Simulated K-by-I Q′-matrices in simulation study. "*" Denotes items used in the $I = 15$ conditions; gray means "1" and blank means "0"; time point is in parentheses.

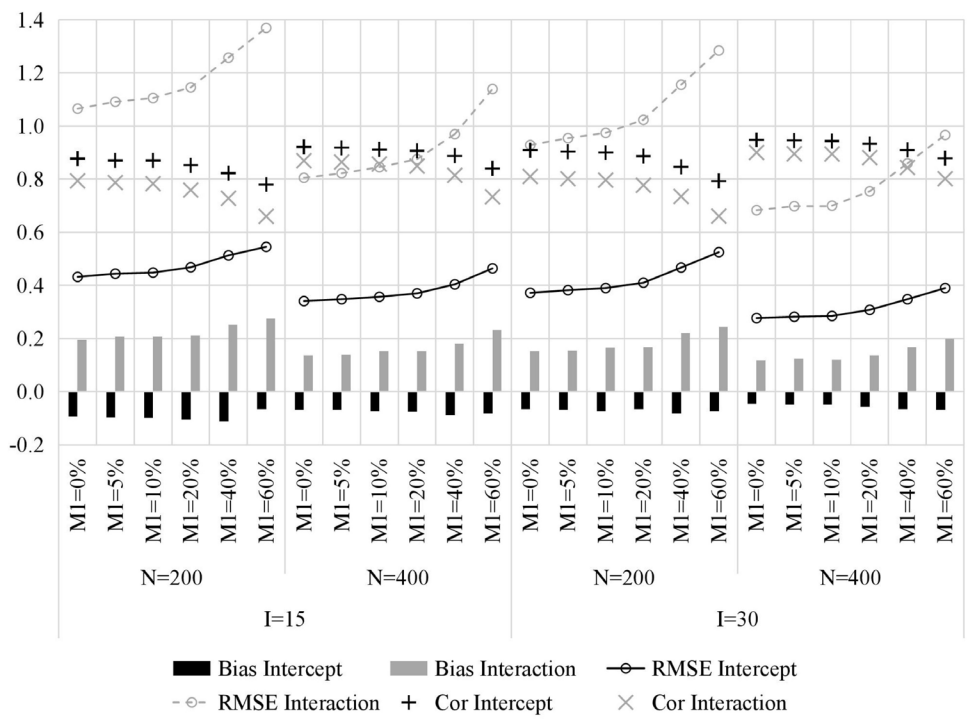


FIGURE 2 | The recovery of item parameters in simulation study. M1, attrition rate; N, sample size; I, test length; Bias, mean bias across all items; RMSE, mean root mean square error across all items; Intercept, item intercept parameter; Interaction, item interaction.

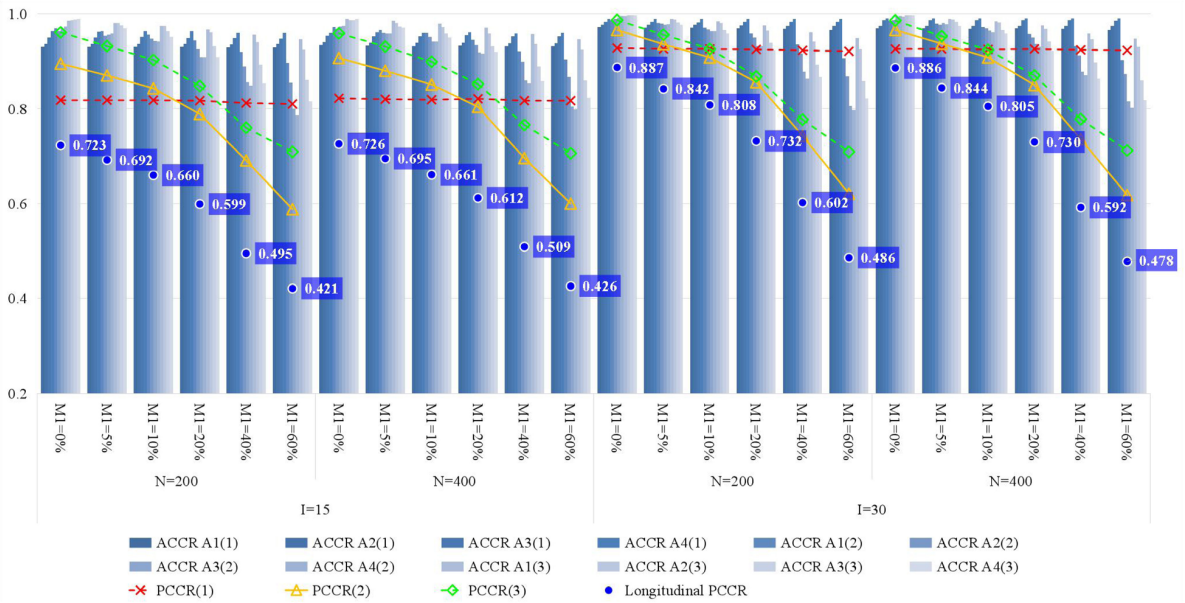


FIGURE 3 | The recovery of attributes in simulation study. M1, attrition rate; N, sample size; I, test length; ACCR, attribute correct classification rate; PCCR, pattern correct classification rate; time point is in parentheses.

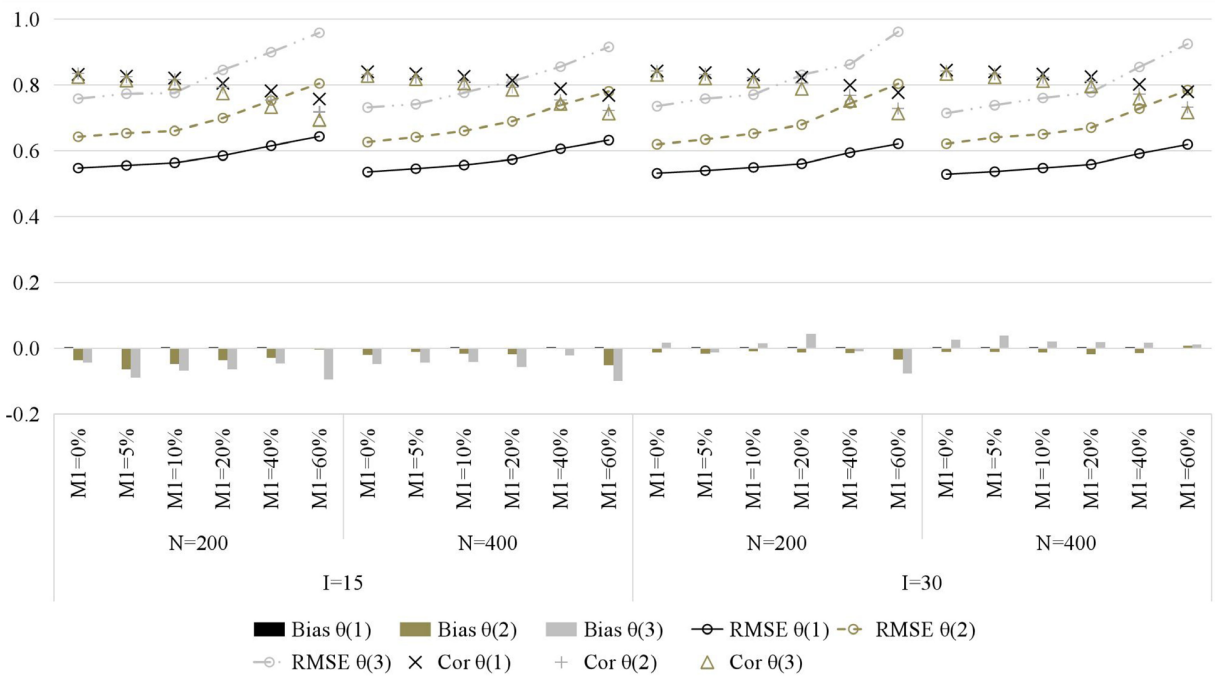


FIGURE 4 | The recovery of general ability parameters in simulation study. M1, attrition rate; N, sample size; I, test length; Bias, mean bias across all persons; RMSE, mean root mean square error across all persons; Cor, correlation between generated and estimated values; time point is in parentheses.

(i.e., sample size) and test length yields better recovery of item parameters, and the former is more influential. Third, intercept parameters were generally estimated more accurately than interaction parameters, mainly because the number of individuals who mastered all required attributes is typically less than the number of individuals who do not master all required attributes.

Figure 3 presents the recovery of attributes. With the increase of the attrition rate, the classification accuracy quickly decreases, particularly for the Longitudinal PCCR. Since there is no attrition

at time point 1, the PCCR of time point 1 is primarily affected by test length. Then, for the PCCR of time points 2 and 3, their downward trend is almost consistent with that of the Longitudinal PCCR. Therefore, if the PCCR is maintained above 80% and the Longitudinal PCCR is maintained above 60%, an attrition rate of less than or equal to around 20% and around 40% is acceptable for short tests and long tests, respectively. In addition, there is a significant result that deserves attention, which is that the classification accuracy of time point 3 is better than that of time point 2; this was also found in the study of Zhan et al. (2019a). Although we currently do not know how to interpret this phenomenon, it is at least not negative for longitudinal learning diagnosis. Furthermore, increasing the number of classes and test length yields higher classification accuracy, but the former has a limited effect.

Figure 4 presents the recovery of general ability parameters. Similarly, with the increase of the attrition rate, the recovery of general ability parameters gradually decreases, which is manifested as higher RMSE and lower Cor (bias is less affected). Compared with item parameters and attributes, the attrition rate has less impact on general ability parameters.

CONCLUSION AND DISCUSSION

This brief research examined the impact of individual-level random attrition on longitudinal learning diagnosis. The results indicate that (1) the recovery of all model parameters decreases with the increase of attrition rate; (2) comparatively speaking, the attrition rate has the greatest influence on the diagnostic accuracy, and the least influence on general ability; and (3) a sufficient number of items is one of the necessary conditions to withstand the negative impact of sample attrition. For relatively short tests (e.g., 15 items), a random attrition rate of 20% or less is necessary to achieve an acceptable longitudinal diagnostic accuracy (i.e., longitudinal PCCR > 0.6); conversely, for relatively long tests (e.g., 30 items), a random attrition rate of 40% or less is necessary.

In summation, the results of this brief study have demonstrated that sample attrition or missing data have a significant impact on diagnostic accuracy of longitudinal learning diagnosis. Therefore, the topics of sample attrition and missing data are worth studying in longitudinal learning diagnosis. As a prolog to future research, the current study only considered some simple cases and left many issues for further discussion. First, this brief research only explores the impact of sample attrition on the sLong-DINA model. Whether the conclusions apply to other longitudinal LDMs is still worth further study in the future. Second, in a different manner from attrition that was focused on this brief research (i.e., monotone missing pattern), a student can be missing at one follow-up time and then measured again at one of the next, resulting in a non-monotone missing pattern. Students' returning indicates that more information is contained in the data. Thus, it can be inferred that the negative influence of the non-monotone missing pattern on longitudinal learning diagnosis is less than that of attrition. However, the specific degree of its impact remains to be determined. Third, the number of simulation conditions in this brief study is still limited. More independent variables (e.g., the number of attributes and the attribute hierarchies) and more complex test situations (e.g., more time points) can be considered in future studies to provide more reference information for practitioners.

Fourth, in practice, students are nested in classes, and classes are further nested in schools. Such a multilevel data structure is not considered in the current study. By utilizing multilevel LDMs (e.g., Huang, 2017; Wang and Qiu, 2019) in future research, the multilevel data structure can be considered and the impact of class-level attrition can also be studied. Fifth, similar to the Andersen's longitudinal Rasch model (Andersen, 1985), for general ability, the sLong-DINA model focuses on the estimates at different time points rather than a specific growth trend (i.e., linear or non-linear). If practitioners focus on the latter, the growth curve LDMs (Huang, 2017; Lee, 2017) can be used. Sixth, only the individual-level random attrition was considered in this brief study, while the impact of other three types of attrition (i.e., class-level random attrition, individual non-random attrition, and class-level non-random attrition) on longitudinal learning diagnosis still remains to be further studied.

Seventh, in further studies, it would be much more interesting to explore the impact of different missing mechanisms upon the parameter recovery of longitudinal LDMs, instead of just generating data based on the missing completely at random scenario (i.e., random attrition), such as the missing at random with respect to both observed outcomes and covariates and the missing at random with respect to covariates only (Muthén et al., 2011; Zheng, 2017). Eighth, in longitudinal assessments, for meaningful comparisons, it is necessary to ensure that the same construct is measured across time points. In the presence of item parameter drift, a special case of differential item functioning, the interpretation of scores across time points or change scores would not be valid. Thus, the consequences of ignoring item parameter drift in longitudinal learning diagnosis is worthy of further attention (cf., Meade and Wright, 2012; Lee and Cho, 2017). Ninth, in Bayesian estimation, the prior distribution reflects the beliefs of the data analyst. The posterior distribution of model parameters will be affected by their prior distribution, particularly for a small sample size or a limited number of items. The choice of prior distribution is also worthy of attention (da Silva et al., 2018; Jiang and Carter, 2019). In practice, we recommend that the data analyst selects appropriate prior distributions based on the actual situation rather than copy those given in the **Supplementary Table S1**.

Last but most important, this brief research is only a superficial study of the missing data in longitudinal learning diagnosis. In the broader field of longitudinal studies, methodologists have been studying missing data for decades and have proposed many methods and techniques to address this issue (see, Daniels and Hogan, 2008; Enders, 2010; Young and Johnson, 2015; Little and Rubin, 2020), such as the traditional imputation methods (e.g., arithmetic mean imputation, regression imputation, and similar response pattern imputation), likelihood-based methods, Bayesian iterative simulation methods, and multiple imputation methods. The performance

of these methods in longitudinal learning diagnosis is well worth further study.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

YP contributed to manuscript drafting. PZ contributed to the conception, design, data analysis, and revising the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 31900795) and the MOE (Ministry of Education in China) Project of Humanities and Social Science (Grant No. 19YJC190025).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01051/full#supplementary-material

REFERENCES

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika* 50, 3–16. doi: 10.1007/BF02294143

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

da Silva, M. A., de Oliveira, E. S., von Davier, A. A., and Bazán, J. L. (2018). Estimating the DINA model parameters using the No-U-Turn Sampler. *Biometr. J.* 60, 352–368. doi: 10.1002/bimj.201600225

Daniels, M. J., and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Milton Park: Taylor & Francis Group.

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

Goodman, J. S., and Blum, T. C. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. *J. Manag.* 22, 627–652. doi: 10.1177/014920639602200405

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *J. Educ. Measure.* 54, 440–480. doi: 10.1111/jedm.12156

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

Jiang, Z., and Ma, W. (2018). Integrating differential evolution optimization to cognitive diagnostic model estimation. *Front. Psychol.* 9:2142. doi: 10.3389/fpsyg.2018.02142

Kaya, Y., and Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. *Educ. Psychol. Measur.* 77, 369–388. doi: 10.1177/0013164416659314

Lee, S. Y. (2017). *Growth Curve Cognitive Diagnosis Models for Longitudinal Assessment*. Unpublished doctoral dissertation. Berkeley: University of California.

Lee, W., and Cho, S.-J. (2017). The consequences of ignoring item parameter drift in longitudinal item response models. *Appl. Measur. Educ.* 30, 129–146. doi: 10.1080/08957347.2017.1283317

Leighton, J. P., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press, doi: 10.1017/CBO9780511611186

Li, F., Cohen, A., Bottge, B, and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

Little, R. J., and Rubin, D. B. (2020). *Statistical Analysis with Missing Data*, 3rd Edn. Hoboken, NJ: John Wiley & Sons, Inc.

Madison, M. J., and Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika* 83, 963–990. doi: 10.1007/s11336-018-9638-5

Meade, A. W., and Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *J. Appl. Psychol.* 97, 1016–1031. doi: 10.1037/a0027934

Muthén, B., Asparouhov, T., Hunter, A. M., and Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: alternative analyses of the STAR* D antidepressant trial. *Psychol. Methods* 16, 17–33. doi: 10.1037/a0022634

Paek, I., Park, H.-J., Cai, L., and Chi, E. (2014). A comparison of three IRT approaches to examinee ability change modeling in a single-group anchor test design. *Educ. Psychol. Measur.* 74, 659–676. doi: 10.1177/0013164413507062

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

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

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

Wang, W.-C., and Qiu, X.-L. (2019). Multilevel modeling of cognitive diagnostic assessment: the multilevel DINA example. *Appl. Psychol. Meas.* 43, 34–50. doi: 10.1177/0146621618765713

Young, R., and Johnson, D. R. (2015). Handling missing values in longitudinal panel data with multiple imputation. *J. Marriage Fam.* 77, 277–294. doi: 10.1111/jomf.12144

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

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

Zhan, P., Jiao, H., Liao, M., and Bian, Y. (2019b). Bayesian DINA modeling incorporating within-item characteristics dependency. *Appl. Psychol. Measur.* 43, 143–158. doi: 10.1177/0146621618781594

Zhan, P., Jiao, H., Man, K., and Wang, L. (2019c). Using JAGS for Bayesian cognitive diagnosis modeling: a tutorial. *J. Educ. Behav. Stat.* 44, 473–503. doi: 10.1093/arclin/acw017

Zheng, X. (2017). *Latent Growth Curve Analysis with Item Response Data: Model Specification, Estimation, and Panel Attrition*. Unpublished doctoral dissertation, University of Maryland, Maryland.

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.

frontiers in Psychology

OPINION
published: 03 July 2020
doi: 10.3389/fpsyg.2020.01185

Longitudinal Learning Diagnosis: Minireview and Future Research Directions

*Peida Zhan**

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

Keywords: longitudinal cognitive diagnosis, learning diagnosis, cognitive diagnostic assessment, assessment for learning, latent transition analysis

INTRODUCTION

The basic premise of "teaching students according to their aptitude" is to have a relatively objective and accurate understanding of the students' current learning statuses (e.g., knowledge mastery level, learning motivation, learning attitude, and learning mode) and the developments/changes they undergo over time (e.g., did the students' knowledge mastery level improve, are the students' learning motivations enhanced). Measuring and improving individual development are topics that are actively tackled in psychological, educational, and behavioral studies.

In the past decades, learning diagnosis, which objectively quantifies students' current learning status and provides diagnostic feedback, has drawn increasing interest (Zhan, 2020). When focusing on fine-grained attributes (e.g., knowledge, skills, and cognitive processes), learning diagnosis can also be regarded as an application of cognitive diagnosis (Leighton and Gierl, 2007) in learning assessment. Although learning diagnosis aims to promote student learning based on diagnostic feedback and the corresponding remedial teaching (intervention), currently, only a few studies have focused on and evaluated the effectiveness of such feedback or remedial teaching (c.f., Tang and Zhan, submitted; Wu, 2019; Wang L. et al., in press; Wang S. et al., 2020). One of the main reasons is that cross-sectional design, which cannot measure individual growth in learning, is adopted by most current learning diagnoses. This issue may also be reflected in current learning diagnosis models (LDMs) or alternatively cognitive diagnosis models (for review, see Rupp et al., 2010; von Davier and Lee, 2019), which are the main tools for data analysis in learning diagnosis. Although various LDMs have been proposed and suggested by previous research, most of them are only applicable to cross-sectional data analysis, such as the deterministic inputs, noisy "and" gate (DINA) model (Junker and Sijtsma, 2001), the deterministic inputs, noisy "or" gate (DINO) model (Templin and Henson, 2006), the log-linear cognitive diagnosis model (LCDM) (Henson et al., 2009), and the generalized DINA (GDINA) model (de la Torre, 2011).

By contrast, longitudinal learning diagnosis evaluates students' knowledge and skills and identifies their strengths and weaknesses over a period of time. The data collected from longitudinal learning diagnosis provide researchers with the opportunities to develop models for learning tracking, which can be used to track individual growth over time as well as to evaluate the effectiveness of feedback. Compared to cross-sectional learning diagnosis, longitudinal learning diagnosis is more helpful when aiming to promote student learning.

Currently, longitudinal learning diagnosis is a new research direction that mainly stays in the model development stage and lacks practical applications and related topic research (e.g., missing data, measure invariance, and linking methods). Moreover, although some longitudinal LDMs have been proposed, these models still have some limitations that need to be further studied. Thus, for the rest of this opinion article, I will first make a minireview of current longitudinal LDMs and then I will elaborate on several future research directions that I believe are worth studying. With this opinion article, I hope to elicit more research attention toward longitudinal learning diagnosis.

OPEN ACCESS

Edited by:
Sergio Machado,
Salgado de Oliveira University, Brazil

Reviewed by:
Shiyu Wang,
University System of Georgia, United States
Claudio Imperatori,
Università Europea di Roma, Italy

***Correspondence:**
Peida Zhan
pdzhan@gmail.com

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

Received: *25 November 2019*
Accepted: *07 May 2020*
Published: *03 July 2020*

Citation:
Zhan P (2020) Longitudinal Learning Diagnosis: Minireview and Future Research Directions. Front. Psychol. 11:1185. doi: 10.3389/fpsyg.2020.01185

MINIREVIEW

To provide theoretical support for data analysis in longitudinal learning diagnosis, longitudinal LDMs are needed. However, the latent variables (namely, attributes) in LDMs are categorical (typically, binary). Therefore, the methods for modeling growth for continuous latent variables (e.g., longitudinal item response theory models) cannot be directly extended to capture growth in the mastery of attributes. For example, the change in the mastery of attributes cannot be directly modeled by the variance-covariance methods when assuming that multiple continuous latent variables follow a multivariate normal distribution (e.g., von Davier et al., 2011).

To this end, in recent years, several longitudinal LDMs have been proposed. They are summarized in **Table 1**. Current longitudinal LDMs can mainly be divided into two categories: the latent transition analysis (Collins and Wugalter, 1992)-based models (e.g., Li et al., 2016; Kaya and Leite, 2017; Chen et al., 2018a; Madison and Bradshaw, 2018; Wang et al., 2018a) and the higher-order latent structural (de la Torre and Douglas, 2004)-based models (e.g., Huang, 2017; Lee, 2017; Zhan et al., 2019a). The diagnostic results of these two model types have high consistency (Lee, 2017).

The latent transition analysis-based methods estimate the transition probabilities from one latent class/attribute to another or the same latent class/attribute. Two main differences exist in these models. First, different measurement models were used. Reduced LDMs, e.g., the DINA model and the DINO model, were used by Li et al. (2016), Kaya and Leite (2017), Chen et al. (2018a), and Wang et al. (2018a), but a generalized LDM, i.e., the LCDM, was used by Madison and Bradshaw (2018). Second, the attribute-level transition probability matrix (i.e., attributes are transited independently from one other) was used by Li et al. (2016) and Wang et al. (2018a), but the attribute pattern-level transition probability matrix was used by Kaya and Leite (2017), Chen et al. (2018a), and Madison and Bradshaw (2018). In addition, different from Li et al. (2016), Kaya and Leite (2017), Chen et al. (2018a), and Madison and Bradshaw (2018), who directly estimated the transition probabilities, Wang et al. (2018a) used a set of covariates, such as a time-invariant general learning ability and intervention indicators, to model the transition probabilities. The effectiveness of different learning interventions was further considered by Zhang and Chang (2019). Additionally, to reduce modeling complexity, Chen et al. (2018a) and Wang et al. (2018a) assumed learning trajectories to be non-decreasing (i.e., respondents did not forget). However, this non-decreasing assumption may only be suitable for short-time interval assessments. Furthermore, by incorporating response times into LDMs (Wang et al., 2018a, 2019; Zhan et al., 2018a; Zhang and Wang, 2018) used response times to assist in measuring students' growth in attribute mastery.

Meanwhile, the higher-order latent structural model-based methods estimate the changes in a higher-order latent ability over time to further infer the changes of lower-order latent attributes. One of the representative models is the longitudinal higher-order DINA (Long-DINA) model (Zhan et al., 2019a), which is a multidimensional extension of the higher-order DINA model (de la Torre and Douglas, 2004). However, multidimensionality does not refer to different general abilities, but rather, the same general ability measured at different time points. As noted by Zhan et al. (2019a), the latent growth curve model instead of the variance-covariance method can also be employed in the third order. Lee (2017) proposed a growth curve DINA model, which can be seen as an alternative of the long-DINA model that incorporates the latent growth curve model but ignores the local item dependence among anchor or repeat items. Furthermore, Huang (2017) proposed a multilevel GDINA model for assessing growth, which can be seen as an extension of Lee's (2017) model in both measurement model part (i.e., from DINA model to GDINA model) and latent structural model part (i.e., from one-level growth curve model to multilevel growth curve model).

TABLE 1 | Summary of longitudinal cognitive diagnosis models.

Basic method	References	Basic model	Learning tracking
Latent transition analysis (LTA)	Li et al., 2016	DINA	LTA with attribute-level transition probability matrix
	Kaya and Leite, 2017	DINA, DINO	LTA with attribute pattern-level transition probability matrix
	Madison and Bradshaw, 2018	LCDM	LTA with attribute pattern-level transition probability matrix
	Chen et al., 2018a	DINA	LTA with attribute pattern-level transition probability matrix
	Wang et al., 2018a	DINA	LTA with modeled attribute-level transition probabilities
Higher-order latent structural model	Lee, 2017	DINA	Latent growth curve model
	Huang, 2017	GDINA	Multilevel latent growth curve model
	Zhan et al., 2019a	Testlet-DINA, DINA	Variance-covariance method

The article collection ended at April 11th 2020; listing only the first article of the proposed model; DINA, deterministic-inputs, noisy "and" gate model Junker and Sijtsma, 2001; DINO, deterministic-inputs, noisy "or" gate model Templin and Henson, 2006; LCDM, log-linear cognitive diagnosis model Henson et al., 2009; GDINA, generalized deterministic-inputs, noisy "and" gate model de la Torre, 2011; Testlet-DINA, deterministic-inputs, noisy "and" gate model for testlet design (see e.g., Zhan et al., 2019b).

FUTURE RESEARCH DIRECTION

Although the utility for analyzing the longitudinal learning diagnosis data of these longitudinal LDMs has been evaluated by some simulation studies and a few applications, these models are not without limitations, which need to be further studied. Based on current research on longitudinal learning diagnosis, I believe that the following are directions that are worthy of further study.

(1) A systematic comparison between different longitudinal LDMs, which can provide theoretical suggestions for practitioners in choosing suitable models.
(2) Only binary attributes (e.g., "1" means mastery and "0" means non-mastery) were considered in all current longitudinal LDMs. However, in actual teaching, it is challenging to use binary attributes to describe the growth of students, as they can be classified into only four categories between two adjacent time points, i.e., $0 \rightarrow 0$, $0 \rightarrow 1$, $1 \rightarrow 0$, and $1 \rightarrow 1$. Some small but existing growths are ignored, which in turn may lead students, especially those with low motivation to learn, to conclude that the current diagnostic feedback is ineffective or to abandon remedial action. Thus, further studies can attempt to extend the current models to handle polytomous attributes (Karelitz, 2004) and probabilistic attributes (Zhan et al., 2018b) because they can describe the learning growth in a more refined way than binary attributes.
(3) Only outcome data (or item response accuracy) were considered in most current longitudinal LDMs. Although a few studies have incorporated item response time into current models (e.g., Wang et al., 2018b), future studies may attempt to introduce other types of process data (e.g., number of trial and error and operation process), or even biometric data (e.g., eye-tracking data; Man and Harring, 2019). Utilizing multimodal data can evaluate the growth of students in multiple aspects, which is conducive to a more comprehensive understanding of the development of students (Zhan, 2019).
(4) All current longitudinal LDMs assumed that attributes are structurally independent, in that mastery of one attribute is not a prerequisite to the mastery of another. However, when attribute hierarchy (Leighton et al., 2004) exists, the development trajectory of students is not arbitrary and should be developed in such hierarchical order. Therefore, incorporating the attribute hierarchy into current longitudinal LDMs is worth trying.
(5) A limited number of attributes at each time point were assumed in current longitudinal LDMs. In practice, a large number of attributes may involve more than 10 or 15 attributes at each time point. In such cases, using current longitudinal LDMs with existing parameter estimation algorithms may lead to unrobust parameter estimation. Thus, more powerful or efficient algorithms or special strategies may need to be introduced.
(6) The simultaneity estimation strategy was adopted by almost all current longitudinal LDMs. This involves the reintegration of response data from multiple time points into one large response matrix, which is then analyzed as a whole (Zhan et al., 2019a). However, this strategy requires subjects to wait until all the tests end before an analysis of the results becomes available. Thus, using this strategy cannot provide timely diagnostic feedback to either students or teachers. In light of the foregoing, new estimation strategies for timely diagnostic feedback should be further studied (e.g., Zhan, 2020).
(7) In addition to theoretical and methodological studies, the corresponding applied studies should also be strengthened. For example, a few studies have focused on and evaluated the effectiveness of diagnostic feedback or remedial teaching in promoting learning (cf. Tang and Zhan, submitted). Moreover, effective and systematic intervention methods based on longitudinal diagnostic feedback are also worth studying.
(8) Adaptive learning system involving LDMs is also worthy of further study (e.g., Chen et al., 2018b; Tang et al., 2019). This system can diagnose an individual's latent attribute profile online while the assessment is being conducted.
(9) Compared with cross-sectional learning diagnosis, the diagnostic accuracy and validity of longitudinal learning diagnosis used to depict the learning trajectories are more worthy of attention by researchers and practitioners. In addition to choosing a suitable longitudinal LDM, many factors such as the quality of the longitudinal test itself, the setting of a cognitive model, students' response attitude, cheating, and missing data will also affect the accuracy and validity of the diagnostic results. The impact of these factors on the longitudinal learning diagnosis and the corresponding compensation or detection methods are also worthy of further discussion.

Overall, there are still many issues related to longitudinal learning diagnosis that are worthy of discussion. In view of the advantages of longitudinal learning diagnosis compared with cross-sectional learning diagnosis, the former is more in line with the idea of assessment for learning (Wiliam, 2011) and the needs of formative assessments.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 31900795) and the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No. 19YJC190025).

REFERENCES

Chen, Y., Culpepper, S. A., Wang, S., and Douglas, J. (2018a). 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

Chen, Y., Li, X., Liu, J., and Ying, Z. (2018b). Recommendation system for adaptive learning. *Appl. Psychol. Meas.* 42, 24–41. doi: 10.1177/0146621617697959

Collins, L. M., and Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behav. Res.* 27, 131–157. doi: 10.1207/s15327906mbr2701_8

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

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

Henson, R., Templin, J., and Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *J. Educ. Meas.* 54, 440–480. doi: 10.1111/jedm.12156

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Karelitz, T. M. (2004). *Ordered Category Attribute Coding Framework for Cognitive Assessments* (Unpublished doctoral dissertation). University of Illinois at Urbana Champaign, Champaign, IL, United States.

Kaya, Y., and Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. *Educ. Psychol. Meas.* 77, 369–388. doi: 10.1177/0013164416659314

Lee, S. Y. (2017). *Growth Curve Cognitive Diagnosis Models for Longitudinal Assessment* (Unpublished doctoral dissertation). University of California, Berkeley, CA, United States.

Leighton, J. P., and Gierl, M. J. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications.* Cambridge University Press. doi: 10.1017/CBO9780511611186

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

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204 doi: 10.1177/0013164415588946

Madison, M. J., and Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika* 83, 963–990. doi: 10.1007/s11336-018-9638-5

Man, K., and Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educ. Psychol. Meas.* 79, 617–635. doi: 10.1177/0013164418824148

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

Tang, X., Chen, Y., Li, X., Liu, J., and Ying, Z. (2019). A reinforcement learning approach to personalized learning recommendation system. *Br. J. Math. Stat. Psychol.* 72, 108–135. doi: 10.1111/bmsp.12144

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

von Davier, M., and Lee, Y.-S. (2019). *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. New York, NY: Springer. doi: 10.1007/978-3-030-05584-4

von Davier, M., Xu, X., and Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika* 76, 318–336. doi: 10.1007/s11336-011-9202-z

Wang, L., Tang, F., and Zhan, P. (in press). Effect analysis of individualized remedial teaching based on cognitive diagnostic assessment: taking "linear equation with one unknown" as an example. *J. Psychol. Sci.*

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

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018a). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87 doi: 10.3102/1076998617719727

Wang, S., Zhang, S., Douglas, J., and Culpepper, S. (2018b). Using response times to assess learning progress: a joint model for responses and response times. *Meas. Interdisciplinary Res. Perspect.* 16, 45–58. doi: 10.1080/15366367.2018.1435105

Wang, S., Zhang, S., and Shen, Y. (2019). A joint modeling framework of responses and response times to assess learning outcomes. *Multivariate Behav. Res.* 55, 49–68. doi: 10.1080/00273171.2019.1607238

Wiliam, D. (2011). What is assessment for learning? *Stud. Educ. Eval.* 37, 3–14. doi: 10.1016/j.stueduc.2011.03.001

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

Zhan, P. (2019). *A Cognitive Diagnosis Model for Analysis Multisource Data in Technology-Enhanced Diagnostic Assessments*. Invited Report at School of Mathematics and Statistics, Northeast Normal University, Changchun, China. Retrieved from: http://math.nenu.edu.cn/info/1063/4271.htm (accessed April 11, 2020).

Zhan, P. (2020). A Markov estimation strategy for longitudinal learning diagnosis: providing timely diagnostic feedback. *Educ. Psychol. Measure.* doi: 10.1177/0013164420912318

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

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

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

Zhan, P., Wang, W.-C., Jiao, H., and Bian, Y. (2018b). Probabilistic-input, noisy conjunctive models for cognitive diagnosis. *Front. Psychol.* 9:997. doi: 10.3389/fpsyg.2018.00997

Zhang, S., and Chang, H. (2019). A multilevel logistic hidden Markov model for learning under cognitive diagnosis. *Behav. Res. Methods* 52, 408–421. doi: 10.3758/s13428-019-01238-w

Zhang, S., and Wang, S. (2018). Modeling learner heterogeneity: a mixture learning model with responses and response times. *Front. Psychol.* 9:2339. doi: 10.3389/fpsyg.2018.02339

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 30 July 2020
doi: 10.3389/fpsyg.2020.01694

Q-Matrix Designs of Longitudinal Diagnostic Classification Models With Hierarchical Attributes for Formative Assessment

Wei Tian, Jiahui Zhang, Qian Peng and Xiaoguang Yang*

Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

Longitudinal diagnostic classification models (DCMs) with hierarchical attributes can characterize learning trajectories in terms of the transition between attribute profiles for formative assessment. A longitudinal DCM for hierarchical attributes was proposed by imposing model constraints on the transition DCM. To facilitate the applications of longitudinal DCMs, this paper explored the critical topic of the Q-matrix design with a simulation study. The results suggest that including the transpose of the R-matrix in the Q-matrix improved the classification accuracy. Moreover, 10-item tests measuring three linear attributes across three time points provided satisfactory classification accuracy for low-stakes assessment; lower classification rates were observed with independent or divergent attributes. Q-matrix design recommendations were provided for the short-test situation. Implications and future directions were discussed.

Keywords: Q-matrix, longitudinal DCMs, hierarchical attributes, TDCM, HDCM

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Lietta Marie Scott,
Arizona Department of Education, United States
Gongjun Xu,
University of Michigan, United States

****Correspondence:***
Jiahui Zhang
nellykim@126.com

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

Received: *02 February 2020*
Accepted: *22 June 2020*
Published: *30 July 2020*

Citation:
Tian W, Zhang J, Peng Q and Yang X (2020) Q-Matrix Designs of Longitudinal Diagnostic Classification Models With Hierarchical Attributes for Formative Assessment. Front. Psychol. 11:1694. doi: 10.3389/fpsyg.2020.01694

INTRODUCTION

Diagnostic cognitive models (DCMs; or cognitive diagnostic models, CDMs) have received increasing attention because the latent variable modeling approach to diagnostic assessment can shed light on the learning process (Rupp et al., 2010). A variety of latent variable models have been proposed in recent decades including specific models (e.g., the Deterministic Input, Noisy "and" Gate, DINA; Junker and Sijtsma, 2001) and generalized frameworks (e.g., the log-linear cognitive diagnostic model, LCDM; Henson et al., 2009). Two recent directions aim to address hierarchical attributes (Gierl et al., 2010; Templin and Bradshaw, 2014) and the mastery of attributes in longitudinal data (Li et al., 2016; Kaya and Leite, 2017; Wang et al., 2017; Madison and Bradshaw, 2018a,b), respectively.

The transition DCM (TDCM), proposed by Madison and Bradshaw (2018a,b), is a longitudinal model combining the LCDM and the latent transition analysis (LTA). The TDCM have been used on tests measuring independent attributes (Madison and Bradshaw, 2018a,b). However, empirical studies have suggested the presence of interdependencies among attributes in many educational cases (e.g., Gierl et al., 2010; Templin and Bradshaw, 2014). The incorporation of attribute hierarchy into the Q-matrix and the model parameterization has become important research topics in recent years. One of the approaches to modeling the attribute relationships is to impose a hierarchical structure in which mastering an attribute could be a prerequisite to mastering another attribute (Tatsuoka, 1983; Leighton et al., 2004; Templin and Bradshaw, 2014). Taking this approach, Templin and Bradshaw (2014) extended LCDM to its hierarchical form—hierarchical diagnostic

classification model (HDCM). Similarly, the longitudinal model TDCM can be constrained to incorporate hierarchical attributes. Following this line of thinking, we proposed the hierarchical transition DCM (H-TDCM) and explored the effects of Q-matrix designs on its classifications in this study.

The Q-matrix design, as a core element of the DCM-based test design, has not been adequately addressed in the context of longitudinal DCMs, since existing research focuses on model development and applications of longitudinal DCMs (e.g., Kaya and Leite, 2017; Madison and Bradshaw, 2018a,b). The Q-matrix links the items and the latent constructs to be measured (i.e., attributes) (Tatsuoka, 1983). Rows of the Q-matrix correspond to items, columns correspond to attributes, and its binary elements indicate whether an item measures an attribute (to put it differently, whether mastery of an attribute is required to succeed on an item). The row vectors of the Q-matrix are also called q-vectors. The Q-matrix plays important roles, both theoretically and statistically. From a theoretical perspective, cognitive theories could have a real impact on testing practice through the Q-matrix. This is especially true when the attributes are related to each other according to the cognitive theory. From a statistical perspective, the Q-matrix plays a significant role in model identification (Xu and Zhang, 2016; Xu, 2017; Köhn and Chiu, 2018; Gu and Xu, 2019a, forthcoming) and classification accuracy (DeCarlo, 2011; Madison and Bradshaw, 2015; Liu et al., 2017; Tu et al., 2019).

The identifiability conditions need to be satisfied for consistent estimation of the model parameters. Gu and Xu (2019a) identified the sufficient and necessary condition for identification of DINA and DINO. It requires that each attribute is measured by at least three items with a Q-matrix in the form $Q = \left(I_K^T, \left(Q'\right)^T\right)^T$ (T denotes transpose), in which any two different columns of the submatrix Q' are distinct (Gu and Xu, 2019a). The indentifiability issue is more complicated for saturated models (e.g., GDINA) and details on strict or generic identification can be found in Gu and Xu (forthcoming). The identification condition for hierarchical DCMs has also been discussed (Gu and Xu, forthcoming).

However, the Q-matrices that lead to identification may provide varying classification accuracy rates (DeCarlo, 2011; Madison and Bradshaw, 2015). To provide guidance for test construction practices based on DCMs, researchers explored the effects of different Q-matrix designs on the classification accuracy. For example, on the effects of Q-matrix designs with independent attributes, DeCarlo (2011) and Madison and Bradshaw (2015) have found that including more items measuring each attributes in isolation could help increase classification accuracy for DINA and LCDM.

When attribute hierarchies are involved, there has not been a consensus on the Q-matrix design regarding whether all q-vectors are eligible (Templin and Bradshaw, 2014; Tu et al., 2019). When a test involves K independent attributes, there are $2^K - 1$ distinct q-vectors. Consider a linear hierarchy with three attributes: $\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$. Attribute α_2 has direct relationships with the other two attributes while Attribute α_1 and α_3 have an indirect relationship. The reachability matrix or R-matrix can be used to capture both direct and indirect relationships (Tatsuoka, 1983; Gierl et al., 2000; Leighton et al., 2004). The R-matrix for three attributes under a linear hierarchy is presented in **Figure 1**. Some researchers argued that an item cannot measure a higher-level attribute without measuring its prerequisite(s) (Leighton et al., 2004; Köhn and Chiu, 2018; Tu et al., 2019), referred to as the restricted Q-matrix approach. According to the restricted Q-matrix approach, only three q-vectors are allowed in the Q-matrix in the case of three linear attributes, which correspond to the three column vectors of the R-matrix. In contrast, some studies use all $2^K - 1 = 7$ q-vectors in the Q-matrix as in an independent-attribute situation (Liu and Huggins-Manley, 2016; Liu et al., 2017), referred to as the unstructured Q-matrix approach.

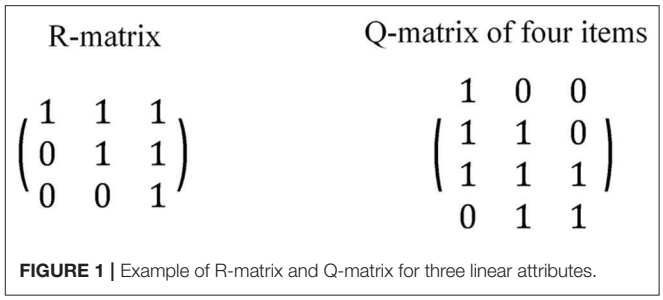


FIGURE 1 | Example of R-matrix and Q-matrix for three linear attributes.

Tu et al. (2019) took the restricted Q-matrix approach in a simulation study and emphasized the importance of containing the transpose of the R-matrix in the Q-matrix. **Figure 1** provided an example Q-matrix containing the transpose of the R-matrix, R^T. Liu et al. (2017), taking the unstructured Q-matrix approach, proposed different approaches to generate Q-matrices with linear, divergent, convergent, or unstructured attributes under the hierarchical diagnostic classification model (HDCM; Templin and Bradshaw, 2014). The adjacent approach (allowing each item to measure at most two attributes with direct relationships) was found to lead to higher classification accuracy in a shorter test (Liu et al., 2017).

To sum up, the purposes of the current study are 2-fold: First, the H-TDCM was defined to incorporate hierarchical attributes in the longitudinal DCM. Second, different Q-matrix designs were explored for TDCM and H-TDCM with a Monte Carlo simulation study. Both longitudinal models are based on LCDM, which is a general framework without limitations of the model fit assumptions. The rest of the paper is organized as follows. The next section briefly introduces LCDM, HDCM, and TDCM before defining the H-TDCM. Then, previous studies on the Q-matrix design are reviewed, followed by a simulation study on Q-matrix designs for TDCM and H-TDCM. The paper is concluded with a discussion of the limitations and educational implications.

MODELS

LCDM, HDCM, and TDCM

The LCDM (Henson et al., 2009) is a general diagnostic model that parameterizes the effects of the attributes measured by the

item on the probability of a correct response given examinee attribute profile. The LCDM subsumes many specific DCMs, including the DINA model (Junker and Sijtsma, 2001) and the DINO model (Templin and Henson, 2006).

Examinee attribute profiles are denoted by vectors $\boldsymbol{\alpha_c} = (\alpha_{c1}, \ldots \alpha_{ck}, \ldots, \alpha_{cK})$, where $c = 1, \ldots, C$ and α_{ck} takes the value of 0 or 1, indicating the non-mastery or mastery, respectively, of the kth attribute. The LCDM classifies examinees into one of the $C = 2^K$ attribute profiles assuming independent attributes. The number of attribute profiles decreases accordingly with hierarchical attributes.

For each item measured on a test, the LCDM item response function models the attributes mastery effects on the item response in terms of an intercept, the main effect for each attribute measured by the item, and the interaction term(s) that correspond to each possible combination of multiple attributes measured by the item. The general form of the LCDM item response function can be expressed as

$$P(X_{ic} = 1|\boldsymbol{\alpha}_c) = \frac{\exp(\lambda_{i,0} + \lambda_i^T \boldsymbol{h}(\boldsymbol{\alpha}_c, \boldsymbol{q}_i))}{1 + \exp(\lambda_{i,0} + \lambda_i^T \boldsymbol{h}(\boldsymbol{\alpha}_c, \boldsymbol{q}_i))} \quad (1)$$

where $\lambda_{i,0}$ is the intercept parameter of item i, λ_i contains all other item parameters including the main effects and interaction terms for item i, $\boldsymbol{q}_i$ denotes the q-vector of item i, the superscript T denotes transpose, and the function $\mathbf{h}$ results in a linear combination of $\boldsymbol{\alpha}_c$ and $\mathbf{q}_i$.

$$\begin{aligned}\lambda_{i,0} + \lambda_i^{\mathbf{T}} \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right) = \lambda_{i,0} + \lambda_{i,1,(k)} \alpha_{ck} q_{ik} + \lambda_{i,2,(l(k))} \alpha_{ck} \alpha_{cl} q_{ik} q_{il} \\ + \lambda_{i,3,(m(l,k))} \alpha_{ck} \alpha_{cl} \alpha_{cm} q_{ik} q_{il} q_{im} + \ldots \quad (2)\end{aligned}$$

Templin and Bradshaw (2014) proposed the hierarchical diagnostic classification models (HDCM) to address hierarchical attributes. Specifically, two changes are made to LCDM. First, the attribute profile space is limited and $\boldsymbol{\alpha}_c$ in Equations (1) and (2) is replaced by $\boldsymbol{\alpha}_c^*$ for notation. When a linear hierarchy is assumed, the number of mastery profiles is reduced from the original $C = 2^K$ to $C = K + 1$. The second change is that model constraints are imposed on LCDM. Specifically, some model parameters of the measurement model are fixed as zero.

Madison and Bradshaw (2018a,b) combined LCDM with latent transition analysis (LTA) to produce TDCM. LTA is a longitudinal latent class model that classifies examinees into latent classes and captures the latent class transitions over time (Collins and Lanza, 2010). As a conventional latent class analysis, it consists of the structural model and the measurement model. It is also a special case of the latent or hidden Markov model (HMM; Baum and Petrie, 1966). LTA parameterizes the probabilities of each latent class transitioning from one latent class to another between each time point in addition to latent class proportions and item parameters (i.e., the parameters estimated in conventional latent class analysis. LCDM serves as the measurement model of LTA. The LTA-DINA (Li et al., 2016) and LTA-DINO (Kaya et al., 2016) can be seen as special cases of the TDCM.

H-TDCM

The proposed H-TDCM combined the features of HDCM and TDCM to deal with hierarchical attributes in longitudinal data. The attribute hierarchy is imposed on TDCM by constraining corresponding item parameters in the measurement model as in HDCM and the structural parameters that are specific to TDCM. Specifically, model parameters for the main effects of nested attributes and some interaction terms are constrained as zero in light of the prerequisite relationships among them. Also, similar constraints are set on the transition parameters and prevalence parameters.

Given the expression of LTA (Collins and Lanza, 2010, p. 198), the probability of an examinee's response vector on I items over T time points is given by

$$\begin{aligned}P\left(\boldsymbol{Y} = \boldsymbol{y}\right) &= \overbrace{\sum_{\boldsymbol{\alpha}_{c_1}^*=1}^{C} \cdots \sum_{\boldsymbol{\alpha}_{c_T}^*=1}^{C} \delta_{\boldsymbol{\alpha}_{c_1}^*} \tau_{\boldsymbol{\alpha}_{c_2|c_1}^*} \cdots \tau_{\boldsymbol{\alpha}_{c_T|c_{T-1}}^*}}^{Structural} \\ &\overbrace{\prod_{t=1}^{T} \prod_{i=1}^{I} \prod_{r_{i,t}=1}^{R_i} \left[\rho_{i,r_{i,t}|\boldsymbol{\alpha}_{c_t}^*, \mathbf{q}_i}\right]^{I\left(y_{i,t}=r_{i,t}\right)}}^{Measurement} \\ &= \overbrace{\sum_{\boldsymbol{\alpha}_{c_1}^*=1}^{C} \cdots \sum_{\boldsymbol{\alpha}_{c_T}^*=1}^{C} \delta_{\boldsymbol{\alpha}_{c_1}^*} \tau_{\boldsymbol{\alpha}_{c_2|c_1}^*} \cdots \tau_{\boldsymbol{\alpha}_{c_T|c_{T-1}}^*}}^{Structural} \\ &\overbrace{\prod_{t=1}^{T} \prod_{i=1}^{I} \prod_{r_{i,t}=1}^{R_i} \left[\frac{\exp(\lambda_{i,0} + \lambda_i^T \mathbf{h}(\boldsymbol{\alpha}_{c_t}^*, \mathbf{q}_i))}{1 + exp(\lambda_{i,0} + \lambda_i^T \mathbf{h}(\boldsymbol{\alpha}_{c_t}^*, \mathbf{q}_i))}\right]^{I\left(y_{i,t}=r_{i,t}\right)}}^{Measurement},\end{aligned} \quad (3)$$

where $i = 1, 2, \ldots, I$; item i has R_i response categories; $y_{i,t}$ is the examinee's response to item i at time point t and $I\left(y_{i,t} = r_{i,t}\right)$ is an indicator function that is equal to 1 when the response is $r_{i,t}$, and equal to 0 otherwise; each sum ranges over each of the C attribute profiles at each time point, the first product is over the T time points, and the second product is over the I items; if the test measures K attributes with a certain hierarchical structure, the attribute profile at Time Point t is $\boldsymbol{\alpha}_{c_t}^* = \left(\alpha_{1_t}, \ldots, \alpha_{k_t}, \ldots, \alpha_{K_t}\right)$, for simplicity, $C_t = C$.

There are three types of parameters to be estimated (similar to the case of TDCM) in Equation (3). The first type includes HDCM item parameters $\lambda_{i,0}$ and λ_i. The second type is the probability of membership in attribute profile c at time point 1, denoted as $\delta_{\boldsymbol{\alpha}_{c_1}}$; and the third is the probability of transitioning between different attribute profiles (from $\boldsymbol{\alpha}_{c_{t-1}}$ to $\boldsymbol{\alpha}_{c_t}$) between time point t−1 to time point t, denoted as $\tau_{\boldsymbol{\alpha}_{c_t}|\boldsymbol{\alpha}_{c_{t-1}}}$, usually expressed as a multinomial regression model (e.g., Reboussin et al., 1998; Nylund, 2007):

$$\begin{aligned}\tau_{\boldsymbol{\alpha}_{c_t}|\boldsymbol{\alpha}_{c_{t-1}}} &= \frac{\exp(a_{c_t} + \mathbf{b}_{c_t|c_{t-1}}^T \mathbf{d}_{c_{t-1}})}{\sum_{c_t=1}^{C} \exp\left(a_{c_t} + \mathbf{b}_{c_t|c_{t-1}}^T \mathbf{d}_{c_{t-1}}\right)} \\ &= \frac{\exp(a_{c_t} + \mathbf{b}_{c_t|c_{t-1}}^T \mathbf{d}_{c_{t-1}})}{1 + \sum_{c_t=1}^{C-1} \exp\left(a_{c_t} + \mathbf{b}_{c_t|c_{t-1}}^T \mathbf{d}_{c_{t-1}}\right)}, t \geq 2; \quad (4)\end{aligned}$$

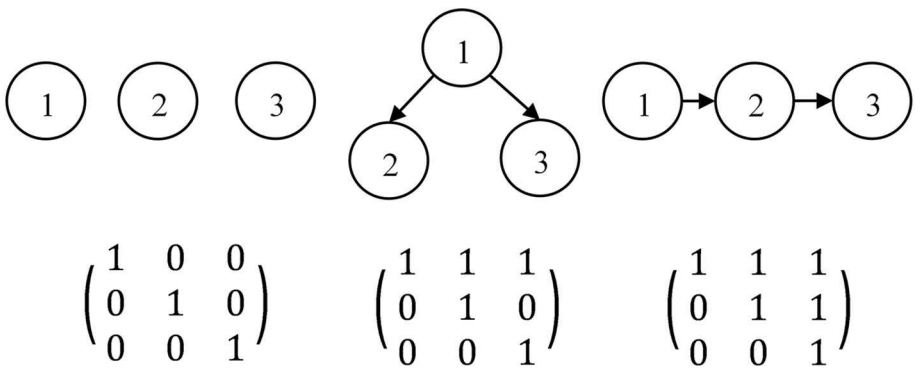


FIGURE 2 | Three attribute hierarchies with three attributes and their R-matrix. **(A)** Independent. **(B)** Divergent. **(C)** Linear.

We take for example a test measuring three linear attributes ($\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$). The $C = 4$ attribute profiles are the rows in

$$\begin{bmatrix}\alpha_{11} & \alpha_{12} & \alpha_{13}\\ \alpha_{21} & \alpha_{22} & \alpha_{23}\\ \alpha_{31} & \alpha_{32} & \alpha_{33}\\ \alpha_{41} & \alpha_{42} & \alpha_{43}\end{bmatrix} = \begin{bmatrix}0&0&0\\1&0&0\\1&1&0\\1&1&1\end{bmatrix}. \tag{5}$$

Four item parameters are to be estimated including the intercept effect $\lambda_{i,0}$, the main effect $\lambda_{i,1,(1)}$, the second-order interaction effect $\lambda_{i,2,(2(1))}$, and the third-order interaction effect $\lambda_{i,3,(3(2,1))}$:

$$\begin{aligned}\lambda_{i,0}+ \quad \lambda_i^{\mathbf{T}}\mathbf{h}\left(\boldsymbol{\alpha}_c^*,\mathbf{q}_i\right) &= \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{c1}q_{i1} + \lambda_{i,2,(2(1))}\alpha_{c1}\alpha_{c2}q_{i1}q_{i2}\\ &+\lambda_{i,3,(3(2,1))}\alpha_{c1}\alpha_{c2}\alpha_{c3}q_{i1}q_{i2}q_{i3}\end{aligned} \tag{6}$$

Note that Equation (3) is a general form of the H-TDCM. The combination of LTA and any other specific hierarchical CDM can be realized by imposing parameter constraints. The H-TDCM, in turn, can be seen as a special case of TDCM, and the two models can be compared with a likelihood-ratio difference test (Collins and Lanza, 2010). When the attribute hierarchy exists, H-TDCM is supposed to provide a more succinct model with a better fit than TDCM (Templin and Bradshaw, 2014).

SIMULATION STUDY

Design

The simulation study aimed to explore the effects of different Q-matrices on the classifications of TDCM with or without an attribute hierarchy. There has been a need for short tests that measure a couple of fine-grained attributes in the classroom setting. The simulation conditions approximated a practical formative assessment over a learning period of 2–4 weeks. A limited number of attributes would be focused on within such a short period, and time for testing is also very limited so short sessions are preferred. This short test is supposed to be administered three times: at the beginning, in the middle, and approaching the end of the learning period. Therefore, the simulations only consider three-attribute tests administered over three time points. Three attribute hierarchies (independent, divergent, and linear) are considered. The three attribute hierarchies with three attributes and the associated R-matrices are presented in **Figure 2**.

As mentioned earlier, there are two general approaches to Q-matrix design with hierarchical attributes—the restricted and the unstructured Q-matrix approaches. The restricted Q-matrix approach only allows q-vectors in the transpose of the R-matrix, denoted as R^T (Leighton et al., 2004; Köhn and Chiu, 2018; Tu et al., 2019), and the general guideline is to contain several R^Ts in the Q-matrix to obtain acceptable classification accuracy (Tu et al., 2019). We took the unstructured Q-matrix approach, which means an item can measure all possible combinations of attributes as in an independent-attribute situation (Liu and Huggins-Manley, 2016; Liu et al., 2017), because there exists no empirical evidence against the possibility of items measuring a higher-level attribute without measuring its prerequisite(s). With three attributes in a test, there are seven q-vectors corresponding to seven item types. However, it remains an open question whether it is still beneficial to contain R^Ts in the Q-matrix even though the unstructured approach was adopted. For each attribute hierarchy, three Q-matrix designs were used. The first Q-matrix design does not contain R^T, denoted as Q_1. The second and third Q-matrix designs include one or two R^Ts, which are denoted as Q_2 and Q_3, respectively. Crossing two factors (i.e., attribute hierarchy and Q-matrix design) led to a total of 9 conditions. The simulation study focused on the Q-matrix design; thus, all Q-matrices were assumed to be correctly specified.

The item parameters are assumed to be time-invariant for the attribute profiles to retain the same meaning over time. Previous studies have shown that the examinee sample size barely has an impact on the classification rates of DCMs (de la Torre et al., 2010; Kaya and Leite, 2017). The effect of sample sizes was explored in Madison and Bradshaw (2018a) with TDCM. Therefore, the sample size was not manipulated but set to be 1,000 in each condition. The attribute profile of examinees

TABLE 1 | Classification rates of three Q-matrix designs.

	Independent			Divergent			Linear		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
PROFILE CORRECT CLASSIFICATION RATES									
Time 1	0.517	0.550	0.557	0.582	0.651	0.671	0.710	0.731	0.725
Time 2	0.522	0.553	0.556	0.595	0.667	0.681	0.725	0.749	0.736
Time 3	0.536	0.577	0.577	0.606	0.680	0.693	0.734	0.761	0.744
Mean	0.525	0.560	0.563	0.594	0.666	0.682	0.723	0.747	0.735
MARGINAL CORRECT CLASSIFICATION RATES									
Time 1 α_1	0.723	0.784	0.821	0.938	0.937	0.917	0.931	0.929	0.901
Time 1 α_2	0.833	0.838	0.809	0.714	0.795	0.840	0.864	0.887	0.904
Time 1 α_3	0.831	0.807	0.810	0.855	0.858	0.860	0.864	0.872	0.885
Mean	0.796	0.810	0.813	0.836	0.863	0.872	0.886	0.896	0.897
Time 2 α_1	0.704	0.774	0.809	0.929	0.925	0.903	0.915	0.913	0.881
Time 2 α_2	0.827	0.835	0.803	0.716	0.804	0.845	0.848	0.875	0.890
Time 2 α_3	0.828	0.796	0.798	0.857	0.859	0.864	0.927	0.932	0.937
Mean	0.787	0.802	0.804	0.834	0.863	0.871	0.897	0.907	0.903
Time 3 α_1	0.713	0.784	0.816	0.927	0.924	0.901	0.912	0.912	0.877
Time 3 α_2	0.835	0.840	0.806	0.724	0.811	0.852	0.849	0.876	0.890
Time 3 α_3	0.834	0.807	0.808	0.864	0.864	0.868	0.944	0.947	0.950
Mean	0.794	0.810	0.810	0.838	0.866	0.874	0.902	0.912	0.906

Three Q-matrix designs Q_1, Q_2, and Q_3 included zero, one, or two R-matrix transposes.

followed a uniform distribution. Ten-item tests were generated under each condition.

To avoid the effects of item quality, we fixed the item parameters over all conditions: The intercept effect was −1, the main effect was 2, and the interaction effect was 1. As a result, $P(X = 1|\boldsymbol{\alpha} = \mathbf{0})$ ranged from 0.1 to 0.3, and $P(X = 1|\boldsymbol{\alpha} = \mathbf{1})$ was between 0.7 and 1.0. There are 8, 5, and 4 attribute profiles under independent, divergent, and linear hierarchies, respectively. With three independent attributes, there were 2^3 attribute profiles: c_1 $(0, 0, 0)$, c_2 $(0, 0, 1)$, c_3 $(0, 1, 0)$, c_4 $(0, 1, 1)$, c_5 $(1, 0, 0)$, c_6 $(1, 0, 1)$, c_7 $(1, 1, 0)$, and c_8 $(1, 1, 1)$. The divergent hierarchy condition had c_1 $(0, 0, 0)$, c_5 $(1, 0, 0)$, c_6 $(1, 0, 1)$, c_7 $(1, 1, 0)$, and $c_8(1, 1, 1)$. Three linear attributes led to four attribute profiles: c_1 $(0, 0, 0)$, c_5 $(1, 0, 0)$, c_7 $(1, 1, 0)$, and $c_8(1, 1, 1)$.

Mplus 7.4 (Muthén and Muthén, 1998–2015) was used to generate and analyze the response data of three time points based on TDCM or H-TCDM via maximum likelihood estimation. We include the Mplus syntax for estimation as an **Supplementary Material**. Evaluation criteria include the marginal correct classification rates (MCCRs) for each attribute and the correct classification rates (CCRs) for each attribute profile. Each simulation condition was replicated 100 times.

RESULTS

The correct classification rates are presented in **Table 1**. The results suggested that including the transpose of the R-matrix in the Q-matrix (i.e., Q_2) increased the profile CCRs and marginal CCRs at each time point for independent, divergent, and linear hierarchies. Including one more transpose of the R-matrix (i.e., Q_3) further slightly increased the CCRs except for the linear hierarchy. Another interesting finding is that the profile CCRs tended to increase with time. The CCRs at Time 3 were the highest. This trend was found under each combination of attribute hierarchy and Q-matrix design. The increase with time was not found in the marginal CCRs for independent attributes. Within the divergent or linear hierarchy, the marginal CCRs of the highest-level attribute (i.e., α_2 and α_3 under the divergent hierarchy and α_3 under the linear hierarchy) increased with time while the lowest-level attribute (i.e., α_1) had decreasing CCRs with time.

Comparing the three attribute hierarchies revealed that the CCRs generally increased as the relationship between attributes became stronger, and meanwhile, the number of attribute profiles became smaller. The profile CCRs were above 0.7, and the marginal CCRs were above 0.85 under the linear hierarchy with 10-item tests. The classifications for the independent attributes were the most difficult.

DISCUSSION

This paper proposed H-TDCM for hierarchical attributes in the longitudinal DCM by imposing model constraints on TDCM. The simulation study explored Q-matrix designs with different numbers of R-matrices. The CCRs generally increased with stronger dependencies between attributes, which is consistent with the findings of Templin and Bradshaw (2014) with LCDM. Ten-item tests for three linear attributes lead to profile CCRs above 0.7 and marginal CCRs above 0.85 at each time point, which might to acceptable for low-stakes classroom assessment. However, longer tests are needed for independent or divergent attributes to obtain acceptable classification rates. The profile CCRs increased with time, which means the attribute profile estimate from the final test would be the most accurate among several tests. The final attribute profile estimation may benefit from information from all the previous tests and provides a relatively accurate picture of the learning outcome, which is a desirable property for the longitudinal model.

Regarding the Q-matrix design, we took the unstructured Q-matrix approach (Liu and Huggins-Manley, 2016; Liu et al., 2017) by allowing all possible q-vectors, but explored Q-matrix designs containing different numbers of R^T. Simulation results showed that including one R-matrix transpose in the Q-matrix increased the CCRs in the case of independent attributes. Note that although the identification issue of CDMs and the Q-matrix design are usually treated as two separate research areas, the identification requirement may not always be satisfied in the Q-matrix design studies, especially for more complicated models and shorter tests.

First, we looked at the results for independent attributes. A closer look at the Q-matrices revealed that the first Q-matrix design (Q_1) did not measure α_1 in isolation; the second Q-matrix design (Q_2) contained only one identity matrix and measured α_1 in isolation only once. This explained the much lower classification rates for α_1 compared with other attributes. This finding with the TDCM agrees with the results of conventional DCMs (DeCarlo, 2011; Madison and Bradshaw, 2015). From the identification perspective, it has been proven that including two identity matrices in the Q-matrix is necessary for a saturated DCM such as LCDM with

TABLE 2 | R^T as a submatrix in the Q-matrix ensures a separable Γ -matrix.

q-vector	Attribute profile			
	000	100	110	111
100	0	1	1	1
110	0	0	1	1
111	0	0	0	1

independent attributes (Gu and Xu, forthcoming). Under Q_1 and Q_2 for independent attributes, the model parameters suffered from the non-identifiability issue and the consequence was reflected in the lower profile CCRs with Q_1 and Q_2 than with Q_3 in **Table 1**. It also explains why the marginal CCRs of α_1 under Q_1 and Q_2 were substantially lower than those under Q_3, while the marginal CCRs of the other two attributes did not differ much between Q-matrix designs.

Including R^T in the Q-matrix also increases the classification rates for the hierarchical cases in this study, which is consistent with the empirical findings from Tu et al. (2019). The results for hierarchical attributes can also be explained from the identification perspective as discussed in Gu and Xu (forthcoming). For a generalized multi-parameter DCM such as LCDM or HDCM, the concept of a separable Γ -matrix was introduced (Gu and Xu, forthcoming). The rows and columns of the Γ -matrix is indexed by the items and the attribute profiles, respectively. An entry of the Γ -matrix equals to 1 if an attribute profile has the highest correct response probability on an item and 0 otherwise. A Γ -matrix is said to be separable if any two column vectors of are distinct. The separability of the Γ -matrix is necessary for strict identification. We show that R^T as a submatrix in the Q-matrix ensures a separable Γ -matrix in **Table 2**. It can be further shown that the matrix of R^T is in the form of

$$\begin{pmatrix} 1 & * & \cdots & * \\ * & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 1 \end{pmatrix}_{K\times K}$$

after some row permutation, in which * takes the value of 0 or 1 and K is the number of attributes. Two R^Ts were contained in Q_3, which led to a separable Γ -matrix. As a result, Q_3 always ensures the identification of the model, while the first design may lead to non-identification issues (Gu and Xu, forthcoming). In contrast, Q_2 contained one R^T and at least one identity matrix instead of two R^Ts, which does not affect the model identification. Therefore, Q_2 and Q_3 showed similar classification rates. One major difference between the two designs is that Q_2 contains more single-attribute items and fewer multiple-attribute items. Under the linear hierarchy, for example, Q_3 has at least two items with q=(111), which has seven item parameters to be estimated. The parameter recovery of such items may be more difficult than single-attribute items, and the classification rate may suffer. As a result, the performance of Q_2 turned out to be better than Q_3 for the linear hierarchy.

This study aimed to demonstrate the classification performance of the H-TDCM with a short test and provide practical guidelines for the applications of this longitudinal model for formative classroom assessment. For the current setting of short tests and only a few attributes, we recommend that the Q-matrix contains (1) two identity matrices for independent attributes, (2) two R^Ts for a divergent hierarchy, and (3) one R^T and one identity matrix for a linear hierarchy. Besides, each attribute should be probed by at least three items. However, it should be noted that the current simulation study assumes that it is possible to develop items of all types of q-vectors with equal easiness, which may not be true for certain subject areas. For example, it may be more difficult to develop items that measure each attribute in isolation.

The formative classroom assessment has received renewed attention recently with the development of curriculum reform. The fusion of curriculum, instruction, and the assessment requires timely and constructive feedback that is closely connected to a curriculum and are based on students' learning history (e.g., Bennett, 2015; Gotwals, 2018; Shepard et al., 2018). Such feedback can be obtained from a diagnostic model that portrays the progression of attribute profiles. To establish the learning progression in terms of attribute profiles, however, is not an easy task. A possible solution could be collecting longitudinal assessment data from multiple classrooms and applying H-TDCM. The model parameters and classification results from H-TDCM can be used to understand the learning process better and to give teachers and students prior information before the learning begins. The current study focused on short tests for classroom applications where the attribute hierarchy is prespecified. Future simulation research can extend to longer tests for the purpose of exploring the learning process by estimating the attribute hierarchy. Those who are interested may refer to the requirement on the Q-matrix design (Gu and Xu, 2019b).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

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

FUNDING

This study is supported by the National Education Sciences Planning Projects Multilevel cognitive diagnostic model: individual and group diagnosis in large-scale educational assessment (CCA150160).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01694/full#supplementary-material

REFERENCES

Baum, L., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* 37, 1554–1563. doi: 10.1214/aoms/1177699147

Bennett, R. E. (2015). The changing nature of educational assessment. *Rev. Res. Educ.* 39, 370–407. doi: 10.3102/0091732X14554179

Collins, L. M., and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ: John Wiley and Sons.

de la Torre, J., Hong, Y., and Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *J. Educ. Measure.* 47, 227–249. doi: 10.1111/j.1745-3984.2010.00110.x

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Appl. Psychol. Measure.* 35, 8–26. doi: 10.1177/0146621610377081

Gierl, M. J., Alves, C., and Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *Int. J. Test.* 10, 318–341. doi: 10.1080/15305058.2010.509554

Gierl, M. J., Leighton, J. P., and Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educ. Measure. Issues Prac.* 19, 34–44. doi: 10.1111/j.1745-3992.2000.tb00036.x

Gotwals, A. W. (2018). Where are we now? Learning progressions and formative assessment. *Appl. Measure. Educ.* 31, 157–164. doi: 10.1080/08957347.2017.1408626

Gu, Y., and Xu, G. (2019a). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika* 84, 468–483. doi: 10.1007/s11336-018-9619-8

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

Gu, Y., and Xu, G. (forthcoming). Partial identifiability of restricted latent class models. *Ann. Stat.*

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Measure.* 25, 258–272. doi: 10.1177/01466210122032064

Kaya, Y., and Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. *Educ. Psychol. Measure.* 77, 369–388. doi: 10.1177/0013164416659314

Kaya, Y., Leite, W. L., and Miller, M. D. (2016). A comparison of logistic regression models for DIF detection in polytomous items: the effect of small sample sizes and non-normality of ability distributions. *Int. J. Assessment Tools Educ.* 2, 22–39. doi: 10.21449/ijate.239563

Köhn, H. F., and Chiu, C. Y. (2018). How to build a complete Q-matrix for a cognitively diagnostic test. *J. Classification* 35, 273–299. doi: 10.1007/s00357-018-9255-0

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

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Measure.* 76, 181–204. doi: 10.1177/0013164415588946

Liu, R., and Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In: *Quantitative Psychology Research,* eds L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, and M. Wiberg (New York, NY: Springer), 243–254.

Liu, R., Huggins-Manley, A. C., and Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educ. Psychol. Measure.* 77, 220–240. doi: 10.1177/0013164416645636

Madison, M. J., and Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educ. Psychol. Measure.* 75, 491–511. doi: 10.1177/0013164414539162

Madison, M. J., and Bradshaw, L. P. (2018a). Assessing growth in a diagnostic classification model framework. *Psychometrika* 83, 963–990. doi: 10.1007/s11336-018-9638-5

Madison, M. J., and Bradshaw, L. P. (2018b). Evaluating intervention effects in a diagnostic classification model framework. *J. Educ. Measure.* 55, 32–51. doi: 10.1111/jedm.12162

Muthén, L. K., and Muthén, B. O. (1998–2015). *Mplus User's Guide*. 7th ed. Los Angeles, CA.

Nylund, K. L. (2007). *Latent transition analysis: Modeling extensions and an application to peer victimization* (doctoral dissertation). University of California, Los Angeles, CA, United States.

Reboussin, B. A., Reboussin, D. M., Liang, K.-Y., and Anthony, J. C. (1998). Latent transition modeling of progression of health-risk behavior. *Multivariate Behav. Res.* 33, 457–478. doi: 10.1207/s15327906mbr3304_2

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

Shepard, L. A., Penuel, W. R., and Pellegrino, J. W. (2018). Classroom assessment principles to support learning and avoid the harms of testing. *Educ. Measure. Issues Pract.* 37, 52–57. doi: 10.1111/emip.12195

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Measure.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin, J., and Bradshaw, L. (2014). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika* 79, 317–339. doi: 10.1007/s11336-013-9362-0

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

Tu, D., Wang, S., Cai, Y., Douglas, J., and Chang, H. (2019). Cognitive diagnostic models with attribute hierarchies: model estimation with a restricted Q-matrix design. *Appl. Psychol. Measure.* 43, 255–271. doi: 10.1177/0146621618765721

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2017). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

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

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

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 August 2020
doi: 10.3389/fpsyg.2020.01714

Growth Modeling in a Diagnostic Classification Model (DCM) Framework–A Multivariate Longitudinal Diagnostic Classification Model

Qianqian Pan[1]*, *Lu Qin*[2] and *Neal Kingston*[1]

[1] Department of Educational Psychology, The University of Kansas, Lawrence, KS, United States, [2] Institutional Research and Assessment, Howard University, Washington, DC, United States

A multivariate longitudinal DCM is developed that is the composite of two components, the log-linear cognitive diagnostic model (LCDM) as the measurement model component that evaluates the mastery status of attributes at each measurement occasion, and a generalized multivariate growth curve model that describes the growth of each attribute over time. The proposed model represents an improvement in the current longitudinal DCMs given its ability to incorporate both balanced and unbalanced data and to measure the growth of a single attribute directly without assuming that attributes grow in the same pattern. One simulation study was conducted to evaluate the proposed model in terms of the convergence rates, the accuracy of classification, and parameter recoveries under different combinations of four design factors: the sample size, the growth patterns, the G matrix design, and the number of measurement occasions. The results revealed the following: (1) In general, the proposed model provided good convergence rates under different conditions. (2) Regarding the classification accuracy, the proposed model achieved good recoveries on the probabilities of attribute mastery. However, the correct classification rates depended on the cut point that was used to classify individuals. For individuals who truly mastered the attributes, the correct classification rates increased as the measurement occasions increased; however, for individuals who truly did not master the attributes, the correct classification rates decreased slightly as the numbers of measurement occasions increased. Cohen's kappa increased as the number of measurement occasions increased. (3) Both the intercept and main effect parameters in the LCDM were recovered well. The interaction effect parameters had a relatively large bias under the condition with a small sample size and fewer measurement occasions; however, the recoveries were improved as the sample size and the number of measurement occasions increased. (4) Overall, the proposed model achieved acceptable recoveries on both the fixed and random effects in the generalized growth curve model.

Keywords: diagnostic classification model, longitudinal data analysis, growth model, cognitive diagnostic assessment, multivariate

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Feiming Li,
Zhejiang Normal University, China
Ren Liu,
University of California, Merced, United States

***Correspondence:**
Qianqian Pan
panqianqian2013@gmail.com

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

Received: *05 October 2019*
Accepted: *22 June 2020*
Published: *07 August 2020*

Citation:
Pan Q, Qin L and Kingston N (2020) Growth Modeling in a Diagnostic Classification Model (DCM) Framework–A Multivariate Longitudinal Diagnostic Classification Model. Front. Psychol. 11:1714. doi: 10.3389/fpsyg.2020.01714

INTRODUCTION

Diagnostic classification models (DCMs; e.g., Rupp et al., 2010), also referred to as cognitive diagnosis models (CDMs; e.g., Leighton and Gierl, 2007), are defined as a family of confirmatory multidimensional latent-variable models with categorical latent variables (Rupp et al., 2010). DCMs evaluate the student's mastery status on each latent variable from a set of narrowly defined latent variables, referred to attributes in the DCM literature, and then classify students into attribute profiles that were determined as a priori (DiBello et al., 1995). DCMs provide fine-grained and multidimensional diagnostic information, which could help educators adjust classroom instruction and improve student learning. Since the traditional scale scores (e.g., IRT scores) have limits in providing enough information to inform classroom instruction and learning (e.g., de La Torre, 2009), DCMs have received growing attention in the educational measurement community as well as from educational practitioners in recent years.

DCMs have been increasingly used for empirical data analysis in recent years. For example, DCMs have been retrofitted to existing large-scale assessments to identify examinees' mastery status of tested skills (e.g., Lee and Sawaki, 2009; George and Robitzsch, 2014; Sedat and Arican, 2015; Ravand, 2016). In addition, some researchers successfully demonstrated the practical uses of DCMs in test development (Bradshaw et al., 2014). DCMs have also been applied in one large-scale assessment program (Dynamic Learning Maps® alternate assessment; DLM®; Dynamic Learning Maps, 2016) to detect distinct patterns of skill mastery for students with significant cognitive disabilities. However, most applications of DCMs are static, meaning that DCMs are used to classify individuals at a single time point. When longitudinal data are modeled, the longitudinal DCM is used to measure the change in the attribute profiles and mastery status over time.

Currently, two types of longitudinal DCMs have been proposed to analyze longitudinal data in the DCM framework. Latent transition analysis (LTA; Collins and Wugalter, 1992)—based longitudinal DCMs (e.g., Li et al., 2016; Kaya and Leite, 2017; Madison and Bradshaw, 2018) estimate the probabilities of transitioning from one latent class to another latent class or staying at the same latent class across two measurement occasions. Higher-order DCM (HDCM; e.g., de la Torre and Douglas, 2004; Templin and Bradshaw, 2014)—based longitudinal DCMs (e.g., Huang, 2017; Zhan et al., 2019) assumes a higher-order continuous factor to predict the mastery status of lower-order attributes so that the changes in the higher-order factor are used to infer the changes of lower-order attributes over time.

These two longitudinal DCM approaches have been evaluated by a few simulation studies and some applied research, which has demonstrated their utility for analyzing longitudinal data in the DCM framework. However, these models are not without limitations. For example, LTA-based longitudinal DCMs are restricted to the balanced data[1] and assume attributes are independent. In addition, LTA-based approach is limited to assessing changes between only two measurement occasions (Huang, 2017). On the other hand, HDCM-based longitudinal DCMs assume all attributes have similar growth trajectories. However, previous studies found attributes could change in different ways (e.g., Li et al., 2016; Madison and Bradshaw, 2018).

So, the overarching goal of the current study is to develop a multivariate longitudinal DCM, improves upon current longitudinal DCMs by (1) being able to incorporate both balanced data and unbalanced data and (2) measuring the growth of multiple attributes that have dissimilar growth trajectories. More specific research questions are presented in the Research Design and Methods section.

LONGITUDINAL DIAGNOSTIC CLASSIFICATION MODELS

Currently, two types of longitudinal DCMs have been developed and applied to measure longitudinal data, including latent transition analysis (LTA; Collins and Wugalter, 1992)-based longitudinal DCMs (e.g., Li et al., 2016; Kaya and Leite, 2017; Madison and Bradshaw, 2018), and Higher-order DCM (HDCM; e.g., de la Torre and Douglas, 2004; Templin and Bradshaw, 2014)—based longitudinal DCMs (e.g., Huang, 2017; Zhan et al., 2019). The definitions, model specifications, and limitations of these two types of longitudinal DCMs are briefly reviewed as follows.

LTA-Based Longitudinal DCMs

Latent class analysis (LCA; e.g., Lazarsfeld and Henry, 1968; Goodman, 1974) is developed for analyzing categorical latent variables. Latent transition analysis (LTA) is the extension of the general LCA for longitudinal data, which enables the estimation of both the latent class membership probability, often called the latent status prevalence in the LTA, and the probabilities of transitions in latent status from one measurement occasion to the next (Lanza et al., 2003, p. 161). LTA-based longitudinal DCMs are a composite of the DCM, as the measurement model to classify individuals into different latent classes at each time point, and the LTA, as the structural model to estimate the transition probability to represent the changes in latent class membership across two measurement occasions.

A few LTA-based longitudinal DCMs have been evaluated in simulation studies as well as applied in empirical studies. For example, Li et al. (2016) used the LTA with DINA(the deterministic-input, noisy-and-gate model; Junker and Sijtsma, 2001) as the measurement model to evaluate the effectiveness of an intervention for four cognitive skills across four measurement occasions for a sample of 109 seventh-grade students. This study provided base-rates of cognitive skills at each measurement occasion and three conditional transition probabilities from Occasion 1 to Occasion 2, Occasion 2 to Occasion 3, and Occasion 3 to Occasion 4, respectively. The results showed that

[1]In the current study, the balanced data refers to equal time intervals and unbalanced data refers to unequal time intervals.

attributes had different base-rates at the beginning and different conditional transition probabilities over time.

Madison and Bradshaw (2018) proposed the transitional diagnostic classification model (TDCM) to measure growth in attribute mastery for pre-test and post-test data, where the LCDM was adopted as the measurement model along with the LTA as the structural model. A simulation study showed that the TDCM could provide accurate and reliable classification and transition probabilities overtime under the variations in the number of attributes, sample size, Q-matrix, pre-test, and post-test base-rates, and marginal mastery transition probabilities. Additionally, the TDCM was applied to two empirical studies. In both studies, four mathematic skills were assessed before and after an intervention. The results showed that the base-rates of all attributes were improved after the intervention. However, the improvement differed by attributes and the groups, e.g., the control group or the intervention groups.

Furthermore, Chen et al. (2018) proposed a family of first-order hidden Markov models (FOHM) to model the learning trajectories with the CDM framework. Compared to the aforementioned LTA-based longitudinal DCMs that estimated the transition probabilities between two measurement occasions, FOHMs could estimate a transition probability matrix across multiple measurement occasions, which shows the probabilities of remaining in the same latent stage or learning some attributes or even losing some attributes from time t to $t + 1$. Such that it could provide an entire learning trajectory across time. Also, Chen et al. (2018) emphasized that there might be different types of learning trajectories, including the unstructured trajectories and non-decreasing trajectories. And, FOHMs are very flexible to estimate not only the most general trajectories but also some more parsimonious trajectories. So, even though the number of parameters in the transition probability matrix increases exponentially with the number of measurement occasions increasing, the restricted learning patterns could reduce the number of parameters.

Higher-Order DCM-Based Longitudinal DCMs

Higher-order DCMs (HDCMs) parameterize the structural model of general DCMs in a certain way to reduce the numbers of structural parameters. Several approaches have been utilized to construct the structural model (e.g., Hartz, 2002; de la Torre and Douglas, 2004; Rupp and Templin, 2008). The majority of HDCM-based longitudinal DCMs are parameterized using the logistic regression models (e.g., Huang, 2017; Zhan et al., 2019), which are composites of two model components. The first component is the HDCM, where a higher-order continuous factor, θ_{rt}, is assumed to predict the mastery statuses of multiple lower-order attributes at time t. The second component is the univariate growth curve models (GCMs; e.g., Raghavarao and Padgett, 2014; Hoffman, 2015), which describes the inter- and intra-individual differences in changes of this higher-order factor over T time points.

Recently, Huang (2017) proposed an HDCM-based longitudinal DCM, where a G-DINA model was used to evaluate the mastery status of attributes at each time point. Then, the Rasch model was utilized to construct the higher-order DCM at each time point. Last, a univariate GCM was applied to describe the growth of the higher-order factor over time. In addition, a set of time-invariant predictors (e.g., gender, age) were included to predict the random intercept and slope. This HDCM-based longitudinal DCM was evaluated in three simulation studies which varied several factors, including the sample size, the test length, the number of attributes, the item difficulty, and the number of measurement occasions. The results showed that a large sample size (1,000 individuals), enough items (30 items), and more measurement occasions (3 measurement occasions) could improve the parameter recovery and classification accuracy. Additionally, this HCDM-based longitudinal DCM was retrofitted to an empirical testing data, which assessed four attributes in a group of 4,177 high school students across three measurement occasions. The results showed that attributes differed in both the initial base-rates and the amount of improvement of the base-rates, for example, the base-rates of the "geometry" attribute were 0.90, 0.89, and 0.92 across three measurement occasions; however, the base-rates of the "number" attribute were 0.36, 0.49, and 0.58 across three measurement occasions. These results indicated different attributes developed different growth rates. Also, Zhan et al. (2019) developed a Long-DINA model, where (1) a DINA model was used to determine the mastery status of attributes at each time point, (2) the examinee's general ability at each measurement occasion was predicted by mastery status of attributes through a 2PL multidimensional higher-order latent structural model, and (3) the mean differences between the general abilities estimated from different measurement occasions represented the growth of examinees. Furthermore, the main improvement of this model was that incorporated specific factors in the DINA model to capture local item dependence due to the repeated measure rather than assuming the measurement invariance across time.

Limitations of Current Longitudinal DCMs

Even though the current longitudinal DCMs have provided a few approaches to analyze longitudinal data in the DCM framework; these longitudinal DCMs have limitations that could restrict their usage with empirical data. As discussed above, LTA-based longitudinal DCMs could estimate the changes of attributes directly over time. However, this method required balanced data. In other words, the time interval between measurement occasions cannot be accounted for in the model. This might result in inaccurately estimated transition probabilities if examinees have a different time interval between administrations. On the other hand, HDCM-based longitudinal DCMs estimate the growth of the higher-order factor via the univariate GCM framework, which could cooperate both balanced and unbalanced data. However, HDCM-based longitudinal models measure the growth of higher-order factors to indicate the growth of lower-order attributes, indicating multiple attributes should have similar growth patterns. While empirical studies' demonstrated attributes had different growth patterns, some attributes were improved over time, and some attributes had a

nearly consistent base-rate over time. For example, Madison and Bradshaw (2018) measured the changes in mastery status for four mathematics skills using pre- and post-test data and found the base-rate of one attribute was almost constant, where the base-rates changed from 0.65 to 0.70. However, base-rates of another three attributes improved more, ranging from 0.38 to 0.58, 0.38 to 0.51, and 0.59 to 0.73, respectively. Therefore, it is not reasonable to assume all attributes have the same growth patterns such that the growth of the higher-order factor cannot represent the changes in lower-order attributes well.

Therefore, there is a need to improve the current longitudinal DCMs. The motivation for the current study is to improve the current longitudinal DCMs by developing a multivariate longitudinal DCM, which could incorporate both balanced and unbalanced data, and measure the growth of attributes directly without assuming that attributes have similar growth patterns.

RESEARCH DESIGN AND METHOD

Multivariate Longitudinal Diagnostic Classification Models

The proposed multivariate longitudinal DCM is a composite of two components, the LCDM as the measurement model component that evaluates the mastery status of attributes at each measurement occasion, and a generalized multivariate growth curve model (e.g., GCM; MacCallum et al., 1997; Goldstein, 2011; Hoffman, 2015) as the structural model component that describes the changes of attributes over time via a logistic link function.

Model Specification

Let x_i denote the item response of item i. Only the binary item response was considered in the current study; however, polytomous item responses could be incorporated as well. Let $t = 1, 2, \ldots,$ T denotes the number of measurement occasions; $k = 1, 2, \ldots,$ K denote the number of attributes; and $\alpha_{rt}^{k} = \alpha_{rt}^{1}, \alpha_{rt,\ldots,}^{2} \alpha_{rt}^{k}$ denote the attribute profile at time t.

A three-level model is considered in the current study; Level 1 is the item level, Level 2 was the within-person level, and Level 3 is the between-person level.

In Level 1, the LCDM estimates the probability of individual r answering item i correct given profile α_r at time t, as shown in Equation (1), where $\lambda_{i,0}$ is the intercept parameter of the LCDM, indicating the logit of guessing the item i correctly without mastering any attributes, λ_i^T is a vector of size $(2^K - 1) \times 1$ with main effect and interaction parameters for item i at Time T, q_i is the set of Q matrix entries for item i, and $h\left(\alpha_{rt}, q_i\right)$ is a vector of size $\left(2^K - 1\right) \times 1$ with linear combinations of the α_{rt} and q_i.

For example, as shown in **Table 2**, the item 4 measures both Attribute 1 and Attribute 2 across all measurement occasions, such that, Equation (1) expresses the probability of a correct response to Item 4 is a function of the intercept ($\lambda_{1,0}$), the simple main effects of attribute 1 ($\lambda_{1,1,(1)}$) and attribute 2 ($\lambda_{1,1,(2)}$), interaction effects between these two attributes ($\lambda_{1,2,(1,2)}$), and the mastery status of two attributes. The intercept represents the log-odds of a correct answer for individuals who did not master any of the attributes. The simple main effects of attributes represent the increase in log-odds for individuals who have mastered only one of the attributes. Moreover, the interaction represents the change in log-odds for individuals who have mastered both attributes. Since the attributes are all dichotomous, $\alpha_1 = 1$ indicates attribute 1 is mastered, while $\alpha_1 = 0$ indicates attribute 1 is not mastered. As mentioned, as a general diagnostic model, the LCDM is able to subsume other frequently used DCMs. Using the same example, when two main effects are fixed to 0, the DINA model is achieved (Bradshaw and Madison, 2016).

$$P\left(X_4 = 1|\alpha_c\right) = \frac{\exp(\lambda_{1,0}+\lambda_{1,1,(1)}(\alpha_1)+\lambda_{1,1,(2)}(\alpha_2)+\lambda_{1,2,(1,2)}(\alpha_1\cdot\alpha_2))}{1+\exp(\lambda_{1,0}+\lambda_{1,1,(1)}(\alpha_1)+\lambda_{1,1,(2)}(\alpha_2)+\lambda_{1,2,(1,2)}(\alpha_1\cdot\alpha_2))} \quad (1)$$

In Level 2, α_{rt}^{k} represents the mastery status of attribute k at time t, Time_{rt} represents the time variable for individual i at time t. Then, the log-odds of $P(\alpha_{rt}^{k} = 1)$, indicating the probability of mastering attribute k at time t, are predicted by the random intercept β_{r0}^{k} and random slope β_{r1}^{k}.

In Level 3, the random intercept β_{0r}^{k} and random slope β_{1r}^{k} are predicted by the average initial level γ_{00}^{k} and average slope γ_{10}^{k}, respectively. u_{0r}^{k} and u_{1r}^{k} represent the individual r's deviations from the average initial level and growth rate for attribute k.

$$\text{Level 1 } \pi_{irt} = P\left(X_{irt} = 1|\alpha_{rt}\right) = \frac{\exp(\lambda_{i,0} + \lambda_i^T h(\alpha_{rt}, q_i))}{1 + \exp(\lambda_{i,0} + \lambda_i^T h(\alpha_{rt}, q_i))} \quad (2)$$

$$\text{Level 2 } logit\left(P\left(\alpha_{rt}^{k} = 1\right)\right) = \beta_{r0}^{k} + \beta_{r1}^{k} Time_{rt} + \epsilon_{rt}^{k} \quad (3)$$

$$R = \begin{bmatrix} \frac{\pi^2}{3} & \cdots & & \\ 0 & \frac{\pi^2}{3} & \cdots & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{\pi^2}{3} & \\ 0 & 0 & \ldots & 0 & \frac{\pi^2}{3} \end{bmatrix} \quad (4)$$

$$\beta_{0r}^{k} = \gamma_{00}^{k} + u_{0r}^{k} \quad (5)$$

$$\beta_{1r}^{k} = \gamma_{10}^{k} + u_{1r}^{k} \quad (6)$$

$$\text{Level 3} \quad \begin{bmatrix} \sigma_{u_0}^{(1)2} & & & \\ \sigma_{u_0^1,u_1^1} & \sigma_{u_1}^{(1)2} & & \\ \vdots & \ddots & & \\ \sigma_{u_0^1,u_0^K} & \sigma_{u_1^1,u_0^K} & \sigma_{u_0}^{(K)2} & \\ \sigma_{u_0^1,u_1^K} & \sigma_{u_1^1,u_1^K} & \sigma_{u_0^K,u_1^K} & \sigma_{u_1}^{(K)2} \end{bmatrix} \quad (7)$$

As shown in Equation (3), ϵ_{rt}^{k} are the Level 2 residuals, which follow a multivariate normal distribution with means of 0 and TK × TK covariance matrix of R, the diagonal elements are $\frac{\pi^2}{3}$, and off-diagonal elements are fixed to 0, indicating there are no covariances among ϵ_{rt} across constructs. In Level 3 variance $[u_{0r}^{k},\ u_{1r}^{k}] \sim MVN(0,\ G)$, G is a $KP \times KP$ covariance matrix, and P is the number of Level 2 random effects (Pan, 2018).

Research Questions

The purpose of the current study is to develop a multivariate longitudinal DCM and evaluate it under several conditions.

This study aims to answer the following research questions:

(1) Does the proposed model provide satisfied classification accuracy under different conditions?
(2) Do the sample size, the growth patterns, and the number of measurement occasions, the G matrix design, and their interactions impact the item parameter recoveries in the measurement model?
(3) Do the sample size, the growth patterns, and the number of measurement occasions, the G matrix design, and their interactions impact the fixed and random effects recoveries in the generalized growth curve model?

Simulation Design

To answer three research questions listed above, a simulation study was conducted, which included four design factors, (1) the sample size; (2) the growth patterns across attributes; (3) the G matrix design; and (4) the number of measurement occasions. Factors including the Q-matrix, the test length, the initial base-rate, and the item parameters were fixed. Simulation conditions are described below.

Design Factors

Sample size

The current study varied the sample size by 100, 200, and 300 to investigate the requirement for the sample size in the proposed model. Previous simulation studies in longitudinal DCMs used to have a large sample size that normally ranged from 500 to 3,000 (e.g., Kaya and Leite, 2017; Zhan et al., 2019; Madison and Bradshaw, 2018). However, the empirical studies usually had a relatively smaller sample size, normally ranging from 100 to 400 (e.g., Li et al., 2016). Therefore, it was useful to investigate the sufficient sample size for the proposed model to detect the growth of attributes over time, which could guide applied researchers to collect adequate participants without a waste of time and money.

Growth patterns across attributes

The proposed multivariate longitudinal DCM improves the current HDCM-based longitudinal DCMs in its potential for estimating the growth of attributes without assuming that attributes have similar growth trajectories. To examine if the proposed model could measure attributes with different growth patterns and attributes with similar growth patterns equally well, two different growth patterns across attributes were considered in the current study: (1) the even growth pattern in which attributes had similar growth patterns over time and (2) the uneven growth pattern in which attributes had different growth patterns over time.

Figure 1 describes these two conditions, where $T1$–$T5$ represent the first to the fifth measurement occasion; $A1$, $A2$, and $A3$ represent Attribute 1, Attribute 2, and Attribute 3, respectively.

Under the even growth pattern condition, the base-rates of all three attributes were improved from the first measurement occasion to the last measurement occasion. Under the uneven growth pattern condition, the base-rates of Attributes 2 and 3 were improved across five measurement occasions, but the base-rates of Attribute 1 kept constant over time.

G matrix design

The G matrix plays an important role in the multivariate GCM, which reflects the relationships between outcomes across time. It is one of the main interests in the longitudinal studies that measure multiple outcomes over time (e.g., Hoffman, 2015).

To examine if the proposed multivariate longitudinal DCM can detect the relationships among attributes, two types of G matrices are considered in the current study: (1) under the equal correlation condition, all attributes had equal correlations between intercept, slopes, and intercept and slope, meaning that attributes are equally correlated, and (2) under the unequal correlation condition, as described in **Figure 1**, Attribute 2 and Attribute 3 had equal correlations between intercept, slopes, and intercept and slope, but Attribute 1 had lower correlations with Attribute 2 and 3. **Table 1** presents the two types of G matrices and corresponding correlation matrices.

Number of measurement occasions

Previous simulation studies in HDCM-based longitudinal DCMs showed inconsistent results in the impacts of the number of measurement occasions on the classification accuracy. Huang (2017) found the number of measurement occasions (e.g., 2 or 3 measurement occasions) did not influence the classification accuracy significantly. However, Zhan et al. (2019) found the classification accuracy slightly increased as the number of measurement occasions increased. For the growth model, more measurement occasions are associated with good parameter recoveries (e.g., Preacher et al., 2008). To examine whether the number of measurement occasions impacted the performance of the proposed multivariate longitudinal DCM, the number of measurement occasions varied between 3 and 5 in the current study.

Fixed Conditions

Test length

A test of 30 binary items was simulated in the current study. The test length fell within the range of applied research as well as simulation studies in the longitudinal DCMs (e.g., Huang, 2017; Kaya and Leite, 2017; Madison and Bradshaw, 2018).

Q-matrix

As discussed above, DCMs are able to incorporate both the simple structure and the complex structure of the Q-matrix. In the current study, a complex structure of the Q-matrix was specified as shown in **Table 2**. Each item measures up to two attributes and attributes were assessed by equal numbers of items. This Q-matrix design was suggested by previous applied research and simulation studies (e.g., Bradshaw and Templin, 2014; Bradshaw et al., 2014; Kaya and Leite, 2017; Madison and Bradshaw, 2018).

Initial base-rates

The initial base-rate was fixed to 0.20, 0.25, and 0.30 for Attribute 1, Attribute 2, and Attribute 3, respectively. The previous empirical studies on measuring growth of attributes found initial base-rates ranged from 0.02 to 0.90 and suggested an easier

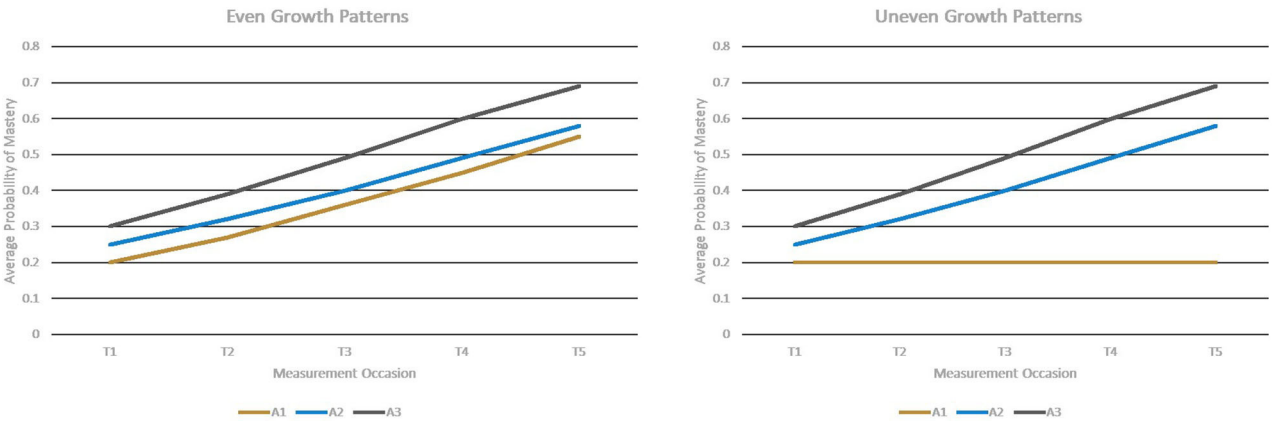


FIGURE 1 | Two patterns of growth across attributes.

TABLE 1 | G matrix specification and corresponding correlation matrix.

	Equal correlation condition						Unequal correlation condition					
	u_0^1	u_1^1	u_0^2	u_1^2	u_0^3	u_1^3	u_0^1	u_1^1	u_0^2	u_1^2	u_0^3	u_1^3
CORRELATION MATRIX												
u_0^1	**1.0**						**1.0**					
u_1^1	0.20	**1.0**					0.10	**1.0**				
u_0^2	0.90	0.10	**1.0**				0.90	0.01	**1.0**			
u_1^2	0.10	0.25	0.20	**1.0**			0.01	0.01	0.20	**1.0**		
u_0^3	0.90	0.10	0.90	0.10	**1.0**		0.10	0.01	0.90	0.10	**1.0**	
u_1^3	0.10	0.25	0.10	0.25	0.20	**1.0**	0.01	0.01	0.10	0.25	0.20	**1.0**
COVARIANCE MATRIX												
	$\sigma^2_{u_0^1}$	$\sigma^2_{u_1^1}$	$\sigma^2_{u_0^2}$	$\sigma^2_{u_1^2}$	$\sigma^2_{u_0^3}$	$\sigma^2_{u_1^3}$	$\sigma^2_{u_0^1}$	$\sigma^2_{u_1^1}$	$\sigma^2_{u_0^2}$	$\sigma^2_{u_1^2}$	$\sigma^2_{u_0^3}$	$\sigma^2_{u_1^3}$
$\sigma^2_{u_0^1}$	**0.1500**						**0.1500**					
$\sigma^2_{u_1^1}$	0.0173	**0.0500**					0.0173	**0.0500**				
$\sigma^2_{u_0^2}$	0.1350	0.0087	**0.1500**				0.1350	0.0009	**0.1500**			
$\sigma^2_{u_1^2}$	0.0087	0.0125	0.0173	**0.0500**			0.0009	0.0005	0.0173	**0.0500**		
$\sigma^2_{u_0^3}$	0.1350	0.0087	0.1350	0.0087	**0.1500**		0.1350	0.0009	0.1350	0.0087	**0.1500**	
$\sigma^2_{u_1^3}$	0.0087	0.0125	0.0087	0.0125	0.0173	**0.0500**	0.0009	0.0005	0.0087	0.0125	0.0173	**0.0500**

u_0^k and u_1^k represent the random intercept and slope for attributes; $\sigma^2_{u_0^k}$ and $\sigma^2_{u_0^k}$ represent the random intercept and slope variance for attributes. Bold values means the correlation of this parameter itself.

attribute might have a base-rate approximately 0.60, a medium attribute might have a base-rate approximately 0.40, and a hard attribute might have a base-rate ∼0.20 (Madison and Bradshaw, 2018); therefore, the base-rates are set to 0.20, 0.25, and 0.30 to mimic the hard, medium-hard, and medium attributes at the first measurement occasion.

Fixed effects (γ_{00}^{k}, γ_{01}^{k})

The linear growth of the log-odds of the probability of mastering attributes was considered in the current study. It should be noted that the linear growth of the log-odds of the probability did not necessarily result in the linear growth of base-rates over time. **Table 3** presents the fixed effects under both even and uneven growth pattern conditions.

Time variables

The current study planned to mimic the context of the interim assessments, which are administered several times within a school year (Great Schools Partnership, 2013). The common interval ranges from 6 to 8 weeks, such that individuals might receive the assessment at different times. Therefore, the current study set the time interval to 8 weeks and the unit of time to 1 week. The mean and standard deviation of time variables at each measurement occasion was fixed to $\mu_{time} = (0, 8, 16, 24, 32)$ and $\sigma_{time} = 1$, such that each individual had his/her own time variable at each measurement occasion to mimic the unbalanced data design.

As shown in **Table 3**, $\gamma_{00} = -1.38$ is the log-odds of the probability of 0.2, meaning at the first measurement occasion, the

TABLE 2 | Q-matrix design.

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

average probability of mastering Attribute 1 is 20%[2]. $\gamma_{01} = 0.05$ is growth rates of Attribute 1 in the log-odds scale, meaning that when time is increasing by one unit, the log-odds of probability of mastering Attribute 1 is increased by 0.05 in average, which is equal to the probability of mastery is increased by 0.008.

Table 4 presents the average base-rates of attributes across five measurement occasions, which was obtained by using the mean of the time variable and fixed effects shown in **Table 3**. Under the even growth pattern condition, the probabilities of mastery of three attributes were improved by 0.35, 0.38, and 0.39, respectively, across the time, and under the uneven pattern condition, the base-rate of Attribute 1 had a constant of 0.20, and the probabilities of mastery were improved by 0.38 and 0.39 for Attributes 2 and 3, respectively. This amount of improvement fell in the range of improvement of base-rates found in the previous studies (Li et al., 2016; Madison and Bradshaw, 2018).

Item parameters

The intercepts of all items were fixed to −1.5 indicating the probability of having a correct answer was 0.18. The simple main effects of all items were fixed to 1.5, indicating the probability of having a correct answer was 0.50 given mastering this attribute. The interaction effects between two attributes were fixed to 0.50, indicating the probability of having a correct answer was 0.88, given mastering two attributes.

Data Generation Procedures

Data were generated in R, version 3.4.2 (R Core Team, 2017). Each condition was replicated 100 times.

Data generation procedures included two stages: first, the probability of mastery was generated for each attribute at five measurement occasions, then the mastery statuses of them was generated; lastly, the item response data was generated, which are proceeded as follows:

Generate the linear predictors of the probability of mastery for each attribute by using the intercept and slope parameters, time variables, and G matrix for each individual;

Convert this linear predictor into the probability of mastery;

A binary mastery status for each attribute is randomly drawn from the binomial distribution with the probability of mastering attributes.

Generate the probability of having a correct answer for each item using a prespecified Q-matrix, item parameters, and person profiles.

A binary item response is randomly sampled from the binomial distribution with the probability obtained from the last step.

Analysis Plan and Outcome Variables

A Markov Chain Monte Carlo (MCMC) algorithm was adopted to estimate model parameters, which was implemented in the JAGS software (Plummer, 2003) by using the *R2jags* package (Su and Yajima, 2015) in the programming environment R (R Core Team, 2017). The JAGS syntax and more details of MCMC analyses can be found in the **Supplementary Material**.

The LCDM was applied to estimate the mastery statuses of attributes at each measurement occasion. For example, as described in the Q-matrix in **Table 2**, item 4 measured both Attribute 1 and Attribute 2. Thus, the probability of providing a correct answer to item 4 given the latent class c at Time t can be expressed as follows:

$$\pi_{4ct} = P(\mathrm{x}_{4\mathrm{ct}} = 1|\alpha_{\mathrm{ct}}) = \frac{\exp(\lambda_{4,0}+\lambda_{4,1,(1)}(\alpha_1)+\lambda_{4,1,(2)}(\alpha_2)+\lambda_{4,2,(1,2)}(\alpha_1\cdot\alpha_2))}{1+\exp(\lambda_{4,0}+\lambda_{4,1,(1)}(\alpha_1)+\lambda_{4,1,(2)}(\alpha_2)+\lambda_{4,2,(1,2)}(\alpha_1\cdot\alpha_2))} \quad (8)$$

For items that only measure one attribute, only the intercept and the main effect of this item were included in the equation.

The generalized multivariate GCM was applied to measure the changes in mastery statuses of attributes over time. First, as suggested by MacCallum et al. (1997), Curran et al. (2012), and Hoffman (2015), a synthesized variable was created, which was a composite of multiple outcome variables (α_{rt}^k in the current study), then a series of dummy variables as exogenous predictors were adopted to control which specific outcomes were referenced within different parts of the model. Let dv_{rt} denote the synthesized variable, which contained individual $r's$ mastery statuses for three attributes across four measurement occasions. A total of three dummy variables, $A1$, $A2$, and $A3$, were included in the model to distinguish which specific element belonged to which specific outcome variables, where $A1$ was equal to 1 for Attribute 1 and $A1$ was equal to 0 for other attributes. Therefore, the probability of mastering attribute α_{rt}^k $(k = 1, 2, 3)$ at time t could be described as follows:

$$\begin{aligned} logit\left(P\left(dv_{rt}=1\right)\right) = A1\Big[\left(\gamma_{00}^1+u_{0r}^1\right)+\left(\gamma_{10}^1+u_{1r}^1\right)Time_{rt}\Big] \\ +A2[(\gamma_{00}^2+u_{0r}^2)+(\gamma_{10}^2+u_{1r}^2)Time_{rt}] \\ +A3[(\gamma_{00}^3+u_{0r}^3)+(\gamma_{10}^3+u_{1r}^3)Time_{rt}] \end{aligned} \quad (9)$$

where the main effects of $A1$, $A2$, and $A3$ represent the initial levels for three attributes, and the interaction effects between

[2]This equation describes the relationships between the log odds of probability and fixed effects. $\log\left(\frac{probability}{1-probability}\right) = \log\left(\frac{0.2}{1-0.2}\right) = -1.38$.

TABLE 3 | Initial level and growth rates of linear predictors.

	Even growth patterns			Uneven growth patterns		
	A1	**A2**	**A3**	**A1**	**A2**	**A3**
γ_{00}	−1.38	−1.10	−0.85	−1.38	−1.10	−0.85
γ_{01}	0.05	0.04	0.05	0	0.04	0.05

A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3.

TABLE 4 | Base-rates of attributes over time.

	T1	T2	T3	T4	T5
EVEN GROWTH PATTERN					
A1	0.20	0.27	0.36	0.45	0.55
A2	0.25	0.32	0.40	0.49	0.58
A3	0.30	0.39	0.49	0.60	0.59
UNEVEN GROWTH PATTERN					
A1	0.20	0.20	0.20	0.20	0.20
A2	0.25	0.32	0.40	0.49	0.58
A3	0.30	0.39	0.49	0.60	0.59

A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3; T1–T5 represent the first measurement occasion to the fifth measurement occasion.

dummy variables and time scores represent the growth rates for attributes.

Once data analysis was finished, the following outcome variables across all 100 replications were obtained for all conditions:

(1) Gelman-Rubin diagnostic ($\hat{R}$) of parameters, including item parameters in the LCDM and both fixed effects and random effects parameters in the generalized growth curve model.
(2) The distribution of estimated parameters, including the mean, standard deviation, and quantiles.

Evaluation Criteria

Convergence rates, the classification accuracy of attributes at each measurement occasion, and the parameter recovery were evaluated in the current study to examine the performance of the proposed model under different conditions.

Convergence Rates

Convergence was assessed by using the Gelman-Rubin diagnostic ($\hat{R}$), also referred to as the "potential scale reduction factor" (Gelman and Rubin, 1992). Suppose there are m independent Markov chains, $\hat{R}$ is given by:

$$\sqrt{\hat{R}} = \sqrt{\frac{n-1}{n} + \frac{1}{n}\frac{B}{W}} \tag{10}$$

where B is the variance between the means of the m chains, W is the average of the m within-chain variances, and n is the number of iterations of the chain after discarding the iterations as burn-in. If the algorithm converges, $\hat{R}$ is approaching 1, indicating a stationary distirbution has been achieved because the marginal posterior variance (weighted combo of between and within-chain variance) are equal to the within-chain variances. In the current study, $\hat{R}$ was calculated for all model parameters, and we adopted the criteria of $\hat{R} < 1.2$ as the indicator of convergence as suggested by the previous study (e.g., Sinharay, 2003).

In one replication, if one or more parameters had the $\hat{R}$ larger than 1.2, this replication was regarded as non-converged. After a total of 100 replications, the convergence rates for this condition was calcualted and reported. Only the results from the converged replications were kept and used in the following analysis.

Classification Accuracy

The classification accuracy was evaluated by using (1) the bias of estimated probability of attribute mastery, (2) the correct classification rates for each mastery status, and (3) Cohen's kappa (Cohen, 1960).

The bias of the estimated probability of attribute mastery was the difference between the estimated and the true probability of attribute mastery. The correct classification rates for each mastery status included (1) the correct classification rates for individuals who truly mastered an attribute, and (2) the correct classification rates for individuals who truly did not master an attribute. Cohen's kappa measures the agreement between the true and the estimated mastery status.

The estimated class membership was obtained by applying 0.5 as the cutpoint, meaning that an individual with an estimated probability larger than 0.5 would be classified as mastery, vice versa.

Parameter Recovery

The bias and mean squared error (MSE) of estimated parameters, including item parameters from the measurement model, intercept and slope parameters, and variance and covariance parameters from the structural model were computed to assess the parameter recovery in each condition.

$$Bias_\theta = \frac{\sum_{r=1}^{R}\sum_{i}^{N}(\hat{\theta}_{ir} - \theta_i)}{RN} = \hat{\theta}_{ir} - \theta_i \tag{11}$$

$$MSE_\theta = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N}\left(\hat{\theta}_{ir} - \theta_i\right)^2}{RN} \tag{12}$$

where θ represents the estimated parameter, which is the mean of the sample distribution obtained from the Bayesian estimation. R is the number of replications; N is the number of elements in the set of θ.

A factorial analysis of variance was adopted to assess the impact of design factors on outcome variables. In all analyses, the α level was controlled at 0.05 level, and partial η^2 was adopted as the measure of effect sizes. According to Cohen (1988) convention, partial η^2 values of 0.01, 0.06, and 0.14 were regarded as small, medium, and large effects.

RESULTS

Convergence Rates

As aforementioned, the Gelman-Rubin diagnostic ($\hat{R}$) of item parameters in the LCDM, fixed effects andrandom effects parameters in the generalized growth curve model were evaluated, and we adopted the criteria of $\hat{R} < 1.2$ as the indicator of convergence as suggested by the previous study (e.g., Sinharay, 2003). When all the parameters, including the item parameters in the LCDM, fixed effects, and the random effects parameters in the generalized growth curve model were converged in one replication, this replication was regarded as converged. Results found that the average convergence rate is 0.95 under the conditions with three measurement occasions ($MO = 3$). And, the average convergence rate is 0.97 under the conditions with five measurement occasions ($MO = 5$). The details in convergence rates can be found in the **Supplementary Material**. Only the converged replications were used in the following analyses.

Classification Accuracy

The classification accuracy was evaluated by using (1) bias of the estimated probability of attribute mastery, (2) correct classification rates for each mastery status, and (3) Cohen's kappa.

The average bias of probability of attribute mastery under the conditions when $MO = 5$ showed that the probability of attribute mastery was recovered well under most conditions. The average bias of the probability of attribute mastery was all close to 0 under most conditions. Similar patterns were found when $MO = 3$. For the sake of page limits, only the average bias from the condition $MO = 5$ in **Table 5**, the summary of $MO = 3$ could be found in the **Supplementary Material**.

Table 6 presents the average correct classification rates for individuals who truly mastered attributes, and **Table 7** presents the correct classification rates for individuals who truly did not master attributes under different conditions when $MO = 5$. The average correct classification rates were very low for individuals who truly mastered the attributes at the first measurement occasion ($T = 1$), but the correct classification rates improved as the number of measurement occasions increased as shown in **Table 6**. For individuals who truly did not master the attributes, **Table 7** shows that the correct classification rates were perfect at the first measurement occasion, and then decreased to about 0.9 at the following measurement occasions.

This pattern might be due to the cut point of 0.5 used in the current study. The true mastery status was randomly generated through a binomial distribution with the true probability of mastery, such that, there is still some probabilities of mastering attributes, even the probability is very low. However, the estimated probability of attribute mastery was very low on the first two measurement occasions; the majority of individuals' probabilities were lower than 0.5. After 0.5 was set as the cut point to classify individuals into mastery or non-mastery classes, most of the individuals were classified into the non-mastery class even they truly mastered the attributes by design. With the increasing of measurement occasions, the estimated probabilities for individuals who truly mastered the attributes were increasing to be larger than 0.5, thus the cut point of 0.5 can classify them correctly. Such that, the correct classification rate was very low on the first two measurement occasion, but it increases as the measurement occasions increase.

The similar patterns were found when $MO = 3$, which could be found in the **Supplementary Material**. In summary, even though the probability of attribute mastery were recovered well, the correct classification rates depended on the individuals' mastery status and the cut point that was adopted to classify individuals.

Cohen's kappa was calculated to evaluate the degree of agreement between the estimated and true mastery status. **Table 8** presents the average kappa under different conditions when $MO = 5$. The calculation of kappa required that both true and estimated mastery status should have at least two levels; however, estimated mastery status only had one level under some conditions, especially at the first measurement occasion. Therefore, kappa was not applicable under some conditions. Results found that kappa values improved as time increased. This pattern might be due to the same reason as discussed above that the estimated probability of mastery was very low for all individuals at the first and second measurement occasions, such that after applying 0.5 as the cutpoint, the most of individuals who truly mastered the attributes were falsely classified to non-mastery. Therefore, kappa values were low at the beginning but improved as the number of measurement occasions increased. Similar patterns were found when $MO = 3$, which could be found in the **Supplementary Material**.

In summary, the agreement between true and estimated mastery status improved as the number of measurement occasions increased, and it was influenced by the cutpoint applied to classify individuals.

Parameter Recovery

The bias and mean square error (MSE) of the estimated parameters were computed to assess the parameter recovery in each condition through the simulation. Then, ANOVA tests were conducted to assess the impact of the design factors on the bias and MSE values of the estimated parameters of the measurement model and the structural model, respectively.

Measurement Model Parameter Recovery

There were three sets of item parameters in the LCDM: the intercept (λ_0), the main effect (λ_{α_k}), and the interaction effect ($\lambda_{\alpha_k\alpha_{k'}}$) parameters. Therefore, the average bias and MSE of all three sets of item parameters were assessed to evaluate the measurement model parameter recoveries.

TABLE 5 | Bias of probability of attribute mastery (MO = 5).

			T1			T2			T3			T4			T5		
			A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
G1	gam1	N100	.	.	0.01	.	.	.	.	.	.	.	.	.	.	.	.
		N200	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		N300	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	gam2	N100	.	.	0.01	.	.	.	.	−0.01	.	.	.	.	.	.	.
		N200	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		N300	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
G2	gam1	N100	.	0.01	.	.	.	.	.	.	.	.	.	.	.	.	.
		N200	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		N300	.	0.01	.	.	.	.	.	.	.	.	.	.	.	.	.
	gam2	N100	0.01	.	0.01	.	−0.01	.	.	−0.01	.	.	−0.01	.	.	.	.
		N200	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
		N300	.	.	.	.	.	.	.	.	.	.	.	.	0.01	0.02	0.01

T1–T5 represent the first to the fifth measurement occasion; A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3; N100, N200, and N300 represent the sample size of 100, 200, and 300, respectively; G1 and G2 represent equal correlation G and unequal correlation conditions of G matrix, respectively; gam1 and gam2 represent the same growth pattern across attributes and unequal growth patterns across attributes, respectively; · represents <0.001.

TABLE 6 | Average correct classification rates for individuals who truly mastered attribute (MO = 5).

			T1			T2			T3			T4			T5		
			A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
G1	gam1	N100	0	0.04	0.06	0.69	0.73	0.76	0.86	0.88	0.88	0.90	0.91	0.92	0.92	0.94	0.94
		N200	0.03	0.02	0.03	0.70	0.73	0.77	0.86	0.87	0.89	0.91	0.91	0.92	0.93	0.93	0.94
		N300	0.02	0.02	0.03	0.70	0.73	0.77	0.87	0.87	0.88	0.91	0.91	0.91	0.93	0.93	0.94
	gam2	N100	0.04	0.04	0.07	0.60	0.71	0.76	0.80	0.86	0.89	0.87	0.90	0.92	0.90	0.92	0.93
		N200	0	0.02	0.03	0.62	0.73	0.77	0.82	0.86	0.89	0.88	0.91	0.92	0.90	0.93	0.94
		N300	0.01	0.02	0.02	0.64	0.73	0.77	0.82	0.88	0.88	0.88	0.91	0.92	0.90	0.93	0.94
G2	gam1	N100	0.05	0.08	0.07	0.69	0.73	0.75	0.87	0.87	0.88	0.91	0.91	0.91	0.93	0.94	0.93
		N200	0	0.03	0.04	0.70	0.73	0.76	0.86	0.87	0.88	0.90	0.91	0.92	0.92	0.93	0.94
		N300	0.02	0.02	0.03	0.69	0.73	0.76	0.86	0.86	0.88	0.90	0.90	0.91	0.91	0.92	0.93
	gam2	N100	0.05	0.04	0.07	0.64	0.70	0.75	0.82	0.86	0.87	0.88	0.91	0.92	0.90	0.92	0.94
		N200	0.02	0.02	0.04	0.63	0.73	0.77	0.83	0.87	0.89	0.87	0.91	0.91	0.90	0.93	0.94
		N300	0.02	0.02	0.02	0.64	0.74	0.76	0.83	0.87	0.88	0.88	0.91	0.91	0.90	0.93	0.93

T1–T3 represent the first to the third measurement occasion; A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3; N100, N200, and N300 represent the sample size of 100, 200, and 300, respectively; G1 and G2 represent equal correlation G and unequal correlation conditions of G matrix, respectively; gam1 and gam2 represent the same growth pattern across attributes and unequal growth patterns across attributes, respectively.

As presented in **Table 9**, the proposed model achieved good parameter recoveries in intercept and main effect parameters, but the interaction parameters had relatively large bias and MSE values under most conditions. However, the recovery of the interaction effect parameters was improved as the sample size and the number of measurement occasions increased.

Since the bias and MSE values of item parameters were not consistent across conditions, ANOVA tests were conducted to examine the impact of design factors on them. When $MO = 3$, results found that the sample size had small to large effects on the recoveries on the intercept and main effects parameters ($\eta^2_{\lambda_{0\text{Bias}}} = 0.05$, $\eta^2_{\lambda_{\alpha\text{Bias}}} = 0.15$; $\eta^2_{\lambda_{0MSE}} = 0.67$, $\eta^2_{\lambda_{\alpha\text{Bias}}} = 0.74$). A large sample size was associated with good recoveries. The recoveries of interaction effect parameters were influenced by the sample size, the G matrix, and the growth pattern. The sample size had large effects on both the bias ($\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,bias}}$=0.66) and MSE ($\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,MSE}}$=0.53). Similarly, a large sample size resulted in better recoveries. Both the growth pattern and the G matrix design had small effects on interaction parameter recoveries (the growth pattern: $\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,bias}}$=0.02, $\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,MSE}}$=0.02; the G matrix: $\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,bias}}$=0.02, $\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,MSE}}$=0.02); the

TABLE 7 | Average correct classification rates for individuals who truly did not master attribute (MO = 5).

			T1			T2			T3			T4			T5		
			A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
G1	gam1	N100	1	1	1	0.87	0.83	0.79	0.88	0.87	0.85	0.92	0.90	0.89	0.94	0.94	0.93
		N200	1	1	1	0.86	0.83	0.79	0.87	0.87	0.86	0.92	0.91	0.90	0.94	0.94	0.93
		N300	1	1	1	0.86	0.82	0.79	0.87	0.87	0.86	0.91	0.91	0.90	0.94	0.94	0.93
	gam2	N100	1	1	0.99	0.91	0.84	0.79	0.91	0.87	0.86	0.94	0.91	0.90	0.96	0.94	0.93
		N200	1	1	1	0.91	0.83	0.79	0.91	0.87	0.85	0.94	0.91	0.89	0.95	0.93	0.93
		N300	1	1	1	0.90	0.82	0.78	0.90	0.87	0.85	0.93	0.91	0.90	0.96	0.94	0.93
G2	gam1	N100	1	1	0.99	0.86	0.82	0.80	0.86	0.86	0.85	0.91	0.90	0.90	0.94	0.93	0.93
		N200	1	1	1	0.85	0.82	0.79	0.87	0.86	0.85	0.91	0.91	0.89	0.94	0.94	0.93
		N300	1	1	1	0.85	0.82	0.79	0.86	0.85	0.85	0.91	0.90	0.89	0.93	0.92	0.92
	gam2	N100	1	1	0.99	0.90	0.85	0.80	0.90	0.86	0.85	0.94	0.91	0.90	0.96	0.94	0.94
		N200	1	1	1	0.90	0.82	0.79	0.90	0.87	0.86	0.94	0.92	0.90	0.95	0.94	0.93
		N300	1	1	1	0.89	0.82	0.79	0.90	0.87	0.85	0.93	0.91	0.89	0.94	0.92	0.91

T1–T3 represent the first to the third measurement occasion; A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3; N100, N200, and N300 represent the sample size of 100, 200, and 300, respectively; G1 and G2 represent equal correlation G and unequal correlation conditions of G matrix, respectively; gam1 and gam2 represent the same growth pattern across attributes and unequal growth patterns across attributes, respectively.

TABLE 8 | Average kappa (MO = 5).

			T1			T2			T3			T4			T5		
			A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3	A1	A2	A3
G1	gam1	N100	.	.	.	0.57	0.55	0.55	0.74	0.75	0.72	0.82	0.81	0.81	0.87	0.88	0.86
		N200	.	.	.	0.56	0.56	0.56	0.73	0.74	0.74	0.82	0.82	0.82	0.87	0.87	0.87
		N300	.	.	.	0.56	0.55	0.56	0.73	0.73	0.74	0.83	0.82	0.82	0.87	0.87	0.87
	gam2	N100	.	.	.	0.54	0.55	0.55	0.72	0.72	0.75	0.82	0.81	0.82	0.86	0.86	0.86
		N200	.	.	.	0.56	0.56	0.56	0.73	0.73	0.73	0.82	0.82	0.81	0.86	0.86	0.87
		N300	.	.	.	0.55	0.55	0.55	0.73	0.74	0.74	0.82	0.82	0.82	0.86	0.87	0.87
G2	gam1	N100	.	.	.	0.56	0.55	0.55	0.72	0.72	0.73	0.82	0.81	0.81	0.87	0.86	0.87
		N200	.	.	.	0.56	0.55	0.55	0.73	0.73	0.74	0.82	0.82	0.81	0.87	0.87	0.86
		N300	.	.	.	0.54	0.55	0.55	0.72	0.71	0.73	0.80	0.80	0.80	0.85	0.84	0.85
	gam2	N100	.	.	.	0.56	0.55	0.55	0.72	0.72	0.73	0.82	0.82	0.82	0.86	0.86	0.87
		N200	.	.	.	0.54	0.55	0.55	0.73	0.74	0.75	0.81	0.83	0.82	0.85	0.87	0.87
		N300	.	.	.	0.54	0.55	0.54	0.73	0.74	0.73	0.81	0.81	0.80	.	.	.

T1–T5 represent the first to the fifth measurement occasion; A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3; N100, N200, and N300 represent the sample size of 100, 200, and 300, respectively; G1 and G2 represent equal correlation G and unequal correlation conditions of G matrix, respectively; gam1 and gam2 represent the same growth pattern across attributes and unequal growth patterns across attributes, respectively; '.' presents the kappa for this condition was not applicable.

growth and the equal correlations conditions resulted in better recoveries.

When $MO = 5$, the item parameter recoveries were mainly influenced by the sample size. The sample size had small to large effects on the recoveries of intercept and main effects ($\eta^2_{\lambda_{0\text{Bias}}} = 0.01$, $\eta^2_{\lambda_{0\text{MSE}}} = 0.33$; $\eta^2_{\lambda_{\alpha\text{Bias}}} = 0.05$, $\eta^2_{\lambda_{\alpha\text{MSE}}} = 0.37$), and large effects on the recoveries of interaction effects ($\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,bias}}$=0.19, $\eta^2_{\lambda_{\alpha_k\alpha_{k'}\,MSE}}$=0.17). The parameter recoveries were improved as the sample size increased. In addition, the recoveries of intercept parameters were influenced by the growth pattern slightly. The non-growth condition had a slightly better intercept parameter recoveries, although the effect sizes were very small. For the sake of page limits, the details of ANOVA results could be found in the **Supplementary Material**.

In summary, the item parameter recoveries were mainly influenced by the sample size, especially for the interaction effect parameters. In general, the larger sample size resulted in the better item parameter recoveries.

Structural Model Parameter Recovery

Recoveries of both fixed effects and random effects in the growth model were evaluated in this study. The fixed effects included the intercept and slope parameters for each attribute ($\gamma^{A_k}_{00}$, $\gamma^{A_k}_{01}$), and the random effects included the variance of

TABLE 9 | Summary of measurement model parameter recoveries.

			Three measurement occasions (MO = 3)						Five measurement occasions (MO = 5)					
			λ_0		λ_{α_k}		$\lambda_{\alpha_k\alpha_{k'}}$		λ_0		λ_{α_k}		$\lambda_{\alpha_k\alpha_{k'}}$	
			Bias	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
G1	gam1	N100	0.03	0.05	−0.09	0.13	0.58	0.60	0.03	0.03	−0.07	0.07	0.31	0.24
		N200	0.02	0.03	−0.06	0.07	0.32	0.27	0.00	0.02	−0.03	0.04	0.15	0.11
		N300	0.02	0.02	−0.04	0.04	0.23	0.17	0.01	0.01	−0.03	0.03	0.11	0.07
	gam2	N100	0.04	0.05	−0.08	0.13	0.62	0.68	0.02	0.03	−0.07	0.07	0.33	0.27
		N200	0.03	0.03	−0.06	0.07	0.36	0.31	0.01	0.02	−0.03	0.04	0.16	0.11
		N300	0.01	0.02	−0.04	0.04	0.24	0.19	0.00	0.01	−0.02	0.03	0.09	0.07
G2	gam1	N100	0.04	0.05	−0.09	0.13	0.62	0.66	0.03	0.04	−0.07	0.07	0.31	0.25
		N200	0.02	0.03	−0.05	0.07	0.36	0.32	0.02	0.02	−0.04	0.04	0.17	0.11
		N300	0.01	0.02	−0.03	0.05	0.25	0.20	0.01	0.01	−0.02	0.03	0.10	0.07
	gam2	N100	0.03	0.05	−0.08	0.13	0.66	0.73	0.02	0.03	−0.06	0.07	0.34	0.28
		N200	0.03	0.03	−0.06	0.07	0.39	0.36	0.02	0.02	−0.04	0.04	0.18	0.13
		N300	0.01	0.02	−0.03	0.05	0.28	0.26	0.01	0.01	−0.02	0.03	0.11	0.08

λ_0, λ_{α_k}, and $\lambda_{\alpha_k\alpha_{k'}}$ represents the intercept, main effect, and interaction effect parameters of the LCDM; A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3; N100, N200, and N300 represent the sample size of 100, 200, and 300, respectively; G1 and G2 represent equal correlation G and unequal correlation conditions of G matrix, respectively; gam1 and gam2 represent the same growth pattern across attributes and unequal growth patterns across attributes, respectively.

intercept and slope parameters for each attribute ($\delta_{u_0^{A_k}}$, $\delta_{u_1^{A_k}}$) as well as the covariance among intercept and slope parameters ($\delta_{u_0^{A_k},u_0^{A_{k'}}}$ $\delta_{u_1^{A_k},u_1^{A_{k'}}}$ $\delta_{u_0^{A_k},u_1^{A_{k'}}}$).

Recovery of the fixed effects

Table 10 presents the summary of average bias and MSE of fixed effects under all conditions when $MO = 5$, which reveals that the proposed model achieved good recoveries on the intercept parameters for Attributes 2 and 3, and slope parameters for all attributes, indicated by the small MSE values and the bias values being close to zero. However, the intercept parameter of Attribute 1 had relatively larger bias than other parameters. When $MO = 3$, similar patterns were found, which can be found in the **Supplementary Material**.

The bias and MSE of intercept parameters were not consistent across different conditions, so ANOVA tests were conducted to investigate if the design factors influenced the intercept parameter recoveries for both $MO = 3$ and $MO = 5$ conditions. As shown in **Table 11**, when $MO = 3$, the sample size had small effects on the MSE values of intercept parameters ($\eta^{2^{A1}}_{\gamma_{00}} = 0.03$, $\eta^{2^{A2}}_{\gamma_{00}} = 0.04$, $\eta^{2^{A3}}_{\gamma_{00}} = 0.03$). A large sample size was associated with small MSE values. However, the bias of fixed effects was not influenced by the design factors.

When $MO = 5$, ANOVA tests found that the sample size had small effects on the MSE values of intercept parameters for Attribute 2 and 3 ($\eta^{2^{A2}}_{\gamma_{00}} = 0.01$, $\eta^{2^{A3}}_{\gamma_{00}} = 0.01$). Similarly, the bias of intercept parameters was not influenced by the design factors.

In summary, the intercept parameters of Attributes 2 and 3 and all the slope parameters were recovered well in the current study, but the intercept parameters of Attribute 1 had a relatively large bias. ANOVA tests found that the sample size had small effects on the MSE values of intercept parameters; a larger sample size resulted in smaller MSE values. However, no design factors were associated with the bias of intercept parameters.

Recovery of the random effects

Regarding the recovery of variance parameters, the average bias and MSE values of the variance of intercept and slope for all attributes were examined, the results reveal that the proposed model achieved good recoveries in both the intercept and slope variance parameters in both $MO = 3$ *and* $MO = 5$. The details of the summary of random variance recoveries could be found in the **Supplementary Material**.

Since bias of intercept variance parameters were not consistent across all conditions, ANOVA tests were conducted to examine the impact of design factors on them. As shown in **Table 12**, when $MO = 3$, results found that the sample size had medium effects on the bias of intercept variance parameters ($\eta^2_{\delta_{u_0^{A_1}}} = 0.14$; $\eta^2_{\delta_{u_0^{A_2}}} = 0.13$; $\eta^2_{\delta_{u_0^{A_3}}} = 0.11$); the large sample size had large bias values.

When $MO = 5$, similar patterns were found. The variance of intercept and slope parameters were recovered well. Since the recoveries of the variance of intercept parameters were varied by conditions, ANOVA tests were conducted to investigate the impact of design factors on them. As showed in **Table 12**, the sample size had small effects ($\eta^2_{\delta_{u_0^{A_1}}} = 0.02$; $\eta^2_{\delta_{u_0^{A_2}}} = 0.02$; $\eta^2_{\delta_{u_0^{A_3}}} = 0.02$); the larger sample size had larger bias values.

In summary, the proposed model achieved good recoveries on the variance of intercept and slope parameters. Moreover, a large

TABLE 10 | Summary of fixed effects recoveries (MO = 5).

			γ_{00}^{A1}		γ_{01}^{A1}		γ_{00}^{A2}		γ_{01}^{A2}		γ_{00}^{A3}		γ_{01}^{A3}	
			Bias	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**	**Bias**	**MSE**
G1	gam1	N100	−0.11	0.08	.	0.01	0.01	0.04	.	0.01	0.02	0.06	.	
		N200	−0.13	0.06	−0.01	.	0.02	0.05	.	.	0.05	0.05	.	.
		N300	−0.10	0.04	.	.	−0.01	0.03	0.01	.	0.03	0.04	−0.01	.
	gam2	N100	−0.12	0.08	0.01	0.01	0.01	0.06	.	0.01	0.01	0.05	.	0.01
		N200	−0.13	0.06	−0.02	0.01	0.02	0.05	0.01	.	0.05	0.06	0.01	0.01
		N300	−0.15	0.06	−0.01	.	.	0.04	0.01	.	0.10	0.04	−0.01	.
G2	gam1	N100	−0.14	0.09	−0.01	0.01	−0.01	0.07	.	0.01	0.07	0.08	0.01	0.01
		N200	−0.16	0.06	0.01	0.01	0.01	0.04	.	.	0.03	0.04	−0.01	0.01
		N300	−0.11	0.06	0.01	0.01	−0.02	0.04	0.01	.	0.01	0.03	−0.01	.
	gam2	N100	−0.12	0.07	−0.01	0.01	−0.03	0.07	−0.01	0.01	0.04	0.07	0.01	0.01
		N200	−0.13	0.07	0.01	.	−0.04	0.05	.	.	0.06	0.05	.	.
		N300	−0.11	0.05	.	.	−0.02	0.03	0.01	.	0.06	0.04	.	.

γ_{00}^{k} and γ_{01}^{k} represents the intercept and slope parameters of attributes; A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3; N100, N200, and N300 represent the sample size of 100, 200, and 300, respectively; G1 and G2 represent equal correlation G and unequal correlation conditions of G matrix, respectively; gam1 and gam2 represent the same growth pattern across attributes and unequal growth patterns across attributes, respectively; represents <0.001.

TABLE 11 | ANOVA results of fixed effects parameter recoveries.

	Three measurement occasions (MO = 3)							**Five measurement occasions (MO = 5)**						
		Bias			**MSE**				**Bias**			**MSE**		
Design factors	**Df**	**F**	η^2	**p**	**F**	η^2	**P**	**df**	**F**	η^2	**p**	**F**	η^2	**p**
γ_{00}^{A1}														
G	1	0.1	.	0.75	0.37	.	0.54	1	0.05	.	0.82	1.05	.	0.31
SZ	2	0.06	.	0.94	14.93	0.03	.	1	2.04	.	0.15	1.11	.	0.29
G×SZ	2	0.42	.	0.66	0.83	.	0.43	1	1.19	.	0.28	0.11	.	0.74
Residuals	1128		0.5			0.5		770		0.5			0.5	
γ_{00}^{A2}														
G	1	0.89	.	0.35	2.78	.	0.1	1	2.57	.	0.11	0.37	.	0.55
SZ	2	0.72	.	0.49	22.14	0.04	.	1	1.02	.	0.31	8.58	0.01	.
G × SZ	2	0.05	.	0.95	0.23	.	0.79	1	0.22	.	0.64	0.08	.	0.78
Residuals	1128		0.5			0.5		770		0.5			0.5	
γ_{00}^{A3}														
G	1	2.31	.	0.13	0.28	.	0.6	1	1.26	.	0.26	0.42	.	0.52
SZ	2	0.03	.	0.97	15.78	0.03	.	1	0.2	.	0.65	6.65	0.01	0.01
G × SZ	2	1.07	.	0.34	1.46	.	0.23	1	0.38	.	0.54	1.4	.	0.24
Residuals	1128		0.5			0.5		770		0.5			0.5	

G represents G matrix design; gamma represents the growth patterns; SZ represents the sample size; · represents <0.001.

sample size was associated with large bias values of the variance of intercept parameters.

Regarding the recovery of covariance parameters, on average, the proposed model achieved good recoveries on the covariance among intercept and slope parameters for both $MO = 5$ and $MO = 3$. However, the covariance between intercepts had a lightly larger bias than other sets of parameters. Details of the summary of covariance parameter recoveries could be found in the **Supplementary Material**.

When $MO = 3$, As shown in **Table 13**, ANOVA tests found that the sample size had medium effects ($\eta^2 = 0.13$) on the bias of covariance between intercept parameters; a large sample size was associated with a large bias.

Similar patterns were found when $MO = 5$, ANOVA tests showed the sample size had medium effects on the bias values of covariance between intercept parameters; a larger sample size was associated with a larger bias value.

On average, the proposed model achieved good recoveries on the covariance among intercept and slope parameters. The bias

TABLE 12 | ANOVA results of random variance parameter recoveries.

	Three measurement occasions (MO = 3)							Five measurement occasions (MO = 5)						
		Bias			MSE				Bias			MSE		
Design factors	***Df***	***F***	η^2	***p***	***F***	η^2	***p***	***Df***	***F***	η^2	***p***	***F***	η^2	***p***
γ_{00}^{A1}														
G	1	0.10	.	0.75	1.53	.	0.22	1	2.21	.	0.14	.	.	0.97
SZ	2	89.68	0.14	.	1.27	0.02	.	1	14.49	0.02	.	2.35	.	0.13
G × SZ	2	2.58	.	0.08	1.12	.	0.33	1	0.01	.	0.94	0.19	.	0.66
Residuals	1128		0.50			0.50		770		0.50			0.50	
γ_{00}^{A2}														
G	1	0.08	.	0.78	2.79	.	0.10	1	.	.	0.97	0.54	.	0.46
SZ	2	84.06	0.13	.	11.79	0.02	.	1	13.35	0.02	.	4.22	0.01	0.04
G × SZ	2	1.78	.	0.17	2.19	.	0.11	1	0.01	.	0.92	1.48	.	0.22
Residuals	1128		0.50			0.50		770		0.50			0.50	
γ_{00}^{A3}														
G	1	0.14	.	0.71	3.54	.	0.06	1	0.64	.	0.43	.	.	1.
SZ	2	72.67	0.11	.	7.26	0.01	.	1	14.08	0.02	.	6.48	0.01	0.01
G × SZ	2	3.01	0.01	0.05	2.37	.	0.09	1	0.08	.	0.77	0.59	.	0.44
Residuals	1128		0.50			0.50		770		0.50			0.50	

G represents G matrix design; gamma represents the growth patterns; SZ represents the sample size; · represents <0.001.

TABLE 13 | ANOVA results of random covariance parameter recoveries.

	Three measurement occasions (MO = 3)							Five measurement occasions (MO = 5)						
		Bias			MSE				Bias			MSE		
Design factors	**Df**	**F**	η^2	**p**	**F**	η^2	**p**	**df**	**F**	η^2	**p**	**F**	η^2	**p**
$\delta_{u_0^{A_k}, u_0^{A_{k'}}}$														
G	1	0.16	.	0.69	3.2	.	0.07	1	0.82	.	0.37	0.1	.	0.75
SZ	2	81.22	0.13	.	6.96	0.01	.	1	5.32	0.01	0.02	5.81	0.01	0.02
G × SZ	2	2.87	0.01	0.06	2.04	.	0.13	1	0.25	.	0.62	0.7	.	0.4
Residuals	1128		0.5				0.5		770		0.5		0.5	

G represents G matrix design; gamma represents the growth patterns; SZ represents the sample size; · represents <0.001.

of covariance among intercept parameters was influenced by the sample size; the larger sample size resulted in larger bias values.

DISCUSSION

Performance of the Multivariate Longitudinal DCM

Model Convergence

Overall, the proposed model achieved satisfactory convergence rates; however, the proposed achieved a slightly higher convergence rates when $MO = 5$ than $MO = 3$, which was reasonable since more measurement occasions would provide more information to help the estimation and the model be converged. Also, as shown in the **Supplementary Material**, the conditions with five measurement occasions had more chains and a longer chain length for each chain than the conditions with three measurement occasions, which might have led to an improvement in the convergence rates. Therefore, the number of chains and the chain length might be not sufficient for the conditions with three measurement occasions.

Classification Accuracy

The bias of the estimated probability of attribute mastery, the correct classification rates for each mastery status, and Cohen's kappa was used to evaluate the classification accuracy of the proposed model.

The probability of attribute mastery was recovered well in the current study consistently across all measurement occasions, which indicated that the proposed model could provide accurate estimates of probabilities of attribute mastery.

Regarding correct classification rates, results found different patterns for individuals who truly mastered the attributes and individuals who truly did not master the attributes. For the individuals who truly mastered the attributes, the correct classification rates improved significantly as the number of

measurement occasions increased. However, for individuals who truly did not master the attributes, the correct classification rates decreased slightly as the number of measurement occasions increased. This pattern might be due to that we adopted the cut point of 0.5 to classify the individuals. Since the estimated probabilities of attribute mastery for most of the individuals were lower than 0.5 on the first two measurement occasions, individuals would be classified into the non-mastery status, even some of them truly mastered the attributes by design. As a result, the correct classification rates were low for individuals who truly mastered the attributes on the first two measurement occasions. As the number of measurement occasions increased, the estimated probability of mastery increased, such that correct classification rates increased. Due to the same reason, Cohen's kappa increased as the number of measurement occasions increased. Therefore, the cutpoint had influenced the correct classification rates and kappa values of the current model.

Parameter Recoveries

The bias and mean square error (MSE) of the estimated parameters were computed to assess the parameter recovery in each condition through the simulation.

Measurement model parameter recoveries

Regarding the item parameter recoveries, conditions with three and five measurement occasions illustrated similar patterns. The proposed model achieved good parameter recoveries in intercept and main effect parameters, but poor interaction effect parameter recoveries. However, the recoveries of the interaction effect parameters were improved as the sample size and the number of measurement occasions increased. In addition, results from the ANOVA tests found the sample size had large impact on the interaction effects recoveries. Nonetheless, this result was expected. Previous research showed that the intercept and main effect parameters were easier to recover than the two-way interaction effect parameters. The recoveries of the interaction effect parameters were problematic when the sample size was <1,000 (e.g., Choi et al., 2010; Kunina-Habenicht et al., 2012). Therefore, these results suggested that a large sample size was necessary to achieve good item parameter recoveries in the LCDM framework, especially for the interaction effect parameters. The maximum sample size ($n = 300$) in the current study was not sufficient for obtaining accurate interaction effect parameters, especially for the conditions with three measurement occasions.

Structural model parameter recoveries

Both the recoveries of fixed effects and random effects in the generalized growth curve model were evaluated.

Regarding the recoveries of the fixed effects, overall, the proposed model achieved good intercept recoveries for Attributes 2 and 3, and slope recoveries for all attributes, but relatively poor recoveries for Attribute 1 intercept. Attribute 1 had relatively small intercept value by design ($\gamma_{00}^{A1} = -1.38$), therefore, the small intercept value might have led to enlarge the bias. To avoid the influence of the small value of the intercept parameter, the time variable could be centered at the medial measurement occasions ($T = 2$ when $MO = 3$, or $T = 3$ when $MO = 5$), such that there would be sufficient information to estimate the intercept parameters.

Regarding the recoveries of the random effects, on average, the proposed model achieved good recoveries on the random effects, including the variance of intercept and slope parameters of each attribute as well as the covariance among intercept and slope parameters within and crossed attributes. To improve the model convergence, the current study adopted the true variance-covariance matrix in the population as the prior of the estimated variance-covariance matrix, which might have led to good recoveries of the random effects.

Conclusion and Recommendations

The current study developed a multivariate longitudinal DCM that could measure growth in attributes over time, and it evaluated this proposed model using a simulation study. The results revealed the following: (1) In general, the proposed model provided good convergence rates under different conditions. (2) Regarding the classification accuracy, the proposed model achieved good recoveries on the probabilities of attribute mastery. For individuals who truly mastered the attributes, the correct classification rates increased as the measurement occasions increased; however, for individuals who truly did not master the attributes, the correct classification rates decreased slightly as the numbers of measurement occasions increased. Cohen's kappa increased as the number of measurement occasions increased. (3) Both the intercept and main effect parameters in the LCDM were recovered well. The interaction effect parameters had a relatively large bias under the condition with a small sample size and fewer measurement occasions; however, the recoveries were improved as the sample size and the number of measurement occasions increased. (4) Overall, the proposed model achieved acceptable recoveries on both the fixed and random effects in the generalized growth curve model.

In summary, a large sample size is recommended for applying the proposed model to the real data. When the sample size is small, the scale with a simple structure of the Q matrix is recommended, because the interaction effects in the LCDM might not be estimated accurately with the small sample size. Also, applied researchers are suggested to center the time variable at the medial measurement occasion to improve the recovery of the intercept parameter in the generalized growth curve model. Additionally, when doing the MCMC analysis, multiple chains with the longer chain length are recommended to achieve satisfied model convergence rates.

Therefore, when practitioners try to measure students' growth in the DCM framework using the proposed model, they should use a larger sample size, an assessment with less complex Q-matrix design, and multiple chains with longer chain length to maximize the convergence rates and the accuracy of parameter estimates.

Contributions and Limitations

In the current study, a multivariate longitudinal DCM was developed to analyze longitudinal data under the DCM framework. It represents an improvement in the current

longitudinal DCMs given its ability to incorporate both balanced and unbalanced data and to measure the growth of a single attribute directly without assuming that attributes grow in the same pattern. The current study had several limitations. First, the true variance-covariance matrix was used as the prior for the random effects parameters in the generalized growth curve model in the current study; however, the true variance-covariance matrix is unknown when fitting the model to the real data. Therefore, future studies could adopt a non-informative variance-covariance matrix as the prior, then evaluate if the proposed model could achieve satisfying recoveries on the random effects as well. Second, local item dependency was not incorporated in the current study. However, in real longitudinal data, repeated measures always have some degree of local item dependence (e.g., Cai, 2010). Therefore, future research could simulate local item dependence with the common items to mimic real data. Third, only three or five measurement occasions were included in the current model. The small number of measurement occasions might have limited the reliability and accuracy of the estimation of the growth curve model (e.g., Finch, 2017). In the future, more measurement occasions could be included to examine the performance of the proposed model comprehensively. Fourth, the definition of the time variable in longitudinal studies is very crucial. In the current study, we follow a conventional way to use the length of time between adjacent measurement occasions as the time variable. However, in reality, students likely have spent different lengths of time learning different attributes. So, in the future, we suggest using the number of hours spent on learning an attribute as the time variable if the data is available. In addition, we applied the cut-score to the average of the post burn-in probability of master to obtain a binary master status of one iteration on each condition, meaning that we cannot obtain a posterior distribution of the mastery status. So, we suggest future researchers applying the cut-score within MCMC analysis to obtain a posterior distribution of mastery status, which should provide a more accurate estimated mastery status. Last but not least, due to the limited data resources, we did not find a real dataset to evaluate the proposed model. We plan to add a real data application if some longitudinal diagnose assessment data is available in the future.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

QP drafted the manuscript, conducted, and interpreted the statistical analyses. LQ reviewed the manuscript and provided expertise on data analyses. NK supervised and reviewed the paper. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

Parts of this study first appeared in QP's dissertation. We thank Dr. Jonathan Templin and Dr. Lesa Hoffman for sharing their wisdom with us in revising the JAGS code and research design, which improved the manuscript greatly.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01714/full#supplementary-material

REFERENCES

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

Bradshaw, L., and Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *Int. J. Test.* 16, 99–118. doi: 10.1080/15305058.2015.1107076

Bradshaw, L., and Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika* 79, 403–425. doi: 10.1007/s11336-013-9350-4

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika* 75, 581–612. doi: 10.1007/s11336-010-9178-0

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

Choi, H.-J., Templin, J., Cohen, A., and Atwood, C. (2010). "The impact of model misspecification on estimation accuracy in diagnostic classification models," in *Paper Presented at the National Council on Measurement in Education (NCME)* (Denver, CO).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. Hilsdale, NJ: Lawrence Earlbaum Associates.

Collins, L. M., and Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behav. Res.* 27, 131–157. doi: 10.1207/s15327906mbr2701_8

Curran, P. J., Mc Ginley, J. S., Serrano, D., and Burfeind, C. (2012). "A multivariate growth curve model for three-level data," in *APA Handbook of Research Methods in Psychology, Vol. 3*, ed H. Cooper (Washington, DC: American Psychological Association), 335–358. doi: 10.1037/13621-017

de La Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Appl. Psychol. Meas.* 33, 163–183. doi: 10.1177/0146621608320523

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

DiBello, L., Stout, W., and Roussos, L. (1995). "Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques," in *Cognitively Diagnostic Assessment*, eds P. Nichols, S. Chipman, and R. Brennan (Hillsdale, NJ: Erlbaum).

Dynamic Learning Maps (2016). *2014–2015 Technical Manual: Integrated Model*. Lawrence, KS: University of Kansas.

Finch, W. H. (2017). Investigation of parameter estimation accuracy for growth curve modeling with categorical indicators. *Methodology* 13, 98–112. doi: 10.1027/1614-2241/a000134

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

George, A. C., and Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychol. Test Assess. Model.* 56, 405–432.

Goldstein, H. (2011). *Multilevel Statistical Models*, *Vol. 922*. New York, NY: John Wiley and Sons. doi: 10.1002/9780470973394

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61, 215–231. doi: 10.1093/biomet/61.2.215

Great Schools Partnership (2013). *Interim Assessment*. Available online at: http://edglossary.org/interim-assessment/ (accessed October 30, 2013).

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality*. Champaign; Urbana: University of Illinois at Urbana-Champaign.

Hoffman, L. (2015). *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change*. New York, NY: Routledge. doi: 10.4324/9781315744094

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *J. Educ. Meas.* 54, 440–480. doi: 10.1111/jedm.12156

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kaya, Y., and Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. *Educ. Psychol. Meas.* 77, 369–388. doi: 10.1177/0013164416659314

Kunina-Habenicht, O., Rupp, A. A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *J. Educ. Meas.* 49, 59–81. doi: 10.1111/j.1745-3984.2011.00160.x

Lanza, S. T., Flaherty, B. P., and Collins, L. M. (2003). "Latent class and latent transition analysis," in *Handbook of Psychology*, eds J. A. Schinka and W. F. Velicer (John Wiley & Sons, Inc.). doi: 10.1002/0471264385.wei0226

Lazarsfeld, P., and Henry, N. (1968). Latent *S*tructure *Analysis*. Boston, MA: Houghton Mifflin Company.

Lee, Y.-W., and Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Lang. Assess. Q.* 6, 239–263. doi: 10.1080/15434300903079562

Leighton, J., and Gierl, M. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511611186

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

MacCallum, R. C., Kim, C., Malarkey, W. B., and Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivar. Behav. Res.* 32, 215–253. doi: 10.1207/s15327906mbr3203_1

Madison, M. J., and Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika* 83, 963–990. doi: 10.1007/s11336-018-9638-5

Pan, Q. (2018). *Growth modeling in a Diagnostic Classification Model (DCM) framework* (Ph.D.). The University of Kansas, Lawrence, KS, United States.

Plummer, M. (2003). "JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (Vienna). Retrieved from: http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/Plummer.pdf

Preacher, K. J., Wichman, A., MacCallum, R. C., and Briggs, N. E. (2008). "Latent growth curve modeling," in *Quantitative Applications in the Social Sciences*, ed V. Knight (Thousand Oaks, CA: Sage), 71–79. doi: 10.4135/9781412984737

R Core Team (2017). *R: A Language and Environment for Statistical Computing (Version 3.4.2)*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed June 22, 2020).

Raghavarao, D., and Padgett, L. (2014). "Growth curve models," in *Repeated Measurements and Cross-Over Designs*, eds D. Raghavarao, and L. Padgett (Hoboken, NJ: John Wiley and Sons, Inc.), 77–104. doi: 10.1002/9781118709153.ch4

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assess.* 34, 782–799. doi: 10.1177/0734282915623053

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

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

Sedat, S., and Arican, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Egitimde ve Psikolojide Ölçme ve Degerlendirme Dergisi* 6, 238–253. doi: 10.21031/epod.65266

Sinharay, S. (2003). Assessing convergence of the Markov Chain Monte Carlo algorithms: a review. *ETS Res. Rep. Ser.* 2003:i-52. doi: 10.1002/j.2333-8504.2003.tb01899.x

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

Templin, J., and Bradshaw, L. (2014). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika* 79, 317–339. doi: 10.1007/s11336-013-9362-0

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

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 September 2020
doi: 10.3389/fpsyg.2020.02230

International Comparative Study on PISA Mathematics Achievement Test Based on Cognitive Diagnostic Models

Xiaopeng Wu[1,2,3†], Rongxiu Wu[4†], Hua-Hua Chang[3], Qiping Kong[2] and Yi Zhang[5*]*

[1] School of Education, Shaanxi Normal University, Xi'an, China, [2] College of Teacher Education, Faculty of Education, East China Normal University, Shanghai, China, [3] College of Education, Purdue University, West Lafayette, IN, United States, [4] College of Education, University of Kentucky, Lexington, KY, United States, [5] School of Mathematic Science, East China Normal University, Shanghai, China

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Yi Zheng,
Arizona State University, United States
Yutong Wang,
National Institute for Education Sciences, China

***Correspondence:**
Xiaopeng Wu
wuxp@snnu.edu.cn;
wxpecnu@outlook.com
Yi Zhang
52185500031@stu.ecnu.edu.cn

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

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

Received: *18 May 2020*
Accepted: *10 August 2020*
Published: *09 September 2020*

Citation:
Wu X, Wu R, Chang H-H, Kong Q and Zhang Y (2020) International Comparative Study on PISA Mathematics Achievement Test Based on Cognitive Diagnostic Models. Front. Psychol. 11:2230. doi: 10.3389/fpsyg.2020.02230

As one of the most influential international large-scale educational assessments, the Program for International Student Assessment (PISA) provides a valuable platform for the horizontal comparisons and references of international education. The cognitive diagnostic model, a newly generated evaluation theory, can integrate measurement goals into the cognitive process model through cognitive analysis, which provides a better understanding of the mastery of students of fine-grained knowledge points. On the basis of the mathematical measurement framework of PISA 2012, 11 attributes have been formed from three dimensions in this study. Twelve test items with item responses from 24,512 students from 10 countries participated in answering were selected, and the analyses were divided into several steps. First, the relationships between the 11 attributes and the 12 test items were classified to form a Q matrix. Second, the cognitive model of the PISA mathematics test was established. The liner logistic model (LLM) with better model fit was selected as the parameter evaluation model through model comparisons. By analyzing the knowledge states of these countries and the prerequisite relations among the attributes, this study explored the different learning trajectories of students in the content field. The result showed that students from Australia, Canada, the United Kingdom, and Russia shared similar main learning trajectories, while Finland and Japan were consistent with their main learning trajectories. The primary learning trajectories of the United States and China were the same. Furthermore, the learning trajectory for Singapore was the most complicated, as it showed a diverse learning process, whereas the trajectory in the United States and Saudi Arabia was relatively simple. This study concluded the differences of the mastery of students of the 11 cognitive attributes from the three dimensions of content, process, and context across the 10 countries, which provided a reference for further understanding of the PISA test results in other countries and shed some evidence for a deeper understanding of the strengths and weaknesses of mathematics education in various countries.

Keywords: PISA, cognitive diagnosis, educational evaluation, international comparison, mathematics education

INTRODUCTION

Initiated by the Organization for Economic Cooperation and Development (OECD) in 1997, the Program for International Student Assessment (PISA) is held every 3 years to assess the fundamental knowledge and critical competencies needed for students approximately 15 years old to participate in society. PISA emphasizes the abilities of students in reasoning from school knowledge and the application of the knowledge to environments outside school (OECD, 2019a). As one of the most influential educational assessment programs globally, PISA has had a large impact on educational practice and reform in many countries by increasing the scopes of tests and strengthening the interpretation of results, thus influencing the decision-making processes for the improvement of national education policies (Breakspear, 2012; OECD, 2013b). For example, the results from PISA 2000 have given rise to a national "PISA shock" in Germany, which has led to massive and rapid educational reforms (Ertl, 2006). Similar educational impacts have also happened in Japan (Takayama, 2008), Denmark (Egelund, 2008), Finland (Dobbins and Martens, 2012), and a number of other European countries (Grek, 2009). The United States, Russia, Japan, and other countries have successively formulated a series of education policies and regulations, forming education quality standards to strengthen the monitoring of the quality of education in the stage of compulsory education. Borrowing from the assessment method of PISA, Singapore has changed the national education assessment model and indicated a new direction for the reform of the national education assessment (Stacey et al., 2015). Mathematics, as one of the core tests in PISA, has also been extensively studied; for instance, educational equity issues have been studied through assessing the opportunities of learning for students (Duru-Bellat and Suchaut, 2005; Luyten, 2017; Hansen and Strietholt, 2018), the gender differences in PISA performance (Steinthorsdottir and Sriraman, 2008; Kyriakides et al., 2014), PISA performance differences in age (Sprietsma, 2010), the relationship between PISA performance and social achievement (Knowles and Evans, 2012), the influence of language on PISA performance (El Masri et al., 2016), the heterogeneity of PISA performance (Wößmann, 2005), etc. However, these studies have focused on either the factors that affect PISA achievements or the impact of PISA achievements on society and education. Few studies have analyzed PISA items, possibly because the PISA items are rarely open to the public. The analyses of the characteristics of mathematics education in different countries through PISA items are of indispensable significance to promote the reform and advancement in mathematics education. To improve the development, mathematics educators, mathematicians, measurement experts, and educational statisticians have been advised to collaborate in research projects to recognize the potential values of concept discussions and secondary analyses that are directly applicable to the existing school systems (Ferrini-Mundy and Schmidt, 2005).

PISA uses item response theory (IRT) in its scaling to overcome the limitations of scoring methods based on number correct or percentage correct. To report the population mean of each subscale, plausible values have been drawn from *a posteriori* distribution by combining the IRT scaling of the test items with a latent regression model using information from the student context questionnaire in a population model (OECD, 2015). Such design is ideal for obtaining accurate rankings for each participating country. However, providing the diagnostic information on the mastery or non-mastery of the examinees of each skill being measured may not be efficient. Under this context, cognitive diagnostic models (CDMs) have risen as advanced psychometric models to support the next-generation assessments aimed at providing fine-grained feedback for students and teachers in the past few decades (Leighton and Gierl, 2007; Templin and Bradshaw, 2014; Chang et al., in press). Researchers have called for additional measurement approaches for reporting and interpreting PISA results (Rutkowski and Rutkowski, 2016). Combining modern statistical methods with cognitive theories, CDMs have been widely utilized in educational and psychological assessment. One of the advantages of using CDMs is their ability to identify the strengths and weaknesses in a set of fine-grained skills (or attributes) when difficulty exists in inferring skill mastery profiles of examinees through traditional methods, such as classical test theory (CTT) and IRT (Choi et al., 2015). Therefore, CDMs have been developed to provide fine-grained information for researchers and educators on the cognitive skills or attributes that are required to solve a particular item, allow applications in various instructional practices, and resolve the limitations that exist in the IRT and CTT models (De La Torre, 2009). By integrating the test objectives into the cognitive process model, CDMs have gained increased attention among the educational and psychological assessments recently (Stout, 2002; Tatsuoka, 2002; Chen and Chen, 2016). Moreover, they can reflect the psychological and cognitive characteristics of the subjects (Templin and Henson, 2010). In the field of mathematics education, diverse cognitive models of mathematics learning and teaching have been developed (Carpenter and Moser, 1982; Greeno, 1991; Rumelhart, 1991; Schneider and Graham, 1992; Zhan et al., 2018) and validated by empirical evidence. It lays a foundation for CDMs that provide the measurement and diagnoses in mathematics educational issues.

The objective of the research is to employ a CDM as an analytic tool to analyze the data set consisting of 10 countries, including China, the United States, Russia, the United Kingdom, Japan, Finland, Singapore, and Australia on the basis of the PISA test contents. The research finding will be based on the mastery levels for the 11 attributes from three aspects, content, process, and context. Through exploring the knowledge states and learning trajectories of the 11 attributes, the study provides new information about mathematics education in the 10 countries regarding the strengths and weaknesses of each the 11 attributes in the study.

COGNITIVE MODEL CONSTRUCTION

Given that PISA tests the fundamental knowledge and key competence necessary for students to participate in the future, the test items are all carried out in specific realistic situations. As

far as the mathematics test items are concerned, students need to apply the mathematical knowledge and skills they have learned to solve a practical problem comprehensively. It has a detailed description of the test items. Therefore, an in-depth cognitive diagnostic analysis of the measurement results can be performed according to the existing coding.

Cognitive Attributes

Attributes play fundamental core roles in cognitive diagnosis measurement. The quality of attributes is directly related to the effectiveness of the cognitive diagnostic evaluation. To some extent, the essence of a cognitive diagnosis is the diagnosis of cognitive attributes. No uniform definition has been given regarding the cognitive attributes in the field of measurement. Attributes are productive rules, project types, program operations, general cognitive tasks (Tatsuoka, 1987), or posited knowledge and thinking skills (Tatsuoka et al., 2004); a description of the procedures, skills, processes, strategies, and knowledge a student must possess to solve a test item (Dogan and Tatsuoka, 2008); or the processing skills and knowledge structure required to complete a certain task (Leighton and Gierl, 2007). The attributes may be of a different nature; they may also be the knowledge, strategies, skills, processes, and methods necessary to complete the task, which is a description of the internal processing of the psychology of students in problem-solving (Cai et al., 2018). Conclusively, the cognitive attribute can be taken as a way of classification to understand the knowledge states of students more precisely on the basis of a certain standard (Wu et al., 2020). According to the definitions of cognitive attributes and the test items provided by the PISA assessment framework, each test item in PISA is defined from three aspects (dimensions), namely, the main subject area involved in the test question, the main mathematical process of problem-solving, and the contexts the test questions are based on (OECD, 2019b). Therefore, the cognitive attributes of PISA test questions can be constructed according to the definition of these three dimensions. We define the term attribute as a mathematical skill or content knowledge that is required to solve a test item. The dimensions, attributes in each dimension, and the corresponding definitions are shown in **Table 1**.

In **Table 1**, the four attributes in the content dimension include almost all the mathematics content in the stage of compulsory education. This division is relatively clear in maintaining a consistent granularity in the various parts. The three attributes of the mathematical process are the same as the reality, mathematization, and recreation described by the famous mathematician Freudenthal (2012). Mathematical operation is the process of recreation in the field of mathematics, and it is an important method in searching for the essential relationship through a superficial phenomenon. The context attributes include each field that students can encounter in the future, and it is an important carrier for training students to see the world with the "eyes" of mathematics.

Q-Matrix

Many test items have been included in PISA so far. However, in terms of mathematical tests, only test items publicized in 2012 are available, and no items can be obtained from other years. Even though PISA 2012 has many items, there are only 12 of them jointly tested by the students in the 10 countries we studied. Therefore, this study has selected 12 test items in PISA 2012 for cognitive diagnostic analysis. In PISA, each mathematics item is intended to target all three attributes in one dimension, which can be considered as a latent construct or dimension (OECD, 2014a). The Q-matrix in the cognitive diagnostic assessment we have constructed is a matrix used to connect test items and cognitive attributes, in which 1 represents the corresponding attribute that is considered in the test item, and 0 is the opposite. The Q-matrix has built a bridge between the observable responses of students and their unobservable cognitive states (Tu et al., 2019). According to the mark of the test item in the PISA 2012 manual, the Q-matrix is obtained, as shown in **Table 2**.

MODEL SELECTION AND INSTRUMENT ANALYSIS

Participants

In this study, the 12 items in PISA 2012 were selected, and the students who completed these 12 items all at the same time were selected as the research objects across the globe. The participants were from the United Kingdom (GBR, 3,811), Finland (FIN, 2,661), and Russia (RUS, 1,666) in Europe; China (CHI, 1,763, including the data selected from Hong Kong, Macau, Shanghai, and other places), Japan (JPN, 1,904) and Singapore (SGP, 1,667) in Asia; the United States (USA, 1,630) and Canada (CAN, 6368) in North America; Australia (AUS, 4,342) in Oceania, and Saudi Arabia in Africa (ALB, 1,402). Given that Brazil, Chile, Colombia, Argentina, and other countries that participated in the PISA 2012 math test in South America did not participate in these 12 tests, no comparable data from South America were available, and no data from Antarctica could be obtained either. The maximum representativeness of the data selection was reached.

Model Selection

Researchers have developed hundreds of measurement models since the cognitive diagnostic assessment theory was proposed. Measurement models are based on different hypotheses, parameters, mathematical principles, and actual situations. Therefore, the comparison and selection of models have played a vital role in the cognitive diagnosis and evaluation process. A large number of cognitive diagnosis practices have shown that choosing an appropriate cognitive diagnostic model is an important prerequisite for an accurate diagnosis or classification of subjects (Tatsuoka, 1984). To obtain a model with a better fit, this study evaluates the parameters of eight models, namely, DINA (Haertel, 1989; Junker and Sijtsma, 2001; De La Torre, 2009), DINO (Templin and Henson, 2006, 2010), RRUM (Hartz, 2002), ACDM (De La Torre, 2011), LCDM (Henson et al., 2009), LLM (Hagenaars, 1990, 1993; Maris, 1999), G-DINA (De La Torre, 2011), and Mixtures Model (von Davier, 2010). Using the LLM and GDINA packages (version 2.8.0) in software R, 2, 451 datasets for model comparisons are selected from 10 countries through the stratified sampling at a ratio of 10:1 in each country.

TABLE 1 | Dimensions of PISA's cognitive attributes.

Dimension	No.	Attribute	Definition
Content	N1	Change and relationships	Use mathematical language such as algebraic expressions, equations, functions, inequalities to describe the relationship between quantity and graphic
	N2	Space and shape	Mainly involves the relationship between planes, points, lines, and planes in space, and the virtual rotation of graphics, etc.
	N3	Quantity	Quantity integrates the quantification of the attributes of objects, relationships, situations, and entities in the world, understands the various manifestations of these quantifications, and judges, interprets, and demonstrates the quantity
	N4	Uncertainly and data	Perception of change, probability and opportunity, representation, evaluation, interpretation of uncertainty-centric data
Process	P1	Mathematization	Use mathematical language to describe and explain problems in real life, and convert relevant information into mathematical quantities
	P2	Mathematical operation	Use mathematical concepts, facts, procedures, and reasoning to identify, calculate, reason, and analyze problems
	P3	Mathematical reality	Ability to apply the results of mathematical solutions to real problems and make assessments and inferences on the results
Contexts	C1	Personal	The project's involvement is based on personal scenarios, mainly focused on the activities of individuals, families or peers
	C2	Occupational	Involving various fields of future work, career scenarios may be related to any level of the workforce, from unskilled jobs to high-level occupational jobs
	C3	Societal	Social issues are concentrated in one's community, the focus of the problem is the community perspective
	C4	Scientific	Problems in the scientific category involve the application of mathematics in nature, as well as problems and topics related to science and technology

TABLE 2 | Q-matrix of 12 test items in PISA.

Items	Attributes										
	N1	N2	N3	N4	P1	P2	P3	C1	C2	C3	C4
PM00QF01	0	0	1	0	0	0	1	1	0	0	0
PM903Q03	1	0	0	0	0	1	0	0	1	0	0
PM918Q01	0	0	0	1	0	0	1	0	0	1	0
PM918Q02	0	0	0	1	0	0	1	0	0	1	0
PM918Q05	0	0	0	1	0	1	0	0	0	1	0
PM923Q01	0	0	1	0	0	1	0	0	0	0	1
PM923Q03	0	1	0	0	0	1	0	0	0	0	1
PM923Q04	1	0	0	0	1	0	0	0	0	0	1
PM924Q02	0	0	1	0	1	0	0	1	0	0	0
PM995Q01	0	1	0	0	0	1	0	0	0	0	1
PM995Q02	0	1	0	0	1	0	0	0	0	0	1
PM995Q03	0	0	1	0	1	0	0	0	0	0	1

The comparison results on parameter statistics, such as deviation, Akaike's information criterion (AIC), and Bayesian information criterion (BIC) are shown in **Table 3** below.

In **Table 3**, the number of parameters represents the load in the operation of the model, which is closely related to the complexity of the Q-matrix and its attributes. The smaller the number, the smaller the load in the model comparisons. Deviation represents how much an indicator deviates from reality in the model. The smaller the deviation, the greater the degree to which the model fits. In the model comparisons, the AIC and the BIC are mainly used as the reference standards. The AIC is for measuring the goodness of statistical model fit, which is based on the concept of entropy and provides a standard that weighs the complexity of the estimated model and the goodness of the fitted data. The smaller the AIC is, the better the data fits the model. Similarly, the smaller the BIC is, the better the data fits the model (Vrieze, 2012). The results in **Table 3** show that the values for Deviation, BIC and AIC of the LLM are the smallest. Therefore, the LLM has a better fit than those in the other models and was preliminarily selected.

Effectiveness Analysis of the Instrument

Reliability

The reliability of the cognitive diagnostic evaluation can be examined from two aspects. One is to treat the test as a common test, and Cronbach's (α) coefficient is calculated under classic evaluation theory (CTT). The other is to calculate the consistency of the retest of attributes. In our study, we

TABLE 3 | Parameter statistics comparison of different models.

Models	Number of parameters	Deviation	AIC	BIC
DINA	2071	29550.05	33692.05	45712.66
DINO	2071	29550.46	33692.45	45713.05
RRUM	2095	28758.58	32948.32	45108.22
ACDM	2095	28791.63	32981.18	45141.09
LCDM	2143	28484.78	32770.78	45209.29
LLM	2095	28229.28	32419.22	44579.13
G-DINA	2143	28498.63	32784.62	45223.13
Mixed Model	2097	28498.02	32691.94	44863.65

*Red rectangular box means the smallest index.

followed Templin and Bradshaw (2013) to estimate the test-retest reliability for our test by simulating repeated testing occasions through repeated draws from an examinee's posterior distribution. A three-step process is usually used for binary attributes, relying upon the correlation of the mastery statuses between two hypothetical independent administrations of the same test. $\alpha = 0.7687 > 0.7$, which is an indication of high reliability under CTT theory. The above index of 11 attributes are 0.8941, 0.8372, 0.9124, 0.8541, 0.8193, 0.8512, 0.8135, 0.7942, 0.8135, 0.9721, and 0.9014 accordingly. Data indicators are obtained through the flexCDMs analysis platform (Tu, 2019). The reliability indexes of these attributes are all greater than 0.7. Therefore, they have a high degree of reliability in general.

Item Discrimination

Cognitive diagnostic assessment measures the accuracy of cognitive attribute analysis and the quality of test items through item discrimination (Wang et al., 2018). The discrimination degree of the cognitive diagnostic test d_j is defined as

$$d_j = P_j(1) - P_j(0),$$

where $P_j(1)$ refers to the probability of mastering all attributes of item j when answering the question. $P_j(0)$ refers to the probability of answering the question correctly without mastering all the attributes of item j. The smaller d_j is, the smaller the impact of mastering attributes on the answer is, and the smaller the difference is. In contrast, the difference is greater. A large degree of discrimination is a sign of high-quality test questions. The item discrimination d_j of the 12 items in this study are in turn equal to 0.902, 0.8497, 0.6901, 0.3174, 0.5758, 0.7457, 0.7716, 0.5213, 0.5912, 0.8078, 0.6142, and 0.5721. All the item discriminations are acceptable except for the fourth item, which is 0.3174. The item discrimination for items 1, 2, 6, 7, and 10 are all greater than 0.7, which has a good discrimination effect.

RESEARCH ANALYSIS AND RESULTS

According to the results of the above model selection, the LLM had the best model fit. Therefore, the LLM was used to evaluate the parameters of the research data. The Bayesian expected *a posteriori* estimation (EAP) was used in the process. The Bayesian method attempts to calculate the posterior mean or median rather than a certain extreme value—the mode, the characteristic of which was to use posterior distribution to summarize the data and determine the inference. The posterior estimation is expected to be simple, efficient, and stable, and it is a better choice in the capacity parameter estimation method (Chen and Choi, 2009). The distribution of these 11 attributes from the 24,512 students were assessed initially. Then, the distribution of the attribute in each country was measured. The results for the proportional distribution of the 11 attributes in the 10 countries are in **Table 4**.

The following discussions are the analyses of the proportional and knowledge states of attribute mastery through the three dimensions of content, process, and context. The proportional distribution of attribute mastery can reflect the differences of attributes in all the countries. Knowledge states can help understand the mastery mode of the attributes of students in different countries and further speculate on the learning trajectories of students.

Comparative Analysis of Attribute Mastery Probability

Content Attribute

The PISA math test involves four content aspects, namely, change and relationships, space and shape, quantity, data and uncertainty, each of which accounts for one quarter of the test (OECD, 2013a). These four overarching ideas ensure the assessment of a sufficient variety and depth of mathematical content and demonstrate how phenomenological categories relate to more traditional strands of mathematical content (OECD, 2010). Almost all content in the junior high school learning has been covered. The probability of mastery of the 10 countries [Saudi Arabia in Africa (ALB), Australia (AUS), Canada (CAN), China (CHI), Finland (FIN), the United Kingdom (GBR), Japan (JPN), Russia (RUS), Singapore (SGP), the United States (USA)] of the four attributes is shown in **Figure 1**.

As can be seen from the distribution in **Figure 1**, China performed best in the three attributes of N1 (change and relationships), N2 (space and shape), and N3 (quantity), and it scored much higher than other countries. In the N4 (data and uncertainty) attribute, Japan performed best, and China was second only to Japan. In contrast, Chinese students still had much room for improvement in the study of N4 (data and uncertainty). Moreover, students from China, Singapore, Japan, Finland, and other countries had advantages in grasping each content attribute compared with those in other countries, such as the United States and Saudi Arabia, who showed evident weakness in the content attribute. The result was also consistent with the overall ranking of PISA (OECD, 2014b). In terms of the distribution of the four attributes, all countries performed better in the N4 (data and uncertainty) attribute than in the other three attributes. The United Kingdom, Finland, Saudi Arabia, and Australia had a low level of mastery of the N1 (change and relationships) attribute, less than 30%, and the probability of mastery was less than half of that of China, Singapore, and other countries. The United States performed relatively poorly on the

TABLE 4 | Proportional distribution of 11 attributes in 10 countries.

Country	Content attribute				Process attribute			Context attribute			
	N1	N2	N3	N4	P1	P2	P3	C1	C2	C3	C4
Saudi Arabia	0.246	0.258	0.237	0.388	0.233	0.259	0.248	0.223	0.183	0.381	0.238
Australia	0.273	0.427	0.466	0.699	0.436	0.445	0.526	0.460	0.244	0.701	0.445
Canada	0.411	0.467	0.523	0.742	0.491	0.500	0.578	0.509	0.330	0.741	0.488
China	0.740	0.743	0.749	0.882	0.744	0.753	0.771	0.754	0.693	0.796	0.737
Finland	0.262	0.492	0.586	0.759	0.521	0.554	0.599	0.535	0.204	0.754	0.527
United Kingdom	0.251	0.441	0.491	0.690	0.451	0.450	0.558	0.486	0.215	0.688	0.453
Japan	0.456	0.556	0.568	0.886	0.538	0.558	0.636	0.585	0.431	0.897	0.504
Russia	0.442	0.451	0.441	0.633	0.442	0.454	0.546	0.479	0.366	0.637	0.430
Singapore	0.635	0.651	0.642	0.791	0.635	0.633	0.661	0.651	0.613	0.721	0.636
United States	0.396	0.372	0.425	0.684	0.396	0.407	0.522	0.416	0.297	0.695	0.392

N2 and N3 attributes, especially in the N2 attribute, which was less than half of that of China. On the basis of the above line graph, the differences in content dimensions of the countries can be drawn, which can provide a reference for the countries to formulate curriculum and learning plans. However, change and relationship, as "one of the most fundamental disciplinary aims of the teaching of mathematics may overlap with other content areas in mathematics as it involves 'functional thinking'" (OECD, 2013a). Across the globe, algebra and measurement questions were significantly more difficult than number, geometry, and data (OECD, 2010). The students from the United States were strong in some content and quantitative reading skills but weak in others, particularly in geometry (Tatsuoka et al., 2004).

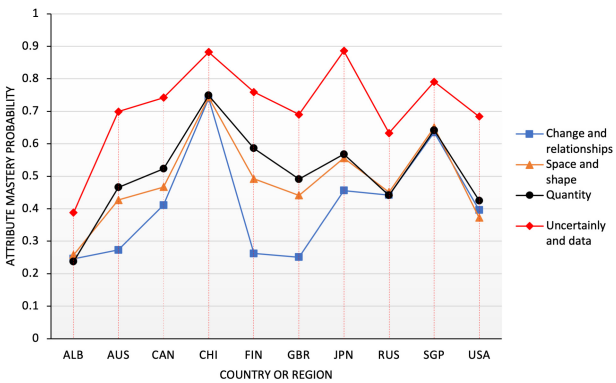


FIGURE 1 | Probability distribution map of content attributes in 10 countries.

Process Attribute

The attributes of mathematical processes involved in the PISA math test consist of three aspects, which are the formation of mathematical scenarios; the concepts, facts, processes, and reasoning of applied mathematics; and the interpretation, application, and evaluation of mathematical results, which account for 25, 50, and 25% (OECD, 2013a), respectively. For interpretation convenience, these three processes are abbreviated as mathematization (P1), mathematical operation (P2), and realization (P3). The probability distribution map of the process attributes in 10 countries are shown in **Figure 2**.

According to **Figure 2**, the performance of each country in the attribute P3 (realization) was better than others in the process attributes, and no big difference was observed in the performance of P1 (mathematization) and P2 (mathematical operation). China, Japan, and Singapore had a better grasp of the process attributes and a relatively balanced performance. It showed that the students in these countries have reached a very good level in mastering the process attributes. The United States and Saudi Arabia had low performance in the mathematical process, and the development was uneven, especially in the mathematization (P1) attribute. Their performance was much lower, only reaching approximately 25% and approximately one-third of that of China. Meanwhile, the mastery of process attributes in other countries was over 40%. Therefore, a considerable number of students had mastered the process attributes. Overall, students were better in the mastery of the process attributes than the content attributes.

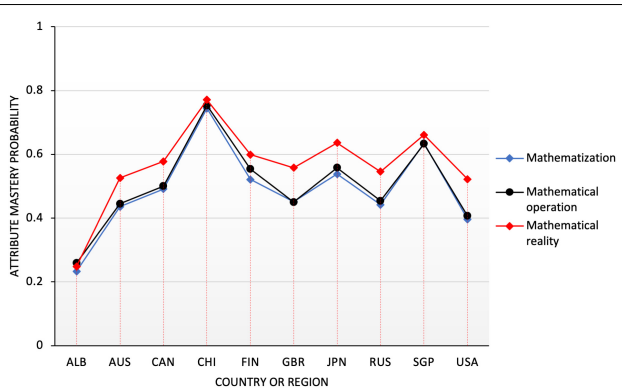


FIGURE 2 | Probability distribution map of process attributes in 10 countries.

Context Attribute

The context questionnaires in the PISA math test involved four parts, namely, the personal (C1), occupational (C2), societal (C3), and scientific contexts (C4). These contexts were necessary for the future student life, and each context accounted for a quarter of the test questions. The context attributes of probability distribution map in 10 countries is shown in **Figure 3**.

As can be seen from **Figure 3**, China performed the best in the personal (C1), occupational (C2), scientific contexts (C3) except

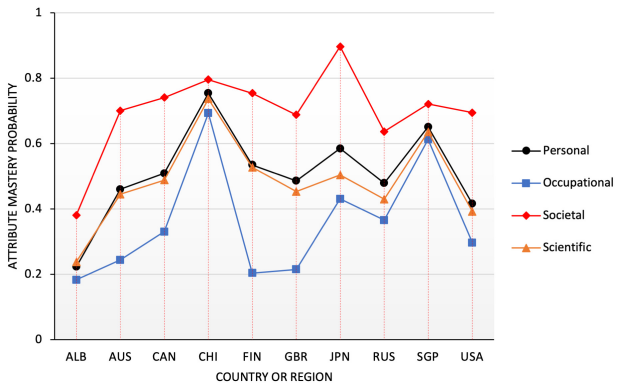


FIGURE 3 | Probability distribution map of 10 countries' context attributes.

for the societal contexts (C3) while Japan performed the best in the societal contexts (C3). Singapore had a relatively balanced and good performance for all the four contexts. Additionally, the United Kingdom, the United States, Russia, and Australia had similar performance in the context attribute, and Saudi Arabia performed relatively lower in all the dimensions. In general, the performance of the societal attribute was superior to the other context attributes and the occupational attribute (C2) was worse than the other three attributes. The differences among the personal context and societal context attributes were not large, which reached a relatively certain and balanced level. The occupational attribute (C2) for Saudi Arabic, Finland and the United Kingdom showed an obvious lower performance than the other countries, which accounted for only 20% approximately. At the same time, the probability of mastering the attribute C3 (social situation) was significantly higher than that of the other three attributes.

Comparative Analysis of Knowledge States

Content Attribute

Knowledge states (KS) refer to a set of arrays consisting of 0 or 1. It represents the mastery of the subject of a certain field of knowledge, skills, etc., where 1 indicates that the subject has mastered the corresponding attributes, and 0 indicates that the subject has not mastered the corresponding attributes (Tatsuoka, 2009). For example, (1111) indicates that the subject has mastered all the attributes, and (0010) indicates that the subject has mastered only the third attribute but not the other three. In this study, through the classification analysis of the attributes of each student, the top five knowledge states of content attributes in the 10 countries were counted, and the proportions of the corresponding knowledge states were calculated.

Table 5 shows that seven countries ranked first (1111) in the knowledge states except for Saudi Arabia, Australia, and the United Kingdom, indicating that a large percentage of students had mastered all content attributes. The proportion of knowledge states (0000) in which no attribute was being mastered was also relatively high. Except for China, all the countries ranked in the top two in this attribute, which indicated that a large number of students in most countries did not have any attributes. The knowledge states (0000) in China ranked third, and the proportion only accounted for approximately 10%. The data from almost all the countries supported that the attribute N4 (data and uncertainty) was a prerequisite for the other attributes in the statistical process of knowledge states. The data from Russia further showed that N3 was the premise of N2, and N2 was the premise of N1. Clearly, a linear learning trajectory of N4 → N3 → N2 → N1 was present. The data for Singapore did not show a clear learning trajectory. The so-called learning trajectories were the hierarchical structure of knowledge states, which characterized the relationship among knowledge states with partial order relationships (Duschl et al., 2011). The structure provided a cognitive sequence for learning the content and supported the effective organization of lesson plans and teaching arrangements. A detailed analysis is provided in 4.4.

Process Attribute

Table 6 summarized the knowledge states and the corresponding proportion of the top five process attributes in the 10 countries. Except for the United States, Russia, Australia and Saudi Arabia, the knowledge states of the countries (111) ranked first, that is, most students had mastered all process attributes. Similar to the content attribute, the knowledge states (000) were ranked at the top two in all the countries, which showed that some students had not mastered any of the process attributes. Notably, the knowledge states that ranked third in all the countries was (001), which showed that the P3 attribute was particularly important in the learning process. Moreover, this attribute became a prerequisite for learning other process attributes. It was also found that almost all data supported a linear learning trajectory such as P3 → P2 → P1.

Context Attribute

Table 7 shows the knowledge states of the top five in the context attributes for the 10 countries. China, Canada, Japan and Singapore ranked the first in the knowledge state (1111), which indicated that a considerable number of students had mastered all the attributes in the contexts. The knowledge states (0000) were ranked at the top two in all the countries, which showed that some students had not mastered any of the context attributes. Additionally, attribute (1011) has a higher percentage among most of the countries, which fully explained that the attribute occupational contexts (C2) is a relatively difficult attribute for most students. More importantly, all the countries except for Singapore supported societal contexts (C3) as a prerequisite attribute of other contexts, which provided a cognitive basis for students to solve the mathematics problems. Students tended to approach the problems related to the societal contexts initially and then deal with the problems related to the other contexts.

Analysis of Learning Trajectories in the Content Area

The biggest advantage of the cognitive diagnostic assessment is that it can grasp the cognitive laws of the subjects more deeply. Then, it can design scientific and reasonable learning

TABLE 5 | Top five knowledge states of content attributes in 10 countries.

Country	1st KS		2nd KS		3rd KS		4th KS		5th KS	
	State	Rate	State	Rate	State	Rate	State	Rate	State	Rate
Saudi Arabia	(0000)	0.583	(1111)	0.220	(0001)	0.147	(0101)	0.012	(1100)	0.011
Australia	(0000)	0.302	(1111)	0.263	(0001)	0.230	(0111)	0.161	(0011)	0.033
Canada	(1111)	0.399	(0000)	0.256	(0001)	0.218	(0111)	0.065	(0011)	0.047
China	(1111)	0.739	(0001)	0.136	(0000)	0.108	(0011)	0.007	(0101)	0.006
Finland	(1111)	0.256	(0000)	0.236	(0111)	0.235	(0001)	0.160	(0011)	0.085
United Kingdom	(0000)	0.300	(1111)	0.250	(0001)	0.204	(0111)	0.182	(0011)	0.049
Japan	(1111)	0.452	(0001)	0.311	(0000)	0.113	(0111)	0.108	(1011)	0.012
Russia	(1111)	0.415	(0000)	0.356	(0001)	0.184	(1011)	0.015	(0111)	0.011
Singapore	(1111)	0.620	(0000)	0.196	(0001)	0.141	(0101)	0.014	(0011)	0.010
United States	(1111)	0.354	(0000)	0.317	(0001)	0.248	(1011)	0.036	(0011)	0.025

TABLE 6 | Knowledge states of the top five process attributes in 10 countries.

Country	1st KS		2nd KS		3rd KS		4th KS		5th KS	
	State	Rate	State	Rate	State	Rate	State	Rate	State	Rate
Saudi Arabia	(000)	0.710	(111)	0.229	(001)	0.045	(011)	0.014	(010)	0.012
Australia	(000)	0.474	(111)	0.431	(001)	0.075	(011)	0.014	(101)	0.006
Canada	(111)	0.485	(000)	0.418	(001)	0.078	(101)	0.010	(010)	0.004
China	(111)	0.716	(000)	0.194	(001)	0.049	(110)	0.024	(101)	0.007
Finland	(111)	0.523	(000)	0.395	(001)	0.051	(011)	0.027	(010)	0.006
United Kingdom	(111)	0.442	(000)	0.442	(001)	0.100	(101)	0.010	(011)	0.007
Japan	(111)	0.539	(000)	0.362	(001)	0.079	(011)	0.019	(010)	0.002
Russia	(000)	0.453	(111)	0.438	(001)	0.089	(011)	0.016	(101)	0.005
Singapore	(111)	0.629	(000)	0.332	(001)	0.032	(110)	0.004	(100)	0.002
United States	(000)	0.478	(111)	0.386	(001)	0.104	(011)	0.021	(101)	0.010

TABLE 7 | Knowledge states of the top five context attributes of 10 countries.

Country	1st KS		2nd KS		3rd KS		4th KS		5th KS	
	State	Rate	State	Rate	State	Rate	State	Rate	State	Rate
Saudi Arabia	(0000)	0.555	(0010)	0.143	(1111)	0.090	(1011)	0.086	(0100)	0.024
Australia	(0000)	0.290	(1111)	0.221	(0010)	0.220	(1011)	0.214	(1010)	0.021
Canada	(1111)	0.287	(0000)	0.244	(0010)	0.197	(1011)	0.187	(1010)	0.031
China	(1111)	0.625	(0000)	0.171	(1011)	0.100	(0010)	0.039	(0110)	0.033
Finland	(1011)	0.310	(0000)	0.238	(0010)	0.194	(1111)	0.185	(1010)	0.034
United Kingdom	(0000)	0.293	(1011)	0.243	(0010)	0.209	(1111)	0.199	(1010)	0.029
Japan	(1111)	0.374	(0010)	0.267	(1011)	0.130	(0000)	0.105	(1010)	0.066
Russia	(0000)	0.316	(1111)	0.288	(0010)	0.152	(1011)	0.131	(0100)	0.035
Singapore	(1111)	0.513	(0000)	0.246	(1011)	0.097	(0110)	0.053	(1010)	0.018
United States	(0000)	0.300	(1111)	0.249	(0010)	0.236	(1011)	0.129	(0110)	0.031

and remedial programs accordingly. The learning trajectories are related to the development of the cognitive laws of learners and the corresponding arrangement of learning knowledge and skills. It is a learning roadmap that strictly follows the cognitive laws of students. The so-called learning trajectories, that is, the hierarchical structure of knowledge states, characterizes the relationship between knowledge states with partial order relationships (Tatsuoka, 2009). In the process of establishing the learning trajectories, the understanding of students regarding the concepts is assumed to follow the order of easiness to difficulty, that is, students first grasp the basic attributes in the attribute hierarchy, and then grasp higher-order attributes, which are more difficult. Therefore, attributes at lower levels should be easy to grasp, and attributes at higher levels should be difficult to grasp. On the basis of this feature, through the cluster analysis of different knowledge states, the learning trajectories can be drawn on the basis of the inclusion relationship shown in **Figure 4**. In this diagram, different

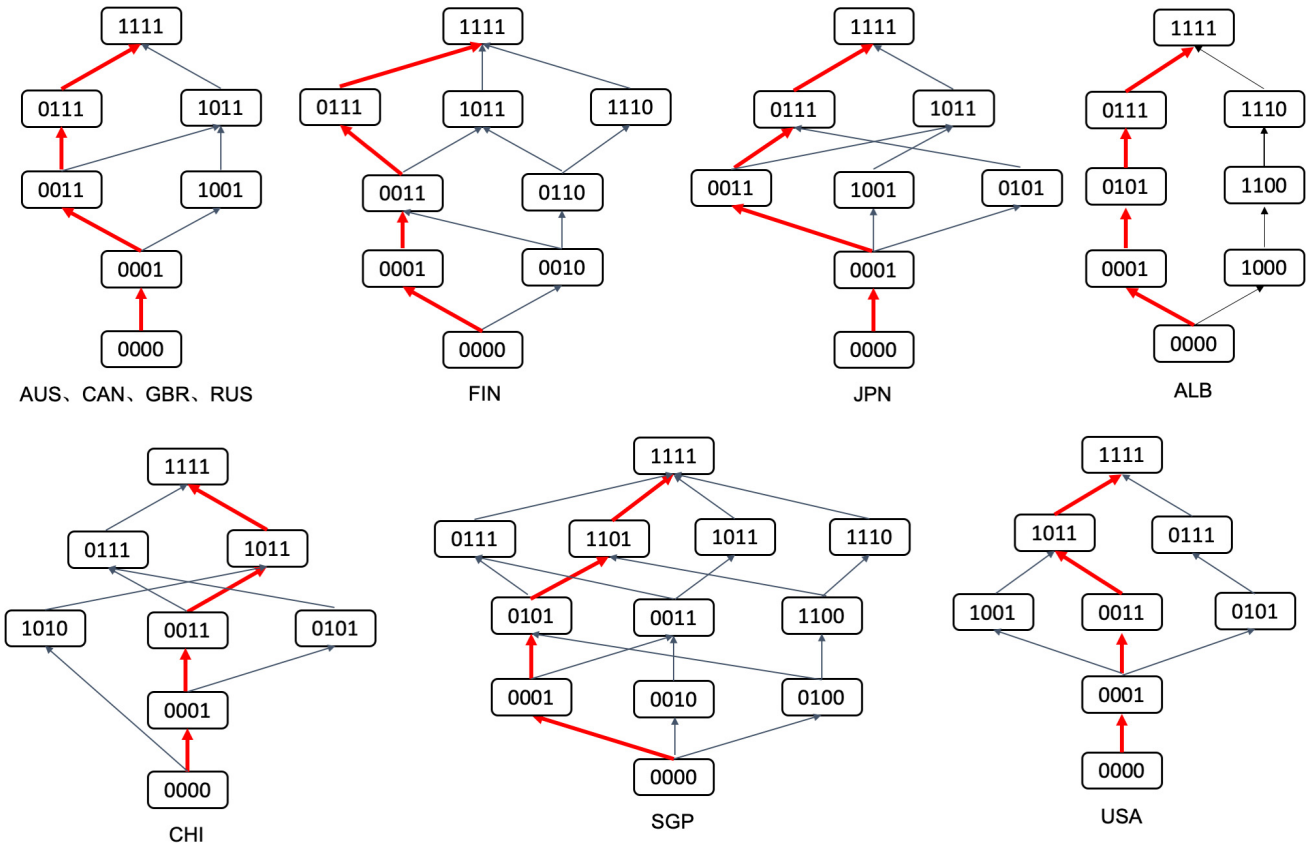


FIGURE 4 | Students' learning trajectories in 10 countries.

learning trajectories can be selected for students with different knowledge states.

In the process of construction of the learning trajectories, the knowledge state of each participant is firstly obtained through parameter evaluation, which is the participant' mastery of each attribute. Then, the participants with the same knowledge state are classified and categorized to establish the trajectory relationship among the knowledge states. The red path is the main trajectory among them, which contains the largest percentage of the participants who own the knowledge states in each level, to some extent, it represents the learning trajectory of a certain group. For instance, in **Figure 4**, compared with (1010) in the attribute mastering mode, the subjects belonging to the knowledge state (1011) have mastered all the attributes belonging to the knowledge states (1010) and other attributes. Therefore, (1010) $\subset$ (1011), which entails an inclusive relationship between the two knowledge states, that is, the trajectory is (1010) $\rightarrow$ (1011). According to CHI in **Figure 4**, (0000) $\subset$ (0001) $\subset$ (0011) $\subset$ (1011) $\subset$ (1111) exists. Therefore, the learning trajectory in red shown in CHI in **Figure 4** is (0000) $\rightarrow$ (0001) $\rightarrow$ (0011) $\rightarrow$ (0111) $\rightarrow$ (1111). According to **Figure 4**, Australia, Canada, the United Kingdom, and Russia have the same learning trajectory. The students have three trajectories to master all the attributes from not mastering any attributes in these countries. However, the most important trajectory is shown in red: (0000) $\rightarrow$ (0001) $\rightarrow$ (0011) $\rightarrow$ (0111) $\rightarrow$ (1111). Most students first obtain N4 (uncertainty and data), then N3 (quantity), then N2 (space and shape), and finally N1 (relationship and change). The learning trajectories of Finland and Japan are more complicated than those of Australia, Canada, the United Kingdom, and Russia. As shown in **Figure 4**, same as their main learning trajectory, they all follow the trajectory of (0000) $\rightarrow$ (0001) $\rightarrow$ (0011) $\rightarrow$ (0111) $\rightarrow$ (1111). The learning trajectory in China is also relatively complex, and it has multiple learning trajectories. The main learning trajectory is (0000) $\rightarrow$ (0001) $\rightarrow$ (0011) $\rightarrow$ (1011) $\rightarrow$ (1111), that is, most students grasp N4 (uncertainty and data) first, then N3 (quantity), and N1 (change and relationships), and finally N2 (space and shape). A difference is observed in the order of obtaining N1 and N2. The learning trajectory in Singapore is the most complicated and has the most diverse learning trajectories. The main learning trajectory is (0000) $\rightarrow$ (0001) $\rightarrow$ (0101) $\rightarrow$ (1101) $\rightarrow$ (1111), that is, most students grasp N4 (uncertainty and data) first, then N2 (space and shape), then N1 (change and relationships), and finally N3 (quantity). The learning trajectories of the United States and Saudi Arabia were comparatively simple and the main learning trajectories of the United States and China are the same. These trajectories are not only directly related to the cognitive order of students but also influenced by factors, such as national curriculum

arrangements and extracurricular tutoring (De Lange, 2007). As can be seen from **Figure 4**, students in different knowledge states can choose different learning trajectories according to their own characteristics and the learning resources around them, which also reflect the diverse choices of learning. The learning trajectories from the low to the top ends represent different ability levels, reflect the ability relationship among knowledge states, describe the development process of students, and shows the clear development of trajectory and direction for students from low-level learning to high-level learning abilities. Therefore, the learning trajectories not only provide students with personalized and refined diagnostic reports but also provide a basis for the remedial teaching of teachers.

DISCUSSION

With the advancement of educational globalization, international understanding and international educational references have enabled us to apply the latest achievements to developmental promotion (Wu et al., 2019). PISA, as a product of the development of educational globalization, provides data support for us to understand basic education worldwide and to compare and learn from one another. PISA also reports the motivation, self-confidence, learning strategies, and the environmental background information of students, including the social, economic, cultural, and educational aspects and population distribution related to knowledge and skills (OECD, 2004). The analysis results in this study provided by PISA have surpassed the ranking comparison among the respective fields of countries. It has also offered a unique globalized perspective on how students attain fine-grained attributes and correct the misconception that "correctly answered" items entail that the examinee has attained all the knowledge required to solve the items.

One fact revealed in the PISA 2012 is that all countries that participated have a sizable share of low performers, including those with the highest performance and equity outcomes. On average, 23% of students are low performers in mathematics across all the participating countries; however, the shares of low performers in mathematics vary significantly from country to country (OECD, 2016). Among the 10 countries participating in this study, the proportions of students with low performance in mathematics in the United States, Russia, and Saudi Arabia are higher than average. Additionally, students from different countries have unbalanced performances in various fields of mathematical study. These conclusions are consistent with the above research. The magnitude of the cognitive ability differences between countries is large, and a likely reason for the difference is the Flynn effect, which massively raised the average IQ in economically advanced countries in the 20th century (Meisenberg and Woodley, 2013). Other studies have suggested that the cognitive disparities between advanced industrial societies and less developed countries have been diminishing (Weede and Kämpf, 2002; Meisenberg and Lynn, 2012). Given the positive correlation between IQ (or IQ growth) and economic growth observed, this trend is probably related to a reduction in the degree of economic inequality among countries (Meisenberg and Woodley, 2013).

In terms of the student performances, these 10 countries have large differences in the attributes. However, the conclusion can be quite similar if the students are examined from the aspect of mathematical literacy or creativity. The creativity of students in mathematics is positively related to their achievement in mathematics at the student level within schools. However, the relationship is the opposite among countries (Sebastian and Huang, 2016). Some researchers and educators have realized that academic performance measured through standardized tests narrowly focuses on a few subjects that emphasize identifying correct answers and avoiding mistakes, which ultimately discourages student creativity and critical thinking (Zhao, 2012; Chomsky and Robichaud, 2014; Darling-Hammond and Turnipseed, 2015). On the basis of the findings, consistent top performers in PISA tests, such as Shanghai, Singapore, Korea, and Japan, have started revising their curriculum to increase their emphasis on creative thinking skills (Shaheen, 2010; Kim, 2011). From the aspect of mathematical literacy, Jablonka (2003) discussed the fundamental nature of mathematical literacy. The contexts may be familiar to some students but not to others. Any attempt to use a single instrument to assess mathematical literacy beyond the most local context appears to be self-defeating. Cultural differences exist among countries, and the invariance of the test items also need to be tested accordingly.

In the analysis of learning trajectories, there are two reasons that we only presented the results of content attributes but not the results of the other two attributes. First, in the current research of mathematics education, the learning trajectories are mostly aimed at the its content but not the context or process. Whether there are any learning paths in the process or context attributes, or whether it meets the assumptions of the learning paths, the conclusions are still to be uncovered (Clements and Sarama, 2004; Daro et al., 2011; Wilson et al., 2013). Second, through data analysis, it is found that there are no rules in the process or context attribute, therefore there is no further analysis of the learning trajectories of these two attributes (Confrey et al., 2014). However, the learning trajectories in educational practice is the concept about change longitudinally–how to trace a student's mastery of the attributes change over time with increasing instruction. The test items in PISA that we studied is cross-sectional due to unavailability of data in the other years. Simply finding relatively large numbers of students in various knowledge states does not imply that individual students move through those states in any specific order. Even if the paths identified reflect reality, there is "correlation does not imply causation" argument to be made; an association between country and these patterns does not imply that specific educational practices lead to those differences. Finally, assuming that the paths are indeed correct, because the data are a single point in time, the silly assertion that these are forgetting paths (i. e., students move from understanding to ignorance) is equally consistent with the data observed. None of this says the learning paths identified are wrong. But any findings from these cross-sectional data are speculative and open to alternative interpretation. They require additional evidence from other sources in order to be evaluated,

for instance, studying longitudinally could be an alternative better examination option (Zhan et al., 2019; Zhan, 2020).

Although this study has conducted a more in-depth analysis of the PISA items and results by using the newly emerging measurement method, many areas still need improvement. The study has provided detailed information on 11 attributes in 10 countries in terms of three dimensions, namely, content, process, and context. The division of attributes depends on the coding of existing test questions in PISA without deeper mining in the fields. Later research can divide the attributes in a more detailed way and compare them latently to obtain the advantages and disadvantages of different countries in finer granularity. Additionally, we suggest that questionnaires be sent out to mine the reasons for the difference further. We also need to admit that measuring the change in student achievement at the country level is more robust than measuring student achievement in any single wave of assessment. More methodological and educational research is required to understand the longitudinal trends at the country level (Klieme, 2016). In the end, analyzing the learning situation of students should focus not only on their test scores but also on their external environment, such as the family and school environments. An analysis of the relationship of test scores to variables external to the test can provide another important source of validity evidence (American Educational Research Association, and National Council on Measurement in Education, 2014). Multilevel hierarchical analysis is an important methodology to be taken into consideration in future research.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

XW designed the study and wrote this manuscript. RW contributed to the manuscript writing and the continued revision provided by the reviewers. H-HC reviewed the manuscript and revised it. QK reviewed the manuscript and provided comments. YZ collected the data and wrote this manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by China Scholarship Council (No. 201906140104); 2020 Academic Innovation Ability Enhancement Plan for Outstanding Doctoral Students of East China Normal University (No. YBNLTS2020-003); and Guizhou Philosophy and Social Science Planning Youth Fund (No. 19GZQN29).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.02230/full#supplementary-material

REFERENCES

American Educational Research Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*, 3rd Edn. Washington, DC: American Educational Research Association.

Breakspear, S. (2012). *The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. OECD Education Working Papers, No. 71*. Paris: OECD Publishing.

Cai, Y., Tu, D., and Ding, S. (2018). Theorems and methods of a complete Q matrix with attribute hierarchies under. restricted Q-matrix design. *Front. Psychol.* 9:1413. doi: 10.3389/fpsyg.2018.01413

Carpenter, T. P., and Moser, J. M. (1982). "The development of addition and subtraction problem-solving skills," in *Addition and Subtraction: A Cognitive Perspective*, eds T. P. Carpenter, J. M. Moser, and T. A. Romberg (Hillsdale, NJ: Lawrence Erlbaurn Associates) 9–24. doi: 10.1201/9781003046585-2

Chang, H-H., Wang, C., and Zhang, S. (in press). Statistical applications in educational measurement. *Annual Review of Statistics and Its Application*.

Chen, H., and Chen, J. (2016). Exploring reading comprehension skill relationships through the G-DINA model. *Educ. Psychol.* 36, 1049–1064. doi: 10.1080/01443410.2015.1076764

Chen, J., and Choi, J. (2009). A comparison of maximum likelihood and expected a posteriori estimation for polychoric correlation using Monte Carlo simulation. *J. Modern Appl. Stat. Methods* 8:32. doi: 10.22237/jmasm/1241137860

Choi, K. M., Lee, Y.-S., and Park, Y. S. (2015). What CDM can tell about what students have learned: an analysis of TIMSS Eighth Grade Mathematics. *Eur. J. Math. Sci. Technol. Educ.* 11, 1563–1577. doi: 10.12973/eurasia.2015.1421a

Chomsky, N., and Robichaud, A. (2014). Standardized testing as an assault on humanism and critical thinking in education. *Radic. Pedag.* 11:3.

Clements, D. H., and Sarama, J. (2004). Learning trajectories in mathematics education. *Math. Think. Learn.* 6, 81–89. doi: 10.1207/s15327833mtl0602_1

Confrey, J., Maloney, A. P., and Corley, A. K. (2014). Learning trajectories: a framework for connecting standards with curriculum. *ZDM* 46, 719–733. doi: 10.1007/s11858-014-0598-7

Darling-Hammond, L., and Turnipseed, S. (2015). Accountability is more than a test score. *Educ. Policy Anal. Arch.* 23, 1–37. doi: 10.14507/epaa.v23.1986

Daro, P., Mosher, F. A., and Corcoran, T. B. (2011). *Learning Trajectories in Mathematics: A Foundation for Standards, Curriculum, Assessment, and Instruction. CPRE Research Reports*. Available at: https://repository.upenn.edu/cpre_researchreports/60 (accessed July 10, 2020).

De Lange, J. (2007). Large-scale assessment and mathematics education. *Second Handbook of Research on Mathematics Teaching and Learning*, 2, 1111–1144.

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

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

Dobbins, M., and Martens, K. (2012). Towards an education approach à la finlandaise? French education policy after PISA. *J. Educ. Policy* 27, 23–43. doi: 10.1080/02680939.2011.622413

Dogan, E., and Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educ. Stud. Math.* 68, 263–272. doi: 10.1007/s10649-007-9099-8

Duru-Bellat, M., and Suchaut, B. (2005). Organisation and context, efficiency and equity of educational systems: what PISA tells us. *Eur. Educ. Res. J.* 4, 181–194. doi: 10.2304/eerj.2005.4.3.3

Duschl, R., Maeng, S., and Sezen, A. (2011). Learning progressions and teaching sequences: a review and analysis. *Stud. Sci. Educ.* 47, 123–182. doi: 10.1080/03057267.2011.604476

Egelund, N. (2008). The value of international comparative studies of achievement–a Danish perspective. *Assess. Educ.* 15, 245–251. doi: 10.1080/09695940802417400

El Masri, Y. H., Baird, J.-A., and Graesser, A. (2016). Language effects in international testing: the case of PISA 2006 science items. *Assess. Educ.* 23, 427–455. doi: 10.1080/0969594X.2016.1218323

Ertl, H. (2006). Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany. *Oxford Rev. Educ.* 32, 619–634. doi: 10.1080/03054980600976320

Ferrini-Mundy, J., and Schmidt, W. H. (2005). International comparative studies in mathematics education: opportunities for collaboration and challenges for researchers. *J. Res. Math. Educ.* 36, 164–175. doi: 10.2307/30034834

Freudenthal, H. (2012). *Mathematics as an Educational Task*. Berlin: Springer Science & Business Media.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *J. Res. Math. Educ.* 22, 170–218. doi: 10.2307/749074

Grek, S. (2009). Governing by numbers: the PISA 'effect'in Europe. *J. Educ. Policy* 24, 23–37. doi: 10.1080/02680930802412669

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Measur.* 26, 301–321. doi: 10.1111/j.1745-3984.tb00336.x

Hagenaars, J. A. (1990). *Categorical Longitudinal Data: Loglinear Panel, Trend, and Cohort Analysis*. Thousand Oaks: Sage.

Hagenaars, J. A. (1993). *Loglinear Models with Latent Variables*. Thousand Oaks: Sage.

Hansen, K. Y., and Strietholt, R. (2018). Does schooling actually perpetuate educational inequality in mathematics performance? A validity question on the measures of opportunity to learn in PISA. *ZDM* 50, 643–658. doi: 10.1007/s11858-018-0935-3

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality*. Doctoral dissertation, ProQuest Information & Learning, Ann Arbor, MI.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Jablonka, E. (2003). "Mathematical literacy," in *Second International Handbook of Mathematics Education*, eds J. Kilpatrick, A. J. Bishop, F. K. S. Leung, M. A. Clements, and C. Keitel-Kreidt (Berlin: Springer), 75–102.

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections. with.nonparametric item response theory. *Appl. Psychol. Measur.* 25, 258–272. doi: 10.1177/01466210122032064

Kim, K. H. (2011). The creativity crisis: the decrease in creative thinking scores on the Torrance Tests of Creative Thinking. *Creat. Res. J.* 23, 285–295. doi: 10.1080/10400419.2011.627805

Klieme, E. (2016). *TIMSS 2015 and PISA 2015. How Are They Related On The Country Level*. Germany: German Institute for International Educational Research.

Knowles, E., and Evans, H. (2012). *PISA 2009: How Does the Social Attainment Gap in England Compare With Countries Internationally*. Available at: http://www.education.gov.uk/publications/ (accessed June 10, 2020).

Kyriakides, L., Charalambous, C. Y., Demetriou, D., and Panayiotou, A. (2014). "Using PISA studies to establish generic models of educational effectiveness," in *Educational Policy Evaluation Through International Comparative Assessments*, eds W. Bos, J. E. Gustafsson, and R. Strietholt (Münster: Waxmann Verlag), 191–206.

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

Luyten, H. (2017). "Predictive power of OTL measures in TIMSS and PISA," in *Opportunity to Learn, Curriculum Alignment and Test Preparation*, Ed. J. Scheerens (Berlin: Springer), 103–119. doi: 10.1007/978-3-319-43110-9_5

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/bf02294535

Meisenberg, G., and Lynn, R. (2012). Cognitive human capital and economic growth: Defining the causal trajectories. *J. Soc. Polit. Econ. Stud.* 37, 141–179.

Meisenberg, G., and Woodley, M. A. (2013). Are cognitive differences between countries diminishing? Evidence from TIMSS and PISA. *Intelligence* 41, 808–816. doi: 10.1016/j.intell.2013.03.009

OECD (2004). *PISA Learning for Tomorrow's World: First Results from PISA 2003*, Vol. 659. New York, NY: Simon and Schuster.

OECD (2010). *Learning Mathematics for Life: A Perspective from PISA*. Paris: OECD Publishing.

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

OECD (2013b). *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed*, Vol. II. Paris: Author.

OECD (2014a). *PISA 2012 Technical Report*. Paris: OECD publishing.

OECD (2014b). *"A Profile of Student Performance in Mathematics." PISA 2012 Results: What Students Know and Can Do (Volume I, Revised Edition, February 2014)*. Paris: OECD Publishing.

OECD (2015). *PISA2015 Technical Report*. Paris: OECD Publishing.

OECD (2016). *Low-Performing Students: Why They Fall Behind and How To Help Them Succeed*. Paris: OECD Publishing.

OECD (2019a). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing.

OECD (2019b). *PISA 2021 Mathematics Framework (Draft)*. Paris: OECD Publishing.

Rumelhart, D. E. (1991). Understanding understanding. *Mem. Emot.* 257:275.

Rutkowski, L., and Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educ. Res.* 45, 252–257. doi: 10.3102/0013189X16649961

Schneider, W., and Graham, D. J. (1992). Introduction to connectionist modeling in education. *Educ. Psychol.* 27, 513–530. doi: 10.1207/s15326985ep2704_7

Sebastian, J., and Huang, H. (2016). Examining the relationship of a survey based measure of math creativity with math achievement: cross-national evidence from PISA 2012. *Int. J. Educ. Res.* 80, 74–92. doi: 10.1016/j.ijer.2016.08.010

Shaheen, R. (2010). Creativity and education. *Creat. Educ.* 1:166. doi: 10.4236/ce.2010.13026

Sprietsma, M. (2010). Effect of relative age in the first grade of primary school on long-term scholastic results: international comparative evidence using PISA 2003. *Educ. Econ.* 18, 1–32. doi: 10.1080/09645290802201961

Stacey, K., Almuna, F., Caraballo, R. M., Chesné, J.-F., Garfunkel, S., Gooya, Z., et al. (2015). "PISA's influence on thought and action in mathematics education," in *Assessing Mathematical Literacy*, eds R. Turner, and K. Stacey (Berlin: Springer), 275–306.

Steinthorsdottir, O. B., and Sriraman, B. (2008). Exploring gender factors related to PISA 2003 results in Iceland: a youth interview study. *ZDM* 40, 591–600. doi: 10.1007/s11858-008-0121-0

Stout, W. (2002). Psychometrics: from practice to theory and back. *Psychometrika* 67, 485–518. doi: 10.1007/BF02295128

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comp. Educ.* 44, 387–407. doi: 10.1080/03050060802481413

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika* 49, 95–110. doi: 10.1007/bf02294208

Tatsuoka, K. K. (1987). *Toward an Integration of Item-Response Theory and Cognitive Error Diagnosis*. Available online at: https://apps.dtic.mil/dtic/tr/fulltext/u2/a183189.pdf

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *J. R. S.tat Soc. C-Appl.* 51, 337–350. doi: 10.1111/1467-9876.00272

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. Abingdon: Routledge.

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

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

Templin, J., and Bradshaw, L. (2014). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika* 79, 317–339. doi: 10.1007/s11336-013-9362-0

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

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

Tu, D. (2019). *flexCDMs*. Available at: http://www.psychometrics-studio.cn (accessed July 5, 2020).

Tu, D., Wang, S., Cai, Y., Douglas, J., and Chang, H.-H. (2019). Cognitive diagnostic models with attribute hierarchies: model estimation with a restricted Q-matrix design. *Appl. Psychol. Measur.* 43, 255–271. doi: 10.1177/0146621618765721

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychol. Test Assess. Models* 52, 8–28.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychol. Methods* 17:228. doi: 10.1037/a0027127

Wang, W., Song, L., and Ding, S. (2018). The index and application of cognitive diagnostic test from the perspective of classification. *Psychol. Sci.* 41, 475–483. doi: 10.16719/j.cnki.1671-6981.20180234

Weede, E., and Kämpf, S. (2002). The impact of intelligence and institutional improvements on economic growth. *Kyklos* 55, 361–380. doi: 10.1111/1467-6435.00191

Wilson, P. H., Mojica, G. F., and Confrey, J. (2013). Learning trajectories in teacher education: supporting. teachers' understandings of students' mathematical thinking. *J. Math. Behav.* 32, 103–121. doi: 10.1016/j.jmathb.2012.12.003

Wößmann, L. (2005). The effect heterogeneity of central examinations: evidence from TIMSS. TIMSS-Repeat and PISA. *Educ. Econ.* 13, 143–169. doi: 10.1080/09645290500031165

Wu, X., Wu, X., and Zhang, Y. (2019). Teacher teaching evaluation system: framework, mechanism and enlightenment:based on the analysis of 18 evaluation systems of seven countries. *Comp. Educ. Res.* 41, 59–67.

Wu, X., Zhang, Y., and Zhang, C. (2020). The construction of cognitive diagnostic assessment methods for key competence. *Modern Educ. Technol.* 30, 20–28.

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

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

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

Zhao, Y. (2012). Flunking innovation and creativity. *Phi Delta Kappan* 94, 56–61. doi: 10.1177/003172171209400111

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 September 2020
doi: 10.3389/fpsyg.2020.02145

Longitudinal Cognitive Diagnostic Assessment Based on the HMM/ANN Model

Hongbo Wen[1], Yaping Liu[1] and Ningning Zhao[2]**

[1] National Assessment Center of Education Quality, Beijing Normal University, Beijing, China, [2] School of Chinese Language and Literature, Beijing Normal University, Beijing, China

Cognitive diagnostic assessment (CDA) is able to obtain information regarding the student's cognitive and knowledge development based on the psychometric model. Notably, most of previous studies use traditional cognitive diagnosis models (CDMs). This study aims to compare the traditional CDM and the longitudinal CDM, namely, the hidden Markov model (HMM)/artificial neural network (ANN) model. In this model, the ANN was applied as the measurement model of the HMM to realize the longitudinal tracking of students' cognitive skills. This study also incorporates simulation as well as empirical studies. The results illustrate that the HMM/ANN model obtains high classification accuracy and a correct conversion rate when the number of attributes is small. The combination of ANN and HMM assists in effectively tracking the development of students' cognitive skills in real educational situations. Moreover, the classification accuracy of the HMM/ANN model is affected by the quality of items, the number of items as well as by the number of attributes examined, but not by the sample size. The classification result and the correct transition probability of the HMM/ANN model were improved by increasing the item quality and the number of items along with decreasing the number of attributes.

Keywords: cognitive diagnostic assessment, longitudinal assessment, hidden Markov model, SSOM neural network, reading comprehension

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Dandan Liao,
American Institutes for Research, United States
Lei Guo,
Southwest University, China

****Correspondence:***
Hongbo Wen
whb@bnu.edu.cn
Ningning Zhao
ningning.zhao@bnu.edu.cn

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

Received: *22 May 2020*
Accepted: *31 July 2020*
Published: *09 September 2020*

Citation:
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

INTRODUCTION

Cognitive diagnostic assessment (CDA) combines cognitive psychology with psychometrics to diagnose and evaluate the knowledge structure and cognitive skills of students (Leighton and Gierl, 2007; Tu et al., 2012). Compared to the traditional academic proficiency assessment, the results of CDA report specific information regarding the strengths and the weaknesses of students' cognitive skills. At present, researchers developed various cognitive diagnostic models (CDMs) to realize the diagnostic classification of cognitive skills. Deterministic inputs, noisy "and" gate (DINA) model (Macready and Dayton, 1977; Haertel, 1989; Junker and Sijtsma, 2001), the deterministic inputs, noisy "or" gate (DINO) model (Templin and Henson, 2006), and other models are representative and widely applied. However, traditional CDMs, such as DINA and DINO, are static models that classify students' cognitive skills on a cross-sectional level. In the education context, students' knowledge and skills are continually developing, and educators

are more concerned with how their cognitive skills develop over time. Notably, traditional CDM cannot model the trajectory of skills development.

In the psychometric field, researchers have used multi-dimensional Item Response Theory (IRT) models to assess the development of students' abilities (Andersen, 1985; Embretson, 1991). These studies utilized multi-dimensional IRT models to measure a single capability at different points in time. With the development of computer algorithms, the hidden Markov model (HMM) can be used to realize the transformation analysis of potential categories (Collins and Lanza, 2010). Currently, DINA and DINO models have been applied as measurement models under the framework of HMM (Li et al., 2016; Kaya and Leite, 2017). Chen et al. (2017) used the first-order HMM to trace learning trajectory. Additionally, Wang et al. (2018) integrated the CDM with a higher-order HMM, which included covariates, to model skill transition and explain individual differences.

This research mentioned above used HMM to realize the transformation analysis of potential states. The methods combine HMM and traditional CDMs, such as DINA and DINO, which are based on the framework of the IRT (Tu et al., 2012). As a result, the methods should satisfy the three basic hypotheses of unidimension, local independence, and monotonicity of capability. The problem is that the data collected in practice can hardly satisfy these three hypotheses. In recent years, researchers have attempted to develop more applicable models to new models to overcome these deficiencies. For instance, Hansen (2013) proposed a unidimensional hierarchical diagnostic model to track the growth of skill, in which local dependence was accounted for through using random-effect latent variables. Furthermore, Zhan et al. (2019) proposed a longitudinal diagnostic classification modeling approach by using a multidimensional higher-order latent structure to explain the relationship among multiple latent attributes, and the local item dependence was well taken into account.

The method mentioned above requires parameter estimation, which involves a large sample size to achieve high accuracy. When the sample size is small, the accuracy of parameter estimation will be affected (Chen et al., 2013), thus seriously influencing the accuracy of the cognitive skill classification of students (Gierl et al., 2008; Cao, 2009; Shu et al., 2013). Some researchers proposed to apply non-parametric methods to classify cognitive skills under a small sample size. For instance, Chiu et al. (2018) used general non-parametric classification method to estimate the student's attribute pattern through minimizing the distance between the observed response and the ideal response when sample sizes are at the classroom level. With the development of artificial intelligence, ANN has been widely applied in various fields. And it is claimed that non-parametric artificial intelligence pattern recognition technology can be utilized to achieve CDA. The advantage is that ANN can perform non-parameter estimation and the bias of the potential classification model can be overcome (Gierl et al., 2008; Cao, 2009; Wang et al., 2015). Additionally, ANN has a relatively high accuracy in small samples, so it can avoid the above-mentioned disadvantages.

Recently, an increasing number of studies have attempted to combine ANNs and CDA (Gierl et al., 2008; Cao, 2009; Shu et al., 2013; Wang et al., 2015, 2016). At present, there are hundreds of artificial neural networks (ANN), among which the supervised self-organizing map (SSOM) is one of the more popular neural networks. SSOM has been widely used in network traffic classification, decoding analyses, and metabolic profiling and demonstrates good classification performance (Wongravee et al., 2010; Hu, 2011; Hu et al., 2011; Lu et al., 2020). The SSOM can activate the network features near the physical location of the neurons according to the similar input mode used to achieve classification, so it has a strong applicability in various fields. Consequently, it is worth further exploring whether it is possible to apply ANN (e.g., SSOM) as the measurement model of HMM so as to achieve the accurate classification of students' cognitive skills while also tracking the development of their skills.

This study aims to explore whether it is possible to establish an HMM/ANN model through using ANN as the measurement model of HMM. This will be used to accurately track the change in students' cognitive skills and to validate the effectiveness of this model in the actual education situation of the small sample.

TECHNICAL BACKGROUND

An Overview of the Artificial Neural Network

Scientists Warren McCulloch and Walter Pitts first proposed the ANN in 1943, which mimics the basic principle of the biological nervous system. It is a network structure system created by a large number of interconnected neurons similar to the neurocyte in the human brain. In ANN, the neurons are usually organized into layers, such as the input layer and the output layer, and information processing is achieved through adjusting the connection between the nodes of each layer (Han, 2006).

The connection between the input and the output layers of ANN can be obtained through performing the training and the testing phases. In the training phase, the input data and the output data (or those only containing input data) of the training set will be applied to train the network. This is done to determine the number of hidden layer neurons as well as the connection weight between layers of neurons. Then, the neural network will be well trained. During the testing phase, a well-trained neural network will be provided a new set of input data that can obtain the output value based on the weight of connections between neurons.

According to the classification of learning paradigm, ANN can be divided into supervised learning and unsupervised learning. The most significant feature of supervised learning is that the data of the input and the output layer of the training set are known, and the output layer is the category label corresponding to the data characteristics of the input layer. The supervised neural network determines the connection weight of the layers through establishing the relationship between the input and the output layers. When there is new data input, the determined connection weights can assist in obtaining the output value. Notably, the characteristic of unsupervised learning is that only the data in

the input layer is known, while the data in the output layer is unknown. Unsupervised neural network reveals the innate law of the data by learning the input data, which is more applicable to cluster analysis.

Supervised Self-Organizing Map

As depicted in **Figure 1**, the SSOM consists of three layers: the input layer, the competition layer, and the output layer. The number of output layers is consistent with the number of classification categories. SSOM is based on the original structure of the self-organizing mapping neural network (Kohonen, 1982, 1990, 2001), adding an output layer to become a supervised neural network, to better realize the classification of data with category labels. In SSOM, the part from the output layer to the competition layer is unsupervised learning, and the competition layer to the output layer is supervised learning. Moreover, the input layer to the competition layer and the competition layer to the output layer are all connected (Zhao and Li, 2012). Notably, it is necessary for the learning and the training of this neural network to adjust the weights from the input layer to the competition layer and from the competition layer to the output layer simultaneously. SSOM can use the existing category marker information to assist clustering and help improve the adjustment rules of neuron weight in the winning neighborhood. This is done so as to make it easier to select winning neurons (Hu et al., 2011).

In the training phase of SSOM, the input training samples $X_i = (X_1, X_2, X_3 \ldots\ldots X_n)$ are known, and n is the number of neurons in the input layer. According to formula (1), the winning neuron g in the competitive layer can be obtained. D_j is the distance from input layer X_i to the neuron j in the competition layer, from which the winning neuron g with the smallest distance from the input layer X_i is found.

$$D_g = \min\left(D_j\right) = \min||X_i - W_{ij}||,\ j = 1, 2, \ldots, m \quad (1)$$

Among them, $||\cdot||$ is the distance function; W_{ij} represents the weighting coefficient between the input layer neuron i and the competition layer neuron j; m is the number of competition layer neurons.

The next step is to adjust the weight, which is mainly divided into three phases:

(1) $Y_k = (Y_1, Y_2, Y_3 \ldots\ldots Y_k)$ is the output value of the input layer X_i, k represents the number of neurons in the output layer, while the output category corresponding to the winning neuron g in the competition layer is O_g .
(2) Calculate the winning neighborhood $N_{c(t)}$ of the winning neuron g.
(3) If $O_g = Y_i$, then the weight coefficient is adjusted according to formula (2, 3) in the winning neighborhood; if $O_g \neq Y_k$,then it is adjusted according to formula (4, 5).

$$W_{ij} = W_{ij} + \eta_1\left(X_i - W_{ij}\right) \quad (2)$$

$$W_{jk} = W_{jk} + \eta_2\left(Y_k - W_{jk}\right) \quad (3)$$

$$W_{ij} = W_{ij} + \mu\eta_1\left(X_i - W_{ij}\right) \quad (4)$$

$$W_{jk} = W_{jk} + \mu\eta_2\left(Y_k - W_{jk}\right) \quad (5)$$

W_{ij} represents the weight coefficient between the input layer neuron i and the competition layer neuron j; W_{jk} represents the weight coefficient between the competition layer neuron j and the output layer neuron k; η_1, η_2 represents the learning rate from the input layer to the competition layer and from the competition layer to the output layer, respectively; μ is the weight coefficient. After adjusting the weights, the output layer becomes an ordered feature graph which reflects the output pattern.

Hidden Markov Model

The HMM is also known as the potential transformation analysis model (Collins and Wugalter, 1992). As depicted in **Figure 2**, the model contains two interconnected random processes. One describes a Markov chain of state transition and the other is a sequence of observations related to states. The reason why it is referred to as the HMM is because, in these two random processes, the first random process, namely, the sequence of state transition, is unobserved and can only be inferred from the observation sequence of the other random process (Rabiner, 1989).

HMM can be described by five parameters:

N: N is the number of states of the Markov chain in the model. The N state is denoted as $S = \{s_1, s_2,\ s_3 \ldots s_N\}$, and the state of the Markov chain at time t is q_t, $q_t \in (s_1, s_2,\ s_3 \ldots\ s_N)$.
M: M is the number of possible observations for each state. M observations are denoted as $V = \{v_1, v_2,\ v_3 \ldots v_M\}$, and the observed value at time t is O_t, $O_t \in (v_1, v_2,\ v_3\ \ldots\ v_N)$.

π: π is initial-state probability, $\pi = (\pi_i,\ i = 1, \ldots, N)$.

$$\pi_i = P(q_1 = s_i) \quad (6)$$

$$0 \leq \pi_i \leq 1,\ \sum_{i=1}^{N}\pi_i = 1 \quad (7)$$

A: A is the state transition probability matrix $(a_{ij})_{N\times N}$, which describes the state transition probability at different points in time. Among them:

$$a_{ij} = P(q_t = s_j|\ q_{t-1} = s_i),\ 1 \leq i \leq N \quad (8)$$

$$0 \leq a_{ij} \leq 1,\ \sum_{j=1}^{N}a_{ij} = 1$$

B: B is an observation probability matrix, namely, the item response probability matrix $(b_{jk})_{N\times N}$. In the educational measurement field, it refers to the probability that the individual in each potential state makes a correct or specific response to each item. Among them:

$$b_{jk} = P(O_t = v_k|\ q_t = s_j) \quad (9)$$

$$1 \leq j \leq N,\ 1 \leq k \leq M \quad (10)$$

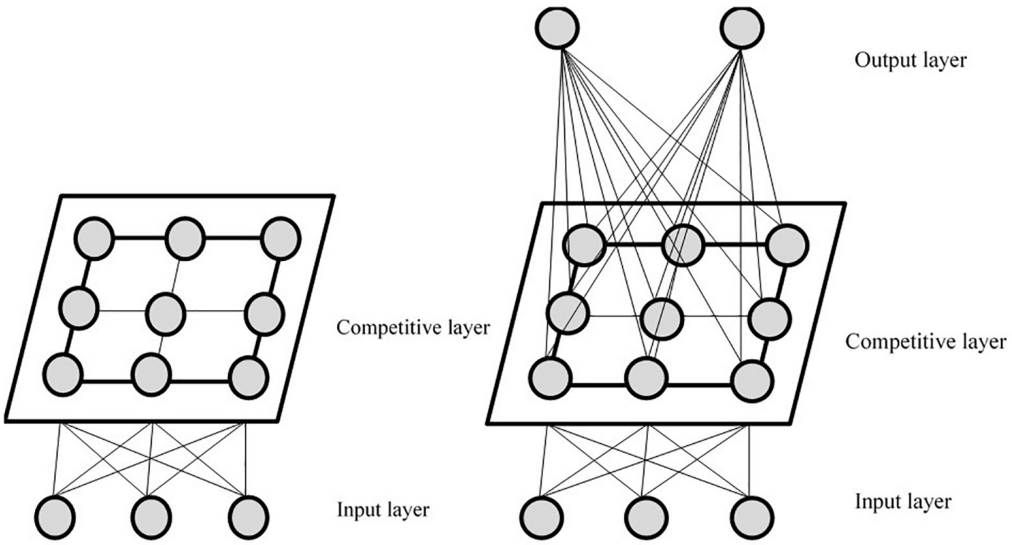


FIGURE 1 | Basic structure of self-organizing map and supervised self-organizing map, respectively.

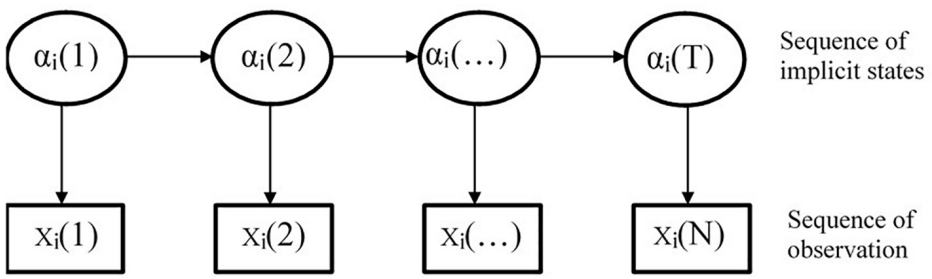


FIGURE 2 | Hidden Markov model.

$$0 \leq b_{jk} \leq 1, \ \sum_{k=1}^{N} b_{jk} = 1 \tag{11}$$

That is, in state j, the probability that the observation is k.

In general, HMM consists of two parts. One of which is a Markov chain (namely, the transition model), which is utilized to describe the change of the hidden state. It is described by the initial state π and the transition probability matrix A, and different π and A determine the different topological structures of the Markov chain and affect the complexity of the model. The other part is the measurement model (namely, the observation probability), which is determined by the observation probability matrix B, connecting the observation score and the hidden state.

HMM/ANN Model

HMM is composed of two parts: the Markov chain and the measurement model. It has a very strong modeling capability of dynamic temporal sequence, which can help us solve issues in timing changes and provide an excellent theoretical framework for realizing the longitudinal tracking of cognitive skills. However, HMM does not have a strong classification ability and cannot be directly used for longitudinal CDA as the measurement model in HMM is not suitable for cognitive diagnosis analysis. The observed probability in HMM represents the probability that the individual in each potential state makes a correct or specific response to each item. HMM can be regarded as an exploratory method to mark the potential state which is based on the item response probability (Nylund et al., 2007). However, CDA is different. This is because the categories of attributes or attribute mastery patterns are known in CDA, and it is necessary to obtain the probability of each category, which is a type of confirmatory process. ANN, on the contrary, has a strong classification ability, which can make up for this shortcoming. Considering the respective superiority of HMM and ANN, it is worth further exploring whether the HMM and the ANN can be combined to realize longitudinal CDA.

At present, there are several ways to combine HMM with ANN, one of which is to calculate the observation probability of the HMM through the ANN model, taking ANN as the measuring model of HMM (Bourlard and Morgan, 1997), as shown in **Figure 3**. The concrete implementation method is

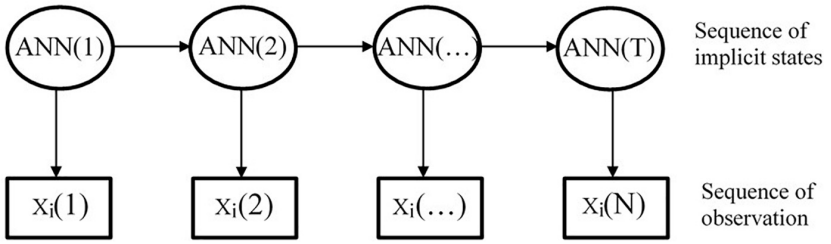


FIGURE 3 | Hidden Markov model/artificial neural network model.

to use ANN model to calculate the observation probability of HMM, that is, to calculate the observation sequence probability value under each state through the ANN model. Then, the transformation probability of the observation sequence is obtained through the Markov chain in HMM.

SIMULATION STUDY

Research Question

The simulation study includes the following question:

> Can the HMM/ANN model accurately track the development of students' cognitive skills?

Method

Data Generation

This study simulates longitudinal data with three time points (T1/T2/T3) based on the DINA model, and several key factors were manipulated, including the number of items (20 or 40), the number of attributes (3 or 6), sample size (200, 500, or 1,000) as well as item discrimination (high or mixed). High item discrimination indicates smaller slip and guess parameters, which are randomly generated on the uniform distribution $U(0, 0.20)$. The mixed discriminability contains both small and large slip and guess parameters, which were randomly generated based on the uniform distribution $U(0, 0.40)$. The selection of these factor levels is based on typical settings in recent simulation studies (e.g., Henson and Douglas, 2005; Rupp and Templin, 2008; Templin et al., 2008; de La Torre and Lee, 2010; Cui et al., 2016). Notably, based on the item and attribute level, four Q matrices were established. The Q matrices, along with item parameters for 20 items, are presented in **Table 1**.

To evaluate whether the HMM/ANN model can accurately track the development of students' cognitive skills, this study fixes the initial mastery probability as well as the transition probability of each attribute. Through combining these two probabilities, the attribute mastery probability and the increase of the attribute mastery probability at time points T2 and T3 can be obtained. Based on previous studies (Madison and Bradshaw, 2018) on the setting of the initial attribute mastery probability, this study sets the initial attribute mastery probability as 0.4, 0.4, and 0.2 under the condition of three attributes. Notably, this study assumes that it is unlikely that students' mastery of the first two attributes will decrease in a relatively short teaching period, while the mastery of the third attribute may decline. Consequently, the transition probability of attribute loss is 0.1, 0.08, and 0.38, respectively. Under the condition of six attributes, the initial attribute mastery probability is 0.4, 0.4, 0.3, 0.3, 0.2, and 0.2. Meanwhile, the attribute loss transfer probability is 0.03, 0.04, 0.13, 0.25, 0.43, and 0.54, respectively. The correlation coefficient between the attributes at the initial time point is fixed at 0.5.

Table 2 depicts the transition probability matrix of each attribute under the conditions of the three and six attributes. In the matrix, 0 indicates that the student did not master the attribute, whereas 1 indicates that the student mastered the attribute. The matrix (from left to right) reflects the probability of students moving from non-mastery to non-mastery, from non-mastery to mastery, from mastery to non-mastery, and from mastery to mastery. **Table 3** illustrates the probability and the growth rate of students' attribute mastery at each time point.

To simulate the observed item response, the students' "true" attribute pattern must be generated. In order to ensure an equal representation of the different attribute patterns, we assumed that student attribute patterns satisfy a uniform distribution. According to the students' "true" attribute pattern, Q matrix, item slid and guess parameter, and students' responses to each item were simulated. To train the neural network, input and output data are essential, that is, students' "true" attribute patterns and their response to items are required, which cannot be obtained through practice. Therefore, the ideal response, the ideal response vector, and its related true attribute pattern are utilized to train the neutral network. Furthermore, since we set the transition probability of each attribute, our simulated data can reflect the growth of cognitive skills. There are $2 \times 2 \times 3 \times 2 = 24$ conditions for each point in time. To obtain stable simulation results, each condition was repeated 30 times. Specifically, the R.3.1.0 (R Development Core Team, 2006) software CDM package was applied to generate data.

SSOM for CDA

The SSOM was used to classify the simulated item response. SSOM is comprised of three layers: the input layer, the competition layer, and the output layer. The number of nodes in the input layer of SSOM is the number of items (20 or 40), and the input data is the students' response to each item. The number of nodes in the output layer represents the number of attribute mastery pattern categories, and three or six skills correspond to the $2^3 = 8$ or $2^6 = 64$ attribute mastery pattern, and the data in the

TABLE 1 | Q matrices and item parameters used in the simulation (20 items).

Item				Discrimination										Discrimination			
	k = 3			High		Mixed		*k* = 6						High		Mixed	
	A1	A2	A3	s	g	s	g	A1	A2	A3	A4	A5	A6	s	g	s	g
1	1	1	0	0.16	0.04	0.33	0.1	1	0	1	1	0	0	0.08	0.17	0.4	0.35
2	0	1	1	0.17	0.15	0.24	0.28	1	1	0	0	0	1	0	0.05	0.35	0.09
3	1	1	0	0.01	0.1	0.26	0.24	1	0	1	1	0	1	0.03	0.15	0.21	0.22
4	1	1	1	0.17	0.05	0.12	0.22	0	1	0	1	1	0	0.11	0.05	0.4	0.21
5	0	1	1	0.1	0.13	0.37	0.15	1	0	0	0	1	0	0.03	0.14	0.39	0.03
6	0	1	0	0.12	0.07	0.35	0.04	0	0	1	0	0	1	0.03	0.13	0.02	0.02
7	1	0	0	0.07	0.15	0.09	0.34	1	1	0	1	0	1	0.19	0.02	0.31	0.16
8	1	0	1	0.08	0.17	0.39	0.14	1	1	1	1	1	0	0.18	0.19	0.03	0.14
9	0	1	0	0.03	0.1	0.19	0.36	1	0	0	1	0	0	0.07	0.06	0.28	0.16
10	0	1	1	0.16	0.07	0.31	0.3	1	1	0	0	0	0	0.03	0.19	0.3	0
11	1	0	1	0.1	0.17	0.11	0.38	0	0	0	1	0	0	0.08	0.16	0.06	0.05
12	0	1	0	0.16	0.11	0.3	0.04	0	0	0	1	0	0	0.06	0.05	0.11	0.32
13	1	0	1	0.13	0.1	0.33	0.19	1	0	0	1	0	0	0	0	0.32	0.21
14	0	1	1	0.03	0.13	0.09	0.18	0	1	0	0	1	1	0.12	0.04	0.33	0
15	1	0	1	0.05	0.17	0.37	0.39	1	1	1	1	1	1	0.05	0.04	0.08	0.11
16	1	1	0	0.05	0.01	0.29	0.06	0	1	0	1	1	1	0.15	0.06	0.12	0.12
17	1	0	1	0.14	0.05	0.39	0.1	0	0	1	0	0	0	0.05	0.16	0.32	0.09
18	1	0	0	0.1	0.17	0.31	0.01	1	1	0	0	0	1	0.07	0.01	0.22	0.04
19	0	1	0	0.11	0.03	0.12	0.19	1	0	0	0	0	1	0.06	0.05	0.32	0.05
20	0	1	1	0.14	0.18	0.18	0.06	1	1	1	1	0	1	0.2	0.04	0.14	0.23

TABLE 2 | Conditional transition probability.

Number of attributes	Attribute		T1–T2			T2–T3	
			0	1		0	1
3	A1	0	0.55	0.45	0	0.66	0.34
		1	0.03	0.97	1	0.06	0.94
	A2	0	0.71	0.29	0	0.82	0.18
		1	0.03	0.97	1	0.05	0.95
	A3	0	0.93	0.07	0	0.9	0.1
		1	0.87	0.13	1	0.54	0.46
6	A1	0	0.45	0.55	0	0.36	0.64
		1	0.03	0.97	1	0.02	0.98
	A2	0	0.56	0.44	0	0.42	0.58
		1	0.04	0.96	1	0.05	0.95
	A3	0	0.61	0.39	0	0.54	0.46
		1	0.13	0.87	1	0.14	0.86
	A4	0	0.72	0.28	0	0.46	0.54
		1	0.25	0.75	1	0.06	0.94
	A5	0	0.78	0.22	0	0.42	0.58
		1	0.43	0.57	1	0.18	0.82
	A6	0	0.83	0.17	0	0.59	0.41
		1	0.54	0.46	1	0.12	0.88

output layer is the attribute mastery pattern. Notably, there are two phases to implement SSOM to estimate the attribute patterns of the simulated item response:

Step 1: Training phase

This study simulates the data of the training set, applying the simulated ideal item response as the input value of the training set and the true attribute pattern as the output value of the training set. The input and the output layers are known, so only the number of neurons in the competing

TABLE 3 | Attribute mastery probability and growth rate.

Number of attributes	Attribute		T1 (%)	T2 (%)	Growth rate	T3 (%)	Growth rate
3	A1	Non-mastery probability	0.6	0.37	0.23	0.23	0.14
		Mastery probability	0.4	0.63		0.77	
	A2	Non-mastery probability	0.6	0.46	0.14	0.34	0.11
		Mastery probability	0.4	0.54		0.66	
	A3	Non-mastery probability	0.8	0.7	0.1	0.62	0.08
		Mastery probability	0.2	0.3		0.38	
6	A1	Non-mastery probability	0.6	0.34	0.26	0.15	0.20
		Mastery probability	0.4	0.66		0.85	
	A2	Non-mastery probability	0.6	0.35	0.25	0.20	0.15
		Mastery probability	0.4	0.65		0.80	
	A3	Non-mastery probability	0.7	0.51	0.19	0.39	0.12
		Mastery probability	0.3	0.49		0.61	
	A4	Non-mastery probability	0.7	0.57	0.13	0.45	0.11
		Mastery probability	0.3	0.43		0.55	
	A5	Non-mastery probability	0.8	0.68	0.12	0.55	0.13
		Mastery probability	0.2	0.32		0.45	
	A6	Non-mastery probability	0.8	0.74	0.06	0.64	0.10
		Mastery probability	0.2	0.26		0.36	

layer needs to be determined. Cui et al. (2016) empirically suggested that the number of nodes in the competing layer should be set to four to 10 times the number of attribute mastery patterns. This study conducted experiments on the number of neurons in the competitive layer under the conditions of three and six attributes and discovered that the number of neurons had no profound impact on the classification accuracy. Consequently, the structure of the neurons in the competitive layer was finally set to 10*10 and 20*20 under the conditions of three and six attributes, respectively.

Moreover, the classification accuracy of SSOM is greatly influenced by the number of iterations. The more iterations, the higher the classification accuracy. As the number of iterations increases, however, so does the elapsed time. Because of this, it is necessary to determine the appropriate number of iterations. This study further explored the classification accuracy of SSOM under different iterations through experiments to determine the iterations. Firstly, the number of iterations was set to 1, and the classification accuracy of the training set was recorded. Then, the number of iterations was increased one by one and the process was repeated until the accuracy of the training set became stable. The accuracy of the training set was stable after two iterations under the condition of three attributes, and the accuracy was 99.5% and 100% for 20 and 40 items. As a result, the number of iterations under this condition is determined to be two. Additionally, the accuracy of the training set became stable after four and seven iterations under the condition of six attributes, 20 and 40 items.

Step 2: Testing phase

After determining the structure and the number of iterations of the SSOM and training the neural network, the well-trained network can perform the diagnostic classification of cognitive skills on the simulated observed item response. If the attribute mastery patterns of the simulation data are estimated, the attribute accuracy rate (ACCR) and the pattern accuracy rate (PCCR) will be calculated through comparing the "true" and the estimated attribute mastery pattern. These two indicators were used as the primary criteria to evaluate the classification accuracy of SSOM. The training and the testing of SSOM were implemented through using the PyCharm software.

The Implementation of the HMM/ANN Model

In this study, the HMM is taken as the overall model, in which SSOM is used for the measurement model to realize the classification of the item response at each time point, while the transition model part is a Markov model. We actually performed two steps to complete the entire model:

Step 1: SSOM is used to calculate the observation probability of HMM

Based on the two phases mentioned in "SSOM for CDA" the SSOM model is used to calculate the probability of the observation sequence in each state, that is, to obtain the information of the attribute mastery pattern at each time point. This is actually the completion of the measurement model of HMM.

Step 2: Calculate the transition probability for HMM

Then, the Markov model in HMM was implemented to obtain information of students' attribute growth. The transition probability of the attribute mastery pattern between time points was calculated by applying the Markov chain. Meanwhile, Matlab was used to calculate the transformation probability. By comparing the true value and the estimated value of attribute

transfer probability, the average correct transformation rates of the HMM/ANN model was evaluated.

Results

Table 4 presents the classification accuracy of SSOM under the three attributes at each time point. Notably, the number of items has a positive influence on the classification accuracy—the larger the number of items, the higher the classification accuracy of SSOM. For instance, compared with 20 items, the classification accuracy of the attribute mastery pattern by SSOM increased from 0.91 to 0.97 under the condition of high discrimination, 500 samples, and 40 items. Furthermore, the discrimination has a positive influence on the classification accuracy. With the decrease of item discrimination, the classification accuracy of each attribute and the attribute mastery pattern also decrease. For instance, the classification accuracy of the attribute pattern by SSOM is between 0.91 and 0.92 under the condition of 500 samples and high discrimination. Meanwhile, in the case of mixed discrimination, the classification accuracy is between 0.74 and 0.80. This is consistent with our expectations, and items with low discrimination are difficult to distinguish as to whether students have mastered or not. Additionally, the influence of sample size is relatively small or absent, with the other conditions unchanged. The classification accuracy of the attribute pattern is between 0.84 and 0.91 under the condition of 200 samples, 20 items, and high discrimination. Moreover, the classification accuracy of the attribute pattern is between 0.91 and 0.92, under the same condition of 500 and 1,000 samples. It can be seen that, under the condition of 200 samples and 20 items, the classification

TABLE 4 | Attribute classification accuracy (k = 3).

Sample size	Number of items	Item discrimination	Time point	Classification accuracy			
				A1	A2	A3	Attribute mastery pattern
200	20	High (0, 0.2)	T1	0.9	0.97	0.96	0.84
			T2	0.93	0.98	0.98	0.87
			T3	0.95	0.98	0.98	0.91
		Mixed (0, 0.4)	T1	0.84	0.89	0.9	0.75
			T2	0.88	0.92	0.92	0.74
			T3	0.91	0.94	0.94	0.79
	40	High (0, 0.2)	T1	0.99	0.98	0.98	0.97
			T2	0.98	0.98	0.99	0.94
			T3	0.98	0.98	0.99	0.94
		Mixed (0, 0.4)	T1	0.93	0.88	0.92	0.75
			T2	0.89	0.84	0.93	0.68
			T3	0.87	0.84	0.94	0.68
500	20	High (0, 0.2)	T1	0.94	0.98	0.98	0.91
			T2	0.94	0.98	0.99	0.92
			T3	0.94	0.98	0.99	0.91
		Mixed (0, 0.4)	T1	0.8	0.89	0.93	0.74
			T2	0.8	0.91	0.94	0.78
			T3	0.8	0.92	0.93	0.8
	40	High (0, 0.2)	T1	0.99	0.99	0.99	0.97
			T2	0.99	0.99	0.99	0.96
			T3	0.99	0.99	0.99	0.98
		Mixed (0, 0.4)	T1	0.89	0.91	0.91	0.71
			T2	0.84	0.93	0.9	0.69
			T3	0.82	0.94	0.9	0.69
1,000	20	High (0, 0.2)	T1	0.94	0.97	0.99	0.91
			T2	0.94	0.98	0.99	0.92
			T3	0.93	0.98	0.98	0.92
		Mixed (0, 0.4)	T1	0.8	0.91	0.94	0.68
			T2	0.81	0.91	0.93	0.65
			T3	0.82	0.92	0.92	0.61
	40	High (0, 0.2)	T1	0.99	0.99	0.99	0.99
			T2	0.99	0.99	0.99	0.99
			T3	0.99	0.99	1	0.99
		Mixed (0, 0.4)	T1	0.87	0.9	0.89	0.71
			T2	0.83	0.92	0.89	0.71
			T3	0.82	0.93	0.91	0.76

accuracy of SSOM is slightly lower than that of 500 or 1,000. Meanwhile, under the condition of 200 samples, 40 items, and high discrimination, the classification accuracy of the attribute pattern is between 0.94 and 0.97, which is very close to the classification accuracy under the condition of sample size 500 and 1,000. Generally, SSOM has a slightly different classification accuracy under the sample size of 200, 500, and 1,000, but its classification accuracy is generally relatively consistent.

Table 5 illustrates the classification accuracy of SSOM under the six-attributes condition. Through comparing **Tables 4**, **5**, it can be seen that when the number of attributes increased from three to six, the classification accuracy of SSOM decreases sharply. For instance, the classification accuracy of the attribute mastery pattern at the first time point is 0.91 under the condition of three attributes, 20 items, 500 samples, and a high degree of discrimination. Under the same condition, however, when the number of attributes is six, the classification accuracy of the attribute mastery pattern is 0.65. This is consistent with previous studies (Cui et al., 2016). When the number of attributes increases, the classification accuracy of both ANN and traditional CDM is poor. Moreover, the influence of sample size, number of items, and item discrimination is consistent with the results under the condition of three attributes, which will not be repeated here.

Table 6 depicts the correct transition rate of the attribute mastering patterns obtained through the HMM/ANN model

TABLE 5 | Attribute classification accuracy (k = 6).

Sample size	Number of items	Item discrimination	Time point	Classification accuracy						
				A1	A2	A3	A4	A5	A6	Attribute mastery pattern
200	20	High (0, 0.2)	T1	0.74	0.71	0.74	0.85	0.82	0.90	0.45
			T2	0.78	0.78	0.79	0.86	0.80	0.62	0.37
			T3	0.86	0.89	0.83	0.88	0.80	0.83	0.56
		Mixed (0, 0.4)	T1	0.64	0.61	0.64	0.72	0.68	0.80	0.35
			T2	0.73	0.73	0.70	0.75	0.71	0.52	0.31
			T3	0.77	0.82	0.76	0.79	0.73	0.71	0.44
	40	High (0, 0.2)	T1	0.71	0.74	0.81	0.81	0.79	0.86	0.43
			T2	0.86	0.87	0.82	0.89	0.90	0.90	0.63
			T3	0.91	0.93	0.89	0.92	0.93	0.96	0.71
		Mixed (0, 0.4)	T1	0.59	0.62	0.74	0.70	0.66	0.81	0.33
			T2	0.76	0.77	0.73	0.79	0.74	0.78	0.39
			T3	0.86	0.85	0.78	0.84	0.76	0.85	0.57
500	20	High (0, 0.2)	T1	0.89	0.87	0.93	0.95	0.98	0.85	0.65
			T2	0.87	0.88	0.93	0.94	0.93	0.88	0.76
			T3	0.9	0.91	0.94	0.94	0.94	0.82	0.72
		Mixed (0, 0.4)	T1	0.77	0.8	0.84	0.8	0.81	0.74	0.36
			T2	0.77	0.8	0.83	0.8	0.82	0.74	0.37
			T3	0.84	0.84	0.86	0.81	0.83	0.71	0.36
	40	High (0, 0.2)	T1	0.88	0.9	0.82	0.92	0.94	0.92	0.7
			T2	0.89	0.92	0.87	0.91	0.94	0.92	0.75
			T3	0.89	0.91	0.9	0.91	0.94	0.94	0.73
		Mixed (0, 0.4)	T1	0.77	0.8	0.74	0.82	0.77	0.8	0.44
			T2	0.79	0.83	0.77	0.78	0.83	0.84	0.45
			T3	0.83	0.84	0.8	0.85	0.78	0.86	0.49
1,000	20	High (0, 0.2)	T1	0.87	0.89	0.91	0.94	0.93	0.88	0.58
			T2	0.87	0.88	0.91	0.94	0.93	0.86	0.56
			T3	0.9	0.92	0.93	0.94	0.93	0.83	0.64
		Mixed (0, 0.4)	T1	0.81	0.77	0.73	0.82	0.83	0.87	0.36
			T2	0.78	0.81	0.81	0.82	0.84	0.74	0.35
			T3	0.86	0.86	0.84	0.81	0.84	0.71	0.39
	40	High (0, 0.2)	T1	0.88	0.9	0.84	0.92	0.92	0.91	0.65
			T2	0.89	0.91	0.85	0.91	0.94	0.93	0.65
			T3	0.89	0.92	0.89	0.91	0.94	0.95	0.68
		Mixed (0, 0.4)	T1	0.84	0.85	0.77	0.85	0.75	0.85	0.45
			T2	0.79	0.82	0.76	0.79	0.82	0.85	0.44
			T3	0.85	0.86	0.81	0.81	0.82	0.89	0.52

TABLE 6 | Correct transition rate of attribute master pattern.

Number of attributes	Sample size	Number of items	Item discrimination	Correct transition rate	
				T1–T2	T2–T3
3	200	20	High	0.95	0.96
			Mixed	0.91	0.92
		40	High	0.98	0.99
			Mixed	0.91	0.93
	500	20	High	0.98	0.97
			Mixed	0.92	0.9
		40	High	0.98	0.98
			Mixed	0.9	0.87
	1,000	20	High	0.98	0.98
			Mixed	0.94	0.95
		40	High	0.99	0.99
			Mixed	0.92	0.94
6	200	20	High	0.98	0.97
			Mixed	0.98	0.98
		40	High	0.99	0.99
			Mixed	0.98	0.97
	500	20	High	0.98	0.97
			Mixed	0.97	0.97
		40	High	0.98	0.99
			Mixed	0.97	0.98
	1,000	20	High	0.98	0.97
			Mixed	0.97	0.96
		40	High	0.98	0.99
			Mixed	0.98	0.98

under each simulation condition. Under the condition of three attributes, the HMM/ANN model demonstrates a high correct transition rate from time 1 to time 2 and time 2 to time 3. The discrimination also has a positive influence on the correct transition rate. Notably, the correct transition rate is higher under the high-discrimination condition than in the mixed-discrimination condition. Under the condition of high discrimination, the HMM/ANN model has a high correct transition rate, which is 0.95–0.99. Under the mixed-discrimination condition, the correct transition rate is reduced, which is 0.87–0.95. The influence of the number of items on the HMM/ANN model is not clear in this simulation study. For example, under the condition of 20 items, the correct transition rate of the HMM/ANN model is 0.90–0.98. Meanwhile, under the condition of 40 items, the correct transition rate was 0.87–0.99. Additionally, the correct transition rate of the HMM/ANN model was also unaffected by the sample size.

Under the condition of six attributes, the correct transition rate of the HMM/ANN model is at a high level, ranging from 0.97 to 0.99, and it was difficult to identify the influence of sample size, the number of items, and the quality of items.

Even in the case of six attributes, the classification accuracy of ANN is reduced, but it does not affect the correct transition rate of the longitudinal model. This may be attributed to the fact that when six attributes are examined, there will be $2^6 = 64$ attribute mastery patterns, forming a 64*64 transfer probability matrix, which is too large, thus affecting the calculation of the correct transition rate.

EMPIRICAL STUDY

Research Question

The empirical study includes the following question:

> What is the effectiveness of the HMM/ANN model in real situations through tracking students' mastery and development of cognitive skills based on actual reading literacy assessment data?

Method

The empirical study analyzed the data of a reading literacy assessment completed by a school in Beijing. There were 190 students who completed the same reading passage—*book* in grade 4 (2015) and grade 5 (2016), which contains a total of eight items. All eight items are scored 0 (incorrect) or 1 (correct). The selected short test examines three skills of acquisition, integration, and evaluation and are examined by two, three, and three questions. The skills examined by each item are displayed in **Table 7**.

The quality of the eight items is good. In terms of the fourth-grade test, the items have medium discriminations

TABLE 7 | The Q matrix of empirical data.

Item	Acquisition	Integration	Evaluation
1	1	0	0
2	1	0	0
3	0	1	0
4	0	1	0
5	0	0	1
6	0	0	1
7	0	0	1
8	0	1	0

between 0.31 and 0.46, except for item 6 which has low discrimination of 0.21. For the fifth-grade test, the discrimination is lower than 0.3 (0.28 and 0.29), except for items 1 and 6. The other items have medium discrimination between 0.31 and 0.43.

The three-layer SSOM network structure was selected in the empirical study. Training of the neural network needs to include both input and output data. In the empirical study, however, we only have the input data in the testing set, namely, the observed item responses of students. To train the SSOM, it is necessary to have both input and output data, that is, the true attribute patterns of students and their response to items, which cannot be obtained in empirical data. Drawing from previous experience, we simulated the ideal item response and true attribute patterns based on the *Q* matrix of empirical data, which are used as the input and the output data of the training set. R.3.1.0 (R Development Core Team, 2006) CDM package was used to generate the training set data, and then PyCharm was used to train and test the SSOM. The determination of the number of nodes in the SSOM competition layer and the number of iterations is the same as in the simulation study. Finally, the network structure of the competition layer was set to 9*9, and the number of iterations was determined to be three. Afterward, the observed responses were classified by the well-trained neural network. Similar to the simulation study, ACCR and PCCR were used as the main criteria to evaluate the classification accuracy of SSOM. As mentioned earlier, we simulated the ideal item response and the true attribute mastery patterns based on the *Q* matrix of the empirical data, so ACCR and PCCR can be successfully calculated by comparing the true and the estimated attribute mastery patterns. Then, Matlab was applied to calculate the transformation probability matrix.

Results

Table 8 reflects the classification accuracy of SSOM for the three attributes examined in the fourth- and the fifth-grade tests as 0.97, 0.98, and 0.90 and 0.98, 0.95, and 0.91, respectively. The classification accuracy of the attribute master pattern is 0.87and 0.85, respectively. Notably, the results of the empirical study are similar to those of the simulation study. SSOM provided an accurate classification at two time points when the tests examined fewer skills and the quality of items was higher.

TABLE 8 | The classification accuracy of supervised self-organizing map.

	Acquisition	Integration	Evaluation	Attribute mastery pattern
Fourth grade	0.98	0.99	0.91	0.87
Fifth grade	0.98	0.95	0.91	0.85

The development of students' reading ability with time is displayed in **Figure 4**. The reading ability of students has improved during year 1. For example, the average reading ability increased from 0.55 to 1.40 from the fourth to the fifth grade.

Figure 5 illustrates the mastery of each attribute in grades four and five. In total, the mastery probability of these three attributes increases with time. For fourth-grade students, the attribute mastery probability is between 0.53 and 0.81, and the average mastery probability is 0.72. The attribute mastery probability is between 0.72 and 0.93 for fifth-grade students, and the average mastery probability is 0.68. Moreover, it can be observed that the mastery probability of acquisition and integration demonstrates the same growth trend. Additionally, the growth trend of evaluation is flatter, and the growth range of the three attributes is between 0.04 and 0.19.

Table 9 depicts the transformation probability matrix of each attribute. The four cells in each 2*2 matrix represent (from left to right) non-mastery to non-mastery, non-mastery to mastery, mastery to non-mastery, as well as mastery to mastery. For the attributes of acquisition, integration, and evaluation, the probability from non-mastery to mastery is 0.80, 0.66, and 0.71, and the probability from mastery to non-mastery is 0.05, 0.19, and 0.14, respectively. This suggests that the majority of students can achieve the transition from non-mastery to mastery at the two time points, and a small percentage of students return from mastery to non-mastery.

Table 10 illustrates the transformation probability matrix of eight attribute mastery patterns. It can be seen that, among students who did not master any skills (000) in the fourth grade, there were still 25% of students who did not even master three skills in the fifth grade, and 38% of the students were able to master all three skills in the fifth grade, which shows a dramatic improvement. For students who fully mastered the three skills (111) in the fourth grade, 71% were still able to master the three skills in the fifth grade. For other categories of attribute mastery patterns, 40–63% of students were able to master all skills in the fifth grade.

Longitudinal CDA can also assist in obtaining information regarding individuals. For example, the student with ID 3410105 scored 0.25 on average on eight items in the fourth grade, and their attribute mastery pattern was "100". In the fifth grade, they scored 0.75, and their attribute mastery pattern was "111," which means that they mastered all three skills. It can be seen that, after a year's study, the students' reading skills have significantly improved and they are able to master integration and evaluation skills. For the student with ID 3430308, they scored 0.38 on average on eight items in the fourth grade, and their attribute mastery pattern was "100". When the student was in the fifth grade, they scored 0.50, but their attribute mastery pattern

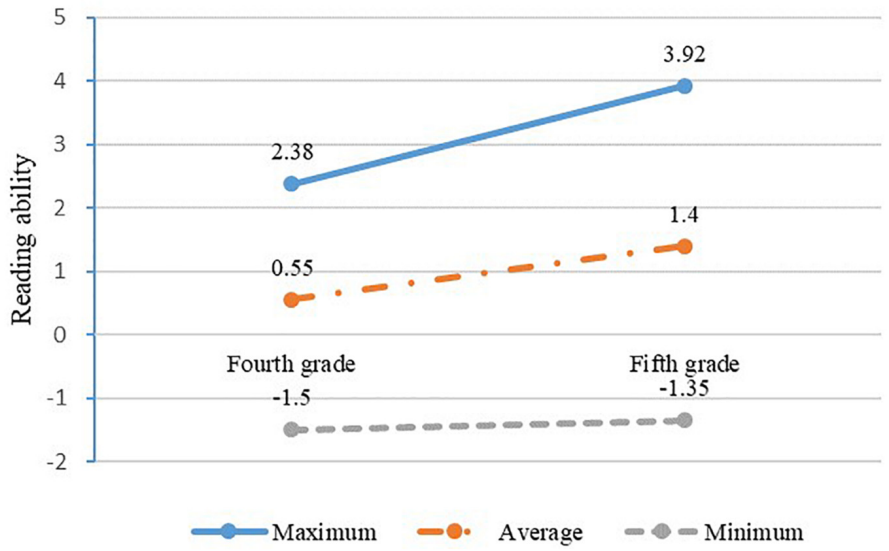


FIGURE 4 | The change in students' reading ability.

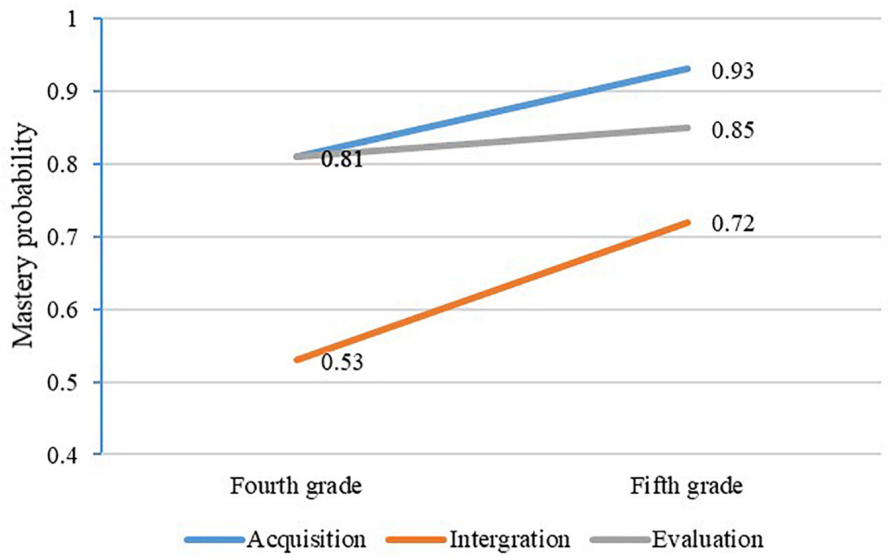


FIGURE 5 | The change of attribute mastery probability.

was also "100". Which indicates that they had not mastered integration and evaluation skills.

DISCUSSION

HMM/ANN Model Achieved Fine-Grained Longitudinal Tracking of Students' Cognitive Skills

Previously, researchers tracked the development of cognitive skills in three ways. Firstly, they used the Multidimensional Item Response Theory (MIRT; Andersen, 1985; Embretson, 1991) to track changes in students' abilities. The second was to integrate traditional CDMs, such as the DINA and the DINO models, within the framework of HMM (Li et al., 2016; Kaya and Leite, 2017; Madison and Bradshaw, 2018; Hung and Huang, 2019). The third was to construct higher-order latent structures for measuring growth to explain the relationship among multiple latent attributes (Hansen, 2013; Zhan et al., 2019). The HMM/ANN model proposed in this study is a further enrichment of longitudinal CDMs. Meanwhile, the second approach is consistent with the overall idea of establishing the HMM/ANN model to realize the longitudinal tracking of the

TABLE 9 | The transition probability matrix of attributes.

Attribute		Transition matrix		
Acquisition			**Fifth grade**	
			0	**1**
	Fourth grade	0	0.20	0.80
		1	0.05	0.95
Integration			**Fifth grade**	
			0	**1**
	Fourth grade	0	0.34	0.66
		1	0.19	0.81
Evaluation			**Fifth grade**	
			0	**1**
	Fourth grade	0	0.29	0.71
		1	0.14	0.86

cognitive skills of students in this study. Notably, both utilized the processing capability of HMM for time-series changes and integrated the diagnostic classification model.

Compared with the MIRT method, the combination of these two models can achieve fine-grained longitudinal tracking of students' cognitive skills. Meanwhile, the method based on MIRT can only obtain how students develop in a single ability. However, students with the same original score may master different skills. Compared with combining HMM with traditional CDMs, the HMM/ANN model proposed in this study has an advantage in the classification accuracy of cognitive skills. As mentioned before, traditional CDM is based on the framework of IRT, and the accuracy of model parameter estimation and classification will be affected when the data are unable to meet the strong hypothesis of IRT or the sample size is small. Because it is not necessary for ANN to perform parameter estimation, it can also obtain a higher classification accuracy when the data do not meet the assumptions of unidimension, local independence, and monotonicity or when the samples are small (Cao, 2009; de La Torre et al., 2010; Chen et al., 2013). This study also supports this. Moreover, the third method takes the problem of local item dependence into account and overcomes the defects of its application in the real educational situation to some extent. However, it still used in the parameter estimation method. In contrast, the HMM/ANN model is non-linear and is not affected by the characteristics of sample distribution and data types. It does not need to meet the strong assumption of IRT or require parameter estimation and is relatively less affected by the sample size. Consequently, the HMM/ANN model is more suitable for data collected in real educational situations and does not need a large scale to obtain good results, and it can also effectively track the changes in the cognitive skills of students in the context of small sample size in schools or classes.

In addition, considering how well the model matches the data, the HMM/ANN model may not always be more powerful than other longitudinal CDMs. For example, based on the DINA model to generate data, the DINA model is used to estimate the model so that the model truly fits the data, and the result will be more powerful than the ANN model; however, when the model and the data are misfit, the advantages of ANN are obvious (Cui et al., 2016).

The Classification Accuracy of SSOM in the HMM/ANN Model Is Affected by Some Factors

The SSOM applied in this study can accurately classify cognitive skills when the test examines three attributes. However, when multiple attributes are incorporated in the test, SSOM demonstrates a lower classification accuracy. This result can be explained because, in the process of algorithm operation, it does not establish a direct mapping relationship between the students' response (input data) and the mastery of each skill (output data). It instead initially obtains the attribute mastery patterns based on the response to the input, and finally outputs the mastery of each skill. As the number of attributes increases, the total number of attribute mastery patterns increases exponentially. When the test examined only three attributes, there were a total of $2^3 = 8$ attribute mastery patterns, and each student was classified into one of the eight attribute mastery patterns. Meanwhile, when the number of attributes increases to six, $2^6 = 64$ attribute mastery patterns are generated. To classify students into the correct attribute mastery patterns, it

TABLE 10 | Transition probability matrix of attribute mastery pattern.

		T2 (fifth grade)							
		000	**001**	**010**	**100**	**110**	**011**	**101**	**111**
T1 (fourth grade)	000	0.25	0	0.13	0.13	0	0.13	0	0.38
	001	0	0.04	0.04	0.22	0	0	0.13	0.57
	010	0	0	0.13	0	0.13	0.13	0	0.63
	100	0	0.07	0	0	0.13	0	0.33	0.47
	110	0	0	0	0	0.3	0	0.3	0.4
	011	0	0	0.13	0	0.13	0.13	0	0.63
	101	0	0.02	0.02	0.02	0.12	0	0.24	0.56
	111	0	0.02	0.01	0.01	0.06	0.02	0.16	0.71

is necessary to identify the students' mastery and non-mastery of each of the six attributes. If one attribute is misclassified, the entire attribute mastery pattern is misclassified as well (Cui et al., 2016).

The classification accuracy of SSOM is also affected by other factors. The greater the number of items and the higher the quality of those items, the higher the classification accuracy of SSOM. The items with low discrimination are not suitable for distinguishing the students' mastery of skills, and it also affects the accuracy of the estimation of attributing mastery (Roussos et al., 2005). The sample size has little influence on the classification accuracy of SSOM. Notably, SSOM's classification results are relatively accurate both in empirical as well as in simulation studies, with sample sizes of 200, 500, or 1,000. Despite so, however, SSOM's classification accuracy is slightly lower than that of 500 or 1,000 samples under the condition of 20 items and 200 samples. However, when the number of questions increased, the classification accuracy of SSOM was relatively consistent under the condition of three sample sizes. Since the ANN does not need parameter estimation, the sample size has little influence on the accuracy of its classification (Gierl et al., 2008; Shu et al., 2013). Furthermore, SSOM is essentially a clustering algorithm, so it is more affected by the number of categories.

The advantage of applying the HMM/ANN model in a real educational situation is that its measurement model is ANN, which does not require parameter estimation under the framework of IRT. When the data cannot meet the hypothesis of IRT and the sample size is small, ANN can assist in obtaining accurate classification results, which makes CDA more widely applied in daily teaching.

Teachers can make use of this model to understand the development of students' cognitive skills, understand the advantages and disadvantages of students' skill mastery, and adjust teaching strategies or teaching priorities in a timely manner. However, the application of this longitudinal cognitive diagnosis model should be conducted cautiously. As the results highlight, ANN does not perform well under all conditions. To obtain a more accurate classification and accurately track the development of students' cognitive skills, teachers and educators should examine the appropriate number of attributes and design high-quality items when applying the HMM/ANN model so as to ensure the accurate classification of cognitive skills.

CONCLUSION, LIMITATIONS, AND FURTHER STUDY

This study constructs a new theoretical model of longitudinal CDA, which combines HMM with ANN, making full use of HMM's advantages to process time series information and the advantages of ANN to process non-linear information to realize the tracking of cognitive skills. This is a useful exploration of the longitudinal CDM and will help promote the technical development of the longitudinal CDA. The purpose of this study was to verify the effectiveness of the proposed HMM/ANN model in longitudinal cognitive diagnostic analysis under different conditions. The results of the simulation and the empirical studies illustrate that the HMM/ANN model can accurately classify cognitive skills and track the development of students' cognitive skills. Consequently, it is reasonable to use the developed model to track the development of students' cognitive skills. Additionally, the classification accuracy of SSOM is better when the number of attributes is low, the number of items is high, and the quality of items is better. Furthermore, the sample size has a slight influence on the classification accuracy of SSOM.

The simulation results demonstrate that the HMM/ANN model has a high correct transition rate under various simulation conditions. When the cognitive skills examined were relatively small, the correct transition rate of HMM/ANN was consistent with the classification accuracy of ANN. When the classification accuracy of ANN was low, the correct transition rate was also relatively low, which is consistent with previous studies (Kaya and Leite, 2017). When a relatively large number of skills were examined, the correct transition rate of the HMM/ANN model was overestimated. Meanwhile, as skills increased, the attribute mastery patterns increased exponentially, forming a larger transformation probability matrix of the attribute mastery pattern. However, the students' attribute mastery pattern is usually concentrated in several categories, so the correct transformation probability has been overestimated. In the empirical study, the HMM/ANN model can assist in obtaining information concerning the students' mastery of each reading skill as well as with the development of reading skills.

This study is a new attempt in using ANN in longitudinal CDA, and there are some limitations and prospects. First, when the number of attributes examined is large, the ANN still cannot achieve better classification results. In future studies, the diagnostic classification model should be further optimized so as to explore the CDMs applicable to classify more attributes. Second, the development of students' mastery of skills in the empirical study does not explain whether the development is caused by teachers' instruction or if it is from the students' natural development. However, this does not affect the results of this study because the focus is to verify whether the proposed HMM/ANN model can accurately track the development of students' cognitive skills. Moreover, it is necessary for the HMM/ANN model to be verified in more educational contexts, so the application of other types of ANNs in CDA can be further tested. Finally, this study did not compare the HMM/ANN model with more longitudinal CDMs because we are more inclined to provide an alternative model rather than to judge whether HMM/ANN has an absolute advantage. In the future, we will further compare this model with a more powerful model.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of the Faculty of Psychology, BNU. Written informed consent to participate in this study was provided by the participants' legal guardian/ next of kin.

AUTHOR CONTRIBUTIONS

HW conceived and designed the study, collected the data, and helped in performing the analysis with constructive discussions. YL and NZ performed the data analyses and wrote the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika* 50, 3–16. doi: 10.1007/BF02294143

Bourlard, H., and Morgan, N. (1997). *Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions*. Berlin: Springer Verlag.

Cao, H. Y. (2009). *Application of Artificial Neural Network in Cognitive Diagnosis*. doctoral dissertation, Jiangxi Normal University, Nanchang.

Chen, J., Kang, C. H., and Zhong, X. L. (2013). Review and prospect of non-parametric item response theory. *China. Exam.* 6, 18–25. doi: 10.19360/j.cnki.11-3303/g4.2013.06.003

Chen, Y., Culpepper, S. A., Wang, S., and Douglas, J. (2017). 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

Chiu, C. Y., Sun, Y., and Bian, Y. (2018). Cognitive diagnosis for small educational programs: the general nonparametric classification method. *Psychometrika* 83, 355–375. doi: 10.1007/s11336-017-9595-4

Collins, L. M., and Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Hoboken, NJ: John Wiley and Sons.

Collins, L. M., and Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behav. Res.* 27, 131–157. doi: 10.1207/s15327906mbr2701_8

Cui, Y., Gierl, M., and Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: an artificial neural network approach. *Educ. Psychol.* 36, 1065–1082. doi: 10.1080/01443410.2015.1062078

de La Torre, J., Hong, Y., and Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *J. Educ. Meas.* 47, 227–249. doi: 10.1111/j.1745-3984.2010.00110.x

de La Torre, J., and Lee, Y. (2010). A note on the invariance of the DINA Model parameters. *J. Educ. Meas.* 47, 115–127. doi: 10.1111/j.1745-3984.2009.00102.x

Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika* 56, 495–515. doi: 10.1007/bf02294487

Gierl, M. J., Cui, Y., and Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns to achievement tests. *J. Mod. Appl. Stat. Methods* 7, 234–245. doi: 10.22237/jmasm/1209615480

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26, 301–321. doi: 10.1111/j.1745-3984.1989.tb00336.x

Han, L. Q. (2006). *Course of Artificial Neural Network*. Beijing: Beijing university of posts and telecommunications press.

Hansen, M. (2013). *Hierarchical Item Response Models for Cognitive Diagnosis*. doctoral dissertation, University of California, Oakland, CA.

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

Hu, T. (2011). *Researchon Network Traffic Classification Method Based on Neural Network*. doctoral dissertation, Guilin university of electronic technology, Guangxi.

Hu, T., Wang, Y., and Tao, X. L. (2011). Network traffic classification method based on SSOM. *Comput. Eng.* 37, 104–106. doi: 10.3969/j.issn.1000-3428.2011.06.036

Hung, S., and Huang, H. (2019). A sequential process model for cognitive diagnostic assessment with repeated attempts. *Appl. Psychol. Meas.* 43, 495–511. doi: 10.1177/0146621618813111

Junker, B., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kaya, Y., and Leite, W. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. *Educ. Psychol. Meas.* 77, 369–388. doi: 10.1177/0013164416659314

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43, 59–69. doi: 10.1007/bf00337288

Kohonen, T. (1990). The self-organizing map. *Proc. IEEE* 78, 1464–1480. doi: 10.1109/5.58325

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd Edn. Berlin: Springer Verlag.

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

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

Lu, Y., Liu, K., Xu, S., Wang, Y., and Zhang, Q. (2020). Identifying flow units by FA-assisted SSOM—An example from the eocene basin-floor-fan turbidite reservoirs in the daluhu oilfield, dongying depression, bohai bay basin, China. *J. Pet. Sci. Eng.* 186:106695. doi: 10.1016/j.petrol.2019.106695

Macready, G. B., and Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *J. Educ. Stat.* 2, 99–120. doi: 10.2307/1164802

Madison, M., and Bradshaw, L. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika* 83, 963–990. doi: 10.1007/s11336-018-9638-5

Nylund, K., Asparouhov, T., and Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a monte carlo simulation stuDY. *Struct. Equ. Modeling.* 14, 535–569. doi: 10.1080/10705510701575396

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286. doi: 10.1109/5.18626

Roussos, L. A., Henson, R. A., and Jang, E. E. (2005). *Simulation Study Evaluation of the Fusion Model System Stepwise Algorithm*. Princeton, NJ: ETS Project Report.

Rupp, A., and Templin, J. L. (2008). The effects of Q-Matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educ. Psychol. Meas.* 68, 78–96. doi: 10.1177/0013164407301545

Shu, Z., Henson, R. A., and Willse, J. (2013). Using neural network analysis to define methods of dina model estimation for small sample sizes. *J. Classif.* 30, 173–194. doi: 10.1007/s00357-013-9134-7

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

Templin, J. L., Henson, R. A., Rupp, A., Jang, E., and Ahmed, M. (2008). Diagnostic Models for Nominal Response Data. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, New York, NY.

Tu, D. B., Cai, Y., and Ding, S. L. (2012). *Cognitive Diagnosis: Theory, Methods and Applications*. Beijing: Beijing Normal University Press.

Wang, L. L., Chen, P., Xin, T., and Zhong, K. D. (2015). Realization cognitive diagnostic computerized adaptive testing based on BP neural network. *J. Beijing Norm. Univ. Nat. Sci.* 51, 206–211. doi: 10.16360/j.cnki.jbnuns.2015.02.019

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

Wang, W. Y., Ding, S. S. L., Song, L. H., Kuang, Z., and Cao, H. Y. (2016). Application of neural networks and support vector machines to cognitive diagnosis. *Psychol. Sci.* 39, 777–782. doi: 10.16719/j.cnki.1671-6981.20160402

Wongravee, K., Lloyd, G., Silwood, C., Grootveld, M., and Brereton, R. (2010). Supervised self organizing maps for classification and determination of potentially discriminatory variables: illustrated by application to nuclear magnetic resonance metabolomic profiling. *Anal. Chem.* 82, 628–638. doi: 10.1021/ac9020566

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

Zhao, J. H., and Li, W. H. (2012). Application of supervised SOM neural network in intrusion detection. *Comput. Eng.* 38, 110–111. doi: 10.3969/j.issn.1000-3428.2012.12.032

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 September 2020
doi: 10.3389/fpsyg.2020.02120

A Semi-supervised Learning Method for Q-Matrix Specification Under the DINA and DINO Model With Independent Structure

Wenyi Wang[1], *Lihong Song*[2]*, *Shuliang Ding*[1], *Teng Wang*[1], *Peng Gao*[1] *and Jian Xiong*[1]

[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, [2] Elementary Education College, Jiangxi Normal University, Nanchang, China

Cognitive diagnosis assessment (CDA) can be regarded as a kind of formative assessments because it is intended to promote assessment for learning and modify instruction and learning in classrooms by providing the formative diagnostic information about students' cognitive strengths and weaknesses. CDA has two phases, like a statistical pattern recognition. The first phase is feature generation, followed by classification stage. A Q-matrix, which describes the relationship between items and latent skills, corresponds to the feature generation phase in statistical pattern recognition. Feature generation is of paramount importance in any pattern recognition task. In practice, the Q-matrix is difficult to specify correctly in cognitive diagnosis and misspecification of the Q-matrix can seriously affect the accuracy of the classification of examinees. Based on the fact that any columns of a reduced Q-matrix can be expressed by the columns of a reachability R matrix under the logical OR operation, a semi-supervised learning approach and an optimal design for examinee sampling were proposed for Q-matrix specification under the conjunctive and disjunctive model with independent structure. This method only required subject matter experts specifying a R matrix corresponding to a small part of test items for the independent structure in which the R matrix is an identity matrix. Simulation and real data analysis showed that the new method with the optimal design is promising in terms of correct recovery rates of q-entries.

Keywords: cognitive diagnostic assessment, Q-matrix, the augment algorithm, the reachability matrix, the conjunctive model, the disjunctive model

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, United States

Reviewed by:
Pasquale Anselmi,
University of Padua, Italy
Feiming Li,
Zhejiang Normal University, China

***Correspondence:**
Lihong Song
viviansong1981@163.com

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

Received: *03 January 2020*
Accepted: *30 July 2020*
Published: *10 September 2020*

Citation:
Wang W, Song L, Ding S, Wang T, Gao P and Xiong J (2020) A Semi-supervised Learning Method for Q-Matrix Specification Under the DINA and DINO Model With Independent Structure. Front. Psychol. 11:2120. doi: 10.3389/fpsyg.2020.02120

INTRODUCTION

In educational assessment, cognitive diagnostic assessment (CDA) that combines psychometrics and cognitive science has received increased attention recently (Leighton and Gierl, 2007; Tatsuoka, 2009; Rupp et al., 2010). This approach potentially provides useful diagnostic information regarding students' strengths and weaknesses, and can facilitate individualized learning (Chang, 2015). Cognitive diagnostic models (CDMs) often utilize a Q-matrix (Embretson, 1984; Tatsuoka, 1990, 1995, 2009). Tatsuoka (2009) pointed out that "Tatsuoka (1990) organized the underlying cognitive processing skills and knowledge that are required in answering test items correctly in a Q-matrix, in which the rows represent attributes and the columns represent items." The entries of a Q-matrix

are 1 or 0, denoted by q_{kj}. If attribute k is involved in correctly answering item j, then $q_{kj} = 1$, and $q_{kj} = 0$ otherwise. The definition of Q-matrix in Tatsuoka (1990) is used in our study. Recently, one common representation of a Q-matrix is that in which the rows represent items and the columns represent attributes (Ma and de la Torre, 2020; Zhan et al., 2020). It should be noted that the representation of the Q-matrix that they used in the study differs from the traditional one.

Cognitive diagnostic assessment has two phases, like statistical pattern recognition and classification methodology. The first phase is feature generation, and then classification stage follows. The specification of Q-matrix corresponds to the feature extractor phase in statistical pattern recognition and classification problems. Feature generation is of paramount importance in any pattern recognition task. So, the Q-matrix plays a very important role in establishing the relation between latent attribute patterns and ideal/latent response patterns.

In practice, the Q-matrix is difficult to specify correctly in cognitive diagnostic assessment (Jang, 2009; DeCarlo, 2011) and misspecification of the Q-matrix can seriously affect the accuracy of both item parameter estimates and the classification of examinees (de la Torre, 2008; Rupp and Templin, 2008). Researchers have proposed several quantitative methods for deriving or refining Q-matrix. These methods can be classified into two categories (Xu and Desmarais, 2018): (a) the unsupervised method, including but not limited to the q-matrix method (Barnes, 2003, 2011), the non-negative matrix factorization technique (Desmarais, 2011; Desmarais et al., 2012; Desmarais and Naceur, 2013) or alternate least-square factorization method (Desmarais et al., 2014; Xu and Desmarais, 2016), the data-driven approach (Liu et al., 2012, 2013), and the exploratory factor analysis method (Barnes, 2003; Close, 2012; Wang et al., 2018b, 2020), and (b) the supervised method, including the sequential EM-based δ method (de la Torre, 2008) and its extension ς^2 method (de la Torre and Chiu, 2016), the Bayesian approach (DeCarlo, 2012), the non-parametric Q-matrix refinement method (Chiu, 2013), the stepwise reduction algorithm (Hartz, 2002), the EM-based methods (Wang et al., 2018a), the residual-based or item fit statistic approach (Chen, 2017; Kang et al., 2018) and so on.

The unsupervised method is deriving a Q-matrix only from test data or item responses. The unsupervised method is very useful because there are many existing tests without specifying the Q-matrix but with test response data. However, it would be difficult to identify the number of latent skills and be slightly more difficult to understand results from real data. A study of Beheshti et al. (2012) found that the number of latent skills estimated from real data is not well-aligned with the assessment of experts.

The supervised method can incorporate the information of experts' Q-matrix and test response data to refine or validate the provisional Q-matrix. If the provisional Q-matrix is unknown for an existing test, the supervised methods cannot be used. Furthermore, this method often needs a high-quality provisional Q-matrix for a whole test. If the provisional Q-matrix is specified by subject matter experts but contains a large amount of misspecification, it will be difficult for the recovery of a high-quality Q-matrix through the supervised method, because the performance of the supervised method relies on the precision of classification of attribute patterns resulting from the provisional Q-matrix (de la Torre, 2008; Rupp and Templin, 2008).

Specifying a Q-matrix for a whole test by experts can be a time-consuming and fatigue process. The purpose of this study is to propose a semi-supervised method for Q-matrix specification in order to check whether only some of items needs to be identified by experts. The semi-supervised method falls between unsupervised and supervised methods.

MODEL AND METHOD

Model

Let K be the number of attributes. Let X_{ij} be a binary random variable to denote the response of examinee i to item j, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, J$. Let $\boldsymbol{\alpha}_i$ be a column vector to denote an attribute mastery pattern or a knowledge state from the universal set of knowledge states. Moreover, Q-matrix that specifies the item-attribute relationship is a $K \times J$ matrix, in which entry $q_{kj} = 1$ if attribute k is required for answering item j correctly; otherwise, $q_{kj} = 0$.

The item response function for the deterministic inputs, noisy "and" gate (DINA) model (Haertel, 1989; Junker and Sijtsma, 2001; Chiu and Douglas, 2013) is as follows:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1 - s_j)^{\eta_{ij}}, \tag{1}$$

where a deterministic latent response $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ki}^{q_{kj}}$ indicates whether or not examinee i possesses all of the attributes required by item j. A value of $\eta_{ij} = 1$ means that examinee i has mastered all of the attributes required by item j, and $\eta_{ij} = 0$ otherwise. The slip parameter s_j refers to the probability of an incorrect response to the item j when $\eta_{ij} = 1$, and the guessing parameter g_j refers to the probability of a correct response to item j when $\eta_{ij} = 0$. Let $\mathbf{B} = (\eta_{ij})$ be a deterministic latent response matrix for the DINA model.

The item response function for the deterministic inputs, noisy "or" gate (DINO) model (Templin and Henson, 2006; Chiu and Douglas, 2013) is as follows:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\boldsymbol{\alpha}_i) = (1 - s_j)^{w_{ij}} g_j^{1-w_{ij}}, \tag{2}$$

where $w_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ki})^{q_{kj}}$ is a deterministic latent response. As in the DINA model, s_j and g_j are the slip and guessing parameters of item j. The DINA and DINO model are conjunctive and disjunctive models (Maris, 1999), respectively. Let $\mathbf{W} = (w_{ij})$ be a deterministic latent response matrix for the DINO model.

A Semi-supervised Learning Approach for the Conjunctive Model

In the rule space method (Tatsuoka, 2009) or the attribute hierarchy method (Leighton et al., 2004), the adjacency matrix denoted by $\mathbf{A}$ represents the direct relationship among attributes. We denote the entry in row k_1 and column k_2 of $\mathbf{A}$ by $a_{k_1k_2}$. If a direct prerequisite relation exists from attribute k_1 to attribute k_2, then $a_{k_1k_2} = 1$, and $a_{k_1k_2} = 0$ otherwise. Let $\mathbf{R}$ denote

a reachability matrix of order (K, K) to specify the direct and indirect relationships among attributes. The $\mathbf{R}$ matrix is given by $\mathbf{R} = (\mathbf{A} + \mathbf{I})^K$ with respect to Boolean operations, where $\mathbf{I}$ is an identity matrix. The reduced Q matrix denoted by $\mathbf{Q_r}$ is obtained by removing the items (columns) that do not satisfy the specified relationships from the incidence Q matrix. The columns of $\mathbf{Q_r}$ and the zero vector forms the student matrix denoted by $\mathbf{Q_s}$ in which the columns forms the universal set of attribute patterns. If K attributes are independent, $\mathbf{A}$ is a zero matrix, $\mathbf{R}$ with K columns is an identity matrix, $\mathbf{Q_r}$ with $2^K - 1$ columns does not include the zero vector, and $\mathbf{Q_s}$ with 2^K columns contains all possible combinations of attribute patterns.

We assume that the cognitive requirement for the multiple skills within an item is conjunctive (Maris, 1999), that is, answering an item correctly requires mastery of all the skills required by that item. For the conjunctive model, Example 1 will show the relationship of latent responses on items with q-vectors corresponding to $\mathbf{R}$ and $\mathbf{Q_r}$.

Example 1 for an independent structure. Let $K = 2$, $\mathbf{R} = \begin{bmatrix}\mathbf{r_1} & \mathbf{r_2}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, $\mathbf{Q_r} = \begin{bmatrix}\mathbf{q_1} & \mathbf{q_2} & \mathbf{q_3}\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 1\end{bmatrix}$, and $\mathbf{Q_s} = \begin{bmatrix}\boldsymbol{\alpha_1} & \boldsymbol{\alpha_2} & \boldsymbol{\alpha_3} & \boldsymbol{\alpha_4}\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 1\\ 0 & 0 & 1 & 1\end{bmatrix}$.
Given $\mathbf{Q_s}$ and a test Q-matrix of $\mathbf{Q_r}$, a latent response matrix $\mathbf{B} = \begin{bmatrix}\boldsymbol{\eta_1} & \boldsymbol{\eta_2} & \boldsymbol{\eta_3}\end{bmatrix} = \begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 1 & 1 & 1\end{bmatrix}$ can be calculated, in which the entry in row i and column j is the deterministic latent response of η_{ij}. If 0 corresponds to F (false) and 1 corresponds to T (true), the logical conjunction and disjunction operators, $\vee$ and $\wedge$, can be applied to two binary vectors of equal length, by taking the bitwise AND or OR of each pair of bits at corresponding positions. It can be observed that $\boldsymbol{\eta_3} = \boldsymbol{\eta_1} \wedge \boldsymbol{\eta_2}$, where $\boldsymbol{\eta_3} = \boldsymbol{\eta_1} \wedge \boldsymbol{\eta_2}$ is the conjunction of $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$. This is because the relationship $\mathbf{q_3} = \mathbf{q}_1 \vee \mathbf{q}_2$ is true, where $\mathbf{q}_1 \vee \mathbf{q}_2$ is the disjunction of $\mathbf{q}_1$ and $\mathbf{q}_2$.

Example 1 illustrates the following fact. For the conjunctive model, consider two latent response matrices denoted by $\mathbf{B}_1$ and $\mathbf{B}_2$ from two tests corresponding two Q-matrices $\mathbf{Q_r}$ and $\mathbf{R}$, where denoted as a reachability matrix. It means that $\mathbf{B}_1$ and $\mathbf{B}_2$ can be generated, respectively from the reduced Q-matrix and the reachability matrix based on the universal set of attribute patterns. From the example above, then any columns of the $\mathbf{B}_1$ can be expressed by the columns of the $\mathbf{B}_2$ under the logical AND operation. This is because the augmented algorithm proposed by Ding et al. (2008, 2009) in the generalized Q-matrix theory (Ding et al., 2015) provided the useful fact that any columns of the reduced Q-matrix can be expressed by the columns of the reachability matrix under the logical OR operation. The argument in Example 1 can be adapted to prove the following theorem.

Theorem 1. For the conjunctive model, if K attributes are independent, then $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$ if and only if $\eta_{ij} = \wedge_{l \in S_j} \eta_{il}$, where $\boldsymbol{\alpha}_i$ is any column of $\mathbf{Q_s}$ and S_j is a subset of $\{1, 2, \ldots, K\}$.

Proof: If $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, we need to consider two cases, when $\eta_{ij} = 1$ and $\eta_{ij} = 0$. If $\eta_{ij} = 1$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $\alpha_{ki} = 1$ for all attributes k with $q_{kj} = 1$ by the definition of the deterministic latent response. That is, examinee i has mastered all the skills required by item j. Since $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, then by the definition of conjunction, we can conclude that $\alpha_{ki} = 1$ for all attributes k with $r_{kl} = 1$ for all $l \in S_j$. We now use the definition of the deterministic latent response to conclude that $\eta_{il} = 1$ for all $l \in S_j$, that is, $\wedge_{l \in S_j} \eta_{il} = 1$. This shows that $\eta_{ij} = \wedge_{l \in S_j} \eta_{il}$ when $\eta_{ij} = 1$. If $\eta_{ij} = 0$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $\alpha_{ki} = 0$ for at least one of attributes with $q_{kj} = 1$ by the definition of the deterministic latent response. That is, examinee i has not mastered all the skills required by item j. Since $q_{kj} = 1$ and $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, there is an item l in S_j such that $r_{kl} = 1$. This means that item l measured attribute k. Since $\alpha_{ki} = 0$, then by the definition of the deterministic latent response, it follows that $\eta_{il} = 0$ for at least one of items in S_j, that is, $\wedge_{l \in S_j} \eta_{il} = 0$. This show that $\eta_{ij} = \wedge_{l \in S_j} \eta_{il}$ when $\eta_{ij} = 0$. Next, we try to prove the converse. First suppose that there exists an attribute $k \in \{1, 2, \ldots, K\}$ such that $\vee_{l \in S_j} r_{kl} = 1$ and $q_{kj} = 0$. Since $\vee_{l \in S_j} r_{kl} = 1$, we know that there exists an item $l \in S_j$ with $r_{kl} = 1$. Due to the arbitrariness of $\boldsymbol{\alpha}_i$, let $\boldsymbol{\alpha}_i = \mathbf{1} - \mathbf{e}_k$, where $\mathbf{1} = (1\ 1\ \ldots\ 1)^T$ and $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's elsewhere. This is a contradiction, because we know that $\eta_{ij} = 1$, while $\wedge_{l \in S_j} \eta_{il} = 0$. Similarly, we assume that there exists an attribute $k \in \{1, 2, \ldots, K\}$ such that $\vee_{l \in S_j} r_{kl} = 0$ and $q_{kj} = 1$. One can still take $\boldsymbol{\alpha}_i = \mathbf{1} - \mathbf{e}_k$. This is also a contradiction, because we know that $\eta_{ij} = 0$, while $\wedge_{l \in S_j} \eta_{il} = 1$. The proof is complete.

The important fact about Theorem 1 is that if a latent response matrix is calculated from a Q-matrix, the relationship between the columns in the Q-matrix can be constructed from the relationship between the corresponding columns in the latent response matrix. It should be noted that an observed item response is a function of an underlying latent response and slip and guessing parameters. In other words, the noise introduced in the process is due to slip and guessing parameters.

Next, we will introduce a semi-supervised learning method for Q-matrix specification for the conjunctive model by using the result of Theorem 1 and considering the noise in item responses. Without loss of generality, we begin by arbitrarily assigning q-vector $\mathbf{q_j}$ to item j. Given a test Q-matrix, written as $\mathbf{Q}_t = [\mathbf{R}_{K \times K}\ \mathbf{q}_j] = [\mathbf{r_1}\ \mathbf{r_2}\ \ldots\ \mathbf{r_K}\ \mathbf{q_j}]$, where $\mathbf{R}$ is a reachability matrix specified by subject matter experts and the remaining $\mathbf{q_j}$ is unknown. Let $\mathbf{U} = [\mathbf{X}_{N \times K}\ \mathbf{Y}_{N \times 1}]$ be an item response matrix on $\mathbf{Q}_t$, where N is the sample size. The estimate of $\mathbf{q_j}$ can be written as

$$\hat{\mathbf{q}}_j = \vee_{r_k \in \hat{S}_j} \mathbf{r}_k, \tag{3}$$

where logical OR is applied to the corresponding entries of the columns in the following set of $\hat{S}_j$

$$\hat{S}_j = \underset{S \in \mathrm{P}(\{r_1, r_2, \ldots, r_K\}) - \varnothing}{\arg\min} (Y_j - \wedge_{r_k \in S} X_k)^T (Y_j - \wedge_{r_k \in S} X_k), \tag{4}$$

where $\mathrm{P}(\{r_1, r_2, \ldots, r_K\})$ is the power set of the set $\{\mathbf{r_1}, \mathbf{r}_2, \ldots, \mathbf{r}_K\}$. The exhaustive method with time complexity $\mathrm{O}(2^K)$ provided a simple way to find a global solution of $\hat{S}_j$.

A Semi-supervised Learning Approach for the Disjunctive Model

For the disjunctive model, the deterministic latent response on an item is correct if and only if an examinee has mastered at least one of the skills required by the item. This is illustrated in Example 2. Similar to what we did in Example 1, Example 2 will show the relationship of latent responses on items with q-vectors corresponding to $\mathbf{R}$ and $\mathbf{Q_r}$.

Example 2 for an independent structure. Let $K = 2$, $\mathbf{R} = \begin{bmatrix}\mathbf{r_1} & \mathbf{r_2}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, $\mathbf{Q_s} = \begin{bmatrix}\boldsymbol{\alpha_1} & \boldsymbol{\alpha_2} & \boldsymbol{\alpha_3} & \boldsymbol{\alpha_4}\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 1\\ 0 & 0 & 1 & 1\end{bmatrix}$, and $\mathbf{Q_r} = \begin{bmatrix}\mathbf{q_1} & \mathbf{q_2} & \mathbf{q_3}\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ 0 & 1 & 1\end{bmatrix}$. From $\mathbf{Q_s}$ and $\mathbf{Q_r}$, a latent response matrix $\mathbf{W}_1 = \begin{bmatrix}\mathbf{w_1} & \mathbf{w_2} & \mathbf{w_3}\end{bmatrix} = \begin{bmatrix}0 & 0 & 0\\ 1 & 0 & 1\\ 0 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$ can be calculated, in which the entry in row i and column j is the deterministic latent response of w_{ij}. It can be observed that $\mathbf{w_3} = \mathbf{w_1} \vee \mathbf{w_2}$. This is because the relationship $\mathbf{q_3} = \mathbf{q}_1 \vee \mathbf{q}_2$ is true.

Consider a latent response matrix, denoted by $\mathbf{W}_2 = \begin{bmatrix}\mathbf{w_1} & \mathbf{w}_2\end{bmatrix}$, corresponding to the $\mathbf{R}$ matrix. The fact illustrated in Example 2 is that any columns of the $\mathbf{W}_1$ can be expressed by the columns of the $\mathbf{W}_2$ under the logical OR operation for the disjunctive model. This is also because the augmented algorithm proposed by Ding et al. (2008, 2009) in the generalized Q-matrix theory (Ding et al., 2015) provided the useful fact that any columns of the reduced Q-matrix can be expressed by the columns of the reachability matrix under the logical OR operation. The following theorem gives the precise statement.

Theorem 2. For the disjunctive model, if K attributes are independent, then $\mathbf{q_j} = \vee_{l\in S_j}\mathbf{r_l}$ if and only if $w_{ij} = \vee_{l\in S_j} w_{il}$, where $\boldsymbol{\alpha}_i$ is any column of $\mathbf{Q_s}$ and S_j is a subset of $\{1, 2, \ldots, K\}$.

Proof: If $\mathbf{q_j} = \vee_{l\in S_j}\mathbf{r_l}$, we need to consider two cases, when $w_{ij} = 1$ and $w_{ij} = 0$. If $w_{ij} = 1$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $\alpha_{ki} = 1$ for at least one of attributes k with $q_{kj} = 1$ by the definition of the deterministic latent response. That is, examinee i has mastered at least one of the attributes required by item j. Without loss of generality, we assume $\alpha_{ki} = 1$ and $q_{kj} = 1$. Since $\mathbf{q_j} = \vee_{l\in S_j}\mathbf{r_l}$, then by the definition of disjunction, we can conclude that $r_{kl} = 1$ is true for at least one of $l \in S_j$. From the definition of the deterministic latent response, it follows that there is at least one item $l \in S_j$ such that $w_{il} = 1$, that is, $\vee_{l\in S_j} w_{il} = 1$. This show that $w_{ij} = \vee_{l\in S_j} w_{il}$ when $w_{ij} = 1$. If $w_{ij} = 0$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $w_{ki} = 0$ for all of attributes with $q_{kj} = 1$ by the definition of the deterministic latent response. That is, examinee i has not mastered any skills required by item j. Since $\mathbf{q_j} = \vee_{l\in S_j}\mathbf{r_l}$, examinee i has not mastered any skills required by any item $l \in S_j$. If we suppose that examinee i has mastered at least one of attributes required by an item $l \in S_j$, then $w_{ij} = 1$, which is a contradiction. It means that item l measured attribute k. It follows that $w_{il} = 0$ for all of items in S_j, that is, $\vee_{l\in S_j} w_{il} = 0$, directly from the definition of the deterministic latent response. This show that $w_{ij} = \vee_{l\in S_j} w_{il}$ when $w_{ij} = 0$. Next, we use a proof by contradiction to prove the converse. First assume that there exists an attribute $k \in \{1, 2, \ldots, K\}$ such that $\vee_{l\in S_j} r_{kl} = 1$ and $q_{kj} = 0$. Since $\vee_{l\in S_j} r_{kl} = 1$, we know that there exists an item $l \in S_j$ with $r_{kl} = 1$. Due to the arbitrariness of $\boldsymbol{\alpha}_i$, let $\boldsymbol{\alpha}_i = \mathbf{e}_k$, where $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's elsewhere. Then, we have $w_{il} = 1$ and $w_{ij} = 0$. Since $w_{ij} = \vee_{l\in S_j} w_{il}$, we know that $w_{ij} = 1$ and arrive at a contradiction. Similarly, we assume that there exists an attribute $k \in \{1, 2, \ldots, K\}$ such that $\vee_{l\in S_j} r_{kl} = 0$ and $q_{kj} = 1$. One can still take $\boldsymbol{\alpha}_i = \mathbf{e}_k$. This is also a contradiction, because we know that $w_{ij} = 1$, while $\wedge_{l\in S_j} w_{il} = 0$. The proof is complete.

The important fact about Theorem 2 is that one can derive the relationship between the columns of a Q-matrix from the relationship between the columns of corresponding latent response matrix. For considering the noise introduced in item responses due to slipping and guessing, we will introduce a semi-supervised learning method for Q-matrix specification for the disjunctive model by using the result of Theorem 2. Without loss of generality, we begin by arbitrarily assigning a q-vector to $\mathbf{q_j}$. Given a test Q-matrix, written as $\mathbf{Q}_t = [\mathbf{R}_{K\times K}\ \mathbf{q_j}] = [\mathbf{r_1}\ \mathbf{r_2}\ \ldots\ \mathbf{r_K}\ \mathbf{q_j}]$, where $\mathbf{R}$ is a reachability matrix specified by subject matter experts and the remaining $\mathbf{q_j}$ is unknown. Let $\mathbf{U} = [\mathbf{X}_{N\times K}\ \mathbf{Y}_{N\times 1}]$ be an item response matrix on $\mathbf{Q}_t$. The estimate of $\mathbf{q_j}$ can be written as

$$\hat{\mathbf{q}}_j = \vee_{r_k\in\hat{S}_j}\mathbf{r}_k, \tag{5}$$

where logical OR is applied to the corresponding entries of the columns in the following set of $\hat{S}_j$

$$\hat{S}_j = \underset{S\in \mathrm{P}(\{r_1, r_2, \ldots, r_K\})-\varnothing}{\arg\min} (Y_j - \vee_{r_k\in S} X_k)^T (Y_j - \vee_{r_k\in S} X_k), \tag{6}$$

where $\mathrm{P}(\{r_1, r_2, \ldots, r_K\})$ is the power set of the set $\{\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_K}\}$. The exhaustive method with time complexity $\mathrm{O}(2^K)$ provided a simple way to find a global solution of $\hat{S}_j$.

A SIMULATION STUDY

Study Design

A simulation study was conducted to investigate the performance of the new method under five factors, such as sample size, item parameters for items corresponding to a reachability matrix, item parameters for new or raw items with unknown q-vectors, two cognitive diagnostic models (the DINA and DINO model), and two designs. Five attributes were considered in the simulation study. Matlab 2015a and R-3.6.1 were used for estimating unknown Q-matrix and analyzing real data below.

In the simulation study, a test Q-matrix $\mathbf{Q_t} = [\mathbf{R}\ \mathbf{Q_r}]$ consists of an identity or a reachability matrix and a reduced Q-matrix, where the reduced Q-matrix with 31 items includes all non-zero possible q-vectors. The number of examinees has 10 levels, such as N =30, 60, …, and 300. Item parameters for $\mathbf{R}$ and $\mathbf{Q_r}$ have 10 levels, such as 0, 0.05, …, and 0.45. In general, for the DINA or DINO model, a high quality or "good" item will have small slip and guessing parameters (Rupp et al., 2010), which means that the noise are small.

Random and optimal designs were considered in the simulation study. For the random design, attribute patterns for

examinees were generated by taking each of the 2^5 possible patterns with equal probability for each sample size. From the proof of Theorem 1 above, we know that the following set of attribute patterns for examinees plays a very important role in discriminating latent response vectors of different q-vectors under the DINA model

$$S_{DINA} = \{\mathbf{1} - \mathbf{e}_1\mathbf{1} - \mathbf{e}_2, \ldots, \mathbf{1} - \mathbf{e}_K\}\{\begin{bmatrix}0\\1\\\vdots\\1\end{bmatrix}, \begin{bmatrix}1\\0\\\vdots\\1\end{bmatrix}, \ldots, \begin{bmatrix}1\\1\\\vdots\\0\end{bmatrix}\} \quad (7)$$

where $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's otherwise. From the proof of Theorem 2 above, another set of attribute patterns for examinees plays a very important role in discriminating latent response vectors of different q-vectors under the DINO model as follows

$$S_{DINO} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}\{\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}, \begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix}, \ldots, \begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix}\}, \quad (8)$$

where $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's otherwise. For the optimal design, attribute patterns for examinees under the DINA or DINA model were randomly drawn with replacement from the set of S_{DINA} or S_{DINO}, respectively. Optimal designs for two models are possible to meet the needs of learners at different stages of skills and knowledge acquisition. For example, the attribute patterns in S_{DINO} containing only one skill. This condition is really improbable for summary assessments in real situations, but is expected to be common for novice learners with respect to the new content to be learned in formative assessments or classroom assessments.

Data Simulation

Simulated data were generated using five attributes. Based on the simulated Q-matrix, item parameters, and attribute patterns, item responses are generated in the following way

$$X_{ij} = \begin{cases} 1, & \text{if } u \leq P_j(\boldsymbol{\alpha}_i), \\ 0, & \text{otherwise}, \end{cases} \quad (9)$$

where u is a random value from a Uniform (0, 1) distribution and $P_j(\boldsymbol{\alpha}_\mathbf{i})$ is the item response function of the DINA or DINO model. A total of 4,000 conditions were simulated (10 sample sizes × 10 item parameters × 10 item parameters × 2 models × 2 designs). Thirty replication data sets were simulated for each condition.

Evaluation Criterion

The performance of the new method is evaluated in terms of the correct recovery rate (CRR) of q-entries. The correct recovery rate equals the ratio of the number of correct q-entries in the estimated Q-matrix to the total number of q-entries (Chiu, 2013)

$$\text{CRR} = \frac{1}{KM}\sum_{k=1}^{K}\sum_{j=1}^{M}\mathbf{I}(\hat{q}_{kj} = q_{kj}), \quad (10)$$

TABLE 1 | Mean and standard deviation (in brackets) of correct recovery rate of q-entries for two models and two designs.

Sample size	The DINA model		The DINO model	
	Random design	Optimal design	Random design	Optimal design
30	0.651 (0.126)	0.720 (0.147)	0.653 (0.126)	0.721 (0.146)
60	0.699 (0.145)	0.769 (0.153)	0.700 (0.144)	0.770 (0.151)
90	0.725 (0.149)	0.796 (0.152)	0.726 (0.149)	0.796 (0.151)
120	0.742 (0.151)	0.815 (0.149)	0.743 (0.151)	0.815 (0.148)
150	0.756 (0.153)	0.827 (0.146)	0.754 (0.153)	0.829 (0.145)
180	0.764 (0.153)	0.839 (0.143)	0.765 (0.153)	0.838 (0.144)
210	0.772 (0.153)	0.847 (0.141)	0.772 (0.151)	0.846 (0.141)
240	0.779 (0.152)	0.854 (0.138)	0.777 (0.152)	0.854 (0.138)
270	0.784 (0.151)	0.858 (0.137)	0.783 (0.152)	0.858 (0.137)
300	0.789 (0.151)	0.863 (0.135)	0.789 (0.152)	0.864 (0.135)
Mean	0.746 (0.154)	0.819 (0.151)	0.746 (0.154)	0.819 (0.150)

where $M = 31$ is the number of columns of the unknown Q-matrix $\mathbf{Q_r}$, q_{kj} is an (k, j)th entry of the simulated $\mathbf{Q_r}$, and $\hat{q}_{kj}$ is an (k, j) entry of the $\hat{\mathbf{Q}}_\mathbf{r}$ estimated from the new method. The mean and standard deviation of the CRR values of the 30 replications were reported for each condition.

Results

Table 1 lists descriptive statistics of correct recovery rate of q-entries for two models and two designs across other conditions. It is clear that the mean of correct recovery rates of q-entries tends to increase as sample size increases, but sample size has slightly affected the standard deviations of correct recovery rates. It should be noted that the mean of correct recovery rates of the optimal design is larger than that of the random design. The semi-supervised learning method for q-matrix specification performed similarly under two cognitive diagnostic models. In addition, since there are 32 possible attribute patterns, no all attribute patterns can be observed in the first sample size condition ($N =$ 30). This might lead to lower rate of correct recovery observed for this condition.

Table 2 shows the correct recovery rates of q-entries from the new method with sample size of 300 for the DINA model under the random design. From correct recovery rates of q-entries, when item parameters for items with known (i.e., the reachability matrix) and unknown q-vectors are $\leq$0.2, most of the average of correct recovery rates of q-entries for the semi-supervised method are larger than or equal to 0.9. From trends of marginal means of last rows and columns in **Table 2**, item parameters of the reachability matrix have a relatively larger impact on the performance of the semi-supervised method than item parameters with unknown q-vectors.

Table 3 presents the correct recovery rates of q-entries from the new method with sample size of 300 for the DINA model under the optimal design. From correct recovery rates of q-entries, when item parameters for items with known and unknown q-vectors are $\leq$0.25, the average of correct recovery

TABLE 2 | The correct recovery rates of q-entries with sample size of 300 for the DINA model and random design.

Item parameters for the R	Item parameters for the new items										
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**M**
0.00	**1.000**	**1.000**	**1.000**	**0.999**	**0.998**	**0.993**	**0.974**	**0.929**	0.837	0.665	0.939
0.05	**1.000**	**1.000**	**0.997**	**0.995**	**0.991**	**0.976**	**0.946**	**0.908**	0.791	0.642	0.925
0.10	**0.994**	**0.992**	**0.984**	**0.984**	**0.967**	**0.954**	**0.927**	0.857	0.767	0.632	0.906
0.15	**0.974**	**0.968**	**0.958**	**0.954**	**0.932**	**0.915**	0.866	0.807	0.724	0.608	0.870
0.20	**0.927**	**0.922**	**0.910**	**0.901**	0.881	0.860	0.826	0.776	0.692	0.591	0.829
0.25	0.866	0.846	0.850	0.839	0.825	0.802	0.776	0.733	0.652	0.567	0.775
0.30	0.791	0.801	0.782	0.793	0.760	0.735	0.727	0.670	0.624	0.563	0.725
0.35	0.728	0.718	0.720	0.709	0.709	0.698	0.683	0.637	0.604	0.546	0.675
0.40	0.673	0.686	0.681	0.680	0.668	0.643	0.620	0.608	0.589	0.527	0.638
0.45	0.647	0.634	0.623	0.620	0.615	0.612	0.604	0.575	0.575	0.537	0.604
M	0.860	0.857	0.851	0.847	0.835	0.819	0.795	0.750	0.686	0.588	0.789

The bold values are larger than 0.9.

TABLE 3 | The correct recovery rates of q-entries with sample size of 300 for the DINA model and optimal design.

Item parameters for the R	Item parameters for the new items										
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**M**
0.00	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.991**	**0.939**	0.780	0.971
0.05	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.997**	**0.981**	**0.917**	0.754	0.965
0.10	**1.000**	**1.000**	**1.000**	**0.999**	**0.998**	**0.995**	**0.981**	**0.955**	0.871	0.714	0.951
0.15	**1.000**	**1.000**	**0.998**	**0.992**	**0.988**	**0.981**	**0.959**	**0.915**	0.841	0.688	0.936
0.20	**0.986**	**0.986**	**0.982**	**0.973**	**0.962**	**0.952**	**0.918**	0.875	0.785	0.661	0.908
0.25	**0.959**	**0.952**	**0.947**	**0.933**	**0.926**	**0.909**	0.879	0.829	0.754	0.624	0.871
0.30	**0.930**	**0.914**	**0.912**	**0.909**	0.883	0.865	0.832	0.786	0.712	0.610	0.835
0.35	0.886	0.888	0.880	0.865	0.847	0.813	0.780	0.724	0.662	0.573	0.792
0.40	0.847	0.834	0.816	0.804	0.774	0.753	0.721	0.679	0.629	0.562	0.742
0.45	0.749	0.738	0.717	0.720	0.688	0.687	0.656	0.619	0.581	0.556	0.671
M	0.936	0.931	0.925	0.920	0.907	0.895	0.872	0.835	0.769	0.652	0.864

The bold values are larger than 0.9.

rates of q-entries for the semi-supervised method are larger than or equal to 0.9. However, item parameters for known q-vectors have slightly larger impact on the performance of the semi-supervised method than for unknown q-vectors, because the row means decreased more quickly than the column means. We need to compare the **Tables 2**, **3** to see which designs are promising. The number of correct recovery rates above 0.9 in **Table 3** were found to be larger than that of **Table 2**. **Tables 4**, **5** show the correct recovery rates of q-entries from the new method with sample size of 300 for the DINO model under the random and optimal design. It can be observed that results for the DINO model are the same as those for the DINA model described above.

REAL DATA ANALYSIS

The purpose of the real data analysis is to examinee whether the proposed method is promising for a non-independent structure under the conjunctive model based on an intuitive fact from the following example.

Example 3 for an unstructured hierarchy under the conjunctive model. Let $K = 3$, $\mathbf{R} = [\mathbf{r_1}\ \mathbf{r}_2\ \mathbf{r}_3] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\mathbf{Q_r} = [\mathbf{q_1}\ \mathbf{q}_2\ \mathbf{q}_3\ \mathbf{q}_4] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$, and $\mathbf{Q_s} = \left[\boldsymbol{\alpha}_0\ \boldsymbol{\alpha_1}\ \boldsymbol{\alpha_2}\ \boldsymbol{\alpha_3}\ \boldsymbol{\alpha_4}\right] = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$. From the ideal response matrix $\mathbf{B} = \left[\boldsymbol{\eta_1}\ \boldsymbol{\eta_2}\ \boldsymbol{\eta_3}\ \boldsymbol{\eta}_4\right] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$, it can be observed that $\boldsymbol{\eta}_4 = \boldsymbol{\eta}_2 \wedge \boldsymbol{\eta}_3$ or $\boldsymbol{\eta}_4 = \boldsymbol{\eta}_1 \wedge \boldsymbol{\eta}_2 \wedge \boldsymbol{\eta}_3$. This is because the relationship $\mathbf{q}_4 = \mathbf{q}_2 \vee \mathbf{q}_3$ or $\mathbf{q}_4 = \mathbf{q}_1 \vee \mathbf{q}_2 \vee \mathbf{q}_3$ is true.

A common data set pertaining to fraction-subtraction data contains 20 items and 536 examines (de la Torre and Douglas, 2004). In our real data analysis, we focused on the analysis of a subset of test items where the expert Q-matrix comes

TABLE 4 | The correct recovery rates of q-entries with sample size of 300 for the DINO model and random design.

Item parameters for the R	Item parameters for the new items										
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**M**
0.00	**1.000**	**1.000**	**1.000**	**0.999**	**0.998**	**0.990**	**0.976**	**0.929**	0.840	0.685	0.942
0.05	**1.000**	**0.999**	**0.998**	**0.996**	**0.991**	**0.980**	**0.946**	0.889	0.796	0.650	0.925
0.10	**0.996**	**0.991**	**0.992**	**0.985**	**0.968**	**0.945**	**0.918**	0.848	0.767	0.639	0.905
0.15	**0.971**	**0.969**	**0.963**	**0.947**	**0.931**	**0.912**	0.864	0.806	0.724	0.604	0.869
0.20	**0.927**	**0.921**	**0.915**	**0.905**	0.885	0.854	0.819	0.763	0.687	0.578	0.825
0.25	0.855	0.854	0.865	0.842	0.831	0.797	0.773	0.733	0.665	0.572	0.779
0.30	0.787	0.795	0.789	0.774	0.763	0.743	0.722	0.677	0.629	0.571	0.725
0.35	0.734	0.721	0.717	0.720	0.720	0.686	0.662	0.634	0.597	0.550	0.674
0.40	0.677	0.678	0.689	0.675	0.660	0.654	0.623	0.624	0.573	0.536	0.639
0.45	0.632	0.630	0.628	0.633	0.611	0.608	0.600	0.586	0.571	0.525	0.603
M	0.858	0.856	0.856	0.848	0.836	0.817	0.790	0.749	0.685	0.591	0.789

The bold values are larger than 0.9.

TABLE 5 | The correct recovery rates of q-entries with sample size of 300 for the DINO model and optimal design.

Item parameters for the R	Item parameters for the new items										
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**M**
0.00	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.992**	**0.939**	0.780	0.971
0.05	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.995**	**0.973**	**0.916**	0.744	0.963
0.10	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**0.994**	**0.983**	**0.954**	0.873	0.715	0.952
0.15	**0.998**	**0.999**	**0.996**	**0.995**	**0.985**	**0.978**	**0.957**	**0.922**	0.836	0.695	0.936
0.20	**0.991**	**0.986**	**0.982**	**0.970**	**0.961**	**0.946**	**0.916**	0.873	0.795	0.666	0.909
0.25	**0.957**	**0.951**	**0.951**	**0.935**	**0.925**	**0.909**	0.885	0.829	0.752	0.628	0.872
0.30	**0.922**	**0.910**	**0.913**	**0.902**	0.894	0.866	0.840	0.772	0.705	0.595	0.832
0.35	0.887	0.882	0.870	0.862	0.842	0.807	0.787	0.731	0.678	0.571	0.792
0.40	0.838	0.830	0.812	0.802	0.788	0.749	0.715	0.675	0.620	0.560	0.739
0.45	0.737	0.733	0.733	0.703	0.690	0.663	0.645	0.619	0.595	0.544	0.666
M	0.933	0.929	0.926	0.917	0.908	0.891	0.872	0.834	0.771	0.650	0.863

The bold values are larger than 0.9.

from **Table 7** both in de la Torre (2008) or DeCarlo (2012). The labels given to the five skills are (A1) performing basic fraction-subtraction operation, (A2) simplifying/reducing, (A3) separating whole numbers from fractions, (A4) borrowing one from whole number to fraction, and (A5) converting whole numbers to fractions.

We assumed the corresponding Q-matrix of items 3, 8, 9, 12, and 10 known since these item parameters are relatively small and the q-vectors of other items are combinations of q-vectors for these five items. Then, the semi-supervised method was applied to estimate q-vectors for the other 10 items. Results in **Table 6** show that the agreement rate of q-entries between the estimate and expert Q-matrix on the 10 items is 84%. The estimated q-entries suggest that items 4, 7, 13, 14, and 15 do not require attribute 2 (simplifying/reducing). Item 4 (similar to item 14) do not required attribute A2, which is consistent with results from DeCarlo (2012). Items 7, 13, and 15 can be answered correctly by using attributes required by item 12. The estimated q-vector of item 1 has largest discrepancy with the expert q-vector. The reason might be that solving item 1 correctly needs to find a common denominator and then performs basic fraction-subtraction operation. The guessing and slip parameter of item 1 are 0.0001 and 0.2769 under the expert q-vector, respectively. The guessing and slip parameter of item 1 are 0.3408 and 0.0716 under the estimated q-vector, respectively. Since item 1 requires an extra attribute (i.e., find a common denominator), the slip parameter for the expert q-vector is relatively large, while the estimated q-vector contains some unnecessary attributes, the guessing parameter is relatively large. In the estimated Q-matrix, attribute A4 has been added to item 11.The guessing probability of item 11 increased sensibly (from 0.10 to 0.48). It indicated that attribute A4 is not necessary for item 11 because this item is different from items 7, 12, and so on.

The generalized DINA model (GDINA; de la Torre, 2011), the DINA model, the linear logistic model (LLM; Fischer, 1995), and

TABLE 6 | The expert and estimated Q-matrix and item parameters estimates of the DINA model for the fractional subtraction data.

No.	Items	The expert Q-matrix and item parameters estimates of the DINA model							The estimated Q-matrix and item parameters estimates of the DINA model						
		A1	A2	A3	A4	A5	$\hat{g}$	$\hat{s}$	A1	A2	A3	A4	A5	$\hat{g}$	$\hat{s}$
1	$\frac{3}{4}-\frac{3}{8}$	1	0	0	0	0	0.00	0.28	1	0	**1**	**1**	**1**	0.34	0.07
2	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1	0	0.21	0.12	1	1	1	1	0	0.21	0.11
3	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0	0	0.14	0.04	1	0	0	0	0	0.10	0.05
4	$3-2\frac{1}{5}$	1	1	1	1	1	0.12	0.13	1	**0**	1	1	1	0.12	0.18
5	$3\frac{7}{8}-2$	1	0	1	0	0	0.34	0.25	1	0	1	0	0	0.35	0.25
6	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1	0	0.03	0.23	1	1	1	1	0	0.03	0.23
7	$4\frac{1}{3}-2\frac{4}{3}$	1	1	1	1	0	0.07	0.08	1	**0**	1	1	0	0.07	0.08
8	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0	0	0.16	0.05	1	1	0	0	0	0.09	0.04
9	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0	0	0.08	0.06	1	0	1	0	0	0.04	0.04
10	$2-\frac{1}{3}$	1	0	1	1	1	0.17	0.07	1	0	1	1	1	0.15	0.09
11	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0	0	0.10	0.10	1	0	1	**1**	0	0.48	0.07
12	$7\frac{3}{5}-\frac{4}{5}$	1	0	1	1	0	0.03	0.13	1	0	1	1	0	0.05	0.14
13	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1	0	0.13	0.16	1	**0**	1	1	0	0.13	0.16
14	$4-1\frac{4}{3}$	1	1	1	1	1	0.02	0.20	1	**0**	1	1	1	0.01	0.24
15	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1	0	0.01	0.18	1	**0**	1	1	0	0.01	0.19

The bold values are the changes.

the reduced reparametrized unified model (R-RUM; Hartz, 2002) were applied to fit the fraction-subtraction data with the expert or estimated Q-matrix. Under the DINA model, the means of the estimates of the guessing and slip parameter for the expert Q-matrix are 0.1080 and 0.1381, respectively, while for the revised Q-matrix, they are 0.1440 and 0.1295, respectively. It means that the estimates of the slip parameter become lower, but the guessing parameters tend to be larger. **Table 7** presents fit results for the fraction subtraction data using the expert and estimated q-matrix. The LLM with the estimated Q-matrix is the best-fitting CDM and the R-RUM with the estimated Q-matrix is slightly worse, whereas the estimated Q-matrix performed worse than the expert Q-matrix only in the DINA model.

CONCLUSION AND DISCUSSION

The supervised methods rely on a provisional Q-matrix for a whole test, the estimates of examinees' attribute patterns and their accuracy. It is not suitable for the case of a provisional Q-matrix with a large amount of misspecification. The purpose of this study is to propose the semi-supervised method under independent structure based on item responses and a reachability R matrix corresponding to a small part of test item specified by subject matter experts. The new method doesn't need to estimate examinees' attribute patterns. The main conclusion of this study is that the new method will play a very important role in assist subject matter experts for Q-matrix specification because it is hard to correctly specify a Q-matrix with a large number of test items by subject matter experts. It may be useful for cognitive diagnostic assessment to facilitate teaching and learning.

The generalized Q-matrix theory has been shown that each column in the reduced Q-matrix can be expressed as a logical disjunction of some of columns of the reachability matrix. With the aid of this theory, this study takes a look inside a latent response matrix and reveals an interesting and useful relationship hidden in its columns. If a latent response matrix is calculated from a Q-matrix under the conjunctive model, a column in the latent response matrix is the conjunction of some other columns in this matrix if and only if the corresponding column of the Q-matrix can be written as the disjunction of their corresponding columns. While for the disjunctive model, the columns of the latent response matrix have exactly the same disjunction relationships as the columns of the Q-matrix. Because any conjunction or disjunction relationship among the columns of a latent response matrix would imply a disjunction relationship among the columns of a Q-matrix, then we are expected that the relationship between the columns in the Q-matrix can be constructed from the relationship between the corresponding columns in an observed response matrix, resulting from the latent response matrix by adding the noise or random errors. Another reason for this expectation is that each entry in the observed response matrix is modeled as a noisy observation of the corresponding entry in the latent

TABLE 7 | Fit results for the fraction subtraction data using the expert and estimated Q-matrix.

Q-matrix	CDM	–LL2	AIC	BIC
Expert Q-matrix	GDINA	6,695	7,133	8,071
Estimated Q-matrix	GDINA	6,548	6,910	7,686
Expert Q-matrix	DINA	6,912	7,034	7,295
Estimated Q-matrix	DINA	7,030	7,152	7,413
Expert Q-matrix	LLM	6,595	6,781	7,179
Estimated Q-matrix	LLM	6,523	6,707	7,102
Expert Q-matrix	R-RUM	6,696	6,882	7,280
Estimated Q-matrix	R-RUM	6,543	6,727	7,122

−2LL, −2 log likelihood; AIC, Akaike's information criterion; BIC, Bayesian information criterion (Chen et al., 2013).

response matrix through slip and guessing parameters (Junker and Sijtsma, 2001) and the discrepancies between the latent and observed response matrices are considered as random errors (Tatsuoka, 1987).

From the key theoretical results above, the semi-supervised method and an optimal design were then proposed for Q-matrix specification based on test response data and a reachability matrix specified by subject matter experts, and the simulation study was conducted to investigate the performance of the new method and the optimal design for examinee sampling in terms of the CRR of q-entries. From the CRR of q-entries, it is clear found that: (a) for the random design, when item parameters for items with known and unknown q-vectors are $\leq$0.20, the average of CRRs of q-entries for the semi-supervised method is larger than or equal to 0.9, (b) for the optimal design, when item parameters for items with known and unknown q-vectors are $\leq$0.25, the average of CRRs of q-entries for the semi-supervised method is larger than or equal to 0.9, and (c) item parameters of the reachability matrix have a larger impact on the performance of the semi-supervised method than item parameters with unknown q-vectors.

Finally, based on the results obtained in this study, some problems worthy of study in the future are put forward. First, how to effectively use the most of data or information on some other items for which experts have also specified q-vectors, because as the increase of the number of item specified q-vectors, the time complexity (more specifically, exponential time) of the exhaustive method grows much faster? If the number of items is increased to double or triple the number of attributes corresponding to the reachability matrix, one should investigate whether choosing a small part of items with high quality will reduce the noise of the responses and improve the estimation of q entries of unknown items. Second, in the simulation study, we know exactly how many attributes all items include. However, in the real situation, some items with unknown Q-matrix may mix additional attributes not specified in the reachability matrix because we haven't reviewed all items. Thus, we should explore a novel or revised method for identifying the possibility of extra attribute(s). Third, if the Q-matrix obtained from the semi-supervised method is taken as an initial matrix or a provisional Q-matrix of the existing supervised methods, is it possible to further improve the recovery of Q-matrix? From the results of the study, it can be seen that item parameters or random errors of item responses have an impact on the recovery of Q-matrix. If there is a method to reduce noise in item responses, the recovery of Q-matrix may be further improved. We only considered the small set of items with known q-vectors and fixed item parameters. Additional work is needed to further examine the impact of not only error patterns for known q-vectors but different item parameters for test items. Fourth, the current study focused on the DINA and DINO model only. In the future, the proposed method should be applied to general families of cognitive diagnostic models such as the generalized DINA model (de la Torre, 2011), the log-linear cognitive diagnostic model (Henson et al., 2009), the general diagnostic model (von Davier, 2008), testlet cognitive diagnosis model (Zhan et al., 2018), or polytomous cognitive diagnosis models (Chen and de la Torre, 2018; Ma, 2019). Lastly, since only the independent attribute structure in the simulation study and hierarchy structures for the conjunctive model in real data analysis were considered, the proposed method for other attribute hierarchies with different cognitive assumptions is worth studying.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.rdocumentation.org/packages/CDM/versions/7.4-19/topics/fraction.subtraction.data.

AUTHOR'S NOTE

Based on the fact that any columns of a reduced Q-matrix can be expressed by the columns of a reachability R matrix under the logical OR operation, a semi-supervised learning approach and an optimal design for examinee sampling were proposed for Q-matrix specification under the conjunctive and disjunctive model. This method only required subject matter experts specifying a R matrix corresponding to a small part of test items. Simulation and real data analysis showed that the new method with the optimal design is promising in terms of correct recovery rates of q-entries.

AUTHOR CONTRIBUTIONS

WW, LS, and TW conducted a design of the study, data analysis, paper writing, and revision. SD revised the paper. PG and JX give some descriptions of data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This research was partially supported by the Key Project of National Education Science Twelfth Five Year Plan of Ministry of Education of China (Grant No. DHA150285).

REFERENCES

Barnes, T. (2011). "Novel derivation and application of skill matrices: the q-matrix method," in *Handbook of Educational Data Mining,* eds C. Romero, Sebastian Ventura, M. Pechenizkiy, and R. S. J. D. Baker (Boca Raton, FL: CRC Press), 159–172.

Barnes, T. M. (2003). *The q-matrix method of fault-tolerant teaching in knowledge assessment and data mining* (Unpublished Doctoral dissertation). Raleigh, NC: North Carolina State University.

Beheshti, B., Desmarais, M., and Naceur, R. (2012). "Methods to find the number of latent skills," in *Proceedings of the 5th International Conference on Educational Data Mining* (Chania).

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika* 80, 1–20. doi: 10.1007/S11336-014-9401-5

Chen, J. (2017). A residual-based approach to validate q-matrix specifications. *Appl. Psych. Meas.* 41, 277–293. doi: 10.1177/0146621616686021

Chen, J., and de la Torre, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Front. Psychol.* 9:1474. doi: 10.3389/fpsyg.2018.01474

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

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Appl. Psych. Meas.* 37, 598–618. doi: 10.1177/0146621613488436

Chiu, C.-Y., and Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classif.* 30, 225–250. doi: 10.1007/s00357-013-9132-9

Close, C. N. (2012). *An exploratory technique for finding the q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data* (Unpublished Doctoral dissertation). Minneapolis, MN: University of Minnesota, Educational Psychology.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/S11336-011-9207-7

de la Torre, J., and Chiu, C.-Y. (2016). A general method of empirical q-matrix validation. *Psychometrika* 81, 253–273. doi: 10.1007/s11336-015-9467-8

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the q-matrix. *Appl. Psych. Meas.* 35, 8–26. doi: 10.1177/0146621610377081

DeCarlo, L. T. (2012). Recognizing uncertainty in the q-matrix via a bayesian extension of the DINA model. *Appl. Psych. Meas.* 36, 447–468. doi: 10.1177/0146621612449069

Desmarais, M. (2011). "Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization," in *The 4th International Conference on Educational Data Mining (EDM 2011),* eds M. Pechenizkiy, C. Conati, S. Ventura, C. Romero, and J. Stamper (Eindhoven).

Desmarais, M. C., Beheshti, B., and Naceur, R. (2012). "Item to skills mapping: deriving a conjunctive q-matrix from data" in *Intelligent Tutoring Systems Lecture Notes in Computer. Science,* Vol. 7315, eds S. A. Cerri, W. J. Clancey, G. Papadourakis, and K. Panourgia (Berlin; Heidelberg: Springer), 454–463.

Desmarais, M. C., Beheshti, B., and Xu, P. (2014). "The refinement of a q-matrix assessing methods to validate tasks to skills mapping," in *Proceedings of the 7th International Conference on Educational Data Mining*, eds J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren (London).

Desmarais, M. C., and Naceur, R. (2013). "A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices," in *Artificial Intelligence in Education (AIED 2013, LNAI 7926),* eds H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik (Heidelberg: Springer), p. 441–450.

Ding, S.-L., Luo, F., Yan, C., Lin, H.-J., and Wang, X.-B. (2008). "Complement to tatsuoka's q matrix theory," in *New Trends in Psychometrics,* eds K. Shigemasu, A. Okada, T. Imaizumi, and T. Hoshino (Tokyo: Universal Academy Press), 417–424.

Ding, S.-L., Zhu, Y.-F., Lin, H.-J., and Cai, Y. (2009). Modification of tatsuoka's q matrix theory. *Acta. Psy. Sinica* 41, 175–181. doi: 10.3724/SP.J.1041.2009.00175

Ding, S. L., Luo, F., Wang, W. Y., and Xiong, J. (2015). "Dichotomous and polytomous q matrix theory," in *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society,* eds L. A. V. D. Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, and M. Wiberg (Beijing: Springer).

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika* 49, 175–186. doi: 10.1007/BF02294171

Fischer, G. H. (1995). "The linear logistic test model," in *Rasch Models: Foundations, Recent Developments, and Applications,* eds G. H. Fischer and I. W. Molenaar (New York, NY: Springer-Verlag), 131–155.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* 26, 301–321. doi: 10.1111/j.1745-3984.1989.tb00336.x

Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). Urbana, IL: University of Illinois at Urbana-Champaign.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/S11336-008-9089-5

Jang, E. E. (2009). Cognitive diagnostic assessment of l2 reading comprehension ability: validity arguments for fusion model application to languedge assessment. *Lang. Test.* 26, 31–73. doi: 10.1177/0265532208097336

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psych. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kang, C., Yang, Y., and Zeng, P. (2018). Q-matrix refinement based on item fit statistic rmsea. *Appl. Psych. Meas.* 43, 527–542. doi: 10.1177/0146621618813104

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

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

Liu, J., Xu, G., and Ying, Z. (2012). Data-driven learning of q-matrix. *Appl. Psych. Meas.* 36, 548–564. doi: 10.1177/0146621612461894

Liu, J., Xu, G., and Ying, Z. (2013). Theory of self-learning q-matrix. *Bernoulli* 19, 1790–1817. doi: 10.3150/12-BEJ430

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

Ma, W., and de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *Br. J. Math. Stat. Psychol.* 73, 142–163. doi: 10.1111/bmsp.12156

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Rupp, A. A., and Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the dina model. *Educ. Psychol. Meas.* 68, 78–96. doi: 10.1177/0013164407301545

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

Tatsuoka, K. K. (1987). Bug distribution and statistical pattern classification. *Psychometrika* 52, 193–206. doi: 10.1007/BF02294234

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

Tatsuoka, K. K. (1995). "Architecture of knowledge structures and cognitive diagnosis: a statistical pattern classification approach," in *Cognitively Diagnostic Assessments,* eds P. D. Nichols, S. F. Chipman, and R. L. Brennan (Hillsdale: Erlbaum), p. 327–359.

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. New York, NY: Taylor and Francis Group.

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

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

Wang, W., Gao, P., Song, L., and Wang, T. (2020). The improved exploratory method of q-matrix specification with noise preprocessing. *J. Jiangxi Normal Univ. (Nat. Sci.)* 44, 136–141.

Wang, W., Song, L., and Ding, S. (2018b). "An exploratory discrete factor loading method for q-matrix specification in cognitive diagnostic models," in *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics and Statistics, Vol. 233*, eds M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar (Cham: Springer).

Wang, W. Y., Song, L. H., Ding, S. L., Meng, Y. R., Cao, C. X., and Jie, Y. J. (2018a). An EM-based method for Q-matrix validation. *Appl. Psych. Meas.* 42, 46–459. doi: 10.1177/0146621617752991

Xu, P., and Desmarais, M. C. (2016). "Boosted decision tree for q-matrix refinement," in *Proceedings of the 9th International Conference on Educational Data Mining*, eds T. Barnes, M. Chi, and M. Feng (Raleigh, NC).

Xu, P., and Desmarais, M. C. (2018). "An empirical research on identifiability and q-matrix design for DINA model," in *11th International Conference on Educational Data Mining (EDM)*, eds K. E. Boyer and M. Yudelson (Raleigh, NC).

Zhan, P., Liao, M., and Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Front. Psychol.* 9:607. doi: 10.3389/fpsyg.2018.00607

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

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 11 September 2020
doi: 10.3389/fpsyg.2020.02205

Measuring Skill Growth and Evaluating Change: Unconditional and Conditional Approaches to Latent Growth Cognitive Diagnostic Models

Qiao Lin[1]*, *Kuan Xing*[2] and *Yoon Soo Park*[3]

[1] Department of Educational Psychology, University of Illinois at Chicago, Chicago, IL, United States, [2] University of Tennessee Health Science Center, Memphis, TN, United States, [3] Department of Medical Education, University of Illinois at Chicago, Chicago, IL, United States

During the past decade, cognitive diagnostic models (CDMs) have become prevalent in providing diagnostic information for learning. Cognitive diagnostic models have generally focused on single cross-sectional time points. However, longitudinal assessments have been commonly used in education to assess students' learning progress as well as evaluating intervention effects. Thus, it becomes natural to identify longitudinal growth in skills profiles mastery, which can yield meaningful inferences on learning. This study proposes longitudinal CDMs that incorporate latent growth curve modeling and covariate extensions, with the aim to measure the growth of skills mastery and to evaluate attribute-level intervention effects over time. Using real-world data, this study demonstrates applications of unconditional and conditional latent growth CDMs. Simulation studies show stable parameter recovery and classification of latent classes for different sample sizes. These findings suggest that building on the well-established growth modeling frameworks, applications of covariate-based longitudinal CDM can help understand the effect of explanatory factors and intervention on the change of attribute mastery.

Keywords: cognitive diagnostic model, covariate extension, latent growth curve, longitudinal analysis, learning progression

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Qianqian Pan,
The University of Hong Kong, Hong Kong
Jiwei Zhang,
Yunnan University, China

***Correspondence:**
Qiao Lin
qlin7@uic.edu;
qiaolin234@gmail.com

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

Received: *21 May 2020*
Accepted: *05 August 2020*
Published: *11 September 2020*

Citation:
Lin Q, Xing K and Park YS (2020) Measuring Skill Growth and Evaluating Change: Unconditional and Conditional Approaches to Latent Growth Cognitive Diagnostic Models. Front. Psychol. 11:2205. doi: 10.3389/fpsyg.2020.02205

INTRODUCTION

Growth of knowledge and skills are important indicators of learning, which commonly results from the implementation of interventions such as course materials, instructional curriculum, teaching methods, and policies. For educational systems and educators, it is important to understand the changes in learning by evaluating the intervention effects. To quantify these effects, longitudinal assessments or pre-and post-test designs have been widely used and the raw score has been examined to reveal the progress in learning. Longitudinal assessment designs involve repeated observations of variables over a period of time while pre-and post-test designs focus on two measurements that are taken before and after a treatment. However, simple comparison between

time points may lack reliability and validity (Linn and Slinde, 1977); therefore, researchers may seek applying psychometric models to measure students' knowledge and aptitude that are characterized as latent constructs. Item response theory (IRT) allows psychometrically specifying students' ability as continues latent variables and has a tradition to employ longitudinal models to assess the growth in ability (Andersen, 1985; Fischer, 1989; Embretson, 1991). The multidimensional IRT models are useful to measure how the unidimensional ability increase over a period of time, yet it is hard to diagnose the increment when the latent constructs are correlated with one another over repeated measures.

In recent years, cognitive diagnostic models (CDMs), also known as diagnostic classification models (DCMs), have drawn increasing attention from researchers, as provides diagnostic information for learning and instruction (Bradshaw and Levy, 2019). Several CDMs and assessments have been developed to evaluate examinees' mastery status on a set of cognitive skills (e.g., DiBello et al., 1995; Bradshaw et al., 2014; Culpepper, 2019; Culpepper and Chen, 2019). Most commonly, CDMs have been used to assess students' skills profiles at a single time point rather than measuring changes in skills proficiency over time. However, it is important for educators to know the students' learning trajectories to achieve learning goals as well as the effects of intervention on the growth of student skills.

In this regard, latent transition analysis (LTA; Collins and Wugalter, 1992) has been incorporated to the recent development of CDM to evaluate changes in skills mastery. For example, Li et al. (2016) employed DINA model as the measurement model in an LTA to demonstrate a means of analyzing change in cognitive skills over time. Similarly, Kaya and Leite (2017) developed a model combining the LTA and deterministic input noisy "and" gate (DINA; Junker and Sijtsma, 2001) and deterministic input noisy "or" gate (DINO; Templin and Henson, 2006) CDMs to address within-individual and between-groups change in follow-up measurements of learning. In addition, Madison and Bradshaw (2018a) proposed the Transition Diagnostic Classification Model (TDCM) that combined log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) and with LTA to provide a more general framework for measuring growth in cognitive diagnostic modeling. Compared to the models proposed by Li et al. (2016) and Kaya and Leite (2017) that assume specific item response structures and place constraints on parameters, TCDM use a general DCM framework that subsume early models and combine it with LTA. They further extended the TDCM to model multiple groups (MG-TDCM), thereby enabling the examination of group differential growth in attribute mastery in pre-and posttest design (Madison and Bradshaw, 2018b).

To model the learning trajectory, Wang et al. (2018) proposed a family of learning models that use higher-order, hidden Markov model (HO-HMM) to model attribute transition and incorporate CDM framework to understand individualized learning trajectory. Furthermore, Chen et al. (2018) proposed a class of dynamic CDM models to trace learning trajectories. They focused on investigating different types of learning trajectories and developed a Bayesian Modeling framework to estimate these learning trajectories. Focusing on modeling the growth in the higher-order latent trait, Lee (2017) proposed longitudinal growth curve cognitive diagnosis models (GC-CDM) that trace changes in the higher-order latent traits to incorporate learning over time into the cognitive assessment framework. Likewise, Huang (2017) embedded a multilevel structure into higher order latent traits and extended the generalized deterministic input, noisy "and" gate (G-DINA) mode to a multilevel higher order CDM, which enable the measurement of changes in the latent trait in longitudinal data. Most recently, Zhan et al. (2019) proposed a longitudinal diagnostic classification modeling approach for assessing learning growth in both repeated measures design and anchor-item design. Different from the LTA-based methods providing attribute-level transition probability matrix, the proposed longitudinal DINA model (Long-DINA) is able to provide quantitative values of overall and individual growth.

Although various longitudinal CDMs have been developed to measure the transition of examinees' attribute mastery statuses over time, fewer studies focus on the intervention effects that drive the changes in skill mastery from the perspectives of covariates extension and latent growth curve model. In this research study, we proposed two latent growth CDMs by using unconditional and conditional approaches to trace changes in latent attributes over time as well as allowing a flexible parameterization to specify covariates that can be meaningful in studying a longitudinal data structure.

Different from other longitudinal CDMs in the literature, the latent growth CDMs proposed in this study is motivated by the well-established growth curve modeling framework that are commonly used in social sciences to measure latent growth (e.g., Duncan et al., 2013). And as such, it becomes important to link the longitudinal CDM framework to existing techniques in the social sciences, to prompt more generalizable and flexible model extensions. Although some previous studies have incorporated growth curve model into CDM, they mainly focused on extending the higher-order latent trait to model the growth in learning, which assumes associations between different latent attributes, that the probability of having the skills depends on a higher order overall ability. Moreover, a focus on attribute-level changes and their impact over time can be handled and specified in the latent growth CDM framework, as proposed in this paper. In this study, we incorporate covariate extensions and latent growth curve model into the attribute-level of CDM framework instead of the higher-order latent trait level. In this way, we can monitor changes in the attribute level directly under the independence assumption for attributes probabilities.

The study consists of three parts: A real-world data analysis and two simulation studies. We first demonstrate the model application using real-world data to motivate the rationale for the latent growth framework and to monitor changes in students' skills mastery and intervention effects. The ensuing sections include two simulation studies conducted separately to examine the parameter recovery of the proposed models. In this manner, the simulation studies with varying longitudinal design components provide comprehensive inference for different number of time points, sample sizes, and covariate specification conditions. Findings from this study could help researchers

apply the latent growth CDMs in practice and promote the development of longitudinal CDMs.

COGNITIVE DIAGNOSTIC MODELS

Cognitive diagnostic models were designed to classify examinees into skill profiles that indicate their mastery in fine-grained skills or attributes based on their performance on a set of items (Rupp et al., 2010). It refers to a class of psychometric models where patterns of attributes mastery have been represented as latent classes. Distinguish from IRT models that latent traits are continuous, CDMs examine categorical latent traits. Different kinds of CDMs have been developed in literature and various generalizations of CDMs have also been proposed including the LCDM (Henson et al., 2009), general diagnostic model (GDM; von Davier, 2008), and the generalized DINA model (G-DINA; de la Torre, 2011).

Reparameterized DINA (RDINA) Model

In this study, we use the DINA model to demonstrate the framework, which can be applied to other CDM families and generalizations of DINA models (e.g., von Davier, 2008; de la Torre, 2011). The DINA model is developed with the idea that in order to answer an item j correctly, the examinee i must have mastered all of the required skills (Tatsuoka, 1985). The binary latent variable $\eta_{ij} = \prod_{k=1}^{k} \alpha_{ik}^{q_{ik}}$ indicate whether the ith examinee has mastered the set of attributes $\underline{\alpha}$ [i.e., attributes, $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{k,})'$] to solve the jth item, where $\eta_{ij} = 1$ means the presence of the necessary set of attributes, and $\eta_{ij} = 0$ otherwise. As specified by the Q-matrix, q_{ik} is either zero or one, indicating whether the attribute k is required for solving item j. This study uses the reparameterized deterministic inputs noisy "and" gate (DeCarlo, 2011) to apply the longitudinal framework, as it facilitates incorporating covariates as intervention effects in latent growth curve model. The RDINA takes the logit of the traditional DINA model

$$\text{logit}\, p(Y_{ij} = 1|\eta_{ij}) = f_j + d_j\eta_{ij} \tag{1}$$

The f_j parameter indicates the log odds of a false alarm that is the probability of getting item j correct given the examinee do not have the requisite skills. The parameter dj provides a measure of how well the item can discriminate an examinee with or without the mastery of required skills. The guessing and slip parameters used in the DINA model can be recovered by exponentiating the RDINA parameters (DeCarlo, 2011)as:

$$g_j = \exp\left(f_j\right) / \left[1 + \exp\left(f_j\right)\right] \tag{1.1}$$

$$s_j = 1 - \exp(f_j + d_j) / \left[1 + \exp(f_j + d_j)\right] \tag{1.2}$$

Covariate Extension to the RDINA Model

Various latent class models have incorporated covariates as extensions (Dayton and Macready, 1988; DeCarlo, unpublished). In the DINA model, covariates can be specified either at the item level and/or attribute level. Park and Lee (2014) proposed a covariate extension to the RDINA model by applying a latent class regression framework. In particular, when a discrete or a continuous covariate, $\underline{Z}$, is introduced into a latent class model, an examinee's response probability can be expressed as

$$p(Y_{i1}, Y_{i2}, \ldots, Y_{ij}|\underline{Z}) = \sum_{\underline{\alpha}} p(\underline{\alpha}|\underline{Z}) \prod_{j} p(Y_{ij}|\underline{\alpha}, \underline{Z}) \tag{2}$$

where $p(Y_{i1}, Y_{i2}, \ldots, Y_{ij}|\underline{Z})$ represents response probability conditioning on covariate $\underline{Z}$, $p(\underline{\alpha}|\underline{Z})$ represents the covariate affecting the attribute probability, $p(Y_{ij}|\underline{\alpha}, \underline{Z})$ represents the covariate affecting the response probability. In particular, the effects of the covariate on the response probability and attribute probability are shown as following:

$$\text{logit}p(Y_{ij}|\underline{\alpha}, \underline{Z}) = f_j + d_j\eta_{ij} + l_j\underline{Z} \tag{3}$$

$$\text{logit}p(\alpha_k|\underline{Z}) = b_k + h_k\underline{Z} \tag{4}$$

where the parameter l_j reflects the changes in the guessing and slip parameter for a unit change in $\underline{Z}$. Similarly, the parameter h_k indicates the changes in the attribute difficulty parameter (b_k), when the covariate $\underline{Z}$ is conditional on the attribute level.

Relationship Between RDINA and General Diagnostic Model

The RDINA model was employed in this study to establish the framework; however, the parameterization used in its covariate extension can be extended and reparameterized as special cases of the GDM (von Davier, 2005; Park and Lee, 2019). In the GDM, the observed response X is modeled for i items, x response categories, and j respondents as follows:

$$P\left(X = x|i, j\right) = \exp\left[f(\lambda_{xi}, \theta_j)\right] / \left\{1 + \sum_{m} \exp\left[f(\lambda_{xi}, \theta_j)\right]\right\} \tag{5}$$

GDM item parameters are the $\lambda_{xi} = (\beta_{xi}, \boldsymbol{q}_{\mathrm{i}}, \gamma_{xi})$, which include slope parameters and the Q-matrix specification, $\underline{q_{\mathbf{i}}}$. In the DINA where attributes are binary, the skill vector for examinee j, $\theta_j = (\alpha_{j1}, \ldots, \alpha_{jk})$, are binary values. As shown in von Davier (2014, p. 58), the DINA can be parameterized as a special case of the GDM as follows:

$$P\left(X_{vi} = 1|\boldsymbol{q}_{i,}^{*}, \boldsymbol{a}^{*}\right) = \frac{\exp(\beta_i + \sum_k \gamma_{ik} a_k^* q_{ik}^*)}{1 + \exp(\beta_i + \sum_k \gamma_{ik} a_k^* q_{ik}^*)} \tag{6}$$

When a covariate Z introduced to Eq. 5, the following h_k parameters are added:

$$P\left(X_{vi} = 1|\boldsymbol{q}_{i,}^{*}, \boldsymbol{a}^{*}, Z\right) = \frac{\exp(\beta_i + \sum_k \gamma_{ik} a_k^* q_{ik}^* + h_k Z)}{1 + \exp(\beta_i + \sum_k \gamma_{ik} a_k^* q_{ik}^* + h_k Z)} \tag{7}$$

Taking the logit simplifies the model to the item-level of covariate extension approach as presented in Eq. 4.

Latent Growth Curve Model

Latent growth curve model has been widely used in longitudinal analysis to estimate growth over time, such as examining

the treatment effects in the pre-post intervention study. As a special case of structural equation model (SEM), latent growth modeling formwork extend SEM to represent repeated measures of dependent variables as a function of time and other measures. Based on the research of Tucker (1958), Rao (1958), Meredith and Tisak (1984) and Meredith and Tisak (1990) furthered SEM to model the interindividual differences in change. To model the changes in a variable over time, latent growth curve model assumes that there is a systematic trajectory of change underlying the repeated measures of the variable. In particular, for i (i = 1,2,. . .,n) subjects measured at j (j = 1,2,. . .,t) occasions, the measurement model of latent growth curve model can be expressed as

$$y_{ij} = \lambda_{0j}\eta_{0i} + \lambda_{1j}\eta_{1i} + \varepsilon_{ij} \tag{8}$$

where y_{ij} is the outcome variable for individual i at time j. η_{0i} and η_{1i} represent latent trajectory parameters: individual's initial level (i.e., intercept) and rate of change over time (i.e. slopes). ε_{ij} represent time-specific error for person i. In the structural model of latent growth curve model, these latent trajectory parameters become outcome variables that can be expressed as:

$$\eta_{0i} = \mu_0 + e_{0i} \tag{8.1}$$

$$\eta_{1i} = \mu_1 + e_{1j} \tag{8.2}$$

where μ_0 represents the sample mean initial level, e_{0i} represents the deviations from mean initial level for individual i; μ_1 represents the sample mean rate of change, e_{1i} represents the deviations from mean rate of change for individual i.

In latent growth curve model, λ_{0j} is fixed at 1 for all j occasions. It should be noted that the equations presented above are considered as an unconditional latent growth curve model because there is no covariate involved. A conditional latent growth curve model that contains covariates can be specified by adding predictors of the outcome variable into Eq 8. The corresponding covariates effects on the latent trajectory parameters could be included in Eqs 8.1 and 8.2.

The Latent Growth Cognitive Diagnostic Model

Motivated from latent growth curve models and RDINA model with covariate extensions, we propose two latent growth curve CDMs (LG-CDMs) using unconditional and conditional approaches to track the changes in examinees' latent attributes as well as evaluating the effects of covariate at the latent attributes level. For the LG-CDMs, an examinee i's response probability $p(Y_{ij} = 1|\eta_{ij})$ was specified by the RDINA model, which was shown in the above Eq. 1.

Unconditional LG-CDM

Motivated from the unconditional latent growth curve model with random intercept and random slope, we develop an Unconditional LG-CDM that includes unconditional latent growth curve to the attribute level of CDM framework as linear model. At the attribute level $\underline{\alpha_k} = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$, we assume a linear relationship between time and attributes to model the changes in attributes over time. When no covariate is specified, Eq. (9) shows the latent growth model with time effect on the probabilities of the K attributes $p(\alpha_k|time)$ as

$$\text{logit}\, p(\alpha_k|time) = b_k + \gamma_k\,(time) + \zeta_k + \varepsilon_{tk}, \quad \varepsilon_{tk} \sim N(0, 1) \tag{9}$$

Equation 9 represents time is conditioned on the attributes probability, which can be viewed as a predictor of the attribute patterns. Following the interpretation of the RDINA model with covariate extension, b_k represents the fixed-effect attributes difficulty parameter. ζ_k represents the random intercept parameter for attribute k, which allows estimation for each attribute, accounting for individual examinee differences at baseline. Similarly, γ_k represents the random slope parameter for attribute k, which allows differences in examinee rates of growth. ε_{tk} represents time-specific error for attribute k. Eq. 9.1 shows the associated equation for the random intercept that follows a normal distribution with mean μ_{0k} and variance of σ_{0k} . Eq. 9.2 shows the association equation for the random slope that also follows a normal distribution with mean μ_{1ki} and variance of σ_{1k}. e_{0k} and e_{1k} represent the deviations from mean initial level and mean rate of change for attribute k, which are random-effect parameters introduced by the latent growth curve model. In this study, the mean and variance of both random effects are fixed to 0 and 1, respectively.

$$\zeta_k = \eta_{0k} = \mu_{0k} + e_{0k} \tag{9.1}$$

$$\gamma_k = \eta_{1k} = \mu_{1k} + e_{1k} \tag{9.2}$$

Where η_0 and η_1 represent latent trajectory parameters: individual's initial level and rate of change over time, which was specified in the latent growth curve model framework.

Conditional LG-CDM

In addition, we also propose a Conditional LG-CDM based on the conditional latent growth curve model with random intercept and random slope to evaluate the effects of covariate (e.g., intervention effect) on changes in the attribute level. In particular, covariate vector $\underline{Z}$, is introduced into the latent growth curve model effecting on the random effects. $\underline{Z}$ could be either discrete or continuous covariate. Equation (10) represents the latent growth model with both time and covariate effects on the attribute probability $p(\alpha_k|time, \underline{Z})$ as

$$\text{logit}\, p(\alpha_k|time, \underline{Z}) = b_k + \gamma_k\,(time) + h_k\,(\underline{Z}) + \zeta_k + \varepsilon_{tk}, \quad \varepsilon_{tk} \sim N(0, 1) \tag{10}$$

The interpretation of Conditional LG-CDM is similar to the Unconditional LG-CDM that b_k represents the attributes difficulty parameter, ζ_k represents the random intercept parameter, and γ_k represents the random slope parameter. Both random effects follow the normal distribution with fixed mean of 0 and variance of 1 for model identification. ε_{tk} represents time-specific error for attribute k. What's more, the parameter h_k represents the regression coefficient of the covariate vector $\underline{Z}$, which reflects the shift in the attribute difficulty b_k, random intercept ζ_k and random slope γ_k when the covariate is present to affect the attribute. Specifically, Eq. 10.1 shows the term for the

respective covariate coefficient (h_{0k}) affecting random intercept parameter for each attribute k. Equation 10.2 shows the term for the respective covariate coefficient (h_{1k}) affecting random slope parameter for each attribute k.

$$\zeta_k = \eta_{0k} + h_{0k}(\underline{Z}) \tag{10.1}$$

$$\gamma_k = \eta_{1k} + h_{1k}(\underline{Z}) \tag{10.2}$$

Likewise, η_{0k} and η_{1k} represent attribute's initial level and rate of change, which can be extended as shown in Eqs 8.1 and 8.2.

From the perspective of multilevel model, Unconditional LG-CDM and Conditional LG-CDM can be viewed as three-level model where the first level is the time level that involves multiple repeated measures of the same examinee (shown in Eqs 9.1 and 9.2; Eqs 10.1 and 10.2). The second level is the item at individual level that the RDINA model was used to specify an examinee i's item response (shown in Eq. 1). And the third level is the attribute at individual level that the latent growth curve model was used to specify the change of attributes over time and the covariate effect (Eqs 9 and 10).

In addition, time effect could be specified at the item level as well. Modeled with the RDINA, we let Y_{ijt} be an examinee *i*'s response for item *j* at time *t*, given the binary latent variable η_{ijt} and time. The response probability of a person *i* getting item *j* correctly at time *t* is shown as following:

$$\text{logit}\, p(Y_{ijt} = 1|\eta_{ijt}, time) = f_j + d_j\eta_{ijt} + \rho_j(time) + \xi_{ijt} + \varphi_{ijt} \tag{11}$$

$$\xi_{ijt} = \eta_{0i} = \mu_0 + e_{0i} \tag{11.1}$$

$$\rho_j = \eta_{1i} = \mu_1 + e_{1i} \tag{11.2}$$

where ξ_{ijt} represents the random intercept parameter that allows variation in baseline for item response probability $p(Y_{ijt} = 1|\eta_{ijt}$, time). The ρ_j represents the random slope parameter that allows growth rate of item response probability vary across time. φ_{ijt} is the error term.

Taken together, this study focuses on applying the latent growth curve model into the attribute-level of CDM to analyze how the attributes mastery change over time. There are several advantages to using the Unconditional LG-CDM and Conditional LG-CDM, including that they directly estimate the learning trajectory parameters on the attribute level but also allow covariates effects involved to estimate the intervention effects simultaneously.

REAL-WORLD DATA ANALYSIS

Methods

Real-world data analysis was conducted to motivate the potential of the LG-CDMs and demonstrate its application. We used the pretest and posttest data of a mathematics test (N = 879) in the real-world analysis. The mathematic test was developed in a large scale education study that investigated the difficulties for the disabled students in solving mathematic problems (Bottge et al., 2014, 2015). Specifically, an instructional method of Enhanced Anchored Instruction (EAI) was employed in the study to help improve the mathematics achievement of disabled students (Bottge et al., 2003). The design of cluster-randomized controlled trial was used to assign clusters of middle school students to a treatment group or a control group. In the treatment group, students received EAI video sessions about mathematics problem-solving and in the control group, students received instruction method as usual. To evaluate effectiveness of EAI instructional method, the mathematics test was administrated before and after the instructional period and the assessment data was collected. Thus, the dataset consisted of 21-item responses to the test measured four attributes over two time points. A Q-matrix was identified for the four attributes that corresponded to the four instructional units, including α_1: ratios and proportional relationships, α_2: measurement and data, α_3: number systems (fractions), and α_4: geometry (graphing). Each test item was mapped to one attribute (**Appendix**). Two latent growth curve CDMs were fit, Unconditional LG-CDM and Conditional LG-CDM.

Data analyses were conducted using Latent GOLD 5.0 (Vermunt and Magidson, 2013).

Two statistics were used to examine attribute classification – (1) proportion correctly classified (P_c) and (2) Lambda (λ) (Clogg, 1995), where both are based on the maximum posterior probability to examine the quality of classification. In particular, Eq. (12) shows the use of estimated posterior probabilities in obtaining an estimate of the expected proportion of cases correctly classified for attribute k (α_k):

$$P_c = \sum_s \left[n_s \times \max\, p(\alpha_k|\underline{Y_{i1}}, \underline{Y_{i2}}, \ldots, \underline{Y_{iJ}}) \right] / N \tag{12}$$

where s represents each unique response pattern, n_s is the frequency of each pattern (i.e., number of cases with a particular pattern), $\max\, p(\alpha_k|\underline{Y_{i1}}, \underline{Y_{i2}}, \ldots, \underline{Y_{iJ}})$ represents the maximum posterior probability for a given response pattern vector $(\underline{Y_{i1}}, \underline{Y_{i2}}, \ldots, \underline{Y_{iJ}})$, N is the total number of cases in a latent class response pattern.

λ makes a correction for classification that can occur by chance, which can be expressed as

$$\lambda = \frac{P_c - \max\, p(\alpha_k)}{1 - \max\, p(\alpha_k)} \tag{13}$$

where $p(\alpha_k)$ represents the latent class size with $p\,(\alpha_k) > 0$. Meanwhile, λ also reflects the relative reduction in classification error (Kruskal and Goodman, 1954; Clogg, 1995).

Both expectation-maximization (EM) and Newton-Raphson algorithms were used to obtain maximum likelihood (ML) or posterior mode (PM) estimates. The PM estimation uses a prior distribution to smooth solutions that are near the boundary of the parameter space. Therefore, this method can avoid a boundary estimation issues that are commonly associated with latent class models (DeCarlo, 2011). In addition, to avoid problems of local maxima, 100 sets of starting values were used to obtain the global maximum. Finally, to check for local identification, the rank of the Jacobian matrix was examined to be of full rank as specified

as a required condition for local identification in latent class regression models (Huang and Bandeen-Roche, 2004).

Results

Model Fit and Classification

Real-world data converged successfully for the two models, unconditional latent growth curve CDM (Unconditional LG-CDM) and conditional latent growth curve CDM (Conditional LG-CDM). **Table 1** shows the classification results. The results show that the P_c estimates are the same for the two LG-CDMs across four mathematics attributes, indicating a satisfactory classification as all statistics are greater than 0.91. For example, with P_c of 0.96 for attribute 2, one would expect that 96% of the cases would be correctly classified into attribute 2. However, it should be noted that if simply classifies all the cases into the attribute with the largest size, then the correctly classification can be achieved but may due to the chance. λ provides a correction for this situation when calculating the proportion correctly classified. The results show that λ are similar for the four attributes across two models and are slightly lower than P_c. For example, the λ classification on attribute 1 in Conditional LG-CDM is 0.70, indicating that the proportion correctly classified increase 70% by using the observers' response pattern over simply classifying all cases into the attribute with the largest size.

Attribute Prevalence

Table 2 shows the attribute prevalence for the four attributes. The attribute prevalence represents the latent class sizes of the four attributes (DeCarlo, 2011). Overall, the attribute prevalence was consistent across the two models. The probabilities of all attributes prevalence are above 0.50, indicating that more than half of the students mastered each of the attribute. The attribute prevalence for Attribute 2 (Measurement and Data) and Attribute 4 (Geometry) had slightly higher probabilities than the attribute prevalence for Attribute 1 (Ratios and Proportional Relationships) and Attribute 3 (Number Systems).

TABLE 1 | Classification: proportion correctly classified (P_c) and Lambda (λ).

Models	Classification	Attributes			
		α_1	α_2	α_3	α_4
Unconditional LG-CDM	P_c	0.91	0.96	0.92	0.93
	λ	0.79	0.90	0.84	0.82
Conditional LG-CDM	P_c	0.91	0.96	0.92	0.93
	λ	0.70	0.90	0.84	0.82

Classification statistics based on Clogg (1995). Conditional LG-CDM considered EAI instruction. Attributes as follows: α_1: Ratios and Proportional Relationships; α_2: Measurement and Data; α_3: Number Systems (Fractions); α_4: Geometry (Graphing).

TABLE 2 | Attribute prevalence.

Model	α_1	α_2	α_3	α_4
Unconditional LG-CDM	0.57 (0.03)	0.62 (0.02)	0.53 (0.03)	0.62 (0.08)
Conditional LG-CDM	0.57 (0.03)	0.62 (0.02)	0.53 (0.03)	0.62 (0.02)

Values in parenthesis are standard errors. Attributes based on the EAI instruction are as follows: α_1: Ratios and Proportional Relationships; α_2: Measurement and Data; α_3: Number Systems (Fractions); α_4: Geometry (Graphing).

Item Parameters

Table 3 shows the item parameters (f_j and d_j) for the Unconditional LG-CDM and Conditional LG-CDM, which are estimated from the RDINA model. Calculating the exponential of the f_j and d_j parameters, one can obtain guessing and slip parameters of the DINA model (shown in Eqs 1.1 and 1.2). Overall, the item parameters are consistent across the two models. Derived from Eqs 1.1 and 1.2, the average guessing and slip parameters estimates for Unconditional LG-CDM were 0.22 and 0.35; for Conditional LG-CDM were 0.22 and 0.36. With respect to each attribute, the average guessing and slip parameters were respectively the same across the two models: for the attribute of Ratios and Proportional Relationships were 0.14 and 0.49; for the attribute of Measurement and Data were 0.28 and 0.23; for the attribute of Number Systems (Fractions) were 0.28 and 0.23; for the attribute of Geometry (Graphing) were 0.25 and 0.36. In particular, for both Unconditional LG-CDM and Conditional LG-CDM, Item 2 (Attribute of Measurement and Data) has the greatest guessing estimates while Item 17 (Attribute of Ratios and Proportional Relationships) has the greatest slip estimates.

Attribute Parameters and Growth

Table 4 summarized the attribute-level parameters for the Unconditional LG-CDM and Conditional LG-CDM, which

TABLE 3 | Item parameters for Unconditional and Conditional LG-CDM.

Item	Attributes	Unconditional LG-CDM		Conditional LG-CDM	
		f_j	d_j	f_j	d_j
Y_1	α_1	−2.09 (0.02)	1.97 (0.17)	−2.10 (0.16)	1.98 (0.17)
Y_2	α_2	−0.24 (0.08)	2.06 (0.13)	−0.24 (0.08)	2.06 (0.13)
Y_3	α_2	−2.61 (0.17)	1.92 (0.19)	−2.62 (0.18)	1.92 (0.19)
Y_4	α_3	−1.87 (0.13)	1.87 (0.14)	−1.88 (0.13)	1.87 (0.14)
Y_5	α_3	−1.14 (0.11)	2.34 (0.13)	−1.14 (0.11)	2.34 (0.13)
Y_6	α_3	−0.74 (0.09)	1.88 (0.12)	−0.74 (0.09)	1.89 (0.12)
Y_7	α_3	−3.73 (0.42)	3.57 (0.41)	−3.75 (0.43)	3.58 (0.42)
Y_8	α_3	−0.96 (0.09)	1.14 (0.12)	−0.96 (0.09)	1.14 (0.12)
Y_9	α_2	−0.76 (0.09)	2.21 (0.12)	−0.76 (0.09)	2.21 (0.12)
Y_{10}	α_2	−0.97 (0.10)	2.84 (0.13)	−0.97 (0.10)	2.84 (0.13)
Y_{11}	α_2	−0.59 (0.09)	3.09 (0.15)	−0.59 (0.09)	3.09 (0.15)
Y_{12}	α_2	−1.38 (0.11)	3.03 (0.14)	−1.38 (0.11)	3.02 (0.14)
Y_{13}	α_1	−0.64 (0.10)	2.01 (0.13)	−0.63 (0.10)	1.99 (0.13)
Y_{14}	α_1	−2.61 (0.23)	2.93 (0.22)	−2.60 (0.22)	2.91 (0.22)
Y_{15}	α_4	−0.99 (0.10)	1.91 (0.13)	−0.99 (0.10)	1.91 (0.13)
Y_{16}	α_4	−0.70 (0.10)	1.82 (0.12)	−0.70 (0.10)	1.82 (0.12)
Y_{17}	α_1	−3.41 (0.27)	1.94 (0.29)	−3.49 (0.28)	2.02 (0.30)
Y_{18}	α_4	−0.27 (0.09)	2.35 (0.14)	−0.27 (0.09)	2.35 (0.14)
Y_{19}	α_4	−1.77 (0.13)	1.96 (0.15)	−1.76 (0.13)	1.96 (0.15)
Y_{20}	α_4	−1.41 (0.11)	1.47 (0.13)	−1.41 (0.11)	1.47 (0.13)
Y_{21}	α_4	−1.92 (0.14)	1.63 (0.15)	−1.93 (0.14)	1.64 (0.15)

Values in parenthesis are standard errors. Attributes based on the EAI instruction are as follows: α_1: Ratios and Proportional Relationships; α_2: Measurement and Data; α_3: Number Systems (Fractions); α_4: Geometry (Graphing).

include attribute difficulty (b_k), intercept and slope of latent growth curve (η_0 and η_1), and regression coefficients (h_k) for the intervention effect. In general, the estimates of random intercept and random slope were very close, indicating that the examinees' initial level and growth rate are similar for the two models. As the Conditional LG-CDM incorporates covariate in addition to the Unconditional LG-CDM, regression coefficients (h_k) of the covariate indicate shifts in the attributes difficulty as a result of the treatment effect (EAI). Results show that the parameter estimates of four attributes difficulty are all lower in the Conditional LG-CDM than the Unconditional LG-CDM, suggesting that all attributes were easier to be mastered when involving the treatment effect. In particular, Attribute 1 (Ratios and Proportional Relationships) and Attribute 4 (Geometry) yielded greater differences between the two models then Attribute 2 (Measurement and Data) and Attribute 3 (Number Systems), with a difference of 0.34 and 0.37 units. While most treatment effects were not significant, Attribute 4 (Geometry) had significant treatment effect, shifting the difficulty parameter by 0.24 units.

SIMULATION STUDY I

Methods

Simulation studies were conducted to evaluate parameter recovery and classification of the two models: (a) Unconditional LG-CDM; (b) Conditional LG-CDM. In simulation study 1, the EAI real-world results were used as generating population (true) values. Following the data structure of the real-world data example, the 21-item response data were generated for two time points; four attributes and Q-matrix were specified as well. Two sample size of 1000 and 2000 were examined across a specification of the four attributes. Therefore, the simulation study includes a total of four simulation conditions (= 2 models × 2 sample size conditions).

Data were generated and fit using Latent GOLD 5.0 (Vermunt and Magidson, 2013). One hundred replications were fitted for each condition. Parameters were estimated using PM estimation and the parameter recovery were evaluated for each condition using three measures: (a) Bias, (b) % Bias, and (c) mean square error (MSE). Here, $Bias\,(x) = \frac{1}{N}\sum_{n=1}^{N}[\widehat{e_n}\,(x) - e(x)]$, $\%\ Bias = |Bias(x)/e(x)| \times 100\%$, $MSE\,(x) = \frac{1}{N}\sum_{n=1}^{N}[\widehat{e_n}\,(x) - e(x)]^2$, where x is an arbitrary indicator of a parameter, e(x) is the generating (true) parameter value, and $\widehat{e_n}\,(x)$ is the nth replicate estimate of parameter x among a total of N = 100 replications. Similar to the real-world data analysis, we set up 100 starting values, and the Jacobian matrix was examined to be of full rank for local identification. EM (expectation-maximization) and Newton-Raphson algorithms were used to avoid a boundary estimation issue using PM estimation. The syntax of unconditional and conditional models is available from the authors upon request.

TABLE 4 | Attribute difficulty, growth curve and intervention effect parameters.

	Unconditional LG-CDM		Conditional LG-CDM	
Parameter	**Estimate**	**p value**	**Estimate**	**p value**
η_0 (Random Intercept)	0.44 (0.11)		0.46 (0.15)	
η_1 (Random Slope)	0.04 (0.00)		0.03 (0.00)	
h_{0k} (Treatment Effect on random intercept)	–	–	0.30 (0.19)	0.13
h_{1k} (Treatment Effect on random slope)	–	–	0.09 (0.00)	<0.001
b_1 (Attribute 1 Difficulty)	−0.11 (0.09)		−0.45 (0.11)	
h_1 (Treatment Effect for Attribute 1)	–	–	0.40 (0.27)	0.13
b_2 (Attribute 2 Difficulty)	0.16 (0.07)		−0.04 (0.09)	
h_2 (Treatment Effect for Attribute 2)	–	–	0.09 (.25)	0.72
b_3 (Attribute 3 Difficulty)	−0.23 (0.08)		−0.44 (0.10)	
h_3 (Treatment Effect for Attribute 3)	–	–	0.02 (0.26)	0.95
b_4 (Attribute 4 Difficulty)	0.16 (0.04)		−0.11 (0.04)	
h_4 (Treatment Effect for Attribute 4)	–	–	0.24 (0.00)	<0.001

(1) Parameter h_k indicates shift in the attributes difficulty due to mastery in treatment effect (EAI), based on the Conditional LG-CDM model. (2) Values in parenthesis represent standard errors. (3) Attributes based on the EAI instruction are as follows: α_1: Ratios and Proportional Relationships; α_2: Measurement and Data; α_3: Number Systems (Fractions); α_4: Geometry (Graphing).

Results

Parameter Recovery

Table 5 shows the parameter recovery results (Bias, % Bias, and MSE) by sample size (1,000 and 2,000) for the two models. Overall, the parameter recovery revealed consistent estimates. Percent bias for item-level parameters were very close in the two models except item discrimination parameter *dj* was slightly higher in the Conditional LG-CDM model. The parameter *dj* provides a measure of how well the item can discriminate an examinee with or without the mastery of required skills. The overall % bias associated with random effects were all less than 2.0% regardless of models. In Conditional LG-CDM, % bias of random effects were lower ($\leq$ 1.0%). The intervention effect on attributes in the conditional model had % bias of 37.4% when sample size is 1000. However, it dramatically dropped to 4% when the sample size increase to 2000. The % bias of attribute difficulty parameter are lower in the Unconditional LG-CDM model (6.7 and 3% at the sample sizes of 1,000 and 2,000).

Classification

Simulation classification indices were summarized in **Table 6**. The results of the two LG-CDM models agree very much. Both P_c and λ showed excellent classification rates on the four attributes in the two models. The classification index was little influenced by the sample sizes.

Cross Fitting of Simulated Data to Other Models

What's more, a cross fitting analysis was conducted to examine the consequences on parameter estimates and classification on fitting incorrect models. Data were generated using the

TABLE 5 | Simulation I results: parameter recovery.

Model	Level	Parameter	Sample n = 1,000			Sample n = 2,000		
			Bias	% Bias	MSE	Bias	% Bias	MSE
Unconditional LG-CDM	Random Effects	λ	0.007	1.8%	0.002	−0.002	0.5%	0.001
	Attribute Difficulty	b_k	−0.002	6.7%	0.002	0.000	3.1%	0.001
	Item	f_j	−0.013	1.6%	0.024	−0.001	0.6%	0.007
		d_j	0.015	0.9%	0.030	0.002	0.4%	0.009
Conditional LG-CDM	Random Effects	λ	0.001	0.3%	0.001	−0.002	0.4%	0.001
		h_k	0.002	0.7%	0.003	−0.003	1.0%	0.004
	Attribute Difficulty	b_k	0.000	18.2%	0.022	0.000	16.9%	0.002
	Intervention Effect	h_k	0.002	37.4%	0.018	−0.001	4.0%	0.018
	Item	f_j	−0.009	1.6%	0.022	−0.002	0.7%	0.007
		d_j	0.011	1.5%	0.030	0.003	0.5%	0.010

Simulation based on 100 data replications.

Unconditional LG-CDM and Conditional LG-CDM and fit with incorrect models one-off, including RDINA model and RDINA model with covariate. The RDINA and RDINA with covariate models do not have a longitudinal component, thereby providing a relative comparison for ignoring the longitudinal component to the model. Model fit indices, P_c and % bias of parameters (random effects, intervention effect, item, and attribute) are presented in the **Table 7**. All the statistics selected the correct models. For the sample size of 2000, the correct models had higher P_c, lower AIC/BIC value as well as lower % bias in both item and attribute parameter estimates. Meanwhile, when fitting with the incorrect models, lower P_c, higher AIC/BIC and higher % bias were shown in the outputs. The greatest impact of fitting incorrect models was found in the % bias of attribute and item parameters. When using Conditional LG-CDM generated data and fit with the RDINA with covariates, % bias was more than 101% for the item parameter and attribute difficulty parameter. Similarly, when fitting generated data with RDINA, % bias were also high for the attribute parameter. It is noticeable when data generated using Conditional LG-CDM were fit using Unconditional LG-CDM, % bias are lower than the incorrect RDINA and RDINA with covariate model.

TABLE 6 | Simulation classification I: proportion correctly classified (P_c) and Lambda (λ).

Models	Classification	α_1	α_2	α_3	α_4
Unconditional LG-CDM (n = 1,000)	P_c	0.91	0.98	0.94	0.94
	λ	0.80	0.95	0.87	0.84
Conditional LG-CDM (n = 1,000)	P_c	0.91	0.98	0.94	0.94
	λ	0.80	0.94	0.87	0.84
Unconditional LG-CDM (n = 2,000)	P_c	0.97	0.99	0.98	0.98
	λ	0.93	0.99	0.97	0.96
Conditional LG-CDM (n = 2,000)	P_c	0.97	0.99	0.98	0.98
	λ	0.93	0.99	0.97	0.96

Classification statistics based on Clogg (1995).

SIMULATION STUDY II

Methods

We conducted an additional simulation study, where three time points were simulated to examine parameter recovery and classification of the proposed Unconditional and Conditional LG-CDMs. As the data of simulation study I were generated using population (true) values derived from the empirical data that was limited to two time points, the condition was expanded to include more time points in simulation study II so that the potentials of the Unconditional and Conditional LG-CDMs can be fully discussed. In simulation study II, to generate data, we referred to the simulation study design conducted by De La Torre and Douglas (2004) to specify the Q-matrix and item parameters, as generating population (true) value: 30 items with five attributes and 1,000 examinees were used across 100 replications. **Table 8** shows the transposed Q-matrix that each attribute appears alone, in pair, or in a triple the same number of times as other attributes. Similar to Simulation Study I, 100 data replications were generated and fit using Latent GOLD 5.0 (Vermunt and Magidson, 2013). Parameters were estimated using PM full name estimation and the parameter recovery were evaluated for each condition using three measures: (a) Bias, (b) % Bias, and (c) mean square error (MSE).

Results

Parameter Recovery

Table 9 shows the parameter recovery results of three time points simulation by sample size 1,000 for the two models. Bias, % bias and MSE were lower for item level parameters (discrimination parameter d_j and false rate parameter f_j) in the two models, indicating model estimates are consistent in the item level. The results of parameter recovery for attribute difficulty are slightly high as well as for the intervention effect on attribute difficulty. Furthermore, the bias and % bias of random effects (random intercept and random slope) are noticeably high in both models, probably because more time points are involved

TABLE 7 | Cross fitting of simulated data.

Data fit	Statistics	Data generating conditions (Sample n = 2,000)	
		Unconditional LG-CDM	**Conditional LG-CDM**
RDINA	AIC	98964.26	98945.77
	BIC	99234.79	99235.29
	P_c	0.87	0.87
	Attribute Difficulty (% Bias)	91.1%	88.4%
	Item (% Bias)	28.1%	28.0%
RDINA with covariates	AIC	–	98847.14
	BIC	–	99161.84
	P_c	–	0.87
	Attribute Difficulty (% Bias)	–	101.1%
	Intervention Effect (% Bias)	–	64.5%
	Item (% Bias)	–	101.2%
Unconditional LG-CDM	AIC	**91921.67**	98728.2
	BIC	**92157.24**	98997.04
	P_c	**0.98**	0.97
	Random Effects (% Bias)	**0.5%**	13.3%
	Attribute Difficulty (% Bias)	**3.1%**	46.1%
	Item (% Bias)	**0.5%**	1.1%
Conditional LG-CDM	AIC	–	**91971.63**
	BIC	–	**92236.65**
	P_c	–	**0.98**
	Random Effects (% Bias)	–	**0.7%**
	Attribute Difficulty (% Bias)	–	**16.9%**
	Intervention Effect (% Bias)	–	**4.0%**
	Item (% Bias)	–	**0.6%**

Results of the correct model fit were bolded.

in the simulation study that ask for more estimations to reach consistent recovery.

Classification

Table 10 summarized classification results of three time points simulation. For both LG-CDMs, P_c estimates suggested satisfactory proportion of cases are correctly classified for all attributes ($P_c > 0.85$). Yet, it should be noted that the classification rates of λ were lower on the second and fourth attributes compared to the other attributes, which may due to the over correction for classification error.

TABLE 8 | The transposed Q-matrix for the simulation study II.

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

TABLE 9 | Simulation II results: parameter recovery.

Model	**Level**	**Parameter**	**Sample n = 1,000**		
			Bias	**% Bias**	**MSE**
Unconditional LG-CDM	Random Intercept	λ_o	−0.592	118.4%	0.354
	Random Slope	λ_1	0.312	62.4%	0.098
	Attribute Difficulty	b_k	−0.311	39.0%	0.123
	Item	f_j	0.073	5.1%	0.059
		d_j	−0.046	2.7%	0.088
Conditional LG-CDM	Random Intercept	λ_o	−0.548	109.6%	0.304
		h_{0k}	0.035	11.8%	0.007
	Random Slope	λ_1	0.315	63.1%	0.101
		h_{1k}	−0.222	74.1%	0.053
	Attribute Difficulty	b_k	−0.297	37.7%	0.122
	Intervention Effect	h_k	0.001	20.6%	0.173
	Item	f_j	0.065	4.7%	0.058
		d_j	−0.038	2.5%	0.082

Simulation based on 100 data replications.

TABLE 10 | Simulation classification II: proportion correctly classified (P_c) and lambda (λ).

Models	Classification	α_1	α_2	α_3	α_4	α_5
Unconditional LG-CDM (n = 1,000)	P_c	0.89	0.86	0.98	0.96	0.95
	λ	0.72	0.47	0.79	0.36	0.76
Conditional LG-CDM (n = 1,000)	P_c	0.90	0.88	0.98	0.97	0.94
	λ	0.76	0.50	0.81	0.35	0.81

Classification statistics based on Clogg (1995).

DISCUSSION AND CONCLUSION

Cognitive diagnostic models have become increasingly important in educational measurement by estimating skill profiles that indicate the examinee's mastery in fine-grained skills based on their performance (Rupp et al., 2010). In most prior studies, CDMs have been applied to single cross-sectional time diagnosis instead of tracking the changes in skills or attributes. However, learning is a process during which students acquire knowledge and improve their skills. As learning progress, students' skills mastery and knowledge could change over time. In addition, the implementation of particular intervention may influence students' learning trajectory, which is important for educators to know in order to evaluate learning and instruction. In this study, we propose two latent growth CDMs, Unconditional LG-CDM and Conditional LG-CDM, to assess students' change in skills mastery over time and evaluate the intervention effect on the growth of skill mastery.

Results from the real-world data analysis showed that the latent growth curve model and covariate extension such as intervention effect, could be used to link with a CDM. The statistics of model classification and attribute prevalence agree very much and are excellent for the two LG-CDMs, indicating both models are well specified to provide consistency results. In particular, although the latent class size of Attribute 1 and Attribute 3 are slightly lower than Attribute 2 and Attribute 4, more than half of the students mastered each attribute. Moreover, results showed that the estimates of attribute difficulty of the Conditional LG-CDM was generally lower than the estimates of the Unconditional LG-CDM. The decrease in attribute difficulty indicated that the attributes have been shifted and implies that it become easier for students to master the attributes when involving educational intervention, which was further confirmed by the results of growth curve and intervention effect parameters. Both the Unconditional LG-CDM and Conditional LG-CDM examined the students' performance at baseline and the growth rate. Although the baseline performance is slightly different, the grow rates are similar across the two models. Thus, the LG-CDMs could inform the researchers and educators that the EAI method has little effect on the growth rate of student' ability. However, with the help of Conditional LG-CDM that incorporate the covariates extension into CDM, we can tell that the treatment EAI method does improve students' mastery on the attribute 1 and attribute 4. In other words, if the students were assigned to the treatment group that receiving the EAI teaching method, they would show progress in math learning, especially in the skills of geometry and ratio/proportional relationships.

The simulation studies showed that the parameters were consistently recovered in general, indicating that incorporating latent growth curve and covariate extension at the attribute level did not affect model estimation. In simulation study I, both attribute and item level estimates were stable with sample size of 1,000 and 2,000 for the Unconditional LG-CDM. For the Conditional LG-CDM, additional attention may be given to the attribute difficulty parameter (b_k) and intervention effect parameter (h_k), percentage bias of 18.2% and 37.4% for the sample size of 1,000. However, with the sample size increase to 2,000, the % bias of intervention effect declines rapidly to 4%. The recovery of random parameters was excellent across two models, with bias of random intercepts are less than 2.0% in the sample size of 1,000 and 2,000.

In Simulation Study II, more time points and items were involved to fully examine the performance of proposed LG-CDMs in terms of parameter recovery. The item level estimates were satisfactory with sample size of 1,000. Although the bias and % bias of attribute difficulty parameter and intervention effect parameters are slightly high, it is expected that they would decrease obviously when the sample size increase. However, it should be noted that the recovery of random effects for data with three time-point specifications were modest and may depend on study design. Thus, latent growth models may need more specifications or constrains on random intercept and random slope parameters to achieve stable recovery. Besides, the classification indices of the both simulation studies were consistent across different conditions. The results obtained from this study help to advance CDMs to better measuring the change in learning over time.

Researchers and educators have long used pre-post assessment to evaluate the effects of new curriculum and teaching method on students' learning. Meanwhile, it is important to know students' learning trajectory to achieve learning goal. Cognitive diagnostic models have provided a diagnostic framework to measure students' mastery in fine-grained skills and different kinds of longitudinal analysis have been incorporate to the CDM to evaluate changes in skills profile mastery (e.g., Li et al., 2016; Kaya and Leite, 2017). Different from other longitudinal CDMs, the LG-CDMs described in this article incorporate well-established latent growth curve model that is more widely used in the social studies. Additionally, covariate extension was integrated to signify the intervention effect. Dayton and Macready (1988) introduced the use of covariate to affect the attribute. Park et al. (2017) included both observed and latent explanatory variables as covariates in the explanatory CDM to inform learning and practice. Thus, this approach is meaningful in the CDM for its diagnostic purpose. In this study, latent growth curve and covariates are both specified at the attribute level. In addition, latent growth curve could be also specified at the item level, contributing to the multilevel studies on CDMs. Meanwhile, more attributes and covariates could be incorporated in specifying the items and explaining the relationship among them. In the empirical study, it is likely to have items that are of high slip and guessing estimates in the test (Lee et al., 2011),

therefore, it is important for the researchers to carefully develop and validate the Q-matrix used for CDM analyses.

For the future studies, different types of variance-covariance structures could be specified in the model. For example, all the parameters could be freely estimated (unstructured) in the covariance structure to explore their relationships in model estimation. Meanwhile, future studies could conduct a comprehensive investigation of the measurement invariance, building on the foundational simulation studies conducted in this paper. For example, additional parameters could be included to examine their effects on model identification and their impact on the measurement invariance. Furthermore, the item level fit statistics could be developed in the future studies for the item-level effects in the LG-CDMs, which provides suitable item level fit information for the studies that involve multiple time points. In the current field of education, individualized learning has been emphasized, which allows students to construct learning progress at own pace. This study provides a flexible framework to diagnose skill mastery as well as advancing the longitudinal CDMs to better measuring the change in learning over time.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The use of datasets requires permission. Requests to access these datasets should be directed to University of Kentucky.

AUTHOR CONTRIBUTIONS

QL contributed to the conceptualization of the manuscript and conducted formal analysis as well as wrote the original draft. KX contributed to the review and edition of the manuscript. YP contributed to the supervision, conceptualization, and review of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Institute of Education Sciences (IES) grant R324A150035.

ACKNOWLEDGMENTS

The opinions expressed are those of the authors and do not necessarily reflect the views of IES.

REFERENCES

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika* 50, 3–16. doi: 10.1007/bf02294143

Bottge, B. A., Heinrichs, M., Chan, S.-Y., Mehta, Z. D., and Watson, E. (2003). Effects of video-based and applied problems on the procedural math skills of average-and low-achieving adolescents. *J. Special Educ. Technol.* 18, 5–22. doi: 10.1177/016264340301800201

Bottge, B. A., Ma, X., Gassaway, L., Toland, M., Butler, M., and Cho, S. J. (2014). Effects of blended instructional models on math performance. *Except. Child.* 80, 237–255. doi: 10.1177/0014402914527240

Bottge, B. A., Toland, M., Gassaway, L., Butler, M., Choo, S., Griffen, A. K., et al. (2015). Impact of enhanced anchored instruction in inclusive math classrooms. *Except. Child.* 81, 158–175. doi: 10.1177/0014402914551742

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

Bradshaw, L., and Levy, R. (2019). Interpreting probabilistic classifications from diagnostic psychometric models. *Educ. Meas. Issues Pract.* 38, 79–88. doi: 10.1111/emip.12247

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

Clogg, C. C. (1995). "Latent class models," in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, eds G. Arminger, C. C. Clogg, and M. E. Sobel (Boston, MA: Springer), 311–359.

Collins, L. M., and Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivar. Behav. Res.* 27, 131–157. doi: 10.1207/s15327906mbr2701_8

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

Culpepper, S. A., and Chen, Y. (2019). Development and application of an exploratory reduced reparameterized unified model. *J. Educ. Behav. Stat.* 44, 3–24. doi: 10.3102/1076998618791306

Dayton, C. M., and Macready, G. B. (1988). "A latent class covariate model with applications to criterion-referenced testing," in *Latent Trait and Latent Class Models*, eds R. Langeheine, and J. Rost (Boston, MA: Springer), 129–143. doi: 10.1007/978-1-4757-5644-9_7

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

De La Torre, J., and Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007\%2FBF02295640

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: the DINA model, classification, latent class sizes, and the Q-matrix. *Appl. Psychol. Meas.* 35, 8–26. doi: 10.1177/0146621610377081

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995). "Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques," in *Cognitively Diagnostic Assessment*, eds P. Nichols, S. Chipman, and R. Brennen (Hillsdale, NJ: Earlbaum), 361–389.

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2013). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Application*. Abingdon: Routledge.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika* 56, 495–515. doi: 10.1007/bf02294487

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika* 54, 599–624. doi: 10.1007/bf02296399

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

Huang, G. H., and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* 69, 5–32. doi: 10.1007/BF02295837

Huang, H. Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *J. Educ. Meas.* 54, 440–480. doi: 10.1111/jedm.12156

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kaya, Y., and Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: an evaluation of model performance. *Educ. Psychol. Meas.* 77, 369–388. doi: 10.1177/0013164416659314

Kruskal, W. H., and Goodman, L. (1954). Measures of association for cross classifications. *J. Am. Stat. Assoc.* 49, 732–764. doi: 10.1007/978-1-4612-9995-0

Lee, S. Y. (2017). *Growth Curve Cognitive Diagnosis Models for Longitudinal Assessment*. Berkeley, CA: UC Berkeley.

Lee, Y.-S., Park, Y. S., and Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *Int. J. Testing* 11, 144–177. doi: 10.1080/15305058.2010.534571

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

Linn, R. L., and Slinde, J. A. (1977). The determination of the significance of change between pre-and posttesting periods. *Rev. Educ. Res.* 47, 121–150. doi: 10.3102/00346543047001121

Madison, M., and Bradshaw, L. (2018a). Assessing growth in a diagnostic classification model framework. *Psychometrika* 83, 963–990. doi: 10.1007/s11336-018-9638-5

Madison, M., and Bradshaw, L. (2018b). Evaluating intervention effects in a diagnostic classification model framework. *J. Educ. Meas.* 55, 32–51. doi: 10.1111/jedm.12162

Meredith, W., and Tisak, J. (1984). "Statistical considerations in Tuckerizing curves with emphasis on growth curves and cohort sequential analysis," in *Annual Meeting of the Psychometric Society*.

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika* 55, 107–122. doi: 10.1007/BF02294746

Park, Y. S., and Lee, Y.-S. (2014). An extension of the DINA model using covariates: examining factors affecting response probability and latent classification. *Appl. Psychol. Meas.* 38, 376–390. doi: 10.1177/0146621614523830

Park, Y. S., and Lee, Y.-S. (2019). "Explanatory cognitive diagnostic models," in *Handbook of Diagnostic Classification Models*, eds M. von Davier, and Y. S. Lee, (Berlin: Springer), 207–222. doi: 10.1007/978-3-030-05584-4_10

Park, Y. S., Xing, K., and Lee, Y.-S. (2017). Explanatory cognitive diagnostic models: incorporating latent and observed predictors. *Appl. Psychol. Meas.* 42, 376–392. doi: 10.1177/0146621617738012

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* 14, 1–17. doi: 10.2307/2527726

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

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *J. Educ. Stat.* 10, 55–73. doi: 10.3102/10769986010001055

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

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika* 23, 19–23. doi: 10.1007/BF02288975

Vermunt, J. K., and Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Res. Rep. Series* 2005, i-35.

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

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Res. Rep. Series* 2014, 1–13. doi: 10.1002/ets2.12043

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *J. Educ. Behav. Stat.* 43, 57–87. doi: 10.3102/1076998617719727

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

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

APPENDIX

TABLE A1 | The Q-matrix for the real-world data analysis.

	Attribute			
Item	α_1	α_2	α_3	α_4
1	1	0	0	0
2	0	1	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	1	0
6	0	0	1	0
7	0	0	1	0
8	0	0	1	0
9	0	1	0	0
10	0	1	0	0
11	0	1	0	0
12	0	1	0	0
13	1	0	0	0
14	1	0	0	0
15	0	0	0	1
16	0	0	0	1
17	1	0	0	0
18	0	0	0	1
19	0	0	0	1
20	0	0	0	1
21	0	0	0	1

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 September 2020
doi: 10.3389/fpsyg.2020.569852

Developing a Learning Progression for Probability Based on the GDINA Model in China

*Shengnan Bai**

School of Mathematics and Statistics, Northeast Normal University, Changchun, China

This research focuses on developing a learning progression of probability for middle school students, and it applies the GDINA model in cognitive diagnosis models to data analysis. GDINA model analysis firstly extracted nine cognitive attributes and constructed their attribute hierarchy and the hypothesized learning progression according to previous studies, curriculum standards, and textbooks. Then the cognitive diagnostic test was developed based on Q-matrix theory. Finally, we used the GDINA model to analyze a sample of 1624 Chinese middle school students' item response patterns to identify their attribute master patterns, verify and modify the hypothesized learning progression. The results show that, first of all, the psychometric quality of the measurement instrument is good. Secondly, the hypothesized learning progression is basically reasonable and modified according to the attribute mastery probability. The results also show that the level of probabilistic thinking of middle school students is improving steadily. However, the students in grade 8 are slightly regressive. These results demonstrate the feasibility and superiority of using cognitive diagnosis models to develop a learning progression.

Keywords: probability, learning progression, GDINA model, attribute hierarchy, learning pathway

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Yutong Wang,
National Institute of Education Sciences, China
Jiahui Zhang,
Beijing Normal University, China

***Correspondence:**
Shengnan Bai
baisn012@nenu.edu.cn

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

Received: *05 June 2020*
Accepted: *04 September 2020*
Published: *23 September 2020*

Citation:
Bai S (2020) Developing a Learning Progression for Probability Based on the GDINA Model in China. Front. Psychol. 11:569852. doi: 10.3389/fpsyg.2020.569852

INTRODUCTION

Learning progression is defined as 'descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time (e.g., 3–5 years)' (National Research Council, 2007). Although different perspectives of concern formed different definitions of learning progression (Catley et al., 2005; Duncan and Cindy, 2009; Mohan et al., 2009), they all focus on the study of core knowledge to investigate students' cognitive development process. It seems that learning progression is an important channel for the dialogue among theoretical researchers, curriculum planners, educational decision-makers and exam examiners, a bridge between learning research and classroom teaching, and a tool with the most potential to connect curriculum standards, teaching and evaluation and promote the consistency of the three.

Quantitative analysis plays an essential part in developing a learning progression. The initial research on learning progression was built on descriptive statistical results. At present, the most effective and widely used method is Rasch measurement theory (Liu and Collard, 2005; Liu and McKeough, 2005; Johnson, 2013; Todd and Romine, 2016), which estimates the item difficulty parameter as the same level as the students' ability parameter (Rasch, 1960/1980). Rasch analysis assumes unidimensionality, that is, a single trait affects the responses of the participants (Wilson, 2005; Chen et al., 2017a). However, because the core concept covers a wide range of attributes, it is

difficult to strictly satisfied the unidimensionality assumption in practice. The learning progression constructed by the above two quantitative analysis methods is a linear step-by-step development process of students as they increase in grade or as time goes by, and the ability level of students is estimated mainly through the total score of the test.

Since the core concepts are directly related to the internal logical structure of the discipline, they are not all linear, so students can understand core concepts through different learning pathways (Alonzo and Steedle, 2008). In recent years, the research on the learning progression of core concepts has been integrated into the process of individual cognitive structure gradually becoming complete. Since the beginning of the last century, psychometrics and cognitive psychology have been increasingly dissatisfied with assessing the ability level of the individual from macro perspective, so a new generation of psychometrics theory has developed a cognitive diagnosis model for the purpose of diagnosing students' cognitive process, processing skills or knowledge structure. Therefore, researchers began to use it as a quantitative analysis method to provide technical support for the construction of learning progression evaluation system, so as to deeply evaluate student' knowledge structure (Derek and Alonzo, 2012; Chen et al., 2017b; Gao et al., 2017).

Compared with traditional methods, cognitive diagnosis models have the following advantages. First, cognitive diagnosis models directly integrate cognitive variables to estimate the attribute mastery pattern (AMP) of each student, thus realizing the measurement and evaluation of individual's cognitive level from the micro perspective. Second, the attributes that students have and have not mastered can be identified from their responses to the test items. These attributes are distributed at different levels of learning progression, which helps to verify and modify the hypothesized learning progression. Third, it is beneficial to promote personalized education. Each level of learning progression based on the cognitive diagnosis models has multiple AMPs, that is, there are multiple learning pathways from the low level to high level, so as to provide targeted teaching according to the individual student's AMP. Generalized Deterministic Inputs, Noisy and Gate (GDINA) model (de la Torre, 2011), as a saturated cognitive diagnosis model, breaks through the assumptions of the previous simplified cognitive diagnosis models on attribute action mechanism, making the model more flexible and widely used. Whereas, there are few studies have been done on learning progression based on the GDINA model.

As one of the most basic core qualities throughout the mathematics curriculum, probability literacy has now become an indispensable quality for every citizen to enter the society (Scheaffer, 1984; Biehler, 1994; Aitken, 2009). However, studies have repeatedly shown that students always have different degrees of cognitive difficulties in the development of probabilistic thinking. Jones et al. (1997, 1999) proposed a framework to describe students' cognition of probability, in which students' understanding of probability concepts is divided into subjective level, transitional level, informal quantitative level and numerical level. English, Fischbein and Lecoutre found that students cannot naturally understand the sample space, because the basic results in different orders should be distinguished and counted as different results (Fischbein and Gazit, 1984; Lecoutre et al., 1990; English, 1993). Whereas, further analysis shows that although previous research on probability investigated all knowledge points, they did not pay enough attention to the core knowledge. Thus, the introduction of learning progression provides a new research perspective for probability.

As shown in the above literature review, from the perspective of students, there are many stubborn misunderstandings and preconceptions in the learning of probability concepts (Green, 1982; Fischbein and Gazit, 1984; Fischbein et al., 1991; Williams and Amir, 1995; Moritz et al., 1996; Potyka and Thimm, 2015). However, Liu and Thompson's research provided a rich description of the kinds of difficulties experienced by teachers in developing coherent and powerful understandings of probability (Liu and Thompson, 2007). From the perspective of empirical research, the existing studies on learning progression ignored the establishment of a cognitive model, so the probabilistic cognitive structure of individual students cannot be systematically described. Additionally, the nature of cognitive diagnosis and learning progression is very consistent, so using it as a measurement tool to construct the learning progression of probability is well worth further exploration.

To address the issues already outlined and to begin to fill the gaps in the previous research, the present study attempts to: (a) judge whether developed measurement instrument is appropriate to evaluate learning progression of students' probability; (b) verify and modify the hypothesized learning progression by the results of the GDINA model analysis; (c) identify what levels of students' AMPs and provide proper learning pathways accordingly.

HYPOTHESIZED LEARNING PROGRESSION

The current course distribution of probability concept is as follows: intuitive perception of probability concepts through experiments, games and other activities is arranged in grades 4 to 6. The systematic study of preliminary probability is set in grade 9, which is mainly the teaching of one-dimensional probability concepts. Further probability knowledge is arranged in grade 11. The curriculum goals are to deeply learn two-dimensional probability concepts and to preliminarily understand relevant probability concepts of finite dimensions. On this basis, the attribute selection, the attribute hierarchy and the hypothesized learning progression are studied one by one.

Attribute Selection

According to the basic process of cognitive diagnosis, the cognitive attributes contained in probability should be extracted first (Tatsuoka, 2009; de la Torre, 2011; Basokcu, 2014; Rupp and van Rijn, 2018). The most common probability concepts in previous research were the following: randomness, sample space, probability of an event, probability comparisons (Fischbein, 1975; Biggs and Collis, 1982; Liu and Zhang, 1985; Jones et al., 1997, 1999; Li, 2003; Piaget and Inhelder, 2014; He and Gong, 2017).

Other studies have also explored students' ability to make probability estimation (Acredolo et al., 1989). However, the components mentioned above do not explicitly indicate the impact of dimensions.

Taking into account previous studies, curriculum standards and textbooks, students' understanding of one-dimensional probability concepts and two-dimensional probability concepts is not synchronized (Liu and Zhang, 1985; Jones et al., 1997; Li, 2003). Hence, when identifying cognitive attributes, the probability was not only divided into randomness, sample space, probability of an event, probability comparisons and probability estimation, but also the effect of dimension was considered. Consequently, we obtained the nine cognitive attributes of probability as follows.

A1: Randomness: distinguish between certain events, random events, and impossible events.
A2: One-dimensional sample space: list all possible outcomes of a one-dimensional event.
A3: Two-dimensional sample space: list all possible outcomes of a two-dimensional event.
A4: One-dimensional probability comparisons: compare the probability of one-dimensional events.
A5: Two-dimensional probability comparisons: compare the probability of two-dimensional events.
A6: Probability of a one-dimensional event: calculate the probability of a one-dimensional event by definition.
A7: Probability of a two-dimensional event: calculate the probability of a two-dimensional event by definition.
A8: Probability estimation of a one-dimensional event: estimate the probability of a one-dimensional event by frequency.
A9: Probability estimation of a two-dimensional event: estimate the probability of a two-dimensional event by frequency.

Attribute Hierarchy

On the basis of attributes selected before, the attribute hierarchy was constructed by considering previous studies and the curricular sequences of the relevant probabilistic content in the curriculum standards and textbooks. Some studies suggested that the understanding of randomness is the starting point for probabilistic thinking, and this ability increases with age (Williams and Amir, 1995; Chan, 1997; Jones et al., 1997). This indicates that randomness is a precondition of sample space and probability estimation. Furthermore, the understanding of sample space is central to understanding probability (Van de Walle et al., 2016). He and Gong (2017) found that students aged 6 to 14 must master the sample space in order to perform well in calculating the probability of an event by definition. Zhang's team demonstrated that students' understanding of the sample space is superior than probability comparisons (Zhang et al., 1985). This indicates that sample space is a premise of probability of an event and probability comparisons.

Considering the impact of dimensions on students' understanding of probability, students who can consistently list all possible outcomes of a one-dimensional event were often inconsistent or unsystematic in listing all possible outcomes of a two-dimensional event (Liu and Zhang, 1985; Jones et al., 1997; He and Gong, 2017). Moreover, probability estimation is an intuitive way to understand the probability of an event through a large number of repeated experiments. Chapin et al. (2003) argued that students in grades 3 to 5 can initially understand the relationship between the frequency and probability of a one-dimensional event. However, interviews with middle school teachers revealed that students also made some errors in estimating the probability of a two-dimensional event, indicating that probability estimation of a one-dimensional event is the prerequisite of probability estimation of a two-dimensional event. As such, the attribute hierarchy was constructed (**Figure 1**). Thereafter, the attribute hierarchy was tested through mathematics curriculum standards, mathematics textbooks, and interviews with teachers. Attribute hierarchy was found to be basically consistent with the curricular sequences and instructional sequences of the related mathematical topics.

Hypothesized Learning Progression

In light of the above analysis, we developed the hypothesized learning progression of probability for middle school students relied on the previous studies, curriculum standards and textbooks. Considering the influence brought by the dimensions, students' understanding of probability was investigated from five aspects: randomness, sample space, probability of an event, probability comparisons and probability estimation. We then used the SOLO (Structure of the Observed Learning Outcome) taxonomy that developed from Piaget's cognitive development phase theory to clarify the learning progression levels (Biggs and Collis, 1982, 1991).

In the hypothesized learning progression of probability (see **Table 1**), Level 1 does not involve any attributes of probability, indicating that the probabilistic thinking of students at Level 1

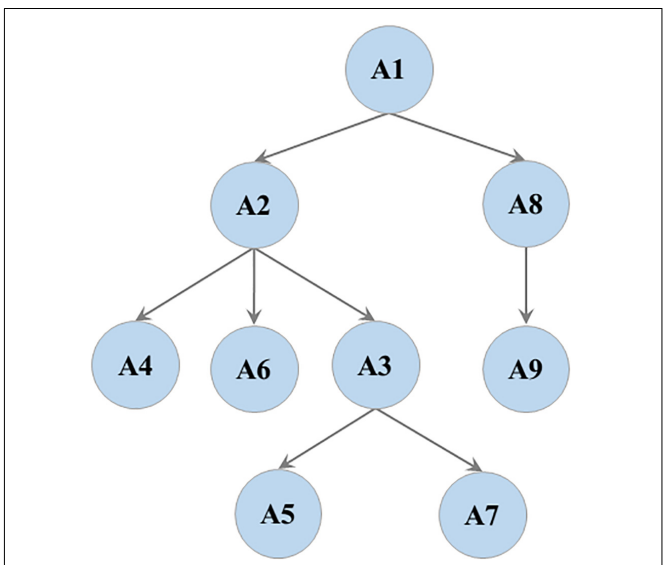


FIGURE 1 | The attribute hierarchy.

TABLE 1 | Hypothesized learning progression of probability.

Level	Content	Attributes
1	Students cannot master any attributes related to probability.	None
2	Students begin to understand the one-dimensional probability concepts, but they cannot transfer their understanding of one-dimensional probability concepts to two-dimensional probability concepts.	At least one of A1, A2, A4, A6, and A8
3	Students can perform two-dimensional sample space and probability estimation of a two-dimensional event.	Further master at least one of A3 and A9
4	Students can understand two-dimensional probability comparisons and probability of a two-dimensional event. Furthermore, they can build a connection between one-dimensional probability concepts and two-dimensional probability concepts.	Further master A5 and A7

has not yet begun to develop. When the students reach Level 2, students begin to understand the one-dimensional probability concepts, indicating that they have mastered at least one of A1, A2, A4, A6, and A8. On the basis of Level 2, students at Level 3 can perform two-dimensional sample space and probability estimation of a two-dimensional event, indicating that students' mastery of A3 and A9. At last, when students reach Level 4, they can understand two-dimensional probability comparisons and probability of a two-dimensional event. This indicates that students have mastered all attributes of probability. So far, we established the correspondence between the hypothesized learning progression levels and the attributes of probability, which will help to verify and modify the hypothesized learning progression through the analysis of GDINA model.

MATERIALS AND METHODS

Item Design

The Q-matrix (**Table 2**), which is established by the selected attributes and their attribute hierarchy, presented the correspondence between each item and each attribute and was used to guide the item design. Q-matrix is based on the design principles proposed by Tu et al. (2012). The first is that the item assessment patterns should include the Reachability Matrix. The second is that each attribute is measured no less than three times. In the Q-matrix, '1' means that the attribute is measured in this item, and '0' means that the attribute is not measured in this item. For instance, '100000000' means that item 1, item 2, item 3, and item 4 only measure A1, and '100000011' means that item 24, item 25 and item 26 measure A1, A8, and A9.

Five mathematics teachers, two subject experts, two mathematics educators and two psychometricians were invited to develop the instrument. The mathematics teachers came from key middle schools in Fujian, Shanxi, Henan and Inner Mongolia, as well as a teaching and research staff from Qinghai Province. The subject experts consisted of a professor and an associate professor who study probability and statistics. The mathematics educators were composed of two professors engaged in mathematics education research. The psychometricians were comprised of a professor and a Ph.D. candidate who do research in psychometrics.

TABLE 2 | Q-matrix of attributes and items.

	Attribute								
Item	A1	A2	A3	A4	A5	A6	A7	A8	A9
1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0
6	1	1	0	1	0	0	0	0	0
7	1	1	0	0	0	0	0	0	0
8	1	1	0	1	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0
10	1	1	0	1	0	0	0	0	0
11	1	1	0	0	0	1	0	0	0
12	1	1	0	0	0	1	0	0	0
13	1	1	0	0	0	1	0	0	0
14	1	1	1	0	0	0	0	0	0
15	1	1	1	0	1	0	0	0	0
16	1	1	1	0	0	0	0	0	0
17	1	1	1	0	1	0	0	0	0
18	1	1	1	0	0	0	0	0	0
19	1	1	1	0	1	0	0	0	0
20	1	1	1	0	0	0	1	0	0
21	1	1	1	0	0	0	1	0	0
22	1	1	1	0	0	0	1	0	0
23	1	0	0	0	0	0	0	1	0
24	1	0	0	0	0	0	0	1	1
25	1	0	0	0	0	0	0	1	1
26	1	0	0	0	0	0	0	1	1

Based on the Q-matrix, curriculum standards and textbooks, the study developed a cognitive diagnostic test of probability, which consists of 26 items and each item corresponds to a specific item assessment pattern (IAP). All items are in multiple-choice or short-answer format. All items are dichotomous, with the correct score of '1' and the wrong score of '0.' **Table 3** presents some example items and their corresponding IAPs.

Participants and Procedure

According to the level of economic development, mainland China can be divided into four types: the most developed areas, the developed areas, the moderately developed areas and the underdeveloped areas. Since the moderately developed areas cover 23 provinces and cities, accounting for a large part of the mainland (Xie and Lu, 2011), the schools and corresponding students in these areas were selected in this study.

In the end, six junior high schools and five high schools were selected from the moderately developed areas. A total of 1624 students participated in this study (**Table 4**). To ease the tension of the students during the test, we informed them that their test results will not affect their academic rankings this semester. The time allocated to the test was 40 min.

TABLE 3 | Example items from the probability test.

Item number	Content	IAP
1	Roll a fair dice and the number rolled is greater than 6. Please determine the type of this event. (A) Certain (B) Random (C) Impossible	100000000
7	Randomly select a number from the set {1, 2, 3, 4, 5}. Please write out how many possible outcomes there are.	110000000
18	Roll two fair dice and observe the number on the up side. Please list all possible outcomes of the numbers rolled by the two dice.	111000000

TABLE 4 | Structure description of the sample.

Grade	7	8	9	10	11
Total number	323	333	302	354	312

Data Analysis

Instrument Functioning

Parscale 4.1, R and SPSS 22.0 were used to investigate the psychometric quality of the developed measurement instrument. First, the rationality of attribute selection and the attribute hierarchy should be attested. Specifically, we performed a linear regression analysis to see if the attributes measured by the item can predict the item difficulty level. We used the hierarchy consistency index (HCI) to measure the degree of matching between the actual item response pattern (IRP) and the expected response pattern under the attribute hierarchy. Second, the test reliability and test validity should be explored. Attribute test–retest reliability was used as the test reliability measure under Cognitive Diagnosis Theory, indicating the internal consistency of each attribute (Templin and Bradshaw, 2013). As for test validity, since our study used a cognitive diagnosis model, the identifiability of the Q-matrix was used as evidence of the test validity. Third, the quality of each item should be explored. This includes the examination of item fitting index, item difficulty and item discrimination. In addition, students with abnormal responses were identified and analyzed by participant fitting index.

GDINA Model Analysis

In cognitive diagnosis assessment, the ability of each student is presented as AMP (attribute master pattern). Attribute refers to the knowledge, skills and strategies required for a student to correctly complete a test item. AMP is a description of whether a student has mastered each attribute. Where, '1' means that the attribute is mastered, and '0' means that the attribute is not mastered.

The GDINA model was used to classify students into different AMPs represented by the observed IRPs. First, the rationality of attribute selection and attribute hierarchy should be verified. Then, the identifiability of Q-matrix and the psychometric quality of the cognitive diagnostic test must be judged. Finally, student's AMP was estimated from his or her IRP through the classical estimation method. Ideally, a student should only correctly answer items that measure the attributes he or she mastered, and incorrectly answer items that measure at least one attribute that he or she did not master. For more information about the GDINA model estimation program, please refer to de la Torre (2011). The above analysis was performed using the GDINA model program in the R package (CDM package). The item response function of the GDINA model is as follows:

$$P(X_{ij}=1|\alpha_{lj}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^{*}} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} + \cdots + \delta_{j12\cdots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk}$$

The function above can be decomposed into the sum of the effects due to the presence of specific attributes and all their possible interactions. δ_{j0} is the intercept of item j, called the baseline probability, that is, the probability that the participant answers the item correctly without mastering all the attributes measured by this item. The value is a non-negative value and can be regarded as the guessing parameter. δ_{jk} is the main effect of attribute k on item j, which is generally a non-negative value. It represents the effect of increasing the probability of answering this item correctly because the participant has mastered the attribute k. The larger the value, the greater the contribution of mastering the attribute to the correct item j. $\delta_{jkk'}$ is the interaction effect of attribute k and attribute k' on item j. $\delta_{j12\cdots K_j^{*}}$ measures the interaction effect between all attributes for item j.

Learning Progression Verification and Modification

Due to the correspondence between the hypothesized learning progression levels and the attributes contained in probability presented in **Table 1**, the attribute mastery probability analyzed by the GDINA model was used to verify and modify the hypothesized learning progression. Students are expected to develop a successively more sophisticated understanding of probability based on the hypothesized learning progression levels.

First, students will master the attributes regarding the one-dimensional probability concepts. Then, students will enter the initial stage of two-dimensional probabilistic thinking, that is, they will continue to learn the sample space and compare the probability of two-dimensional events. It ends with students being able to build a connection between one-dimensional probability concepts and two-dimensional probability concepts. If the hypothesized learning progression is reasonable, the attributes at higher levels are generally more difficult to master than the attributes at lower levels.

RESULTS

Instrument Functioning

In this study, a cognitive diagnostic test was developed under the guidance of the GDINA model. Tu et al. (2012) suggested to first attest the rationality of the attribute selection and the attribute hierarchy. For attribute selection, the result of linear regression analysis with the item difficulty as the dependent variable and the

columns of the Q-matrix as the independent variables shows that the adjusted R^2 value is 0.875. This means the explanatory power of the selected attributes to the item difficulty is 87.5%, which verifies the attribute selection. For attribute hierarchy, Cui and Leighton (2009) proposed that it is feasible to use HCI index to test the rationality of attribute hierarchy. Wang and Gierl (2007) pointed out that if the mean value of HCI index is greater than 0.6, the attribute hierarchy has good rationality. Based on the current data, the mean value of HCI index is 0.90, which proves that the attribute hierarchy is reasonable.

Regarding the quality of cognitive diagnostic test, the reliability and validity needs to be checked. Based on attribute test–retest reliability, the internal consistency value of each attribute ranges from 0.88 to 0.99, indicating that each attribute has good reliability. Then the test was prepared according to the design principles of Q-matrix proposed by Tu et al. (2012), which can confirm the validity of the test.

As for the quality of each item, the item fitting index RMSEA for all items is less than 0.08, with an average of 0.03, indicating that each item has a good fit to GDINA model. The item difficulty index under CTT shows that the difficulty value of most items is between 0.37 and 0.84, with an average of 0.62, and only seven items have difficulty values higher than 0.84. The estimation of item difficulty under IRT shows that the difficulty range is between −3.41 and 0.95. As for the item discrimination, when the discrimination is greater than 0.4, the item is considered excellent (Ray and Margaret, 2003; Tu et al., 2019), and all items meet the standard.

According to the participant fitting index, if the index is greater than −2, the participant's response is in good agreement with the model. In this study, 94.6% of the students' responses have a good fit.

Learning Progression Verification and Modification

GDINA Model Analysis

The results of the GDINA model analysis show that 1624 students are classified into 34 AMPs (**Table 5**). All students mainly concentrated in the following six AMPs: AMP 1, 4, 11, 13, 31, and 34.

Further analysis of these AMPs reveals that 94.6% of students can develop a perception of randomness because they have mastered A1. 93.11% of students are able to list all possible outcomes of a one-dimensional event due to their proficiency in A1 and A2. 84.62% of students can calculate the probability of a one-dimensional event, in view of their mastery of A1, A2, and A4. 84.56% of students know how to compare the probability of one-dimensional events, which stems from their mastery of A1, A2, and A6. 63.36% of students can estimate the probability of a one-dimensional event because of their mastery of A1 and A8. 61.71% of students can form good one-dimensional probabilistic thinking as they have mastered A1, A2, A4, A6, and A8.

By shifting the discussion of students' probabilistic thinking from one-dimensional to two-dimensional, 60.91% of students can build a connection from one-dimensional to two-dimensional on the probability estimation (A8, A9), with a slightly reduced proportion of the latter. The percentage of students who can migrate from one-dimensional sample space (A2) to two-dimensional sample space (A3) drops significantly to 58.19%. The number of students able to progress from one-dimensional probability comparisons (A4) to two-dimensional probability comparisons (A5) decreases from 84.62 to 50.06%. The proportion of students who can calculate the probability of a two-dimensional event by definition (A7) is 40.16%. However, only 30.79% of students have mature probabilistic thinking (A1–A9). In summary, we can find that middle school students have basically formed a good one-dimensional probabilistic thinking, but the development of students' two-dimensional probabilistic thinking is not optimistic.

In terms of the classification of students in each grade, students in grade 7 are mainly concentrated in AMP 1, 4, 6, 11, 13, 29, 31 and 34, which indicates that they have a good mastery of A1, A2, A4, and A6. Students in grade 8 are mainly concentrated in AMP 1, 2, 4, 6, 9, 11, 13, 31 and 34, and the AMP 34 showed that students in grade 7 had more percentage than grade 8. This phenomenon may be due to the fact that after learning probability concepts in primary school, students in grade 8 have not been exposed to probability concepts for a longer period of time than students in grade 7, so their performance is somewhat backward. Students in grade 9 are mainly concentrated in AMP 11, 13, 31, 32 and 34, indicating that they have further mastered A3, A8, and A9 on the basis of grades 7 and 8. Students in grades 10 are mainly concentrated in AMP 11, 13, 25, 31, 32 and 34, which means that they have made great progress in probability estimation (A8, A9). Students in grades 11 are mainly concentrated in AMP 13, 25, 31 and 34, which shows that they can master almost all the attributes, and the proportion of students with mature probabilistic thinking increases from 7.43 to 58.65%.

Learning Progression Verification and Modification Process

The attribute mastery probability, which can be estimated by GDINA model analysis, is used to verify the hypothesized learning progression. If the hypothesized learning progression can truly reflect the development of students' probabilistic thinking, the attribute mastery probability should be directly affected by the level of the attributes in hypothesized learning progression. That is, the attributes at a higher level should be more difficult to master, while the attributes at a lower level should be easier to master.

The GDINA model analysis shows that the order of attribute mastery probability from high to low is A1 (0.94), A2 (0.93), A6 (0.85), A4 (0.84), A8 (0.63), A9 (0.61), A3 (0.58), A5 (0.50), A7 (0.41). According to this result, in the one-dimensional probability concepts, A1, A2, A4, and A6 are relatively easy to master, except for the probability estimation of a one-dimensional event (A8). Then, A3, A8, and A9 are at a moderate difficulty level. Moreover, A5 and A7 are more difficult to master. This indicates that the attribute mastery probability levels are basically consistent with the hypothesized learning progression levels. However, the attribute mastery probability of A8 at the Level 2 of the hypothesized learning progression is lower than expected.

TABLE 5 | Classification of students' AMPs.

	AMP	Total (%)	Grade 7 (%)	Grade 8 (%)	Grade 9 (%)	Grade 10 (%)	Grade 11 (%)
1	000000000	5.54	11.46	13.21	2.98	0.00	0.00
2	100000000	1.29	1.86	4.20	0.33	0.00	0.00
3	100000010	0.06	0.31	0.00	0.00	0.00	0.00
4	110000000	4.86	8.36	14.11	1.32	0.28	0.00
5	110000011	0.06	0.00	0.30	0.00	0.00	0.00
6	110001000	2.22	4.64	5.11	0.99	0.28	0.00
7	110001010	0.12	0.00	0.60	0.00	0.00	0.00
8	110001011	0.62	1.24	1.20	0.33	0.28	0.00
9	110100000	2.65	3.41	6.61	2.98	0.28	0.00
10	110100011	0.25	0.62	0.60	0.00	0.00	0.00
11	110101000	9.91	15.48	13.51	15.56	4.80	0.64
12	110101010	0.74	1.86	1.20	0.66	0.00	0.00
13	110101011	13.49	18.89	14.71	6.95	14.12	12.18
14	111000000	0.06	0.31	0.00	0.00	0.00	0.00
15	111001000	0.31	0.31	0.30	0.33	0.28	0.32
16	111001111	0.06	0.00	0.00	0.00	0.28	0.00
17	111010000	0.06	0.00	0.30	0.00	0.00	0.00
18	111011011	0.12	0.31	0.00	0.00	0.28	0.00
19	111100000	0.06	0.00	0.00	0.33	0.00	0.00
20	111101000	1.54	1.86	1.80	2.65	1.13	0.32
21	111101010	0.06	0.00	0.00	0.00	0.00	0.32
22	111101011	1.29	1.55	0.30	1.66	1.98	0.96
23	111101100	0.74	0.62	0.90	1.66	0.28	0.32
24	111101110	0.25	0.62	0.30	0.00	0.00	0.32
25	111101111	3.76	0.31	0.90	1.32	8.19	7.69
26	111110000	0.18	0.62	0.00	0.00	0.28	0.00
27	111110011	0.12	0.00	0.00	0.66	0.00	0.00
28	111110111	0.25	0.00	0.00	0.00	0.85	0.32
29	111111000	3.76	7.74	3.90	3.31	2.54	1.28
30	111111010	0.37	0.00	0.00	0.33	1.13	0.32
31	111111011	10.10	8.36	6.01	7.95	14.97	12.82
32	111111100	3.45	1.55	2.40	6.62	4.24	2.56
33	111111110	0.86	0.31	0.60	1.32	1.13	0.96
34	111111111	30.79	7.43	6.91	39.74	42.37	58.65

So as to modify the hypothesized learning progression, we first determine the transition points of the hypothesized learning progression levels from the perspective of the attribute mastery probability. According to the GDINA model analysis, the mastery probability of all attributes is within the range of 0.4 to 0.95. Meanwhile, 5.5% of the students still cannot master any attributes related to probability (see **Table 5**). Therefore, the Level 1 of learning progression is set so that students cannot master any attributes. Next, we divide 0.4 to 0.95 into three parts, and each part corresponds to a learning progression level. Based on the perspective of attribute mastery probability, the modified learning progression is presented in **Table 6**.

Students' Understanding Levels of Probability

Regarding the modified learning progression, middle school students are classified into Level 2, Level 3 and Level 4, with more students at Level 3 and Level 4. This implies that the one-dimensional probabilistic thinking of middle school students is basically mature, and the development of two-dimensional probabilistic thinking (A3, A5, A7, A9) is relatively slow, which is consistent with the three stages of the probabilistic cognitive development proposed by Piaget and Inhelder (2014).

As for the learning progression levels of students in different grades, students in grade 7 are classified into four levels on average, with more students at Level 2 and Level 3, but still 11.46% of the students cannot master any attributes. This implies that although the vast majority of students can recognize the concept of randomness, there are still a few students in the embryonic stage of probabilistic thinking, which confirms the previous research conclusions (Moritz et al., 1996; Chan, 1997; He and Gong, 2017).

Students in grade 8 are mainly classified into Level 2, and averaged at the other three levels. However, there are more students at Level 1 and Level 2 than in grade 7, suggesting that the probabilistic thinking levels of students in grade 8 are slightly degraded compared with grade 7. This may be because the teaching of probability concepts is mainly set in grade 9, while students in grade 8 have not learned the probability concepts

TABLE 6 | Modified learning progression of probability.

Level	Content	Attributes
1	Students cannot understand any attributes related to probability.	None
2	Students begin to understand the four one-dimensional probability concepts (randomness, sample space, probability of an event and probability comparisons). But they cannot perform well in probability estimation of a one-dimensional event, and they cannot transfer their understanding of probability concepts from one-dimensional to two-dimensional.	At least one of A1, A2, A4, and A6
3	Students can perform probability estimation and two-dimensional sample space. And they cannot integrate all the two-dimensional probability concepts.	Further master at least one of A3, A8, and A9
4	Students can understand two-dimensional probability comparisons and probability of a two-dimensional event. Furthermore, they can build a connection between one-dimensional probability concepts and two-dimensional probability concepts.	Further master A5 and A7

for a long time, which leads to their backward thinking. This suggests that the teaching of probability should be properly arranged for each grade.

Students in grade 9 are mainly classified into Level 3 and Level 4, which indicates that they have basically mastered all one-dimensional probability concepts, and the two-dimensional probabilistic thinking is also developing steadily. Students in grades 10 and 11 are mainly classified into Level 4, with 42.37 and 58.65% of them reaching the AMP (111111111), which implies that about half of them have not yet formed a mature probabilistic thinking. That is to say, although students have mastered the two-dimensional sample space, they are unable to effectively establish a connection between two-dimensional probability concepts, which makes the previous studies confirmed again (Liu and Zhang, 1985; Li, 2003; He and Gong, 2017).

In short, middle school students develop a successively more sophisticated understanding of the concepts involved in the learning progression levels, but the reasons for the decline of students in grade 8 still need to be further explored.

DISCUSSION

The current study aimed to develop a learning progression for probability. To this end, we built a new measurement instrument based on cognitive diagnosis theory for data collection and data analysis. The findings will be discussed from the learning progression for probability, the types probability AMPs for students, learning progression verification and modification and practical implications.

Learning Progression for Probability

The learning progression for probability, based on the cognitive diagnosis theory, is presented in **Table 6**. Students at Level 1 cannot master any attributes. In particular, 5.5% of the students are at Level 1, and the number of students at Level 1 decline as the grade progresses. This confirms the research conclusion of He and Gong, there are still a small number of middle school students who do not understand the concepts related to probability (He and Gong, 2017). Thus, although the curriculum of junior high school should be spiraling upward in primary school, those students whose probabilistic thinking has not yet sprouted cannot be completely ignored.

Students at Level 2 can master at least one of one-dimensional probability concepts, with the exception of probability estimation, which combines the Uni-Structural level with the Multi-Structural level in Li's research (Li, 2003). By contrast, contemporary students have made progress in probability, suggesting that formal teaching in the early stage of secondary school has achieved good results. It is worth noting that the probability estimation found by Li's research is out of step with the development of other one-dimensional probability concepts (Li, 2003). This may be due to the dispersion of the probability content in the junior high school. Some suggestions are also put forward for the classroom setting, which implies that we should pay more attention to the cultivation of probability estimation thoughts, and should not ignore the importance of probability estimation as the foundation of statistics learning in the future.

Students at Level 3 can further master probability estimation of a one-dimensional event, two-dimensional sample space and probability estimation of a two-dimensional event. Logically speaking, mastering one-dimensional probability concepts is the prerequisite for continuing to learn two-dimensional probability concepts. Meanwhile, the construction of sample space is the prerequisite for probability calculation and probability comparisons. This is similar to the conclusions of earlier studies by Lecoutre, Fischbein, and English that students cannot naturally understand the sample space, because the basic results in different orders should be distinguished and counted as different results (Lecoutre et al., 1990; Fischbein et al., 1991; English, 1993).

Students at Level 4 can further develop probability of a two-dimensional event and two-dimensional probability comparisons. These two attributes belong to the last stage of probabilistic cognitive development proposed by Piaget and Inhelder — the stage of formal operation, thus verifying the setting of Level 4 (Piaget and Inhelder, 2014). In discipline logic, the sample space is the basis of probability of an event and probability comparisons. However, not all students who mastered the two-dimensional sample space can enter Level 4, and the reasons are worth exploring. Referring to the answers of these students, some students have not formed a stable understanding of the sample space and are in a wandering stage and some students have a lack of calculation formula or calculation ability in the process.

The Types Probability AMPs for Students

There are 34 AMPs for students based on the GDINA model analysis. As can be seen from **Table 6**, students' AMPs for probability can be summarized into two types through the correspondence between learning progression levels and AMPs.

The first type of AMPs is to master all the attributes at a lower level and then develop the attributes at the next level, such students account for 86.89% of the total. For instance, AMP 12 (110101010) indicates that students in this category have mastered A1, A2, A4, A6, and A8, that is, after mastering all the attributes at Level 2, they have developed A8 at Level 3.

It is worth noting that there is another type of AMPs. After developing partial attributes at a low level, students develop the attributes at the next level, reciprocating cycles, and finally forming mature probabilistic thinking. Such students account for 13.11% of the total, mainly at Level 3 and Level 4. Looking at the attribute hierarchy (see **Figure 1**), because the probability estimation A8 and A9 are independent of the other attributes, these two attributes may be the hardest for some students to master. For example, AMP 32 (111111100) indicates that students in this category have mastered A1–A7, but have not mastered A8 and A9, that is, they have mastered all attributes of Level 2 and Level 4, but for Level 3, they have only mastered A3, and we believe that they have reached Level 4.

To sum up, most students develop probabilistic thinking in a spiraling manner, while a few develop probabilistic thinking by learning each knowledge point independently. This result indicates that the curriculum, teaching and evaluation should attach importance to the cultivation and investigation of core knowledge and ability, and further thinking is still needed on how to form a good cognitive structure for students around core knowledge.

Learning Progression Verification and Modification

As the levels of learning progression correspond to the attributes contained in probability, the results of GDINA model analysis are used to verify and modify the hypothesized learning progression. On the one hand, the order of attribute mastery probability is basically consistent with the levels of hypothesized learning progression, that is, the attributes at a low level are easier to master, while the attributes at a high level are more difficult to master. However, the attribute mastery probability of probability estimation of a one-dimensional event is lower than expected, which implies that students who can perform well on other one-dimensional probability concepts (A1, A2, A4, and A6) still perform poorly on probability estimation of a one-dimensional event (A8). On the other hand, the AMPs of each grade students can also be used to verify and modify the hypothesized learning progression. The AMPs of students in grades 7 and 8 show that they can perform well in A1, A2, A4 and A6, and slightly worse on A8. The AMPs of students in grade 9 indicate that they have further mastered A3, A8, and A9 on the basis of grades 7 and 8. Students in grades 10 and 11 can master almost all the attributes. The above analysis means that students' understanding of all one-dimensional probability concepts is not completely synchronized in junior high school.

From the attribute hierarchy, the probability estimation of a one-dimensional event (A8), which is an approximation of the probability of an event from the experimental perspective, is independent of A2–A7. Therefore, A8 may be more difficult to master than A1, A2, A4, and A6. This finding may be due to the fact that students in the second learning phase (grades 4 to 6) have already begun the initial study of probability, but the curriculum standards and textbooks for this phase focus on one-dimensional probability concepts and do not formally introduce probability estimation. It is not until the third learning phase (grade 9) that students begin to systematically contact the idea of probability estimation. This result shows that it is unreasonable to put A8 at the Level 2 and adjust it to the Level 3.

Compared with the hypothesized learning progression, the modified learning progression has obvious advantages. From the macro perspective, the modified learning progression combines the experience of subject experts, front-line teachers, and the students' actual learning conditions, which is closer to the development characteristics of students' probabilistic thinking in each grade. From the micro perspective, each student's path from a lower level to a higher level is not unique. Starting from the student's current AMP and taking AMP 34 (111111111) as the learning target, a path can be selected to match the learning progression and attribute hierarchy. This suggests that the learning progression constructed by the GDINA model includes both macroscopic and microscopic observations, which can improve the theoretical nature of teaching decision-making, enhance the operability of teaching practice, and provide possibility for students' self-improvement, so as to promote the integration of curriculum, teaching and evaluation.

Practical Implications

Through the GDINA model analysis, this study used cross-sectional data to construct a learning progression for probability. Although no longitudinal data has been collected for verification, the attribute hierarchy, learning progression and the student's AMP can still be helpful to front-line teachers. Before teaching, the results of this study can provide a more scientific analysis of learning situations for teaching design. After teaching, the cognitive diagnostic test in this study can be used to check the

TABLE 7 | Individual information of three students.

ID	Score	AMP	IRP	Non-mastered attributes and remedy pathway
67	14	110101000	111111111111100001000000000	A3, A5, A7, A8, A9 A8→A9→A3→A5→A7
179	14	110101011	110111101111100000000001101	A3, A5, A7 A3→A5→A7
529	14	110100011	111111111100000000000001111	A3, A5, A6, A7 A6→A3→A5→A7

learning effect of students, thereby providing a plan for teaching review and teaching remedy.

Many researchers pointed out that students' learning pathways are not unique (Baroody et al., 2004), and teachers can find several remedy pathways for students with specific AMP to master all attributes by combining the learning progression and attribute hierarchy. In addition, teachers can gather students with specific AMP, which is more effective. For example, for students with AMP (110100000), the remedy pathway may be A8→A9→A6→A3→A5→A7 or A6→A8→A9→A3→A5→A7. In the first remedy pathway, students will first learn probability estimation, then learn probability of a one-dimensional event, and finally learn the sample space, probability comparison, and probability of a two-dimensional event. In the second remedy pathway, students will first develop a good one-dimensional probabilistic thinking, and then gradually develop a mature two-dimensional probabilistic thinking.

Furthermore, a student's individual diagnostic report, including individual test score, IRP, AMP and non-mastered attributes, can be used to conduct an in-depth analysis of the student's knowledge state and provide personalized remedial suggestions. For example, **Table 7** shows the individual information of three students. Even if they have the same score, they may have different IRPs, AMPs, and remedy pathways. This directly demonstrates the significant advantages of using cognitive diagnosis assessment to develop learning progression.

Limitations and Future Work

Although this study has the above findings and implications, there are still some limitations. First, this study used cross-sectional data to construct a learning progression for probability, but learning progression itself is a developmental concept, so longitudinal data can be collected for more in-depth exploration in the future. In addition, some scholars have recently explored longitudinal cognitive diagnosis theory (Li et al., 2016; Zhan et al., 2019; Zhan, 2020), so longitudinal tracking data can be collected under the guidance of longitudinal cognitive diagnosis theory to build learning progression that can reveal more about the laws of education. Second, the effect of the constructed learning progression is not fully explored in this study, so that future research can use remedy pathways to examine the validity of the cognitive diagnosis results. For example, students can be divided into an experimental group and a control group. Courses and teaching are arranged for the students in the experimental group according to the learning progression, while the students in the control group follow the normal teaching plan. If there is a significant difference in performance between the two groups at the end of the course, we believe that the learning progression is effective. Further exploration can group students with a specific AMP and select different remedy pathways to find the most effective way for these students. In addition, realizing the computerization of students' diagnostic reports and targeted remedial suggestions is also the direction of future development. That is, computer programs need to be programmed to report results automatically, which can help students achieve self-remedy learning.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Northeast Normal University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SB participated in the design of the study, test preparation, data collection and analysis, analysis of the results, and writing the manuscript, and agreed to submit the final version of the manuscript.

ACKNOWLEDGMENTS

I would like to thank the students, teachers, and experts for participating in and assisting with the present research.

REFERENCES

Acredolo, C., O'Connor, J., Banks, L., and Horobin, K. (1989). Children's ability to make probability estimates: skills revealed through application of Anderson's functional measurement methodology. *Child. Dev.* 60, 933–945. doi: 10.2307/1131034

Aitken, C. G. G. (2009). Some thoughts at the interface of law and statistics. *Law. Probab. Risk.* 8, 73–83. doi: 10.1093/lpr/mgp019

Alonzo, A., and Steedle, J. T. (2008). Developing and assessing a force and motion learning progression. *Sci. Educ.* 93, 389–421. doi: 10.1002/sce.20303

Baroody, A. J., Cibulskis, M., Lai, M. L., and Li, X. (2004). Comments on the use of learning trajectories in curriculum development and research. *Math. Think. Learn.* 6, 227–260. doi: 10.1207/s15327833mtl0602_8

Basokcu, T. O. (2014). Classification accuracy effects of Q-Matrix validation and sample size in DINA and G-DINA models. *J. Educ. Pract.* 5, 220–230.

Biehler, R. (1994). Probabilistic thinking, statistical reasoning, and the search for causes—Do we need a probabilistic revolution after we have taught data analysis. *Res. Pap. ICOTS* 4, 20–37.

Biggs, J. B., and Collis, K. F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York, NY: Academic Press.

Biggs, J. B., and Collis, K. F. (1991). "Multimodal learning and the quality of intelligent behavior," in *Intelligence: Reconceptualization and measurement*, ed. H. A. Rowe (Hillsdale: Lawrence Erlbaum), 57–76.

Catley, K., Lehrer, R., and Reiser, B. (2005). *Tracing a Prospective Learning Progression for Developing Understanding of Evolution*. Washington, DC: National Academy of Sciences.

Chan, W. S. (1997). *16 to 18 Year Old Students' Errors and Misconceptions in Learning Probability*. Dissertation, Nanyang Technological University, Singapore.

Chapin, S., Koziol, A., MacPherson, J., and Rezba, C. (2003). *Navigating Through Data Analysis and Probability in Grades 3-5*. London: National Council of Teachers of Mathematics.

Chen, F., Yan, Y., and Xin, T. (2017a). Developing a learning progression for number sense based on the rule space model in China. *Educ. Psychol. U.K.* 37, 128–144. doi: 10.1080/01443410.2016.1239817

Chen, F., Zhang, S., Guo, Y., and Xin, T. (2017b). Applying the rule space model to develop a learning progression for thermochemistry. *Res. Sci. Educ.* 47, 1357–1378. doi: 10.1007/s11165-016-9553-7

Cui, Y., and Leighton, J. P. (2009). The hierarchy consistency index: evaluating person fit for cognitive diagnostic assessment. *J. Educ. Meas.* 46, 429–449. doi: 10.1111/j.1745-3984.2009.00091.x

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

Derek, C. B., and Alonzo, A. C. (2012). "The Psychometric modeling of ordered multiple-choice item response for diagnostic assessment with a learning progression," in *Proceedings of the Learning Progressions in Science Conference*, Iowa City, IA.

Duncan, R. G., and Cindy, E. H. S. (2009). Learning progressions: aligning curriculum, instruction, and assessment. *J. Res. Sci. Teach.* 46, 606–609. doi: 10.1002/tea.20316

English, L. D. (1993). Children's strategies for solving two-and three-dimensional combinatorial problems. *J. Res. Math. Educ.* 24, 255–273. doi: 10.2307/749347

Fischbein, E. (1975). *The Intuitive Sources of Probabilistic Thinking in Children*. Dordrecht: Reidel Publishing Company. doi: 10.1007/978-94-010-1858-6

Fischbein, E., and Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? *Educ. Stud. Math.* 15, 1–24. doi: 10.1007/BF00380436

Fischbein, E., Nello, M. S., and Marino, M. S. (1991). Factors affecting probabilistic judgements in children and adolescents. *Educ. Stud. Math.* 22, 523–549. doi: 10.2307/3482209

Gao, Y., Chen, F., Xin, T., Zhan, P., and Jiang, Y. (2017). Applying psychometric models in learning progressions studies: theory, method and breakthrough. *Adv. Psychol. Sci.* 25, 1623–1630. doi: 10.3724/sp.j.1042.2017.01623

Green, D. R. (1982). *Probability Concepts in 11-16 Year Old Pupils*. Dissertation, Loughborough University of Technology, Loughborough.

He, S. Q., and Gong, Z. K. (2017). Learning Progressions of Probability in Children of 6 to 14 Years. *Curric. Teach. Mater. Method* 37, 61–67.

Johnson, P. (2013). ""How students' understanding of particle theory develops: a learning progression?," in *Concepts of Matter in Science Education*, eds T. Georgios and S. Hannah (Dordrecht: Springer), 47–67. doi: 10.1007/978-94-007-5914-5_3

Jones, G. A., Langrall, C. W., Thornton, C. A., and Mogill, A. T. (1997). A framework for assessing and nurturing young children's thinking in probability. *Educ. Stud. Math.* 32, 101–125. doi: 10.1023/a:1002981520728

Jones, G. A., Langrall, C. W., Thornton, C. A., and Mogill, A. T. (1999). Students' probabilistic thinking in instruction. *J. Res. Math. Educ.* 30, 487–519. doi: 10.2307/749771

Lecoutre, M. P., Durand, J. L., and Cordier, J. (1990). A study of two biases in probabilistic judgments: representativeness and equiprobability. *Adv. Psychol.* 68, 563–575. doi: 10.13140/2.1.4664.8324

Li, F., Cohen, A., Bottge, B., and Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educ. Psychol. Meas.* 76, 181–204. doi: 10.1177/0013164415588946

Li, J. (2003). *Teaching and Learning of Probability in Primary and Secondary Schools*. Shanghai: East China Normal University Press.

Liu, F., and Zhang, Z. J. (1985). *Cognitive Development and Education of Children*. Beijing: People's Education Press.

Liu, X., and Collard, S. (2005). Using the Rasch model to validate stages of understanding the energy concept. *J. Appl. Meas.* 6, 224–241.

Liu, X., and McKeough, A. (2005). Developmental growth in students' concept of energy: analysis of selected items from the TIMSS database. *J. Res. Sci. Teach.* 42, 493–517. doi: 10.1002/tea.20060

Liu, Y., and Thompson, P. (2007). Teachers' understandings of probability. *Cogn. Instruc.* 25, 113–160. doi: 10.2307/27739856

Mohan, L., Chen, J., and Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *J. Res. Sci. Teach.* 46, 675–698. doi: 10.1002/tea.20314

Moritz, J. B., Watson, J. M., and Pereira-Mendoza, L. (1996). "The language of statistical understanding: an investigation in two countries," in *Proceedings of the Annual Conference of the Australian Association for Research in Education*, Camberwell.

National Research Council (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: The National Academies Press.

Piaget, J., and Inhelder, B. (2014). *The Origin of the Idea of Chance in Children (Psychology Revivals)*. London: Psychology Press. doi: 10.4324/9781315766959

Potyka, N., and Thimm, M. (2015). "Probabilistic reasoning with inconsistent beliefs using inconsistency measures," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, Menlo Park, CA.

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: The University of Chicago Press.

Ray, A., and Margaret, W. (eds) (2003). *PISA Programme for international student assessment (PISA) PISA 2000 technical report: PISA 2000 technical report*. Paris: OECD Publishing.

Rupp, A. A., and van Rijn, P. W. (2018). GDINA and CDM packages in R. *Meas. Interdiscip. Res. Perspect.* 16, 71–77. doi: 10.1080/15366367.2018.1437243

Scheaffer, R. L. (1984). "The role of statistics in revitalizing precollege mathematics and science education," in *Proceedings of the Section on Statistical Education*, (Washington, DC: American Statistical Association), 19–21.

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method*. New York: Routledge. doi: 10.4324/9780203883372

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

Todd, A., and Romine, W. L. (2016). Validation of the learning progression-based assessment of modern genetics in a college context. *Int. J. Sci. Educ.* 38, 1673–1698. doi: 10.1080/09500693.2016.1212425

Tu, D. P., Cai, Y., and Ding, S. L. (2012). *Cognitive Diagnosis: Theory, Methods and Applications*. Beijing: Beijing Normal University Press.

Tu, D. P., Cai, Y., Gao, X. L., and Wang, D. X. (2019). *Advanced Cognitive Diagnosis*. Beijing: Beijing Normal University Press.

Van de Walle, J. A., Karp, K., and Bay-Williams, J. M. (2016). *Elementary and Middle School Mathematics: Teaching Developmentally*, 9th Edn. London: Pearson Education Inc.

Wang, C., and Gierl, M. J. (2007). "Investigating the cognitive attributes underlying student performance on the SAT critical reading subtest: an application of the attribute hierarchy method," in *Proceedings of the annual meeting of the National Council on Measurement in Education*, Washington, DC.

Williams, J. S., and Amir, G. S. (1995). *11-12 Year Old Children's Informal Knowledge and Its Influence on their Formal Probabilistic Reasoning*. ERIC ED387256. London: ERIC.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9781410611697

Xie, J. S., and Lu, Y. S. (2011). The quantitative classification of the level of economic development of Chinese provinces and cities. *Modern Finan. Econ.* 31, 96–99.

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

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

Zhang, Z. J., Liu, F., and Zhao, S. W. (1985). A study of the development of 5 to 15-year-old's concept of probability. *J. Psychol. Sci.* 6, 3–8. doi: 10.16719/j.cnki.1671-6981.1985.06.001

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fpsyg.2020.570365

Cognitive Diagnostic Models for Random Guessing Behaviors

Chia-Ling Hsu[1], Kuan-Yu Jin[2] and Ming Ming Chiu[1]*

[1] Assessment Research Centre, The Education University of Hong Kong, Tai Po, Hong Kong, [2] Hong Kong Examinations and Assessment Authority, Wan Chai, Hong Kong

Many test-takers do not carefully answer every test question; instead they sometimes quickly answer without thoughtful consideration (*rapid guessing*, RG). Researchers have not modeled RG when assessing student learning with *cognitive diagnostic models* (CDMs) to personalize feedback on a set of fine-grained skills (or attributes). Therefore, this study proposes to enhance cognitive diagnosis by modeling RG via an advanced CDM with item response and response time. This study tests the parameter recovery of this new CDM with a series of simulations via Markov chain Monte Carlo methods in JAGS. Also, this study tests the degree to which the standard and proposed CDMs fit the student response data for the Programme for International Student Assessment (PISA) 2015 computer-based mathematics test. This new CDM outperformed the simpler CDM that ignored RG; the new CDM showed less bias and greater precision for both item and person estimates, and greater classification accuracy of test results. Meanwhile, the empirical study showed different levels of student RG across test items and confirmed the findings in the simulations.

Keywords: response time, rapid guessing, G-DINA model, DINA model, DINO model

OPEN ACCESS

Edited by:
Peida Zhan,
Zhejiang Normal University, China

Reviewed by:
Kaiwen Man,
University of Maryland, College Park, United States
Gongjun Xu,
University of Michigan, United States

****Correspondence:***
Chia-Ling Hsu
clhsu@friends.eduhk.hk

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

Received: *07 June 2020*
Accepted: *07 September 2020*
Published: *25 September 2020*

Citation:
Hsu C-L, Jin K-Y and Chiu MM (2020) Cognitive Diagnostic Models for Random Guessing Behaviors. Front. Psychol. 11:570365. doi: 10.3389/fpsyg.2020.570365

INTRODUCTION

Cognitive diagnostic models (CDMs) assess whether test-takers have the skills needed to answer test questions (*attributes*), so that their test results can give them diagnostic feedback on their strengths and weaknesses in these attributes (Rupp et al., 2010). Specifically, a CDM analysis determines whether a person shows mastery (vs. non-mastery) of a set of fine-grained attributes (*latent class*). Teachers, clinicians and other users of test scores can use such specific information on each student or client to adapt and improve their instructions/interventions more effectively, compared to a simple, summative score.

However, some test-taking behaviors can distort current CDM results and thereby jeopardize the validity of their assessments. Recently, researchers have proposed different approaches to account for test-taking behaviors when assessing test-taker performance and item characteristics. In this study, we focus on two frequently-observed test-taking behaviors during actual tests: solution attempt and *rapid guessing* (RG; Wise and Kong, 2005). In a solution attempt, test-takers carefully try to find answers to test questions. By contrast, RG refers to test-takers quickly answering test questions without thoughtful consideration (e.g., Wise and DeMars, 2006). For instance,

Meyer (2010) integrated a two-class mixture Rasch model (Rost, 1990) to classify a test-taker as either making a solution attempt or RG, but not allowing different behaviors by the same person during a test. To address this limitation, Wang and Xu (2015) proposed a model with a latent indicator to allow each test-taker to engage in either a solution attempt or RG on each item. Furthermore, the indicator can depend on either a test-taker's RG propensity or on an item-level feature (Wang et al., 2018). As RGs are typically much shorter than solution attempts, CDMs can use a test-taker's reaction time (RT) to each test question to properly model RGs and distinguish them from solution attempts (but not necessarily pre-knowledge answers, e.g., Wang et al., 2018). As no published study has proposed and tested a CDM that models RG, we do so in this study.

This study proposes a new framework of CDMs to recognize different test-taking behaviors by using RT and item responses simultaneously. This new class of CDMs: (1) models two test-taking behaviors (RG vs. solution attempt) for each item-person concurrence, (2) allows multiple switch points between RG and solution attempts among the items for each test-taker, (3) thereby yields person and item estimates with greater accuracy, and (4) generalizes to available CDMs, RT functions and other kinds of dissimilar behaviors.

The generalized DINA model (G-DINA, de la Torre, 2011) conceptualizes and shows the utility of this framework. Specifically, the two special cases of the G-DINA model, the deterministic input, noisy "and" gate (DINA) model (Junker and Sijtsma, 2001) and its counterpart, the deterministic input, noisy "or" gate (DINO) model (Templin and Henson, 2006) are simple to compute, estimate, and interpret, so they serve as illustrations. Nevertheless, researchers can extend this approach to other CDMs, especially G-DINA-liked formulation CDMs, such as the general diagnostic model (GDM; von Davier, 2005) and the linear logistic model (Maris, 1999).

After we present the functions for describing RT and item response, we specify the new model. Next, our simulation study illustrates the new model's performance, followed by its application to real data. Lastly, we discuss the implications of this study for identifying test-taking behaviors and improving the estimation accuracy of both person and item parameters.

A NEW CDM FRAMEWORK

The new model requires distinct functions to separately specify two fundamentals for an item, RT and item response, while two main facets, person and item, affect the observed RT and item response. This section describes the adopted RT and item response functions, before specifying the new model.

The Lognormal RT Model

As cognitive test data typically resemble a lognormal distribution more closely than a normal distribution, we use a lognormal function to characterize RT (van der Linden, 2006, 2007). Let RT_{ij} be the observed RT of person i ($i = 1, 2, \ldots, I$) to item j ($j = 1, 2, \ldots, J$). In the lognormal function, the two parameters of *person speed* and *time intensity,* respectively, represent the two facts, person and item, as follows,

$$log\left(RT_{ij}\right) \sim N(\beta_j - \tau_i, 1/\kappa_j^2) \tag{1}$$

where τ_i indicates the average speed of test-taker i on a test (person speed); β_j indicates the mean time that the population needs to resolve item j (time intensity); and κ_j^2 indicates the dispersion of the logarithmized RT distribution (time discrimination parameter) of item j.

The G-DINA Model

The G-DINA model loosens some restrictions of the DINA model and its saturated form is equivalent to other general CDMs via link functions (de la Torre, 2011). Hence, the G-DINA model can (a) present different CDMs with similar formulations via various constraints and (b) substantially reduce the number of latent classes for an item – especially for models with more than five attributes. The original G-DINA model with identity link can be expressed as

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik} \tag{2}$$

For test-taker i, the reduced attribute vector α_{ij}^* has the required attributes for item j. The intercept for item j, δ_{j0} represents the probability of a correct response without the required attributes (*baseline probability*). The main effect δ_{jk} reflects the extent to which mastery of a single attribute α_k changes the probability of a correct response. The interaction effect $\delta_{jkk'}$ indicates the extent to which mastery of both attributes α_k and $\alpha_{k'}$ changes the probability of a correct response. The interaction effect $\delta_{j12\cdots K_j^*}$ reflects the extent to which mastery of all the required attributes $\alpha_1, \alpha_2, \cdots,$ and $\alpha_{K_j^*}$ changes the probability of a correct response.

Like most CDMs, the G-DINA model requires a $J \times K$ Q-matrix (Tatsuoka, 1983), in which K knowledge attributes are required to correctly answer J items. $K_j^* = \sum_{k=1}^{K} q_{jk}$ is the number of required attributes for item j, where $q_{jk} = 1$ if the correct response to item j requires attribute k; and 0 otherwise. As the number of required attributes for item j is smaller than that of the all attribute vectors ($K_j^* < K$), the G-DINA model can reduce the number of required latent classes ($2^{K_j^*} < 2^K$) for an item. To illustrate the G-DINA-like formulations, we use two common cases: the DINA and DINO.

The DINA Model

In the non-compensatory DINA, individuals are classified into one of two latent classes for an item: (a) the attribute vectors have all of an item's required attributes (*mastery*) or (b) the attribute vectors are missing at least one of the item's required attributes (*non-mastery*). The two latent classes' corresponding probabilities for a correct response entail that (a) mastery individuals do not slip, or (b) non-mastery individuals guess

correctly (Junker and Sijtsma, 2001). Thus, the DINA model can be re-formed by setting to zero, all G-DINA model parameters except δ_{j0} and $\delta_{j12\cdots K_j^*}$

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik} \tag{3}$$

In Eq. 3, $\delta_{j0} = g_j$ is the probability of a correct response to item j for a non-mastery test-taker i, where g_j is the guessing parameter for item j; $\delta_{j0} + \delta_{j12\cdots K_j^*} = 1 - s_j$ is the probability of a correct response to item j for a mastery test-taker i, where s_j is the slipping parameter for item j. In the DINA, a mastery test-taker with all the required attributes (K_j^*) for item j generally answers it correctly and other test-takers generally answer it incorrectly. Like the DINA, Eq. 3 shows that except for the attribute vector $\alpha_j^* = 1_{K_j^*}$ (in which $1_{K_j^*}$ is a vector of ones with length K_j^*), other latent classes ($2^{K_j^*} - 1$) have the same probability of correctly answering item j. As shown in Eq. 3, this probability increases only after mastering all the required attributes. Under the DINA model assumption (Junker and Sijtsma, 2001), the G-DINA has two parameters per item (see Eq. 3).

The DINO Model

Unlike the non-compensatory DINA, the compensatory DINO only entails at least one of the required attributes to answer an item, so the parameters in G-DINA are set to

$$\delta_{jk} = (-1)\,\delta_{jk'k''} = \cdots = (-1)^{K_j^*+1}\,\delta_{j1,2,\cdots,K_j^*} \tag{4}$$

where $k = 1, \cdots, K_j^*$, $k' = 1, 2, \cdots K_j^* - 1$, and $k'' > k', \cdots, K_j^*$. The orders of the interactions vary the alternating sign, and the quantities of the main effects and interactions have the same value.

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \delta_{jk}\alpha_{ik}. \tag{5}$$

For a test-taker i with at least one of the required attributes, the probability of answering item j without slipping (s_j') is $\delta_{j0} + \delta_{jk} = 1 - s_j'$. Likewise, for a test-taker i with none of the required attributes, the probability of correctly answering item j is the guessing parameter, $\delta_{j0} = g_j'$. Unlike the DINA, all latent classes except for the attribute vector $\alpha_j^* = 0_{K_j^*}$ (a vector of zeros and of length K_j^*) have the same probability of correctly answering item j. Like the DINA, the DINO only needs two parameters for an item (Eq. 5, Templin and Henson, 2006).

To use both information of RT and item response, two functions must be specified. Hence, RT-GDINA, RT-DINA and RT-DINO jointly model RT and item response with the lognormal distribution (Eq. 1) and G-DINA, DINA and DINO, respectively.

New Class of CDMs

We introduce a new class of G-DINA to account for varying test-taking behaviors. RT (RT_{ij}) and item response (Y_{ij}) are modeled individually. As test-takers can switch between RG and solution behaviors, like Wang and Xu (2015), a latent indicator (ξ) is employed, where if test-taker i tries to solve item j, $\xi_{ij} = 1$ (0 otherwise; RG is specified in this study). Incorporating this latent indicator into the lognormal RT model extends Eqs 1–6

$$\begin{cases} log\left(RT_{ij}\right) \sim N\left(\beta_j - \tau_i, 1/\kappa_j^2\right), & \text{if } \xi_{ij} = 1; \\ log\left(RT_{ij}\right) \sim N\left(\beta_0, 1/\kappa_0^2\right), & \text{if } \xi_{ij} = 0. \end{cases} \tag{6}$$

indicates that the logarithmized RT is normally distributed as Eq. 1 if test-taker i solves item j from solution attempt ($\xi_{ij} = 1$), and it is normally distributed with mean time intensity β_0 and time discrimination κ_0^2 if test-taker i responds to item j with a RG ($\xi_{ij} = 0$). For a RG on item j by test-taker i, RT is constant.

Likewise, adding ξ_{ij} to the G-DINA yields

$$P\left(\alpha_{ij}^*\right) = \xi_{ij}\Bigg(\delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}\Bigg) + \left(1 - \xi_{ij}\right)\delta_j^* \tag{7}$$

The G-DINA model is the underlying model for a solution attempt on item j by test-taker i ($\xi_{ij} = 1$). We assume that a RG on item j by test-taker i ($\xi_{ij} = 0$) yields δ_j^*. For simplicity, like Wise and DeMars (2006), we assume that test-taker i has the same probability of correctly answering item j both by RG and by guessing with none of the required attributes ($\delta_j^* = \delta_{j0}$), that is, guessing randomly for all options. Hence, Eq. 7 can be re-written as

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \xi_{ij}\Bigg(\sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}\Bigg) \tag{8}$$

The latent indicator ξ_{ij} in Eqs 6–8 is a binary result of test-taker i on item j's behavior (solution attempt vs. RG). It can be modeled by a Bernoulli distribution with π_j, the marginal probability of the solution attempt. Using the DINA and DINO to identify RGs and solution attempts, Eqs 3 and 5 are re-written, respectively, as Eqs 8 and 9.

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \xi_{ij}\delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik} \tag{9}$$

$$P\left(\alpha_{ij}^*\right) = \delta_{j0} + \xi_{ij}\delta_{jk}\alpha_{ik} \tag{10}$$

If test-taker i tries to solve item j ($\xi_{ij} = 1$), Eqs 6, 8–10 reduce, respectively, to Eqs 1–3, and 5. Thus, the lognormal, G-DINA, DINA, and DINO are special cases of our proposed new CDM framework. Likewise, jointly modeling RT and item response with a latent indicator via Eqs 6 and 8–10 are, respectively, represented as RT-GDINA-RG, RT-DINA-RG and RT-DINO-RG.

To illustrate this approach, we combine the lognormal RT distribution and the easy-to-understand DINA and DINO models, with the latent indicator ξ (RT-DINA-RG, RT-DINO-RG) and without it (RT-DINA, RT-DINO). We estimate their parameters via the Bayesian method with the Markov chain Monte Carlo (MCMC) algorithm in the freeware JAGS (Plummer, 2017). For the JAGS code and the priors for the estimated parameters of the RT-DINA-RG and RT-DINO-RG models (see **Appendix**).

SIMULATION STUDY 1: PARAMETER RECOVERY OF RT-DINA-RG

Design

In simulation study 1, we evaluated the parameter recovery of the RT-DINA-RG for a test of 30 dichotomous items measuring five non-compensatory attributes. See the artificial Q-matrix in **Table 1**. The guessing (g_j) and slipping (s_j) parameters were randomly generated, respectively, from the uniform distributions of $U(0.05, 0.3)$ and $U(0.05, 0.2)$, which reflect a high quality test. This data-generating procedure for the 30 simulated items yielded item discrimination indices (IDI) that ranged from 0.51 to 0.88, indicating a test with high measurement quality (Lee et al., 2012).

TABLE 1 | Specified Q-matrix and item parameters in simulation 1.

Item	q_1	q_2	q_3	q_4	q_5	π_j
1	1	0	0	0	0	0.9
2	0	1	0	0	0	0.9
3	0	0	1	0	0	0.9
4	0	0	0	1	0	0.9
5	0	0	0	0	1	0.9
6	1	0	0	0	0	0.8
7	0	1	0	0	0	0.8
8	0	0	1	0	0	0.8
9	0	0	0	1	0	0.8
10	0	0	0	0	1	0.8
11	1	1	0	0	0	0.9
12	1	0	1	0	0	0.9
13	1	0	0	1	0	0.9
14	1	0	0	0	1	0.9
15	0	1	1	0	0	0.9
16	0	1	0	1	0	0.8
17	0	1	0	0	1	0.8
18	0	0	1	1	0	0.8
19	0	0	1	0	1	0.8
20	0	0	0	1	1	0.8
21	1	1	1	0	0	0.9
22	1	1	0	1	0	0.9
23	1	1	0	0	1	0.9
24	1	0	1	1	0	0.9
25	1	0	1	0	1	0.9
26	1	0	0	1	1	0.8
27	0	1	1	1	0	0.8
28	0	1	1	0	1	0.8
29	0	1	0	1	1	0.8
30	0	0	1	1	1	0.8

We manipulated two conditions. In the RG condition, the marginal probability of RG ($1 - \pi_j$) was set for items at two levels: 0.1 and 0.2 (Wang et al., 2018). To describe the dynamic latent indicator of person i on item j in the RG condition, the ξ-parameter was generated from a Bernoulli distribution with probability either of 0.8 or 0.9. In the RT-DINA-RG, mean item time intensity (β_0) and item discrimination (κ_0) were (a) set, respectively, at 2 and 1.6 for rapid guessers ($\xi_{ij} = 0$) and (b) generated, respectively, from $U(2, 4)$ and $U(0.15, 2)$ for normal test-takers ($\xi_{ij} = 1$). In non-RG condition, RG never occurs, and the RT-DINA served as the data-generating model, yielding parameters similar to the RT-DINA-RG. Mean item time intensity and item discrimination can be generated to accommodate various test situations (e.g., Man et al., 2018), but they do not affect the use of the proposed model. Therefore, we leave this interesting topic for further study.

We simulated 1,000 test-takers across conditions, and each test-taker had generated five latent attributes with positive correlations, following Henson and Douglas (2005) procedure. Specifically, we randomly generated 1,000 vectors with five values, $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \alpha_{i4}, \alpha_{i5})'$, $i = 1, 2, \ldots, 1{,}000$, from a multivariate normal distribution with no interaction, $MVN(\mathbf{0.5}, \Sigma)$ with Σ diagonal elements of 1.0 and others of 0.5. A cut-off value of 0.253 ($z_{0.6}$) indicated mastery of the attribute (if $\alpha_{ik} > 0.253$, $\alpha_{ik} = 1$; otherwise, $\alpha_{ik} = 0$), yielding ∼60% mean mastery of each attribute, which generally ranged from easy to moderate. The person speed parameter (τ_i) was generated from $N(0, 0.3^2)$. Each condition was replicated 100 times from an R script.

Both the RT-DINA and RT-DINA-RG were fit to these data to test three hypotheses: (1) with some RG, the RT-DINA-RG efficiently recovers item and person estimates; (2) ignoring RG via the RT-DINA yields biased item parameter estimates, less accurate classification of attribute mastery, and less reliable person speed estimates; and (3) with no RG, the RT-DINA-RG performs as well as the RT-DINA. To evaluate the recovery of item parameters, the bias and root mean squared error (RMSE) were computed as dependent variables:

$$\text{Bias}\left(\hat{\nu}\right) = \sum_{r=1}^{100}\left(\hat{\nu}_r - \nu\right)/100 \tag{11}$$

$$\text{RMSE}\left(\hat{\nu}\right) = \sqrt{\sum_{r=1}^{100}\left(\hat{\nu}_r - \nu\right)^2/100} \tag{12}$$

where ν and $\hat{\nu}_r$ indicate respectively, true and estimated values in the r-th replication of an item parameter. We examined test-takers' true and estimated latent classes to evaluate the classification accuracy of each attribute. The reliability of the person speed parameter was computed as:

$$\text{Reliability}\left(\hat{\tau}\right) = \text{Correlation}\left(\hat{\tau}, \tau\right)^2 \tag{13}$$

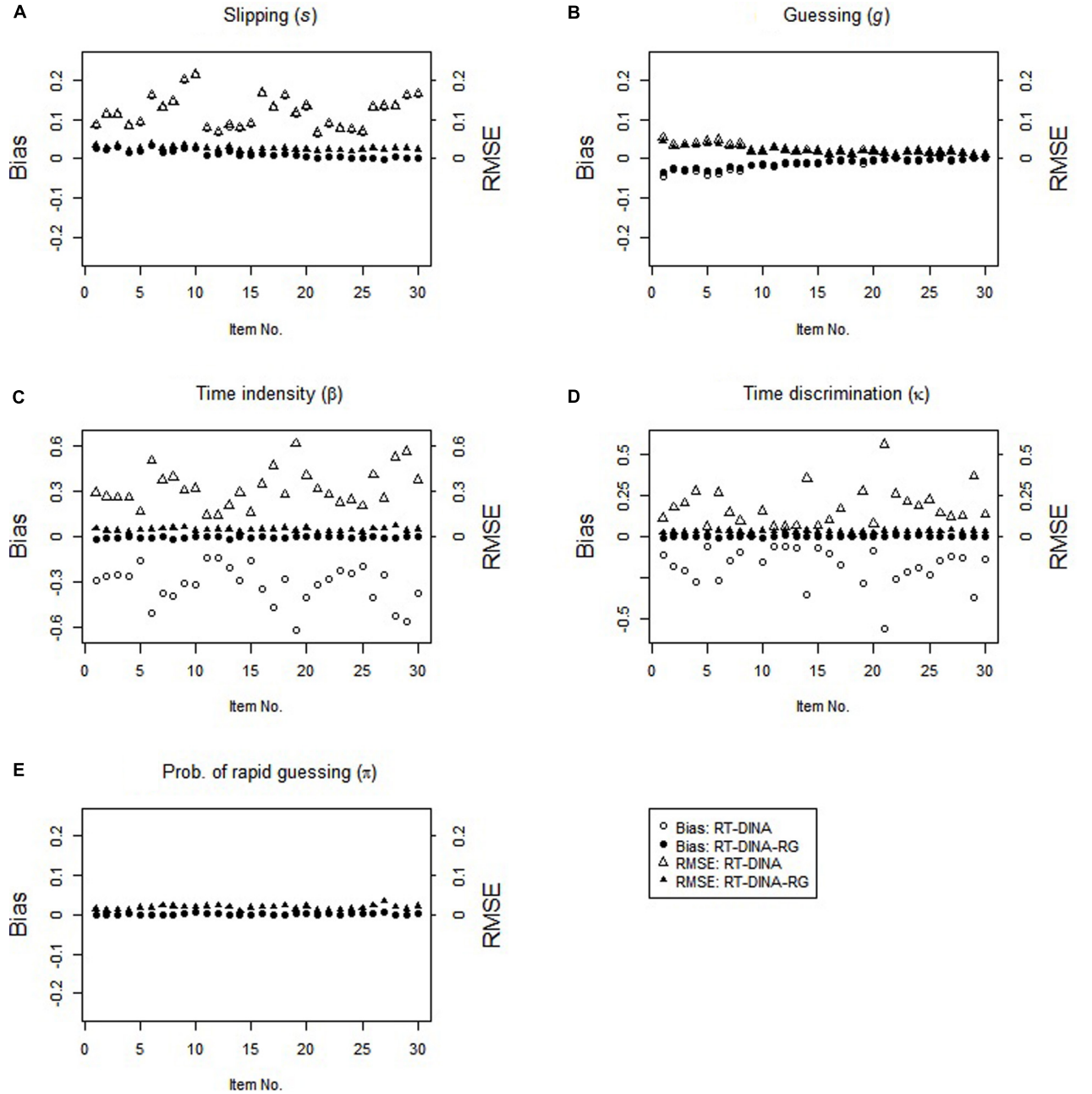


FIGURE 1 | Parameter recovery for the RT-DINA and RT-DINA-RG under the RG condition in simulation 1.

Results

In the RG condition, the RT-DINA-RG generally yielded unbiased parameter estimates, whereas the RT-DINA overestimated the slipping parameters and underestimated the item intensity, guessing, and time discrimination parameters (see **Figure 1**). Greater RG increased the severities of slipping overestimation and time intensity underestimation. For test-takers without the required attributes, RG did not influence the success rate, so ignoring RG did not substantially influence estimation of the guessing parameters. Across five attributes and 100 replications, mean classification accuracy was higher for the RT-DINA-RG than the RT-DINA (0.936 > 0.924), suggesting that ignoring RG reduces the accuracy of attribute classification. Also, the RT-DINA-RG outperformed the RT-DINA on reliability of the person speed parameter (M: 0.66 > 0.57).

In the non-RG condition, both RT-DINA and RT-DINA-RG recovered the parameters well (see **Figure 2**). The bias and RMSE for the π-parameter in the RT-DINA-RG were nearly zero. Also, both models yielded practically identical classification accuracy (M = 96.6%) and reliability of person speed parameter

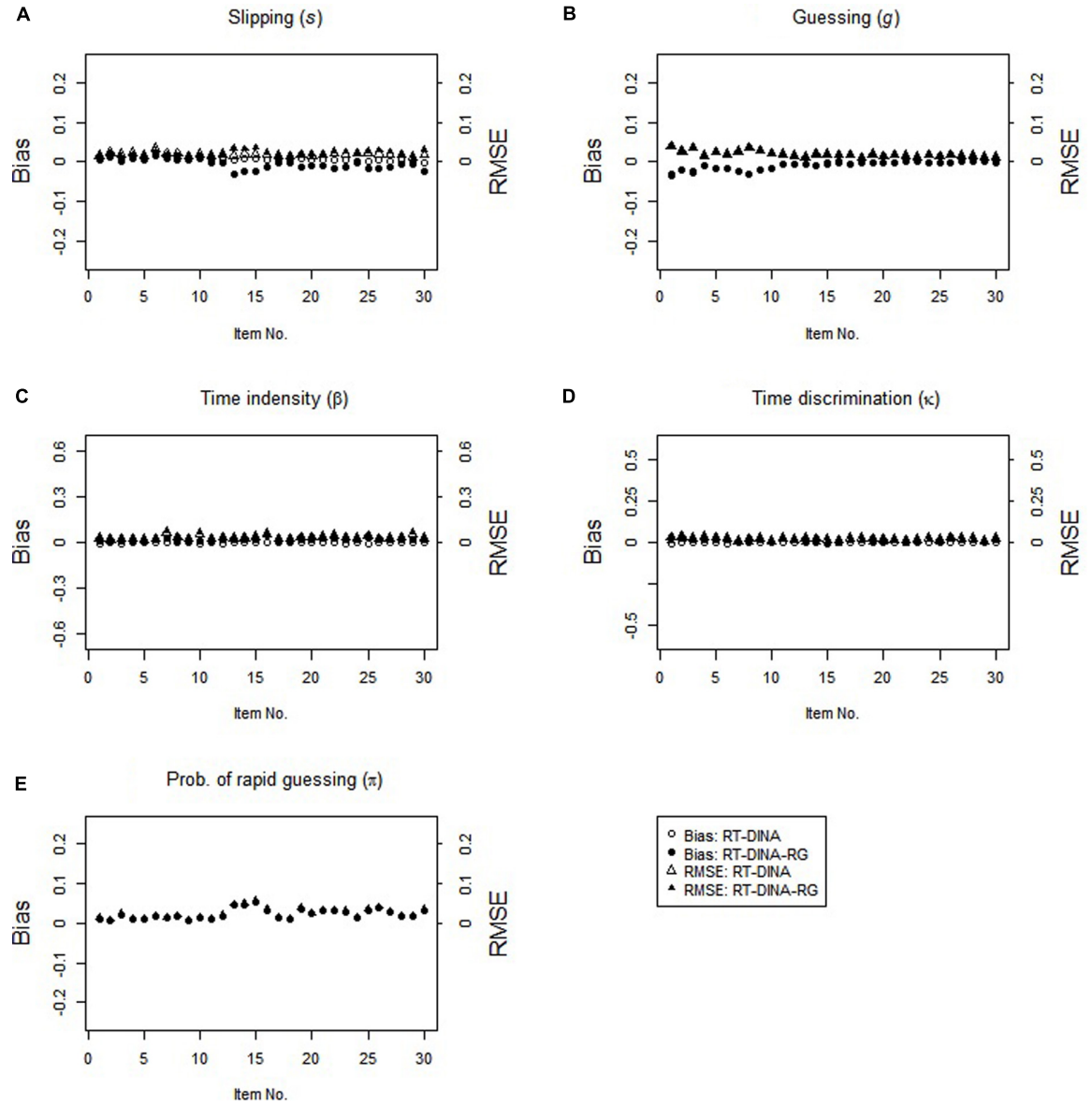


FIGURE 2 | Parameter recovery for the RT-DINA and RT-DINA-RG under the non-RG condition in simulation 1.

(M = 0.76) across 100 replications. Hence, overfitting the RT-DINA-RG to data without RG showed no significant harm. In brief, the simulation results supported our three hypotheses.

SIMULATION STUDY 2: PARAMETER RECOVERY OF RT-DINO-RG

Design

Study 2 simulated compensatory attributes and analyzed parameter recovery by the RT-DINO and RT-DINO-RG. The item responses and RTs were generated for (a) the RG condition with the RT-DINO-RG and (b) the non-RG condition with the RT-DINO. The parameters, data generation and evaluation criteria were the same as those in simulation study 1. Paralleling study 1, we test three hypotheses: (1) with some RG, the RT-DINO-RG efficiently recovers item and person estimates; (2) ignoring RG via the RT-DINO yields biased item parameter estimates and less accurate classification of attribute mastery; and (3) with no RG, the RT-DINO-RG performs as well as the RT-DINO.

Results

The study 2 results resemble the study 1 results (see **Figure 3**). In the RG condition, the RT-DINO-RG recovered the parameters well, whereas the RT-DINO overestimated the slipping parameters and underestimated the item intensity, guessing, and time discrimination parameters. Greater RG

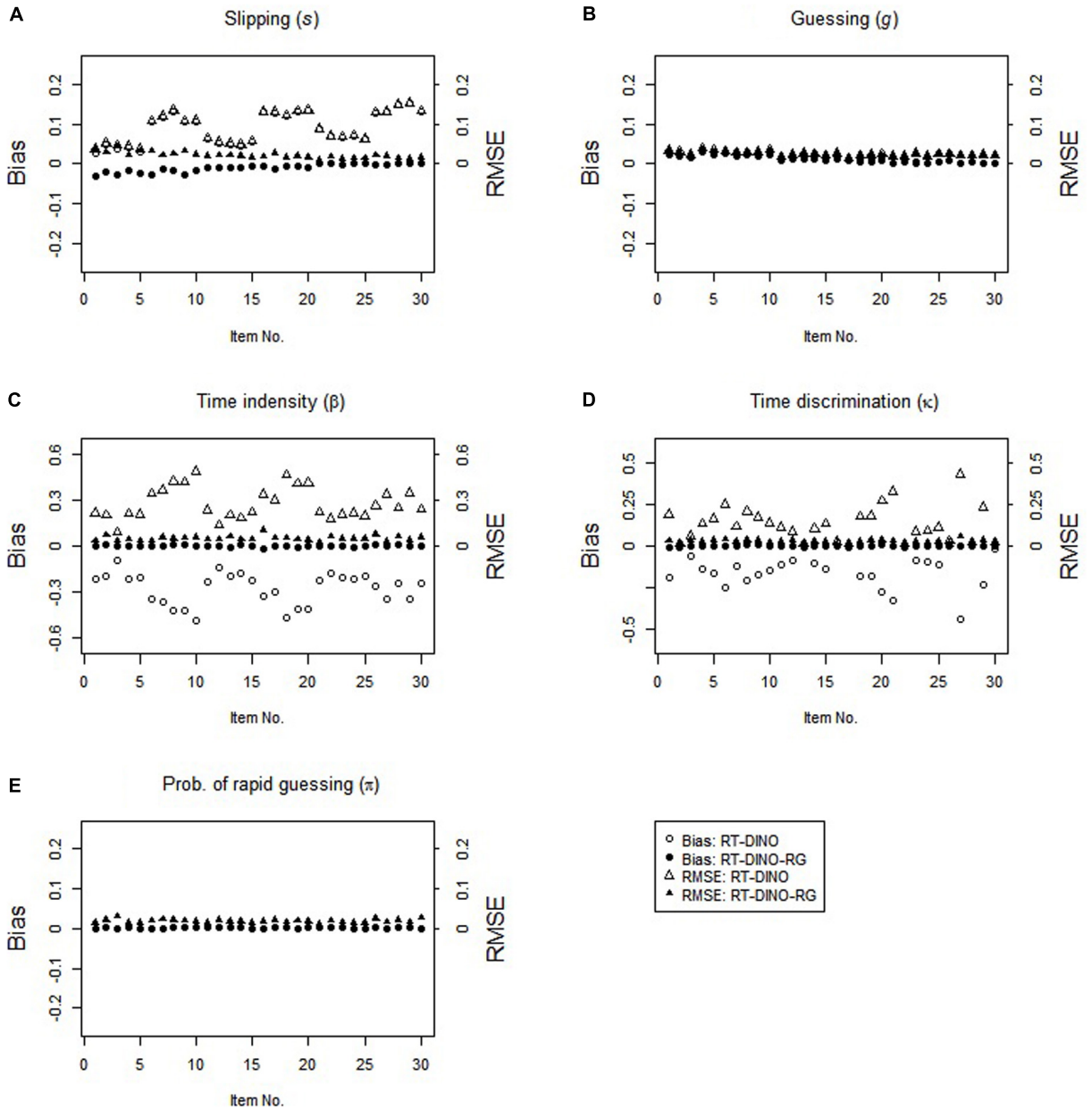


FIGURE 3 | Parameter recovery for the RT-DINO and RT-DINO-RG under the RG condition in simulation 2.

increased the severities of slipping overestimation and time intensity underestimation. The RT-DINO-RG outperformed the RT-DINO on both mean classification accuracy (0.946 > 0.922) across five attributes and reliability of the person speed parameter (M: 0.64 > 0.57).

In the non-RG condition, both RT-DINO and RT-DINO-RG recovered the item parameters well (see **Figure 4**). The bias and RMSE for π-parameter in the RT-DINO-RG model were very small. Also, both models had practically identical classification accuracy (M = 98.4%) and reliability of person speed parameter (M = 0.71) across replications. In sum, these simulation results supported our three hypotheses.

REAL DATA ANALYSIS

To illustrate a RT-GDINA-RG application, we analyzed a PISA 2015 mathematics test with 22 questions. After screening out students with missing responses, we analyzed 5,158 students' responses. The PISA 2015 mathematics assessment framework (OECD, 2017a,b) and the released computer-based mathematics items' log-file databases covered eight attributes: change and relationships (α_1), quantity (α_2), space and shape (α_3), uncertainty (α_4), occupational (α_5), societal (α_6), scientific (α_7), and personal (α_8). The Q-matrix for the mathematics test shows two cognitive attributes for each item (see **Table 2**). We fit

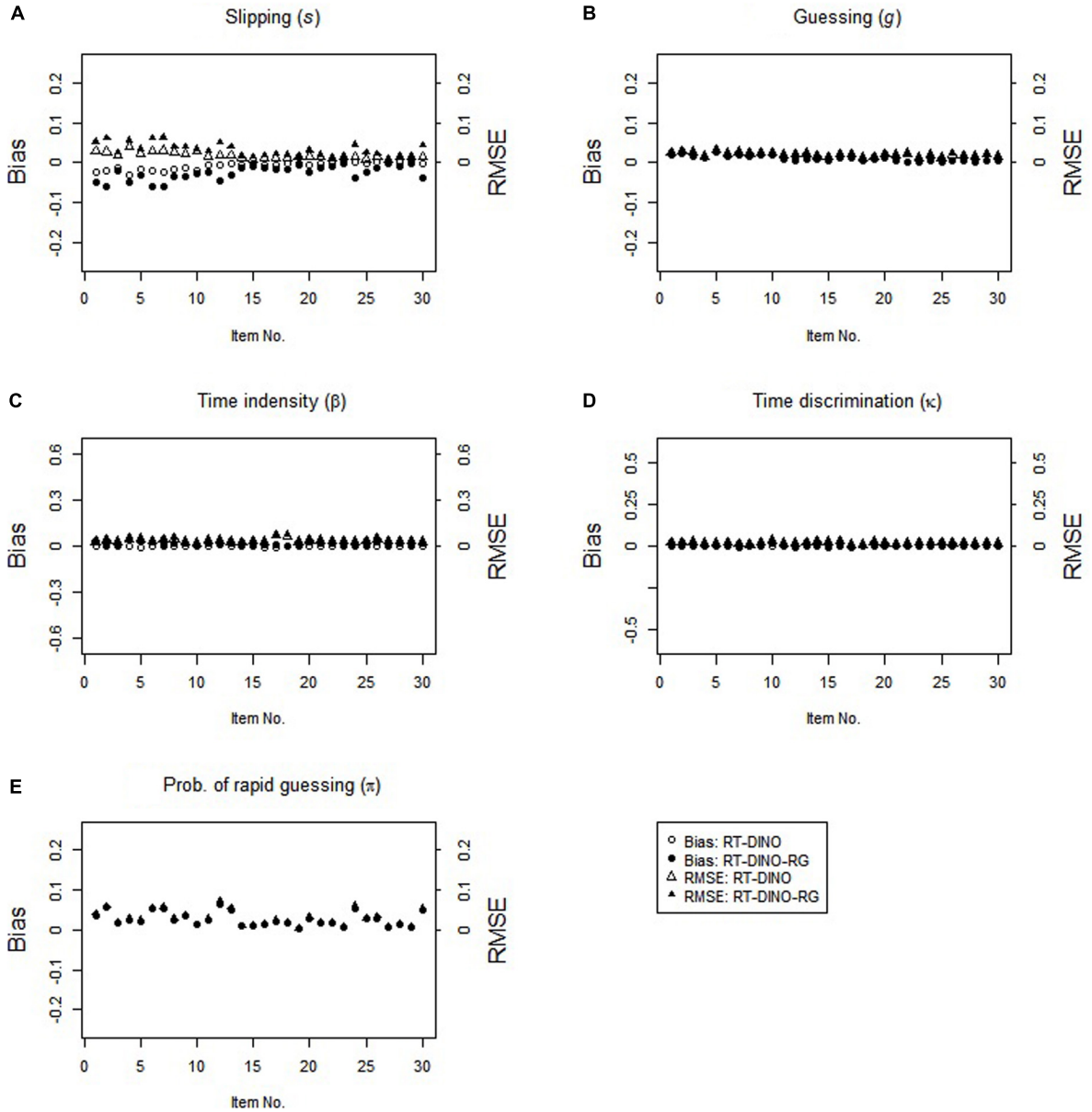


FIGURE 4 | Parameter recovery for the RT-DINO and RT-DINO-RG under the non-RG condition in simulation 2.

the four CDM models (RT-DINA, RT-DINO, RT-DINA-RG, RT-DINO-RG) to these data. Superior models have lower deviance information criteria (DIC; Spiegelhalter et al., 2002).

The results indicate both compensatory attributes and RG. DICs showed that the compensatory models outperformed the non-compensatory ones (RT-DINO < RT-DINA: 1,351,697 < 1,433,173; and RT-DINO-RG < RT-DINA-RG: 1,327,068 < 1,360,978) suggesting that the eight attributes' relationships were more compensatory than non-compensatory. Also, the RG models outperformed the simpler models (RT-DINO-RG < RT-DINO: 1,327,068 < 1,351,697; and RT-DINA-RG < RT-DINA: 1,360,978 < 1,433,173), showing substantial RG. As the data indicated both compensatory attributes and RG, the RT-DINO-RG showed the best fit. Hence, we examine the RT-DINO and RT-DINO-RG results in greater detail.

Like study 2, the RT-DINO estimated higher slipping parameters and lower guessing parameters, compared to the RT-DINO-RG (slipping: $M_{\text{RT-DINO}} > M_{\text{RT-DINO-RG}}$: 0.27 > 0.22; guessing: $M_{\text{RT-DINO}} < M_{\text{RT-DINO-RG}}$: 0.28 < 0.30). Also, the mean discrimination power of RT-DINO-RG exceeded that of RT-DINO ($\text{IDI}_{M(\text{RT-DINO-RG})} > \text{IDI}_{M(\text{RT-DINO})}$: 0.47 > 0.45). Ranging from 0.74 to 0.99, RT-DINO-RG's RG estimates (π) moderately correlated (r = 0.49) with the difference in the slipping parameters of RT-DINO and RT-DINO-RG (see **Figure 5**), supporting the simulation study 2 finding of

TABLE 2 | Specified Q-matrix for the real data.

Item	Label	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
1	CM033Q01S	0	0	1	0	0	0	0	1
2	CM474Q01S	0	1	0	0	0	0	0	1
3	DM155Q02C	1	0	0	0	0	0	1	0
4	CM155Q01S	1	0	0	0	0	0	1	0
5	DM155Q03C	1	0	0	0	0	0	1	0
6	CM155Q04S	1	0	0	0	0	0	1	0
7	CM411Q01S	0	1	0	0	0	1	0	0
8	CM411Q02S	0	0	0	1	0	1	0	0
9	CM803Q01S	0	0	0	1	1	0	0	0
10	CM442Q02S	0	1	0	0	0	1	0	0
11	DM462Q01C	0	0	1	0	0	0	1	0
12	CM034Q01S	0	0	1	0	1	0	0	0
13	CM305Q01S	0	0	1	0	0	1	0	0
14	CM496Q01S	0	1	0	0	0	1	0	0
15	CM496Q02S	0	1	0	0	0	1	0	0
16	CM423Q01S	0	0	0	1	0	0	0	1
17	DM406Q01C	0	0	1	0	0	1	0	0
18	DM406Q02C	0	0	1	0	0	1	0	0
19	CM603Q01S	0	1	0	0	0	0	1	0
20	CM571Q01S	1	0	0	0	0	0	1	0
21	CM564Q01S	0	1	0	0	0	1	0	0
22	CM564Q02S	0	0	0	1	0	1	0	0

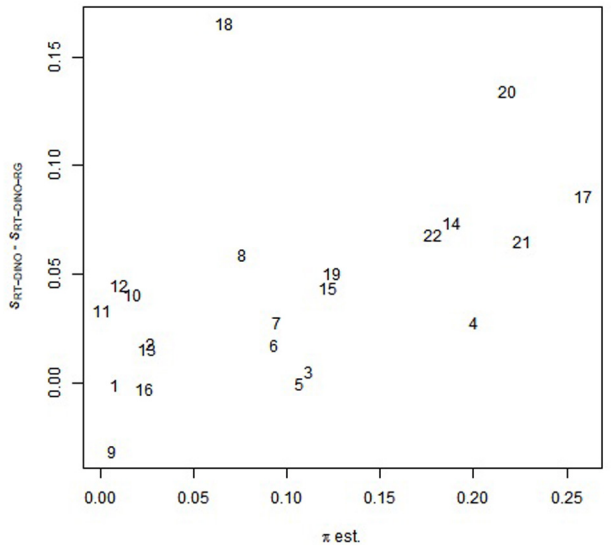


FIGURE 5 | Relationship between π-parameter in the RT-DINO-RG and the difference in the slipping parameters between the RT-DINO the RT-DINO-RG models. Numbers are item identifiers; πEst is RG estimates; and $S_{RT-DINO} - S_{RT-DINO-RG}$ is the difference in the slipping parameters of RT-DINO and RT-DINO-RG.

overestimated slipping parameters when ignoring RGs. Also, the πs of items 1–11 were generally lower than those of items 12–22. If these items appeared on the test in this sequence (item position information was not publicly available), these π results

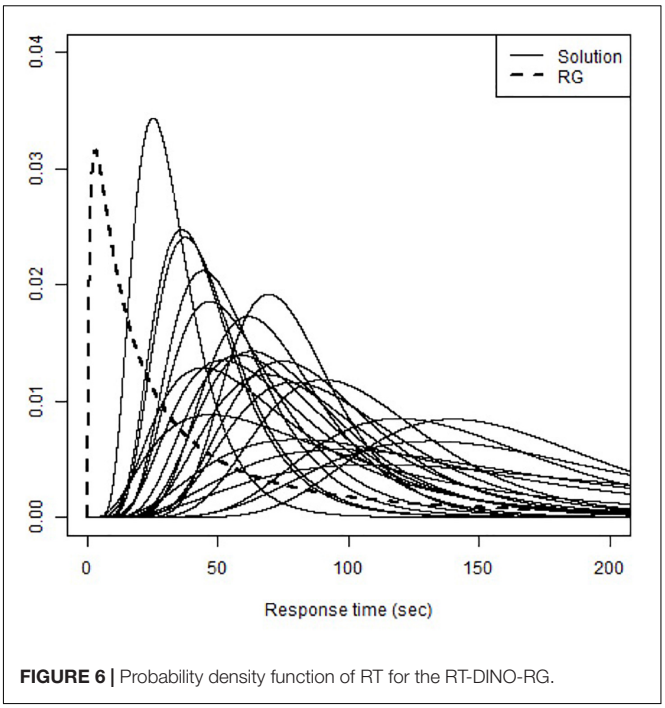


FIGURE 6 | Probability density function of RT for the RT-DINO-RG.

suggest that test-taker accuracy depended on their completion speed (*speededness*).

The RT-DINO-RG also uses response time to recognizes RGs and solution attempts, showing estimated mean time intensity (β_0) of 3.21 and time discrimination (κ_0) of 0.70. The various probability density functions of response time for RGs and solution attempts in the RT-DINO-RG (see **Figure 6**) suggest that students used varied answering strategies to spend more time on some items and less time on others (including RGs). The RT-DINO and RT-DINO-RG did not consistently classify mastery of the eight attributes [Cohen's κ ranged from 0.48 (quantity) to 0.98 (occupational), see **Table 3**]. Notably, few students had knowledge of the third attribute (space and shape). The simulation studies suggest that the RT-DINO-RG classifications are more reliable than the RT-DINO ones.

DISCUSSION

CDMs assess whether test-takers have the needed skills (*attributes*) to answer each test question and give suitable diagnostic feedback, but they have not adequately modeled RG vs. solution attempts with reaction times. Hence, this study

TABLE 3 | Mastery of attributes for the RT-DINO and RT-DINO-RG.

	Attributes							
	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
RT-DINO	0.090	0.381	0.057	0.133	0.173	0.235	0.366	0.349
RT-DINO-RG	0.234	0.359	0.036	0.151	0.168	0.166	0.254	0.361
Cohen's *k*	0.483	0.861	0.644	0.899	0.975	0.762	0.721	0.885

developed a new class of CDMs based on the G-DINA model (de la Torre, 2011), namely RT-GDINA-RG, with a latent indicator to jointly utilize both item responses and RTs to model RG and solution attempts to enhance cognitive diagnosis. We propose two models based on the DINA and DINO models, namely RT-DINA-RG and RT-DINO-RG.

The RT-DINA-RG and RT-DINO-RG were evaluated via (a) simulation studies with Markov chain Monte Carlo methods in JAGS and (b) real data analysis by analyzing the PISA 2015 computer-based mathematics test. Complementing Wang and Xu (2015) person-level manipulation of RG, this study manipulated RG at the item level (Wang et al., 2018). The simulation results and real data analysis showed that the RT-DINA-RG and RT-DINO-RG recovered parameters well and assessed test-takers' diagnostic results more accurately. In contrast, ignoring RGs by fitting simpler models yielded biased parameters, less reliable person speed parameter, and less classification accuracy of test results.

Hence, this study extends research showing how analyses of RT improves cognitive assessments of test-takers (e.g., van der Linden, 2008; Lee and Chen, 2011; Wang and Xu, 2015). When test-takers rapidly guess, the RT-GDINA-RG yields greater accuracies in person parameters, item parameters, and cognitive results. Therefore, researchers or users should use the RT-GDINA-RG to depict a data if RGs might occur. The choice of RT-GDINA-RG model (i.e., RT-DINA-RG or RT-DINO-RG) depends on the nature of the test items. If a test item's needed underlying constructs can compensate for one another, then RT-DINO-RG is suitable. If the underlying constructs cannot compensate for one another, then RT-DINA-RG is suitable.

Moreover, the person and item parameters of the RT-GDINA-RG were assumed to be, respectively, independent in this study. As attributes and item parameters of a CDM are often related in practice, we capture the relations between them with correlational structures (e.g., van der Linden, 2007). Note that the commonly-used multivariate normal distribution to specify the relations among person parameters is not feasible for the discrete feature of attributes in CDMs. Following Zhan et al. (2018), one can address this problem by using a higher-order latent trait to link the correlated attributes (de la Torre and Douglas, 2004), and then assuming that the person parameters (i.e., the higher-order latent trait and person speed) follow a bivariate normal distribution.

In addition, this study assumes the same probability of correctly answering an item by a RG as by guessing with none of the required attributes for the sake of simplicity. Such a naïve assumption can be further explored as in Wang and Xu (2015). Further, the RT-GDINA-RG distinguishes between solution attempt and RG for cognitive diagnosis via a latent indicator. In addition to RG, RT-GDINA-RG can be easily extended to adapt diverse test-taking behaviors and various tests' requirements. For example, we can extend CDMs to include other test-taking behaviors such as *prior knowledge/pre-knowledge* (Wang et al., 2018; Man et al., 2019) or *nonresponses* (Ulitzsch et al., 2019) if and only if the probabilities of a correct response from different latent indicators (or classes) can be clearly defined. In a high-stakes test, individuals often use pre-knowledge to correctly answer items with extremely short RT (unlike solution attempts with relatively long RT and unlike RGs with often wrong answers and short RT). Furthermore, we can adapt the functions for depicting RT and item response to the testing contexts, such as linear transformation (Wang et al., 2013), a gamma distribution to depict RT for mental rotation items (Maris, 1993), etc. (De Boeck and Jeon, 2019). Also, the item response function can be replaced by other CDMs, such as the GDM (von Davier, 2005) or the linear logistic model (Maris, 1999). Future studies can investigate these approaches.

In addition, ignoring RGs can harm the development and application of cognitive assessments (for both high- and low-stakes tests), distort test results, or invalidate inferences. For example, greater precision of test parameters via the RT-GDINA-RG ensures the quality of item bank construction and assembly of tests, especially for large-scale assessments. Their greater precision also reduces the number of necessary test items to accurately assess a test-taker's domain knowledge, thereby enabling more subdomains to be assessed. The RT-GDINA-RG results regarding time can also inform designers of timed tests regarding the time needed for different solution approaches to a test question. For example, for a timed test, items have frequent RG might because test-takers perceive that they lack sufficient time to attempt a solution. Thus, such information can provide the users of test scores to set a suitable time (e.g., increasing the response time) for completing the test. In addition, greater accuracy in the estimation of test scores increases users' confidence in the results and their subsequent inferences.

When using RT-GDINA-RG to estimate more precise person and item parameters during RG, Q-matrix is an essential component in CDM contexts. An identifiable Q-matrix ensures the consistency of a CDM estimation. In this study, the simulation studies used an identifiable Q-matrix (Xu and Zhang, 2016; Xu, 2017), and the real data analysis adopted a partially identifiable Q-matrix (Gu and Xu, 2020). To enable consistent CDM estimation, checking the identifiability of the Q-matrix in advance is crucial. Besides, for ease of use, a tutorial to introduce the RT-GDINA-RG in JAGS can be developed in future work (cf. Curtis, 2010; Zhan et al., 2019).

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

All authors contributed to the article and approved the submitted version.

REFERENCES

Curtis, S. M. (2010). BUGS code for item response theory. *J. Stat. Softw.* 36, 1–34. doi: 10.18637/jss.v036.c01

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

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

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

Gu, Y., and Xu, G. (2020). Partial identifiability of restricted latent class models. *Ann. Statist.* 48, 2082–2107. doi: 10.1214/19-AOS1878

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

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258—-272. doi: 10.1177/01466210122032064

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

Lee, Y.-S., de la Torre, J., and Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. *Asia Pac. Educ. Rev.* 13, 333–345. doi: 10.1007/s12564-011-9196-3

Man, K., Harring, J. R., Ouyang, Y., and Thomas, S. L. (2018). Response time based nonparametric kullback-leibler divergence measure for detecting aberrant test-taking behavior. *Int. J. Test.* 18, 155–177. doi: 10.1080/15305058.2018.1429446

Man, K., Harring, J. R., and Sinharay, S. (2019). Use of data mining methods to detect test fraud. *J. Educ. Meas.* 56, 251–279. doi: 10.1111/jedm.12208

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables and their application to psychometric models for response times. *Psychometrika* 58, 445–469. doi: 10.1007/BF02294651

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/BF02294535

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Appl. Psychol. Meas.* 34, 521–538. doi: 10.1177/0146621609355451

OECD (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving*. Paris: OECD.

OECD (2017b). *PISA 2015 Technical Report*. Paris: OECD.

Plummer, M. (2017). *JAGS version 4.3 User Manual.*

Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis. *Appl. Psychol. Meas.* 14, 271–282. doi: 10.1177/014662169001400305

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–616. doi: 10.1111/1467-9868.00353

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20, 345—-354. doi: 10.1111/j.1745-3984.1983.tb00212.x

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

Ulitzsch, E., von Davier, M., and Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *Br. J. Math. Stat. Psychol.* 2019:e12188. doi: 10.1111/bmsp.12188

van der Linden, W. J. (2006). A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

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

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *J. Educ. Behav. Stat.* 33, 5–20. doi: 10.3102/1076998607302626

von Davier, M. (2005). *A General Diagnostic Model Applied to Language Testing Data*. ETS Research Report RR-05–16. Princeton, NJ: Educational Testing Service.

Wang, C., Chang, H., and Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

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

Wang, C., Xu, G., Shang, Z., and Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: a comparison of mixture modeling method and residual method. *J. Educ. Behav. Stat.* 43, 469–501. doi: 10.3102/1076998618767123

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

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

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

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

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

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.

APPENDIX

JAGS Code for the RT-DINA-RG and RT-DINO-RG Models

```
###################
### RT-DINA-RG###
###################
RT-DINA-RG.model
{
    for (i in 1:N) {
      for (k in 1:5) { # five attributes
            alpha[i,k] ~ dbern(ap[k])}
      tau[i] ~ dnorm(0, Inv_sigma2)
      for (j in 1:J) {
            xi[i,j] ~ dbern(pii[j]) # 0 = RG, 1 = normal
            eta[i,j] < -
                  pow(alpha[i,1],Q[j,1])*pow(alpha[i,2],Q[j,2])*pow(alpha[i,3],Q[j,3])*p
                  ow(alpha[i,4],Q[j,4])*pow(alpha[i,5],Q[j,5])
            prob[i,j] < - pow(1-slip[j],eta[i,j]*xi[i,j])*pow(guess[j],1-eta[i,j]*xi[i,j])
            r[i,j] ~ dbern(prob[i,j])
            rt.mu[i,j] < - (1-xi[i,j])*beta.0 + xi[i,j]*(beta[j] - tau[i])
            rt_kappa[i,j] < - (1-xi[i,j])*kappa2.0 + xi[i,j]*kappa2[j]
            RT[i,j] ~ dlnorm(rt.mu[i,j], rt_kappa[i,j])}}
    # Priors
    for (k in 1:5) {
      ap[k] ~ dunif(0, 1)}
    for (j in 1:J) {
      pii[j] ~ dunif(0, 1)
      slip[j] ~ dunif(0, 0.5)
      guess[j] ~ dunif(0, 0.5)
      beta[j] ~ dnorm(3, 0.1)
      kappa2[j] ~ dgamma(0.1, 0.1)
      kappa[j] < - sqrt(kappa2[j])}
    Inv_sigma2 ~ dgamma(0.1, 0.1)
    sigma < - 1/sqrt(Inv_sigma2)
    beta.0 ~ dnorm(0, 0.1) %_% I(,min(beta))
    kappa2.0 ~ dgamma(0.1, 0.1)
    kappa.0 < - sqrt(kappa2.0)}
###################
### RT-DINO-RG###
###################
RT-DINO-RG.model
{
    for (i in 1:N) {
      for (k in 1:5) { # five alphaibutes
            alpha[i,k] ~ dbern(ap[k])}
    tau[i] ~ dnorm(0, Inv_sigma2)
    for (j in 1:J) {
            xi[i,j] ~ dbern(pii[j]) # 0 = RG, 1 = normal
            eta[i,j] < - 1-pow(1-alpha[i,1],Q[j,1])*pow(1-alpha[i,2],Q[j,2])*pow(1-
                  alpha[i,3],Q[j,3])*pow(1-alpha[i,4],Q[j,4])*pow(1-
                  alpha[i,5],Q[j,5])
            prob[i,j] < - pow(1-slip[j],eta[i,j]*xi[i,j])*pow(guess[j],1-eta[i,j]*xi[i,j])
            r[i,j] ~ dbern(prob[i,j])
            rt.mu[i,j] < - (1-xi[i,j])*beta.0 + xi[i,j]*(beta[j] - tau[i])
            rt_kappa[i,j] < - (1-xi[i,j])*kappa2.0 + xi[i,j]*kappa2[j]
            RT[i,j] ~ dlnorm(rt.mu[i,j], rt_kappa[i,j])}}
```

```
    # Priors
    for (k in 1:5) {
      ap[k] ~ dunif(0, 1)}
    for (j in 1:J) {
      pii[j] ~ dunif(0, 1)
      slip[j] ~ dunif(0, 0.5)
      guess[j] ~ dunif(0, 0.5)
      beta[j] ~ dnorm(3, 0.1)
      kappa2[j] ~ dgamma(0.1, 0.1)
      kappa[j] < - sqrt(kappa2[j])}
    Inv_sigma2 ~ dgamma(0.1, 0.1)
    sigma < - 1/sqrt(Inv_sigma2)
    beta.0 ~ dnorm(0, 0.1) %_% I(,min(beta))
    kappa2.0 ~ dgamma(0.1, 0.1)
kappa.0 < - sqrt(kappa2.0)}
```

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 February 2021
doi: 10.3389/fpsyg.2020.579199

Integrating a Statistical Topic Model and a Diagnostic Classification Model for Analyzing Items in a Mixed Format Assessment

Hye-Jeong Choi[1], Seohyun Kim[2], Allan S. Cohen[1], Jonathan Templin[3] and Yasemin Copur-Gencturk[4]*

[1] Georgia Center for Assessment, Department of Educational Psychology, University of Georgia, Athens, GA, United States, [2] Department of Psychology, University of Virginia, Charlottesville, VA, United States, [3] Department of Psychological and Quantitative Foundations, University of Iowa, Iowa City, IA, United States, [4] Rossier School of Education, University of Southern California, Los Angeles, CA, United States

Selected response items and constructed response (CR) items are often found in the same test. Conventional psychometric models for these two types of items typically focus on using the scores for correctness of the responses. Recent research suggests, however, that more information may be available from the CR items than just scores for correctness. In this study, we describe an approach in which a statistical topic model along with a diagnostic classification model (DCM) was applied to a mixed item format formative test of English and Language Arts. The DCM was used to estimate students' mastery status of reading skills. These mastery statuses were then included in a topic model as covariates to predict students' use of each of the latent topics in their written answers to a CR item. This approach enabled investigation of the effects of mastery status of reading skills on writing patterns. Results indicated that one of the skills, Integration of Knowledge and Ideas, helped detect and explain students' writing patterns with respect to students' use of individual topics.

Keywords: text analysis, mixed format test, diagnostic classification model, structural topic model, statistical topic models

OPEN ACCESS

Edited by:
Feiming Li,
Zhejiang Normal University, China

Reviewed by:
Hao Song,
Other, United States
Jiwei Zhang,
Yunnan University, China
Kirk Becker,
Pearson (United States), United States

****Correspondence:***
Hye-Jeong Choi
hjchoi1@uga.edu

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

Received: *01 July 2020*
Accepted: *21 December 2020*
Published: *09 February 2021*

Citation:
Choi H-J, Kim S, Cohen AS, Templin J and Copur-Gencturk Y (2021) Integrating a Statistical Topic Model and a Diagnostic Classification Model for Analyzing Items in a Mixed Format Assessment. Front. Psychol. 11:579199. doi: 10.3389/fpsyg.2020.579199

INTRODUCTION

Selected response (SR; e.g., multiple choice or true–false) items and constructed response (CR; e.g., short answer, long answer essay, or performance) items are often found in the same test. An important benefit of SR items is their efficiency in being scored quickly with minimal potential for raters' bias. CR items, on the other hand, have been shown to be appropriate for assessing certain types of higher order knowledge, as this type of item can be used to require students to construct their answers and frequently show their reasoning in their answers (Brookhart, 2010).

While SR and CR items are used together, existing psychometric approaches do not benefit from both data sources efficiently. Most psychometric models, including item response theory models and diagnostic classification models (DCMs), have been developed for focusing on item scores, i.e., correctness of the responses. This is true for CR items as well. The partial credit model (Masters, 1982) and the general diagnostic model (von Davier, 2008), for example, can be used for CR items, but these models only focus on item scores and do not directly include analysis of students' constructed responses, when estimating

model parameters. As a result, any additional information contained in the text of students' answers is ignored.

Statistical topic models (Blei, 2012), on the other hand, are designed to detect the latent thematic structure in the textual data. In education, topic models have recently used. For example, Daenekindt and Huisman (2020) used a topic model to investigate trends in research topics in higher education by analyzing journal abstracts. Moretti et al. (2015) explored the use of different topics on teacher evaluation policy by examining research articles found on the internet. Kim et al. (2017) investigated growth and change in use of academic vocabulary as a result of an instructional intervention, and Duong et al. (2019) found that students' differential use of topics in their CR answers reflected differences in students' reasoning associated with differences in the instructional training of their teachers.

In this study, we present an approach in which results from a DCM were used in a topic model as covariates to understand the relationship between students' mastery status of reading skills and the latent thematic structure in students' writing to answer to a CR item. Specifically, a log-linear cognitive diagnostic model (LCDM; Henson et al., 2009) was used as a DCM and a structural topic model (STM; Roberts et al., 2013) was used as a topic model. This combined use of the two models enabled direct investigation of the relationships between mastery of reading skills and use of latent topics. In the next section, we describe the LCDM and the STM.

LOG-LINEAR COGNITIVE DIAGNOSIS MODEL

Diagnostic classification models (Rupp et al., 2010) are probabilistic models developed to obtain information regarding students' mastery status on a set of pre-determined skills. DCMs predict response patterns for individual mastery profiles based on the attribute structure given in the Q-matrix for a test. In this way, the DCM provides a deterministic confirmatory framework for the assessment. The DCM also include the capability of accounting for uncertainties in examinees' behavior on a test, such as guessing or slipping. Several models have been proposed by imposing different conditions for determining the probability of answering the item correctly and handling these kinds of sources of uncertainty.

As a general frame of reference for a DCM, in the LCDM, the probability of getting a correct answer is modeled as a function of item (j) parameters and the mastery status of the individual (i) given the Q-matrix as follows (Henson et al., 2009):

$$P(Y_{ij} = 1|\alpha_i, q_j) = \frac{\exp\left[\lambda_{j0} + \lambda_j^{\mathrm{T}} h\left(\alpha_i, q_j\right)\right]}{1 + \exp\left[\lambda_{j0} + \lambda_j^{T} h\left(\alpha_i, q_j\right)\right]},$$

where λ_{j0} indicates the intercept, λ_j represents a vector of coefficients indicating effects of the mastery of attributes on the response for item j, and $h\left(\alpha_i, q_j\right)$ is a vector of linear combinations of the α_i and q_j, which specifies an effect structure of the model. $h\left(\alpha_i, q_j\right)$ can include main effect of each attribute, two-way interactions, three-way interactions, etc., depending on how many attributes there exist in the test. For instance, if the effect structure includes only main effects and two-way interactions, the model can be represented as

$$\lambda_j^T h\left(\alpha_i, q_j\right) = \sum_{s=1}^{S} \lambda_{js}\left(\alpha_s q_{js}\right) + \sum_{s=1}^{S}\sum_{u>s}^{S} \lambda_{jsu}\left(\alpha_s \alpha_u q_{js} q_{ju}\right)$$

where λ_{js} represents the main effects of attribute s on item j and λ_{jsu} represents the two-way interaction effects between the combination of attributes s and u on item j. As indicated earlier, this can be extended to three-way or more interaction terms, if needed. Due to the flexibility of this effect structure, the LCDM provides a general framework for DCMs. Further, one can investigate whether the relationship among attributes is compensatory or non-compensatory. For example, using a significance test for λ_{jsu} without predetermining the magnitude of the relationship of the two attributes s and u on item j, the relationship between attributes s and u on item j can be tested.

STRUCTURAL TOPIC MODEL

Topic models are statistical models designed to extract the latent topic structure in a collection of documents (Blei et al., 2003; Griffiths and Steyvers, 2004). Latent Dirichlet allocation (LDA; Blei et al., 2003) is one of the simplest topic models. It assumes that each document in a corpus is a mixture of topics, and each topic is assumed to have a multinomial distribution over a fixed vocabulary of words. A topic is defined as a mixture over words, where each word has a separate probability of belonging to each topic in the model and each document is assumed to consist of a mixture of topics. In LDA, the topics are latent variables to be inferred from the words in a corpus which are the observed variables. In LDA, the order of the words and the grammatical role of the words in the text are ignored. This is called the "bag of words" assumption (Blei et al., 2003).

Roberts et al. (2013) proposed the STM as an extension of the LDA in which a document-level covariate structure can be included to help detect the latent topics in the corpus of textual data. In the STM, one or more covariates can be added to predict the topic proportions or the word probabilities, or both. In the current study, we focused on the use of covariates for predicting topic proportions. To this end, the generative process for estimating topic proportions with an STM is defined to include a covariate structure for the topic proportions for the document (θ) as follows (Roberts et al., 2013):

- For each document, d:
  - Draw the topic proportions for the document $(\theta_d) \sim$ LogisticNormal (μ, Σ)
    - $\mu_{d,k} = X_d \gamma_k$
    - $\gamma_k \sim N\left(0, \sigma_k^2\right)$
- For each word in the document, $[n \in (1, \cdots, N_d)]$
  - Draw word's topic assignment $(z_{d,n}) \sim$ Multinomial (θ_d)

- Conditioning on the topic chosen, draw an observed word from that topic ($w_{d,n}$) ~ Multinomial ($\beta_{k=z_{d,n}}$)

where X, γ, and Σ are covariates, coefficients, and the covariance matrix, respectively. The coefficients for topic k (γ_k) follow normal distributions (mean = 0 and variance = σ_k^2). θ_d denotes a vector for topic proportion for a document, $\beta_{k=z_{d,n}}$ denotes a vector for word probabilities, and d denotes a document that is a sequence of N words ($w_{d,n}$). The inclusion of one or more covariates allows the model to borrow strength from documents with similar covariate values for estimating the document proportion (Roberts et al., 2013). In the current study, we investigated the relationship between students' reading ability and students' writing ability by using an STM in which students' mastery status of reading skills was used as covariates to help explain the use of topics in writing.

For the current study, the model was set to run for a maximum of 500 EM iterations and convergence was monitored by setting convergence tolerance 0.00001. We used the default options for priors for γ and Σ. **Figure 1** depicts the model used in the current study.

READING AND WRITING ASSESSMENT

Integrated assessments have been used in assessing English language proficiency to enhance the authenticity and validity of assessment (Read, 1990; Feak and Dobson, 1996; Weigle, 2004; Plakans, 2008; Weigle and Parker, 2012). In a typical integrated assessment, students read one or more passages and use the information from the passages as source material to respond to the item. Some borrowing of material is considered appropriate (e.g., used as source material for the answer) but simply copying is not considered appropriate (Weigle and Parker, 2012).

Reading interventions have been shown to help improve students' writing performance (Graham et al., 2018). Reading and writing skills, although connected, are cognitively separate (Fitzgerald and Shanahan, 2000; Deane et al., 2008; Schoonen, 2019). In this study, the STM topic model along with the LDCM was used to investigate the relationships between reading attributes and writing ability.

MATERIALS AND METHODS

Data and the Q-Matrix

The data consisted of responses of 2,323 Grade 8 students' responses to the argumentative genre of an English and Language Arts (ELA) test. The test was designed to provide formative information on how well students understood concepts and could demonstrate their knowledge in reading and writing.

Skills Measured

The test consisted of five items: three multiple choice items, one short answer (SA) item, and one extended response (ER) item to measure reading and writing ability. Two scores were assigned for the ER item. A confirmatory factor analysis supported this two-factor model: the multiple choice and SA items formed one factor, reading ability, and the two scores for the ER item measured the other factor, writing ability. A non-linear internal consistency estimate (Green and Yang, 2009; Kim et al., 2020) for this two-factor assessment was 0.83, suggesting acceptable reliability (Kline, 2000, p.13).

The multiple choice and SA items were designed to measure three skills: identifying key ideas (*Idea*), identifying the structure of a text (*Structure*), and integrating knowledge of ideas (*Integration*). These three skills were used to create the entries in the Q-matrix shown in **Table 1**. Three items required a single attribute to answer and one item required two attributes to answer. For the item designed for two attributes, the main effect of each attribute and two-way interactions between these two attributes were identified in the effect structure in the LCDM. *Mplus version 7.4* (Muthén and Muthén, 1998–2017) was used to estimate the LCDM.

To measure writing ability, the ER item consisted of two passages: one passage was about environmental facts and the other was about economic facts. Students were instructed to write an argumentative essay indicating whether their congressional representative should allow the protected forest to be developed into commercial timberland and to support their argument with information from each of the passages. The rubric based score of this item ranged 0–7 points. Partial credit was awarded if part of the response was correct (See **Appendix A** for the rubric). In the current study, students' written responses to this item were used to estimate the latent topic structure using the STM as described in more detail in the next section.

Fitting the Topic Model

The STM topic model was used to identify latent topics in students' written responses to the ER item and investigate the relationship between reading and writing ability. The first step in applying any topic model is to preprocess the text. This is

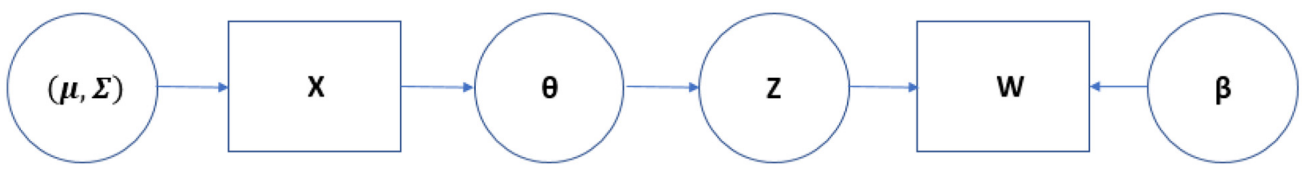


FIGURE 1 | Schematic presentation of structural topic model for this study. Note. The squares indicate the observed variables [i.e., students' mastery status of reading skills (X) and words in the corpus of students' writing (W)]; the circles indicate the parameters in the model.

done to help the estimation process and improve interpretability of the resulting model (Schofield et al., 2017). Preprocessing consists of (1) removing stopwords and (2) stemming words. Stop words are high-frequency but low-information words such as *a, the, that, it, be* (*am, are, is, were, have been, etc.*), *but, or, etc.* Stemming consists of converting words to their root form. For instance, all verbs were converted to the present tense, plural forms were converted to singular form, words that have similar morphology (e.g., do, doing, and done) were converted to a root form such as do, and typographical errors were corrected.

After stemming words and removing stopwords, words with a frequency of less than 10 and documents with less than 15 words were excluded. In addition, documents with a score of 0 were excluded as this indicated the responses were not scorable. As shown in **Appendix A**, reasons for non-scorability included being blank, simply copying from the passages, answers were written in a language other than English, and answers were too limited, off topic or generally non-responsive to the prompt. The final data set included 2,108 students' responses with a total of words 270,405 in the corpus. The number of unique words was 891 and the average answer length was 128.3 words (SD = 76.4 words).

The next step was to determine how many latent topics appeared in the data. This is an exploratory analysis. That is, we estimated STM models with from 2 to 20 topics as candidate models. For the STM, students' mastery statuses on each attribute were included as a set of document-related covariates for predicting the use of topics. There is no single best method for determining the best fitting topic model. Roberts et al. (2014) suggested use of semantic coherence (Mimno et al., 2011) and exclusivity (Bischof and Airoldi, 2012). These two measures are complementary. These indices were used in this study to inform the selection of the best fitting topic model. In addition, the cosine similarity (Cao et al., 2009) between topics was estimated. The lower cosine similarity indicates better fit as this indicates topics are distinct each other. The R package *stm* (Roberts et al., 2019) was used to estimate the STM.

RESULTS

Students' Reading Skill Profiles

For item 4, as no significant interaction effect for attributes 1 and 3, the interaction term was dropped from the effect structure in the final LCDM model. **Table 2** presents item parameter estimates for the final model. All main effects were significant at $p < 0.01$. Intercepts for items 1 and 3 were significant ($p < 0.01$), but the intercepts for items 2 and 4 were not. **Table 3** presents students' mastery profiles of the reading skills, the marginal proportions, and reliabilities for each of the skills. Skill reliabilities were relatively low, reflecting the small number of items measuring each skill. The correlation between *Idea* and *Structure* was 0.86, the correlation between *Idea* and *Integration* was 0.67, and the correlation between *Structure* and *Integration* was 0.57. These indicated substantial relationships between skills. Eight different mastery profiles are possible for the three skills in the Q-matrix. Results in **Table 3**, however, indicate that only four of the eight profiles were detected. These included students who had mastered none of three skills (0,0,0), students who had mastered only *Integration* (0,0,1), students who had mastered *Idea* and *Integration* (1,0,1), and students who had mastered all three skills (1,1,1). Students' mastery statuses for each attribute obtained by this analysis were included in the STM as covariates to predict the use of topics.

TABLE 1 | Q-Matrix of three reading skills for the multiple-choice and short response items.

Item	Idea	Structure	Integration
Multiple-choice item 1	x		
Multiple-choice item 2		x	
Multiple-choice item 3	x		
Short answer item 4	x		X

TABLE 2 | Item parameter estimates for the log-linear cognitive diagnostic model for students' reading skills.

Item	Intercept	Main effect		
		Key idea	Craft and structure	Integration
Multiple-choice item 1	−0.613	1.557	–	–
Multiple-choice item 2	*	–	3.370	–
Multiple-choice item 3	0.434	1.967	–	–
Short answer item 4	*	6.004	–	0.924

**indicates no significance with a significance level of 0.01 and – indicates not applicable given the item.*

TABLE 3 | Students' mastery status of reading skills and reliability of each skill.

Profile*	Key ideas	Craft and structure	Integration of knowledge and ideas	Count (%)
1 (000)	0	0	0	323 (13.90)
2 (001)	0	0	1	296 (12.74)
3 (010)	0	1	0	0 (0.00)
4 (011)	0	1	1	0 (0.00)
5 (100)	1	0	0	0 (0.00)
6 (101)	1	0	1	146 (6.28)
7 (110)	1	1	0	0 (0.00)
8 (111)	1	1	1	1,558 (67.07)
Marginal proportion**	66%	59%	51%	2,323 (100.00)
Skill reliability	0.69	0.62	0.51	

**0 indicates being classified non-mastery and 1 indicates being classified mastery.*
***Marginal proportion of students who have mastered each skill.*

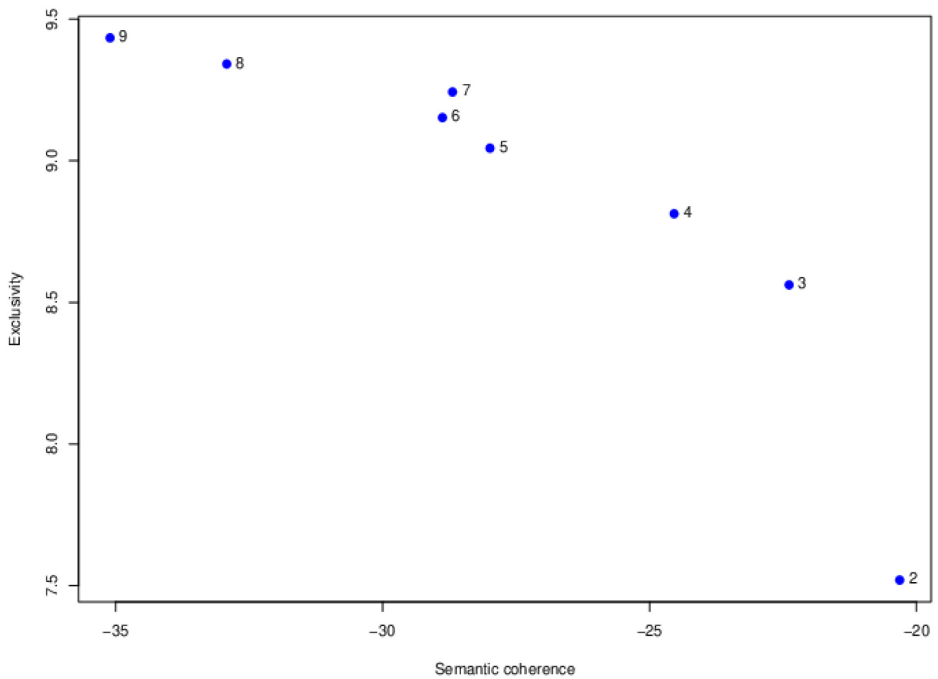


FIGURE 2 | Comparisons semantic coherence and exclusivity among 2- to 9-topic structural topic models.

Selection of the Topic Model and Interpretation of Topics

To detect the number of topics, STM models with from 2 to 20 topics were fit to the data as an exploratory analysis. As described in Methods section, semantic coherence, exclusivity, and cosine similarity were used to determine the number of topics. **Figure 2** presents the results of semantic coherence and exclusivity for each of the model with from two to nine topics. The horizontal axis is semantic coherence and the vertical axis is exclusivity. Models in the upper right corner would be models that are higher in both semantic coherence and exclusivity. The best models based on these two indices would be the three- and four-topic models. Cosine similarity results suggested the four-topic model was a better fit than the three-topic model. Based on these results, the four-topic model was selected as the best-fit model.

One way to help interpret and characterize each topic in the model is to examine (1) written responses of students who were the highest probability users of each topic and (2) the highest probability words for each topic. The 15 highest probability words in each topic for the four-topic STM are listed in **Table 4**. The answer of the student who was the most frequent user of words from each topic is presented below. **The bold and underlined words are the highest frequency words for the given topics**.

In the first topic, the highest frequency words were *pollution*, *paper*, *mill*, *industry*, *coastal*, and *water* (*Pollution* was used as a stemming word for *pollution*, *polluter*, and *pollutant*). These words come from the prompt (i.e., either the two passages in the prompt or the stem of the SR items). Students had been instructed to use information from the passages to support their arguments. This topic was labeled *Integrative Borrowing* as it reflected this use of the terms in the prompt. The following is the answer of the student who was the most frequent user of words from this topic.

> (Integrative Borrowing) **Paper mills** are having a negative effect. Passage A says "**Paper mills** are the third largest **polluters** in the United States., releasing **pollutants** into

TABLE 4 | The 15 high frequent words in each topic detected from the 4-topic STM.

Topic 1	Topic 2	Topic 3	Topic 4
Integrative borrowing	Everyday language	Copying from passage	Copying from stem
Pollution	Tree	Georgia	Forest
Paper	Down	Timber	Protect
Mill	For	Acre	Timber
Industry	Cut	More	Should
Coastal	If	Forest	Commercial
Plain	Animal	Coastal	Animal
Water	Make	Plain	Plant
Georgia	Can	For	Species
Passage	More	Pine	Representative
Fish	Because	Commercial	Allow
Cause	People	Industry	Because
Environment	Need	Grow	Develop
Due	Go	Year	Congress
Provide	Get	Passage	Destroy
Forest	Land	Land	Live

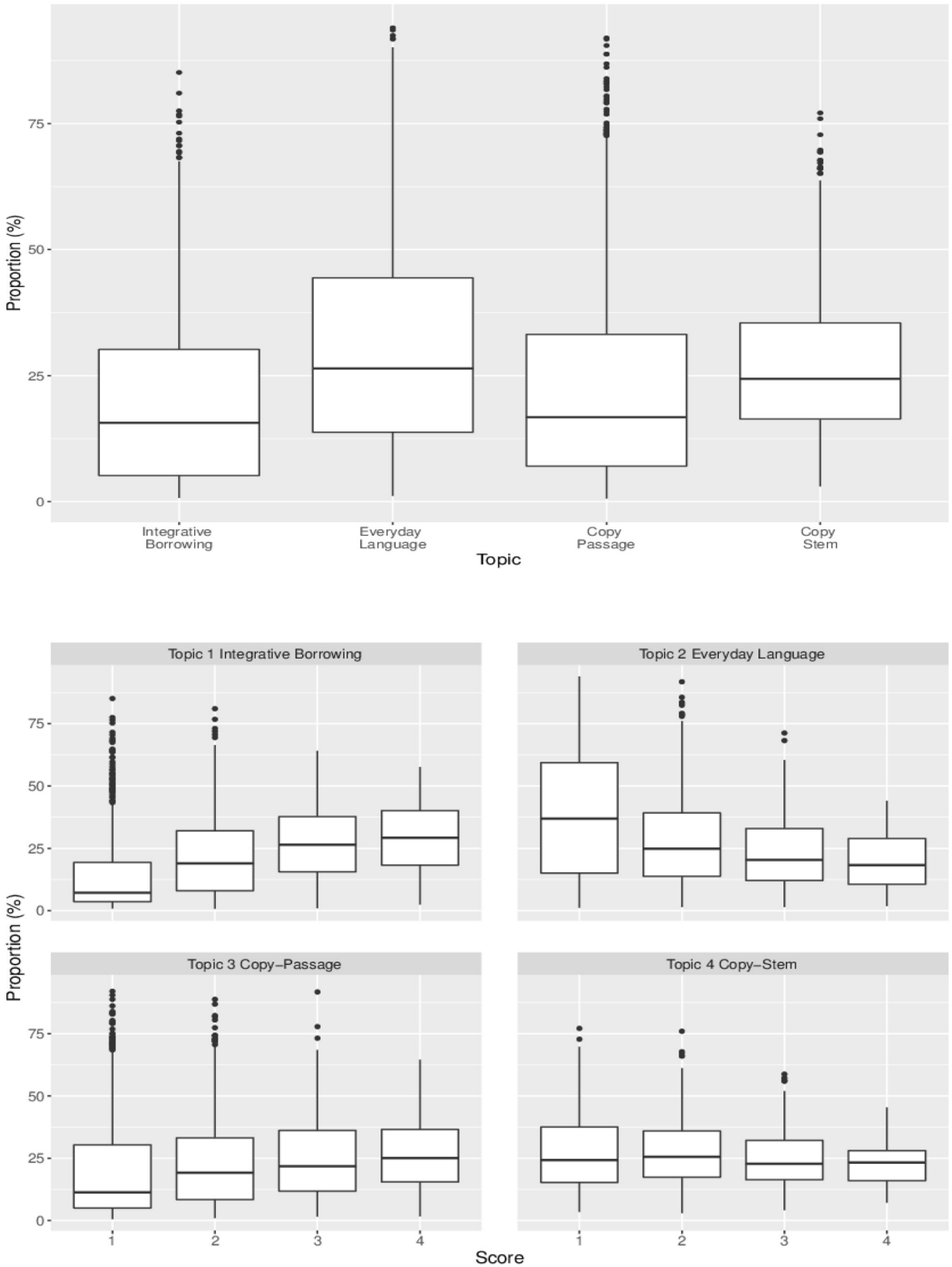


FIGURE 3 | Boxplot for topic proportion distribution. The plot on the upper panel presents the distributions of marginal topic proportions. The plots on the lower two panels present score distributions for each topic. For the plot on the lower panel, the X-axis indicates each score point and the Y-axis indicates the proportions of use of each topic. The whiskers on the boxes indicate variability outside the upper and lower quartiles and the horizontal lines in the boxes indicate the mean usage of the topic for the given score point.

> the air, **water**, and soil." Passage A also say that "many **paper mills** are working to reduce the amount of **pollutants** they produce today." But they are letting it out, and it also effecting in passage B it says that "the fishing **industry** decreases due to **pollution** caused by **paper mills**." That why I think **paper mills** are having a negative effect.

The highest frequency words in the second topic were *tree*, *animal*, *cut*, and *people*. These words reflect use of everyday language but not directly related to the question. This topic was labeled as *Everyday Language*. The following is the answer of a student who was a high frequency user of words from topic 2.

> (Everyday Language) No because they are killing all the plants and taking the **animals** homes away so how would you feel if someone just took your house away and built something else and just put your family out on the street with nowhere to go, that's how the **animals** feel. Your destroying our plants life to that we need the plants and **animals** were there first and they really don't have any other home to go to besides a zoo why do that when they can just be free without the **people** harming them.

The highest frequency words in the third and fourth topics were both borrowed directly from either the stem or the passages. The words in the third topic were copied from the passages (*Georgia*, *timber*, *acre*, *coastal*, *plain*, and *pine*). (*Timber* was used as a stemming word for *timber*, *timberland*, and *timberwood*). The words in the fourth topic were copied from the stem (*forest*, *protect*, *timber*, *should*, *commercial*, *representative*, *allow*, and *congress*). The followings are answers of students who were the highest frequency users of words from the third and the fourth topics, respectively. Characteristic of users of topics 3 and 4 is that these words were simply copied from the passage or stem without any clear effort to integrate the words into the argument.

> (Copying from Passages) I think that the small protected forest should not be developed into commercial **timberland** because you don't have a lot of land. The text states in passage B that "Sixty percent of **Georgia's coastal** plain is covered in forest. The forest is one of the most diverse ecosystems in America and includes forest, grassland, sandhill, marsh, swamp, and **coastal** habitats. Several varieties of pine and oak are the most common trees. The growth of the ground under the long leaf **pine** forest contains 150–300 plant species per **acre**, more birds than any other **Georgia** forest type, and 60% of the amphibian and reptile species found in the Southeast. The **Georgia** state reptile, the gopher tortoise, lives in **pine** forest habitats and is a key species in the ecosystem. Though once an endangered species, the American alligator is now very common, numbering an estimated 2 million in the Southeast." This shows that the forest has already been occupied by one of the most diverse ecosystems in America and includes many plants and many amphibian and reptiles. In conclusion this is why I feel like the small protected forest should not be developed into commercial **timberland**.

> (Copied from Stem) The **representative should** not **allow** the **protected forest** to be developed into the **commercial timberland**. They **shouldn't** because, in passage B it states that the soil isn't suitable for any kind of **forest**. The **timberland** is worth an average of $97 a year because the land isn't suitable for the tree's and soil. That is why you **shouldn't** allow them to put the **protected forest** in the **timberland**.

Figure 3 presents box plots of students' use of individual topics. The plot on the upper panel indicates that overall, students used 20, 31, 22, and 27% of Topics 1, 2, 3, and 4, respectively. The plots on the lower two panels show the rubric based score distribution for each topic. There are two distinct patterns in the **Figure 3**: (1) students who used more *Integrative Borrowing* in their answers tended to have higher scores and (2) students who used more *Everyday Language* in their answers tended to have lower scores.

What Is the Relationship Between Students' Mastery Status of Reading Skills and the Use of the Latent Topics in Writing?

Table 5 presents results for the effects of students' mastery status of reading skills on their use of each of the four topics in the STM.

TABLE 5 | Results of STM for predicting the use of topics by mastery status of reading skills.

	Estimate	SE	*t*-test	Pr(> | t|)
Topic 1: Integrative borrowing				
(Intercept)	0.11	0.012	9.01	0.00
Key ideas	0.03	0.020	1.32	0.19
Craft and structure	0.02	0.017	0.98	0.33
Integration of knowledge and ideas	0.07	0.018	3.77	0.00
Topic 2: Everyday language				
(Intercept)	0.49	0.015	32.54	0.00
Key ideas	−0.04	0.023	−1.73	0.08
Craft and structure	−0.03	0.020	−1.45	0.15
Integration of knowledge and ideas	−0.15	0.020	−7.27	0.00
Topic 3: Copying from passage				
(Intercept)	0.16	0.014	11.76	0.00
Key ideas	0.02	0.023	0.90	0.37
Craft and structure	0.01	0.020	0.35	0.73
Integration of knowledge and ideas	0.05	0.019	2.60	0.01
Topic 4: Copying from stem				
(Intercept)	0.24	0.011	21.77	0.00
Key ideas	−0.01	0.018	−0.42	0.67
Craft and structure	0.01	0.015	0.34	0.74
Integration of knowledge and ideas	0.03	0.015	2.03	0.04

The values in **Table 5** indicate the coefficients for the intercept and for each of the three skills estimated from the DCM. The intercept can be interpreted as the expected use of the topic when students do not master any skills at all, and other coefficients can be interpreted as the expected use of the topic when students master individual skills.

The results indicate mastery status of either *Key Ideas* or *Craft and Structure* did not have a significant impact on students' use of the topic. *Integration of Knowledge and Ideas* (*Integration*) was the only skill that had a significant effect on the use of each topic at $p < 0.05$. As seen in **Figure 3**, *Integrative Borrowing* and *Everyday Language* tended to be related to the rubric based score. The results in **Table 5** show similar patterns. This suggests that when students master the *Integration*, their probability of using the integrative borrowing topic increases by 0.07, their probability of using the copying from passage topic increases by 0.05, their probability of using the copying from stem topic increases by 0.05, but their probability of using the everyday language topic decreases by 0.15.

CONCLUSION

In this study, an approach was described a topic model to obtain the latent thematic structure in students' written answers to an ER item. In the topic model, results from a DCM applied to the item scores (i.e., the correctness of students' answers) were included as covariates to predict students' use of the topics. Although three skills were identified in the Q-matrix, only four of the eight possible mastery profiles were present in the data. The four-topic STM was found to be the best fit to the textual data from the students' answers to the test questions along with students' reading skills as covariates. The results showed that mastery status of *Integration of Knowledge and Ideas* was the pivotal skill for the use of each of the four topics. That is, as students mastered *Integration of Knowledge and Ideas*, they tended to use more of the *Integrative Borrowing* topic in their writing and less of the *Everyday Language* topic. CR or ER items are often used to assess higher-order thinking skills. Rubric-based scores provide useful information regarding students' knowledge status with respect to the objectives being measured on the test. There is also information about students' thinking and reasoning as reflected in their answers, however, that can be missed by the rubric-based scores alone (Cardozo-Gaibisso et al., 2020). For example, each topic could represent a set of possible misconceptions (Shin et al., 2019) or writing style.

The assessment used in this study was a formative assessment and was not specifically designed to fit a DCM model. Due to the small number of items in the assessment, the skill reliabilities were relatively low, which is a possible limitation of this study. Even with this limitation, however, results demonstrate that combining results from a DCM with a topic model enables the possibility of investigating the relationship between the knowledge as measured by the multiple choice items and cognitive skills used in answering to the CR items. Topic modeling is relatively new in educational research, but it has been found to provide a useful set of methodological tools for extracting this added information in the text of answers to CR items.

Some of current techniques developed in natural language processing or machine learning may not be applicable for the text in education as the text in education may have different characteristics from the text in social networks or publications. Further studies would be helpful to address important issues in this area, such as what could be the effects of stemming methods on latent topic structure or what methods could be used for selecting the best fitting topic model.

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 Human Subjects Office, University of Georgia. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was partially supported by the National Science Foundation (Award No. DRL-1813760).

REFERENCES

Bischof, J., and Airoldi, E. (2012). "Summarizing topical content with word frequency and exclusivity," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, Edinburgh, 201–208.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* 55, 77–84. doi: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Brookhart, S. M. (2010). *How to Assess Higher-Order Thinking Skills in Your Classroom*. Alexandria, VA: ASCD.

Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing* 72, 1775–1781. doi: 10.1016/j.neucom.2008.06.011

Cardozo-Gaibisso, L., Kim, S., Buxton, C., and Cohen, A. (2020). Thinking beyond the score: multidimensional analysis of student performance to inform the next generation of science assessments. *J. Res. Sci. Teach.* 57, 856–878. doi: 10.1002/tea.21611

Daenekindt, S., and Huisman, J. (2020). Mapping the scattered field of research on higher education. A correlated topic model of 17,000 articles, 1991–2018. *High. Educ.* 80, 571–587. doi: 10.1007/s10734-020-00500-x

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., and Bivens-Tatum, J. (2008). Cognitive models of writing: writing proficiency as a complex integrated skill. *ETS Res. Rep. Series* 2008, i–36. doi: 10.1002/j.2333-8504.2008.tb02141.x

Duong, E. V., Mellom, P., Weber, J., Gokee, R., and Cohen, A. S. (2019). Examining the impact of instructional conversation on students' writing through topic model analysis. *Paper Presented at the Annual Meeting of the American Educational Research Association*, Toronto, CA.

Feak, C., and Dobson, B. (1996). Building on the impromptu: a source-based academic writing assessment. *Coll. ESL* 6, 73–84.

Fitzgerald, J., and Shanahan, T. (2000). Reading and writing relations and their development. *Educ. Psychol.* 35, 39–50. doi: 10.1207/s15326985ep3501_5

Graham, S., Liu, X., Bartlett, B., Ng, C., Harris, K. R., Aitken, A., et al. (2018). Reading for writing: a meta-analysis of the impact of reading interventions on writing. *Rev. Educ. Res.* 88, 243–284. doi: 10.3102/0034654317746927

Green, S. B., and Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: an alternative to coefficient alpha. *Psychometrika* 74, 155–167. doi: 10.1007/s11336-008-9099-3

Griffiths, T. L., and Steyvers, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. U.S.A.* 101(Suppl. 1), 5228–5235.

Henson, R., Templin, J., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., and Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *J. Writ. Anal.* 1, 82–102.

Kim, S., Lu, Z., and Cohen, A. S. (2020). Reliability for tests with items having different numbers of ordered categories. *Appl. Psychol. Meas.* 44, 137–149. doi: 10.1177/0146621619835498

Kline, P. (2000). *The Handbook of Psychological Testing*. Hove: Psychology Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/bf02296272

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (Stroudsburg, PA: Association for Computational Linguistics), 262–272.

Moretti, A., McKnight, K., and Salleb-Aouissi, A. (2015). *Application of Sentiment and Topic Analysis to Teacher Evaluation Policy in the US*. Available online at: https://www.educationaldatamining.org/EDM2015/uploads/papers/paper_310.pdf (accessed January 20, 2021).

Muthén, L. K., and Muthén, B. O. (1998–2017). *Mplus (Version 7) [Computer Software]*. Los Angeles: Muthén & Muthén.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writ.* 13, 111–129. doi: 10.1016/j.asw.2008.07.001

Read, J. (1990). Providing relevant content in an EAP writing test. *English Specific Purp.* 9, 109–121. doi: 10.1016/0889-4906(90)90002-t

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). stm: an r package for structural topic models. *J. Stat. Softw.* 91, 1–40. doi: 10.18637/jss.v091.i02

Roberts, M. E., Stewart, B. M., Tingley, D., and Airoldi, E. M. (2013). *The Structural Topic Model and Applied Social Science. Workshop Presented at the NIPS Workshop on Topic Models: Computation, Application and Evaluation*. Available online at: https://scholar.princeton.edu/files/bstewart/files/stmnips2013.pdf (accessed January 20, 2021).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., et al. (2014). Structural topic models for open-ended survey responses. *Am. J. Polit. Sci.* 58, 1064–1082. doi: 10.1111/ajps.12103

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

Schofield, A., Magnusson, M., and Mimno, D. (2017). "Understanding text pre-processing for latent Dirichlet allocation," in *Proceedings of the 15th conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2, (Stroudsburg, PA: Association for Computational Linguistics), 432–436.

Schoonen, R. (2019). Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. *Read. Writ.* 32, 511–535. doi: 10.1007/s11145-018-9874-1

Shin, J., Guo, Q., and Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Front. Psychol.* 10:825. doi: 10.3389/fpsyg.2019.00825

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writ.* 9, 27–55. doi: 10.1016/j.asw.2004.01.002

Weigle, S. C., and Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *J. Sec. Lang. Writ.* 21, 118–133. doi: 10.1016/j.jslw.2012.03.004

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.

APPENDIX

Appendix A: Rubric for the extended response item

Appendix A1 |The rubric has two traits: Idea Development, Organization, and Coherence; and Language Usage and Conventions. The scale for Idea Development, Organization, and Coherence ranges from 0 to 4, and the scale for Language Usage and Conventions ranges from 0 to 3.

Trait 1: Idea development, organization, and coherence.

If the student scored 4 points. . .

- Effectively introduced a claim or argument.
- Effectively organized the reasons, using logical reasons and evidence.
- Provided clear, relevant reasons/evidence to support the opinion.
- Acknowledged and developed counter-claims, as appropriate.
- Used linking words and phrases effectively to connect opinions and reasons.
- Maintained a formal style appropriate for the task.
- Provided a strong concluding statement or section.

If the student scored 3 points. . .

- Introduced a claim or argument.
- Included an organizational structure that supported the reasons and evidence.
- Provided reasons, facts, and evidence to develop the claim.
- Attempted to introduce a counter-claim, as appropriate.
- Used some linking words to connect opinions and reasons.
- Used a formal style fairly consistently appropriate for the task.
- Provided a concluding statement or section that follows the argument.

If the student scored 2 points. . .

- Attempted to introduce an opinion or a claim.
- Attempted to provide some organization, but structure sometimes impeded the reader.
- Attempted to provide reasons and facts that sometimes support the opinion, but the reasoning is unclear.
- Made no or little attempt to introduce a counter-claim.
- Used few linking words to connect opinions and reasons.
- Used a formal style inconsistently or the style was inappropriate for the task.
- Provided a weak concluding statement or section that does not support the argument.

If the student scored 1 point. . .

- The student did not include a claim or claims, or the claim must be inferred.
- The organizational structure was not evident, not appropriate, or was formulaic.
- There may not have been sufficient support for the claim (if stated).
- The student made no attempt to introduce a counter-claim.
- Very few, if any, linking words and phrases were used.
- Used an informal style not appropriate for the task.
- There was no conclusion, or the conclusion was not related to the essay.

If the student scored 0 points. . .

- The response was blank, copied, or too brief to score.
- The response was illegible, incomprehensible, or was written in another language.
- The response was off topic, off task, or was offensive.

(Continued)

Appendix A1 |Continued

Trait 2: Language usage and conventions.

If the student scored 3 points. . .

- There was a variety of sentence types for meaning and interest, and sentences were clear and complete.
- Conventions and language were used appropriately.
- Errors in usage and conventions were infrequent and did not interfere with the meaning of the response.

If the student scored 2 points. . .

- There was some variety of sentence types, and most were complete.
- Demonstrated some knowledge of conventions and language.
- Minor errors in usage did not significantly interfere with the meaning of the response.
- If the student scored 1 point. . .
- There were fragments, run-ons, and other sentence structure errors.
- Conventions and language were not appropriate.
- Frequent errors in usage interfered with the meaning of the response.

If the student scored 0 points. . .

- The response was blank, copied, or too brief to score.
- The response was illegible, incomprehensible, or was written in another language.
- The response was off topic, off task, or was offensive.

frontiers in Psychology

METHODS
published: 06 May 2021
doi: 10.3389/fpsyg.2021.509844

Automated Test Assembly for Multistage Testing With Cognitive Diagnosis

*Guiyu Li[1,2], Yan Cai[1], Xuliang Gao[3], Daxun Wang[1] and Dongbo Tu[1]**

[1] School of Psychology, Jiangxi Normal University, Nanchang, China, [2] Department of Curriculum and Instruction, East China Normal University, Shanghai, China, [3] School of Psychology, Guizhou Normal University, Guiyang, China

Computer multistage adaptive test (MST) combines the advantages of paper and pencil-based test (P&P) and computer-adaptive test (CAT). As CAT, MST is adaptive based on modules; as P&P, MST can meet the need of test developers to manage test forms and keep test forms parallel. Cognitive diagnosis (CD) can accurately measure students' knowledge states (KSs) and provide diagnostic information, which is conducive to student's self-learning and teacher's targeted teaching. Although MST and CD have a lot of advantages, many factors prevent MST from applying to CD. In this study, we first attempt to employ automated test assembly (ATA) to achieve the objectives of MST in the application of CD (called CD-MST) via heuristic algorithms. The mean correct response probability of all KSs for each item is used to describe the item difficulty of CD. The attribute reliability in CD is defined as the test quantitative target. A simulation study with the G-DINA model (generalized deterministic input noisy "and" gate model) was carried out to investigate the proposed CD-MST, and the results showed that the assembled panels of CD-MST satisfied the statistical and the non-statistical constraints.

Keywords: cognitive diagnosis, computer multistage test, automated test assembly, cognitive diagnosis modules, heuristic algorithms

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Manqian Liao,
Duolingo, United States
Fabrizio Stasolla,
Faculty of Law, Giustino Fortunato University, Italy

****Correspondence:***
Dongbo Tu
tudongbo@aliyun.com

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

Received: *04 November 2019*
Accepted: *29 March 2021*
Published: *06 May 2021*

Citation:
Li G, Cai Y, Gao X, Wang D and Tu D (2021) Automated Test Assembly for Multistage Testing With Cognitive Diagnosis. Front. Psychol. 12:509844. doi: 10.3389/fpsyg.2021.509844

INTRODUCTION

The computer multistage adaptive test (MST), as a "balanced compromise" between CAT and P&P, not only can provide high measurement accuracy as CAT (Kim et al., 2015) but also can meet the need of test developers to manage test forms and keep test forms parallel. CAT is an item-level adaptive test; however, MST sets a module to manage items and to be adaptive at the module level. MST allows subjects to modify the item answers in the current stage, which is beneficial to reduce the examinees' test anxiety and improve the measurement accuracy. Compared with CAT, MST has many inherent advantages: (1) CAT does not allow examinees to modify item answers, which leads to the lack of test control and generates test anxiety for the examinees. MST can allow examinees to modify their item answers in the current stage, which helps alleviate test anxiety while avoiding measurement mistakes caused by errors. (2) CAT pursues the items with the maximum information during an adaptive stage, which will result from overexposure of items with high information. In contrast, MST can effectively enhance the use rate of item bank and control item exposure rate by constructing several parallel panels. (3) CAT is not good at balancing the non-statistical characteristics of the test [e.g., content constraints,

item types, enemy item (there are clues to the answers between the items), word count, etc.]. MST can manage both statistical and non-statistical characteristics, which can greatly improve content validity and measurement precision. (4) Compared with CAT online testing, MST preassembles a test before performing the test administration, which can help test developers better manage a test. Because of these benefits, many high-stake tests have switched from the CAT mode to the MST mode (Wang et al., 2015), such as the United States National Education Progress Assessment (NAEP), the US Graduate Entrance Examination (GRE), the Program for the International Assessment of Adult Competencies (PIAAC), and other large examinations (Yamamoto et al., 2018).

Currently, the classical test theory (CTT) and the item response theory (IRT) have been widely used in education, psychology, psychiatry, etc. However, both the CTT and the IRT mainly focus on the examinees' trait or competency level, and therefore, they cannot provide further information on the internal psychological processing, processing skills, and cognitive structures hidden behind the results of the test scores (Embretson and Yang, 2013). Unlike the CTT and IRT, which can only provide an examinee's score, cognitive diagnosis (CD) can further report the examinee's knowledge states (KSs), cognitive structures, and other diagnostic information. This feature of CD can help teachers carry out targeted teaching and promote education development. Currently, CD, as a representation of the new generation testing theory, has widely attracted the attention of researchers and practitioners and has become an important area of psychometrics research.

Recently, researchers consider that the cognitive diagnostic model can be applied to the MST (von Davier and Cheng, 2014). It is called CD-MST, a new test mode that combines the advantages of CD and MST. First, it can present items with the function of CD and help test developers to manage a CD test before administering it. Second, CD-MST can provide rich diagnostic information to each examinee and guide students and teachers to self-study, adaptive study, individual teaching, remediation teaching, etc. Third, CD-MST is adaptive in modules, where examinees can review and revise item answers. That is closer to the examination scene and helps to reduce examinees' test anxiety. Finally, the adaptive CD-MST can use fewer items to provide immediate and accurate cognitive diagnostic feedback information, and the advantages of CD-MST are especially highlighted in classroom assessment or practice.

Although CD-MST has many advantages, some problems make its assembly infeasible: (1) Item difficulty index. In MST with the IRT, the item difficulty parameter b can accurately indicate the examinees' traits value θ because they are in the same scale. At this point, MST can use the b parameter to divide the item bank and assemble modules based on item difficulty. However, there is no item difficulty parameter in CD, and item parameters and examinee parameters are not set on the same scale. Even if the reduced reparameterized unified model (R-RUM; Hartze, 2002) has a completeness parameter based on the attribute, it is difficult to describe the item difficulty and to explain the relationship between the attribute master pattern and the item difficulty. Therefore, the key for CD-MST is to develop a new item difficulty index in CD. (2) Information or measurement precision index. MST with the IRT focus on a continuous variable. Fisher information, a typical statistic curving continuous variable, is used to ensure measurement precision and to control measurement errors, but CD measures discrete multidimensional variables, Fisher information is not suitable. In order to ensure the test reliability, accuracy, or to control measurement errors, selecting another robust statistical information index of CD is worth further study.

This study aimed to address this aforementioned issue and to develop a CD-MST framework. The rest of the paper is organized as follows. The MST framework is briefly introduced first. Then, the CD-MST framework is proposed, where two indexes, namely, the item difficulty index and the information (or measurement precision) index based on CD, and the automated test assembly (ATA) method for CD-MST are also proposed. Furthermore, the simulation study and the results were carried out to verify the proposed CD-MST framework. Finally, we discuss the limitations of this study and the further directions of CD-MST.

MST FRAMEWORK

Multistage Adaptive Test

MST is built on several parallel panels. A complete panel includes the following elements: module, stage, and pathway, as shown in **Figure 1**. In MST, the test has three adaptive stages, and each stage contains several modules. Modules are composed by items that are according to certain test specifications and of different levels of item difficulty. In **Figure 1**, 1Medium indicates that the item difficulty of the first stage is moderate; 2Easy, 2Medium, and 2Hard indicate that the item difficulty of the second stage is easy, moderate, and difficult, respectively; and 3Easy, 3Medium, and 3Hard are analogous for the third stage. Panels 1, 2, and N represent the parallel test panels. When the test starts, examinees are randomly assigned to a pre-assembled test panel, and then according to their responses in the first stage, examinees are adapted to the module in the next stage that matches their ability. A series of modules responded by examinees is used to construct a response pathway. Each panel

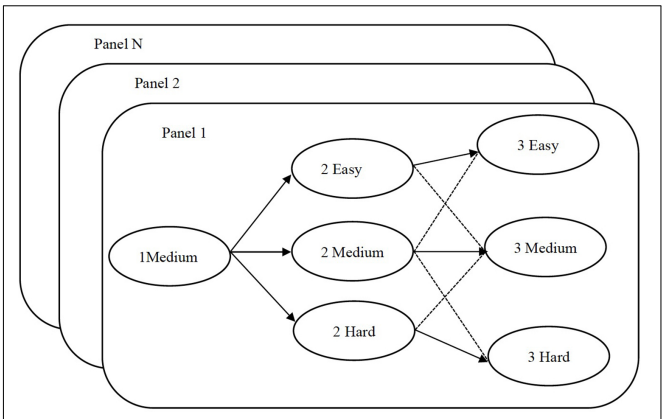


FIGURE 1 | Three-stage multistage adaptive test (MST) of multiple parallel panels.

has seven test pathways, as shown in **Figure 1** (see the arrow's direction in **Figure 1**). Among them, the solid line arrows (e.g., 1Medium + 2Easy + 3Easy, 1Medium + 2Medium + 3Medium, and 1Medium + 2Hard + 3Hard) denote the three primary pathways that examinees are most likely to adapt, whereas the dotted lines denote the four secondary pathways (Luecht et al., 2006).

Parallel test panels are the core of MST. It needs to meet the requirements of the test specifications. Test specifications include both the statistical targets (e.g., test information) and the non-statistical targets (e.g., content constraint), which ensure that each test panel has precise reliability and validity. In MST, the statistical and non-statistical targets mainly relate to the target test information function (TTIF), the test length, the number of stages and modules, the content balance, the exposure control, etc. However, these factors are not independent from each other when building panels, but rather are tightly integrated into the MST architecture (Yan et al., 2014). Like in linear tests, to ensure the safety of the test and the use rate of the item bank, MST researchers hope to set up multiple parallel panels (Samejima, 1977). In linear tests, an item is preliminarily formed into a fixed test form. When test information and other measurement targets are sufficiently similar, it can be assumed that these pre-assembled test forms are parallel. Test pathways in MST are the same as test form in the linear test. However, modules in MST have different difficulty levels; pathways constituted by modules are often not parallel in statistical information. Automated test assembly is a way to achieve parallelism between tests and to meet the test specifications. We build parallel panels according to specific test specifications. When two different pathways in two different panels are parallel, the panels can be viewed as parallel (Yan et al., 2014). It is important to note that when parallel pathways are set up for the test specification, it is not necessary to have parallelism between the modules (Yan et al., 2014).

MST assembly should meet the following three goals: (1) the module has a clear information curve enough to distinguish between the different stages of tests; (2) the information of corresponding pathways between panels is similar to ensure that the panel is parallel; and (3) each pathway of each panel satisfies non-statistical constraints (Yan et al., 2014).

Multistage Adaptive Test Design

The MST design includes the number of panels, the number of stages in panels, the number of modules in stages, the number of items in modules, the level of item difficulty, etc. (Yan et al., 2014). It also involves the assembly strategies and the assembly methods. The assembly strategies determine the item difficulty levels, the content balance, and other elements parallel in modules or pathways. The ATA method ensures that these elements (statistical and non-statistical constraints) are parallel on panels. Statistical constraints are initially determined by the item difficulty and discrimination of the CTT (Gulliksen, 1950), and now, the test information function (TIF) has become the main form of statistical characteristics. The target TIF of IRT usually uses the Fisher information, which was described in detail by Luecht and Burgin (2003). Besides, the statistical constraints of the target TIF need to consider whether the item bank meets test specifications. For example, the quality and the number of items in an item bank are required to provide a great TIF.

Multistage Adaptive Test Assembly Strategies

After the MST design is completed, the parallel panels need to be assembled by using MST assembly strategies, which involve a bottom–up strategy and a top–down strategy (Luecht and Nungester, 1998).

In the top–down strategy, parallel panels are based on the pathways. Several parallel panels are constructed from an item bank, and the corresponding pathways in different panels are parallel. Here, parallel includes the statistical constraints (target TIF) and the non-statistical constraints. The parallel pathways contain two types of pathways, namely, the three primary parallel pathways (see the three thick line pathways in **Figure 1**) and all the parallel pathways. When the three primary parallel pathways are used, the test specification is divided into three primary pathways, and other pathways randomly assemble with the item difficulty. Because the three primary pathways represent the majority answer pathways of examinees, the panels only need to ensure that the three primary pathways are parallel in different panels. When all parallel pathways are used, test specifications are divided into all possible pathways. When building parallel MST panels with a top–down strategy, we set the target TIF for the entire test and assign the non-statistical constraint to the pathways.

In the bottom–up strategy, parallel panels are based on the modules. The assembly of parallel modules is parallel between the statistical constraints (target TIF) and the non-statistical constraints. When the modules are parallel, we can mix the parallel modules to assemble multiple parallel panels. As the modules are parallel to each other, the corresponding pathways of panels will automatically be parallel. When using the bottom–up strategy to set up parallel MST panels, we set different target TIF to modules with different item difficulties. In contrast, non-statistical constraints are allocated to each module (Yan et al., 2014).

THE CD-MST FRAMEWORK

Cognitive Diagnosis Combined With Multistage Adaptive Test

As mentioned above, CD-MST combines the advantages of both CD and MST. Similar to MST, CD-MST also includes similar elements or parts, such as the panel, module, stage, pathway, CD-MST design, assembly strategies, and assembly methods. The main difference between MST and CD-MST is that the latter can provide additional rich diagnostic information for each examinee. The information can provide insight on self-study, adaptive learning, and remediation teaching.

In the *Introduction* section, we noted some indexes in the test assembly for MST, such as the item difficulty and the Fisher information describing continuous variables and reflecting the measurement precision. They may not be suitable for CD-MST framework because CD mainly focuses on the multidimensional

and discrete cognitive attributes or KSs. To develop a CD-MST framework, we propose a new assembly method for CD-MST as below.

CD-MST Assembly Strategy

The ATA method is the main algorithm for MST, which currently contains heuristic methods, linear programming methods (Zheng et al., 2012), and Monte Carlo methods (Belov and Armstrong, 2005). The linear programming algorithm can successfully complete the test requirements and strictly meet all the test assembly constraints (e.g., content constraints and enemy items) (Zheng et al., 2012). However, solving the 0–1 linear programming problem is very complex (Theunissen, 1989) and time consuming. With the test constraint complexity increasing, the limited item bank cannot meet all the test constraints. It will induce infeasible problems about overconstrained optimization and lead to test assembly failure.

According to the heuristic algorithms, the test assembly is decomposed into a series of local optimization problems. Each local optimization problem is chosen as a single item for tests (Ackerman, 1989; Lord, 1977). It uses statistical information as a central function (such as the TIF) and considers non-statistical constraints. Heuristic algorithms are less computationally intensive and always effectively complete the test assembly (Zheng et al., 2012); therefore, we used heuristic algorithms to assemble a test for CD-MST in this study.

Item Difficulty Index for Cognitive Diagnosis

In this study, the mean correct response probability of all KSs of one item was used to indicate the item's difficulty. The attribute mastery pattern in an item is finite and known when the Q-matrix is fixed. Therefore, the mean correct response probability of all KSs can reflect this item's difficulty levels, and it is expressed as:

$$Diff_j = \frac{\sum_{c=1}^{2^K} P_j(\alpha_c)}{2^K}, \tag{1}$$

where $Diff_j$ is the difficulty parameter of item j on CD, K is the number of attributes, and $P_j(a_c)$ is the correct response probability on item j for individuals with the KS of $a_c \cdot P_j(a_c)$, which can be calculated by the item response function of CD models (such as the G-DINA model, see Equation 16). The lower the value of $Diff_j$ is, the more difficult item j is.

To investigate whether this index can represent item difficulty, we compared $Diff_j$ and the item difficulty parameter estimated by the IRT model (such as the Rasch model). We used the G-DINA model (for details, see Equation 16) to generate the response data (including 100 items, 1,000 individuals, and five independent attributes), and then we used the G-DINA model and the Rasch model to estimate the same response data, respectively. We calculated each item difficulty on CD via Equation 1 and the item difficulty parameter on Rasch model. The correlation coefficient of item difficulty between CD and IRT reached a value above 0.85 ($p < 0.001$), which clearly shows that the item difficulty based on CD had a significantly high correlation with the item difficulty on IRT. Therefore, the mean correct response probability of all KSs can be viewed as an item difficulty index under the CD framework.

Reliability of Cognitive Diagnosis

Templin and Bradshaw (2013) proposed an empirical reliability index for CD. The reliability index defined the recalculation consistency using the tetrachoric correlation coefficient. They used the following steps to estimate the attribute reliability. (1) Calculate the marginal mastery probability of attribute k for examinee e $\hat{p}_{ek}$ by using CD models. (2) Establish the replication contingency table. For the binary attribute, four elements are calculated as follows:

$$P(\alpha_{\cdot k1} = 1; \alpha_{\cdot k2} = 1) = \frac{\sum_{e=1}^{N} \hat{p}_{ek}\hat{p}_{ek}}{N}, \tag{2}$$

$$P(\alpha_{\cdot k1} = 1; \alpha_{\cdot k2} = 0) = \frac{\sum_{e=1}^{N} \hat{p}_{ek}(1-\hat{p}_{ek})}{N}, \tag{3}$$

$$P(\alpha_{\cdot k1} = 0; \alpha_{\cdot k2} = 1) = \frac{\sum_{e=1}^{N} (1-\hat{p}_{ek})\hat{p}_{ek}}{N}, \tag{4}$$

$$P(\alpha_{\cdot k1} = 0; \alpha_{\cdot k2} = 0) = \frac{\sum_{e=1}^{N} (1-\hat{p}_{ek})(1-\hat{p}_{ek})}{N}, \tag{5}$$

The attribute reliability was calculated by the tetrachoric correlation coefficient of $\alpha_{\cdot k1}$ and $\alpha_{\cdot k2}$,which also represents the re-test reliability of attribute k. More details can be found in Templin and Bradshaw (2013).

Quantitative Targets for CD-MST

Quantitative targets include the test target reliability of CD, item difficulty, etc. In this study, the attribute reliability of the cognitive diagnostic model proposed by Templin and Bradshaw (2013) was used as a metric of the test reliability. This index provides attribute reliability to each cognitive attribute. In the study, the reason for using reliability to assemble the test is that a good reliability can reduce the measurement error and improve the reliability for the test. Reliability or information has always been used to measure the test reliability of both CTT and IRT. In CTT, the reliability coefficient was used to control test error. In IRT, information was used to control test error, but in CDM, the attribute mastery patterns are discrete variables. Based on the characteristics of CDM, Templin and Bradshaw (2013) proposed attribute reliability to control test error and ensure reliability. On the other hand, mainstream assembly algorithms in MST use test information function (TIF) to assemble test pathways, for example, Yang and Reckase (2020) used the Fisher information to assemble test for optimal item pool design in MST, and Xiong (2018) used the Fisher information in a hybrid strategy to construct MST. Yamamoto et al. (2019) used test characteristic curves (TCCs) in MST test design for PISA 2018. Whether these studies use reliability or information, the purpose is to control test errors and provide a greater reliability. Therefore, borrowing the ideas from the previous studies, we used attribute reliability to assemble tests and to control test errors because of the characteristics of CDM and MST.

THE NORMALIZED WEIGHTED ABSOLUTE DEVIATION HEURISTIC FOR CD-MST

The normalized weighted absolute deviation heuristic (NWADH; Luecht, 1998), a popular heuristic algorithm, has been applied to the MST assembly. The weighted deviations of constrained targets are used in this algorithm, and the deviation of each constraint is standard with the same scale (van der Linden and Glas, 2000). They also are compatible with multiple contents or classification dimensions, multiple quantitative targets, multiple test modules, and other complex test group issues, such as the enemy items (Luecht, 1998). Therefore, the NWADH is employed for the test assembly in CD-MST.

In NWADH, both statistical and non-statistical constraints are combined to set the objective function and to meet the current test requirement. With the selection of each item, the objective function is updated according to the measurement characteristics of the selected item, which is done until the test assembly is completed (Luecht, 1998). A well-designed test has a clear test specification so that measurement properties, quantitative targets, and other constraints should be considered in the test assembly. The statistical and non-statistical constraints for a test specification will be described in detail below.

Let T_k denote the target reliability of attribute k with test. u_k^j denotes the observed reliability of attribute k in the test with a length of J items, which can be calculated by the tetrachoric correlation coefficient, and the difference of attribute reliability between the target attribute reliability and the observe attribute reliability can be calculated as follows:

$$d^J = \sum_{k=1}^{K} \left| T_k - u_k^J \right| \Big/ K \tag{6}$$

In Equation 6, J denotes the selected items in the test, and d^J represents the mean absolute deviation between the target attribute reliability and the observe attribute reliability with J items. When the new item was added to the test with J items, the test length is $J+1$ items. At this time, the difference of attribute reliability between the target attribute reliability and the observed attribute reliability can be calculated as Equation 7:

$$d_i^{J+1} = \sum_{k=1}^{K} \left| T_k - u_k^{J+1} \right| \Big/ K; i \in R_J \tag{7}$$

In Equation 7, R_J refers to the remaining items in the item bank after selecting J items. The item i is selected from R_J. In order to meet statistical constraints, in CD-MST, the next item i of R_J that makes d_i^{j+1} with the smallest values was selected.

At the same time, in order to optimize the NWADH algorithm, we can transform the minimizing of the absolute deviation function in Equation 6 into the maximization, as follows:

$$MAX(e_i) \tag{8}$$

where e_i is the "priority index" and is expressed as:

$$e_i = 1 - \frac{d_i^{J+1}}{\sum_{i \in R_J} d_i^{J+1}}; i \in R_J \tag{9}$$

In Equation 8, e_i denotes the priority index of item i. That means that CD-MST priority selects the items to make e_i with the maximum values in the remaining item bank R_J.

Equations 6 and 9 are the NWADH algorithms (Luecht, 1998) when only considering the statistical quantitative target. However, a complete CD-MST also needs to consider non-statistical constraints such as content balance, item type, item answer, and other constraints. The NWADH algorithm can merge multiple content constraints (Luecht, 1998). When considering the content constraints, it is necessary to give a certain weight to constraints based on the test specifications. In general, the weight values depend on the test specifications that can be obtained by the pre-simulation (Luecht, 1998). The NWADH algorithm (Equation 9) contains the content constraints as follows:

$$e_i^* = \left[1 - \frac{d_i^{J+1}}{\sum_{i \in R_J} d_i^{J+1}} \right] + \frac{c_i}{\sum_{i \in R_J} c_i}; i \in R_J \tag{10}$$

where:

$$c_i = v_{ig} W_g + (1 + v_{ig}) \underline{W}_g, \tag{11}$$

$$\underline{W}_g = W^{[\max]} - \frac{1}{G} \sum_{i=1}^{G} W_g. \tag{12}$$

In Equation 10, c_i denotes the content constraint weight for each unselected item in the remaining item bank. In Equation 11, g denotes the total number of content constraints $g = 1, \ldots, G$. $v_{ig} = 0$ indicates that item i does not contain the content constraint g, whereas $v_{ig} = 1$ indicates otherwise. W_g represents the weight of each content constraint g. $\underline{W}_g$ represents the mean weight of each content constraint g. In Equation 12, $W^{[max]}$ represents the maximum weight values of G kinds of content constraints. In this study, the weight of the non-statistical constraints was according to the method proposed by Luecht (1998). The non-statistical constraints in the study were set as follows:

$$if \sum_{i \in R_{j-1}} v_i \geq Z_g^{[\max]}, then\, W_g = 1, \tag{13}$$

$$if \sum_{i \in R_{j-1}} v_i < Z_g^{[\min]}, then\, W_g = 2, \tag{14}$$

subject to the constraints,

$$\sum_{i=1}^{I} v_{ig}, g = 1, \ldots, G. \tag{15}$$

Let $Z_g^{[max]}$ represent the maximum constraint values of constraint g. $Z_g^{[min]}$ represents the minimum constraint values of constraint g. Therefore, when tests contain non-statistical constraints, e_i in Equation 9 was instead replaced by e_i^* in Equation 10.

Test Assembly Procedure

After all experimental conditions are set up, the program of test assembly, written under the NWADH (see Equations 6–15), was run to assemble test. We briefly describe the assembly procedure step-by-step as follows:

First: Take the hard pathway as an example; the test assembly program is based on the initial items in the first stage to find the new item in item bank. The new item needs to have the largest e_i^* value in the remaining item bank, and e_i^* was calculated by Equation 10.

Second: When the item with the largest e_i^* was selected to the hard pathway, we will select the next new item based on the new item and initial item of the first stage. The next new item also needs to have the largest e_i^* in the remaining item bank.

Third: Repeat the above two steps until the test length meets the experimental requirements. It should be noted that each item was selected only once, which means that the selected new item needs to be removed from the remaining item bank.

THE GENERAL COGNITIVE DIAGNOSIS MODELS: THE G-DINA MODEL

Cognitive diagnosis models play an important role in CD. They connect examinees' external response and internal knowledge structure. We need to select the appropriate cognitive diagnostic models for the test to ensure the accuracy and effectiveness of the test.

Generalized DINA (G-DINA; de la Torre, 2011) is an expansion of the DINA model (Deterministic-in-put, noisy-and-gate model; Haertel, 1984; Junker and Sijtsma, 2001). It considers that examinees with different attribute mastery patterns have different probability attributes. For G-DINA, $K_j^* = \sum_{k=1}^{K} q_{jk}$, where K_j^* is the number of attributes k of item j. The G-DINA model divides examinees into $2^{k_j^*}$ categories and let a_{lj}^* denote the reduced attribute mastery patterns based on the measurement attributes of item j, $l = 1, 2, \ldots, 2^{k_j^*}$. The G-DINA model has different mathematical expressions depending on the function. The three main link functions are the identify link function, logit link function, and log link function. de la Torre (2011) pointed out that the G-DINA model based on the identify link function is a more general form of the DINA model, and its mathematical equation is:

$$P(X_{ij} = 1|\alpha_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk'} + \ldots + \delta_{j12\ldots k^*} \prod_{k=1}^{K_j^*} \alpha_{lk}. \quad (16)$$

δ_{j0} denotes the intercept of item j. That is, if examinees do not master all the attributes measured by an item, the value is a non-negative value. δ_{jk} is the main effect for α_k. $\delta_{jkk'}$ is the interaction effect between α_k and $\alpha_{k'}$. $\delta_{j12\ldots k_j^*}$ denotes the interaction effect from $\alpha_1,\ldots,\alpha_{k_j^*}$.

SIMULATION STUDY

Simulation Design

Generated Item Bank

In the simulation study, the number of attributes and the test length were set to five attributes and 21/25 items, respectively. The number of panels were fixed to five or 10 panels. Therefore, there were 2 (the test length) × 2 (the number of panels) = 4 total conditions for this study. Across the conditions, we generated an item bank with 1,000 items. For both IRT and CDM, the measurement of reliability requires a certain test length to ensure that the test reliability can be accurately measured. The test length in the study is based on the CAT and MST. In general, 21 items can provide a good test information in CAT. At the same time, the test is usually divided into three or four stages in MST, and each stage with five or seven items. Therefore, the test length was set to 21 and 25 items in the study.

Divided the Item Difficult

For the item difficult level of divide, we referred to the approach of MST. In MST, the item difficult level is divided by the theta parameters because the item difficult parameters and the theta parameters are in the same scale in IRT framework. More specifically, the method is averaged to divide the theta value from large to small into three intervals, and three different intervals represent three different difficulty pathways: easy, medium, and hard pathways. So, we used the same method to divide the difficulty in CD-MST. In the study, item difficulty called $Diff_j$ was described as the mean correct response probability of all KSs of one item. The $Diff_j$ is a probability between 0 and 1. According to the value of $Diff_j$ from item bank, three cut-points were averaged and generated from max $Diff_j$ 0.74 to min $Diff_j$ 0.24 (see Equation 1). We can classify items into easy (0.58–0.74), medium (0.42–0.57), and difficult (0.24–0.41) intervals for CD-MST. The difficult interval with a low value represents the hardest item set. The easy interval with a large value represents the easiest item set.

Set Reliability Criteria

Templin's attribute reliability index is a probability between 0 and 1. Educational Testing Service (2018) proposed 0.9 as representing a very good reliability in CDM. In order to guarantee the test reliability, we chose a high value of 0.9 as the reliability criteria. Therefore, the attribute reliability higher than 0.90 was set as the target reliability value for each attribute.

Set the First Stage

In the study, each panel contained three stages. The number of items in each stage is listed in **Table 1**. It is worth noting that items in the first stage only measured one attribute, whose purposes are to prove the parameters identifiability of CD models (Xu and Zhang, 2016) in the early stage and to improve the classification accuracy of attributes.

Set Quantitative Targets

Quantitative targets are defined as the target attribute reliability proposed by Templin and Bradshaw (2013). The target attribute reliability of each attribute was set to 0.90. The non-statistical constraints in each panel are listed in **Table 2**, and it should be noted that the test assembly needed to meet the minimum limit constraints. For example, the content balance was divided into four categories, where each category contained at least four items after the test was completed.

Set Assembly Strategy

The top–down strategy was used to assemble the panels, so the non-statistical constraints and quantitative targets would remain parallel between the pathways. R (Version 3.5.1 64-bit; R Core Team, 2018) was used to write the test assembly program under the NWADH.

Simulation Process

Step 1: Knowledge states. In the study, the test included five independent attributes, and all possible KSs were $2^5 = 32$. The KS of 1,000 examinees was randomly generated from 32 KSs.

Step 2: Q-matrix. The item bank included 1,000 items, and the Q-matrix was randomly generated from 25 to 1 = 31 item attribute patterns.

Step 3: Item parameters. It was generated by the GDINA package (Version 2.1.15; Ma and de la Torre, 2017) in R (Version 3.5.1 64-bit; R Core Team, 2018). According to de la Torre (2011), the item parameters of the G-DINA model are simulated according to $P_j(0)$ and 1-$P_j(1)$, and $P_j(0)$ represents the probability of examinees who do not master any attribute required by item j and correctly respond to item j, 1-$P_j(1)$ represents the probability of examinees who master all the attributes required by item j with wrong response to item j. Here, the parameters $P_j(0)$ and 1–$P_j(1)$ were randomly generated between uniform (0, 0.25). This simulation study was replicated 100 times.

Step 4: Test assembly. After all experimental conditions are set up, the program of test assembly, written under the NWADH (see Equations 6–15), was run to assemble the test.

TABLE 1 | Number of items in each stage.

	21 items for test length			25 items for test length		
Pathway	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 1**	**Stage 2**	**Stage 3**
Easy	5	8	8	5	10	10
Medium	5	8	8	5	10	10
Hard	5	8	8	5	10	10

TABLE 2 | Number of non-statistical constraints in test assembly.

Constraints group	Categories	Constraints
Content balance	4	4
Item types	2	8
Answer balance	4	4
Enemy items	1	0
The number of each attribute	5	3

Evaluation Criteria

For this simulation study, some criteria were computed to evaluate the target attribute reliability violated and the number of constraints violated on each test pathway. The index of the target attribute reliability violated is expressed as:

$$D_{ik} = R_{ik} - T_{ik}, \tag{17}$$

where R_{ik} is the observed reliability of attribute k on pathway i, $T_{ik}$is the target reliability of attribute k on pathway i, and D_{ik} represents the difference between the observed reliability and the target reliability.

The number of constraints violated on each constraint is computed as:

$$V = \sum_{i=1}^{N} V_i, \tag{18}$$

where V_i represents the number of constraints violated, N is the constraint number of each test pathway, and V is the constraint number for the test pathway.

Other criteria were reported in the results, for example, the item difficulties based on CD, the item difficulties based on the Rasch model, the expected number-correct score based on CD, and the Cronbach α coefficient based on the CTT on each test pathway.

Results

Figures 2–5 documented the results of the difference between the observed and the target attribute reliability (i.e., D_{ik}; see Equation 17) under four experimental conditions. In **Figures 2–5**, the points D_{ik} represent the difference values between the target attribute reliability and the experimental reliability value, and the lower D_{ik} value indicates a smaller test error. It means that the observed reliability is closer to the target reliability 0.9. Three lines represent different difficulty pathways. We also presented the difference value under different experimental conditions in **Figures 2–5**.

Figure 2 shows the experimental condition results for five attributes, 21 items, and five panels, and A1–A5 represent attributes 1–5, respectively. Each attribute reliability in each main pathway reached about 0.9, and all the differences between the observed and target reliability were within +0.2. It indicated that the quantitative targets were satisfied. The results of the three other experimental conditions (see **Figures 3–5**) were very similar to the above experimental condition. Besides, the attribute reliabilities (see **Figures 2–5**) had slight differences under different item lengths. More detailed, the attribute reliability with 25 items was slightly higher than 21 items, which indicated that the item length affected the attribute reliability, and this result verifies that the test length also affects reliability in CD-MST.

Table 3 summarizes the item statistics for the three primary pathways in different experimental conditions. First, we show the item difficulty of different pathways. The results indicated that item difficulty, in the same simulation data, was very different among three primary pathways in the CDM and the IRT Rasch model. More specifically, the hard pathway with more difficult items has lower $Diff_j$ values (mean correct response probability of all KSs) than those

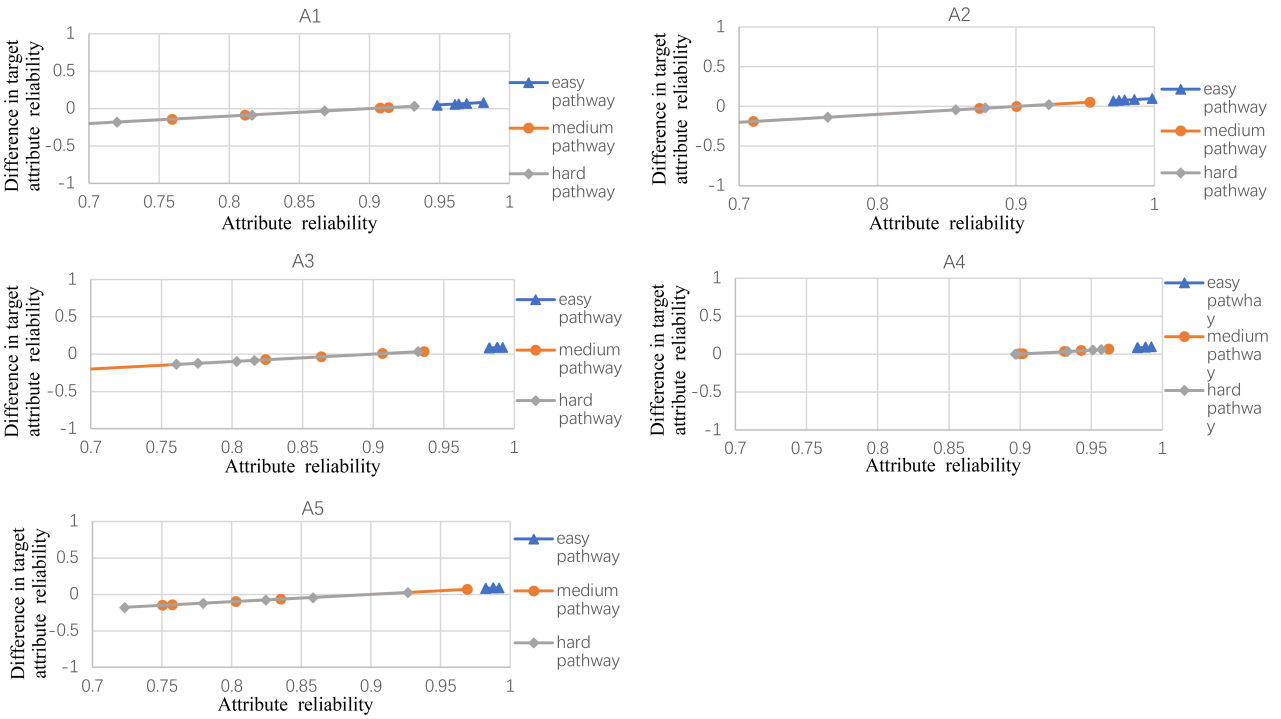


FIGURE 2 | The difference between observed and target reliability with five-attribute, 21-item, and five-panel conditions.

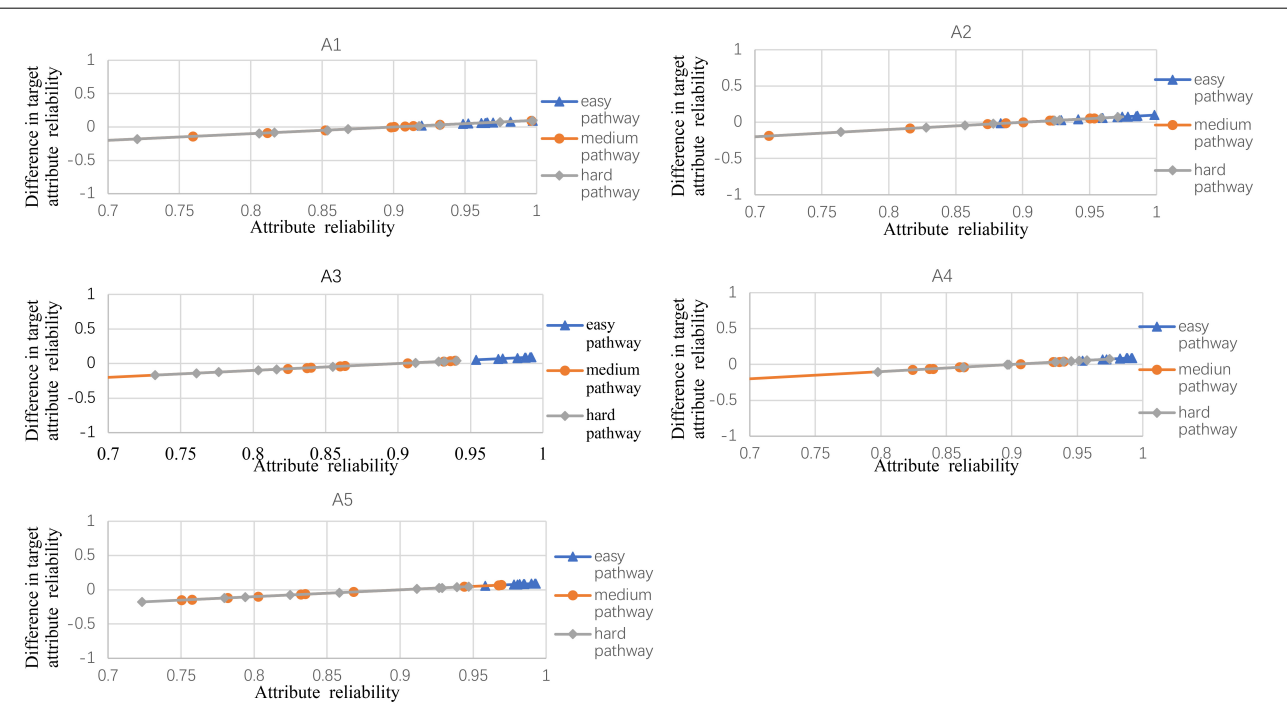


FIGURE 3 | The difference between observed and target reliability with five-attribute, 21-item, and 10-panel conditions.

of medium and easy pathways. The medium pathway had a lower value of $Diff_j$ than that of the easy pathway. It should be noted that the lower $Diff_j$ values represent the harder item difficulty in this study. Moreover, the three pathways also show the difference of item difficulty in IRT Rasch model. For example, the item difficulty in the hard

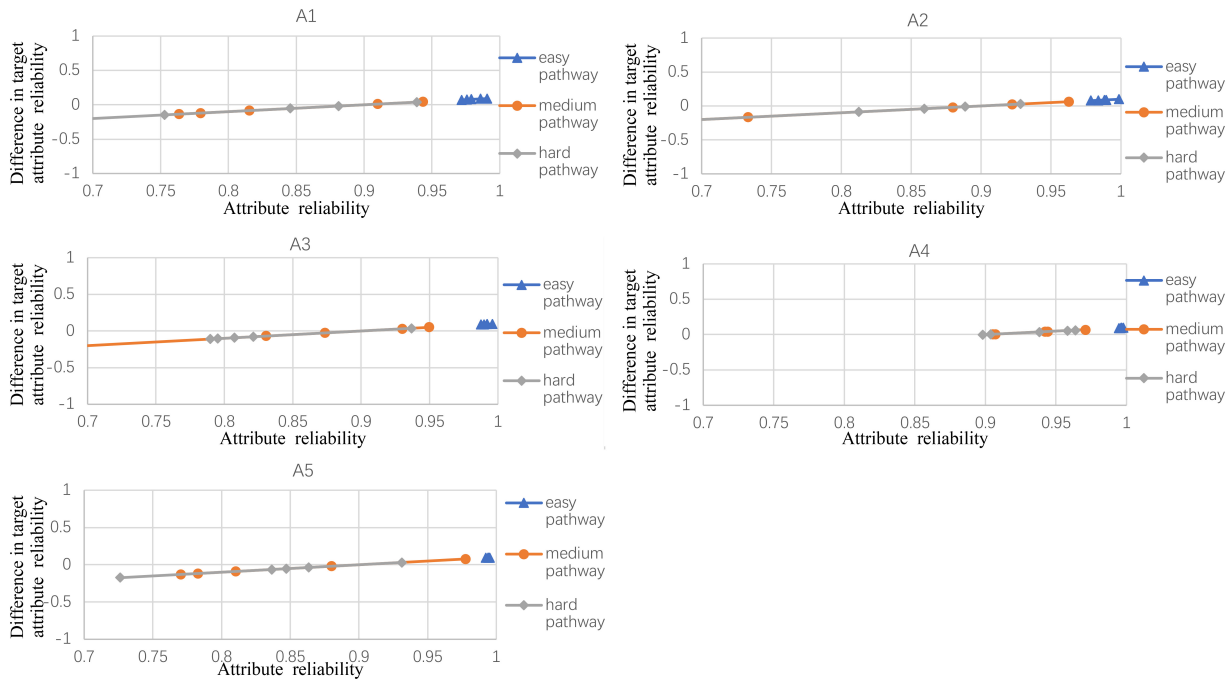


FIGURE 4 | The difference between observed and target reliability with five-attribute, 25-item, and five-panel conditions.

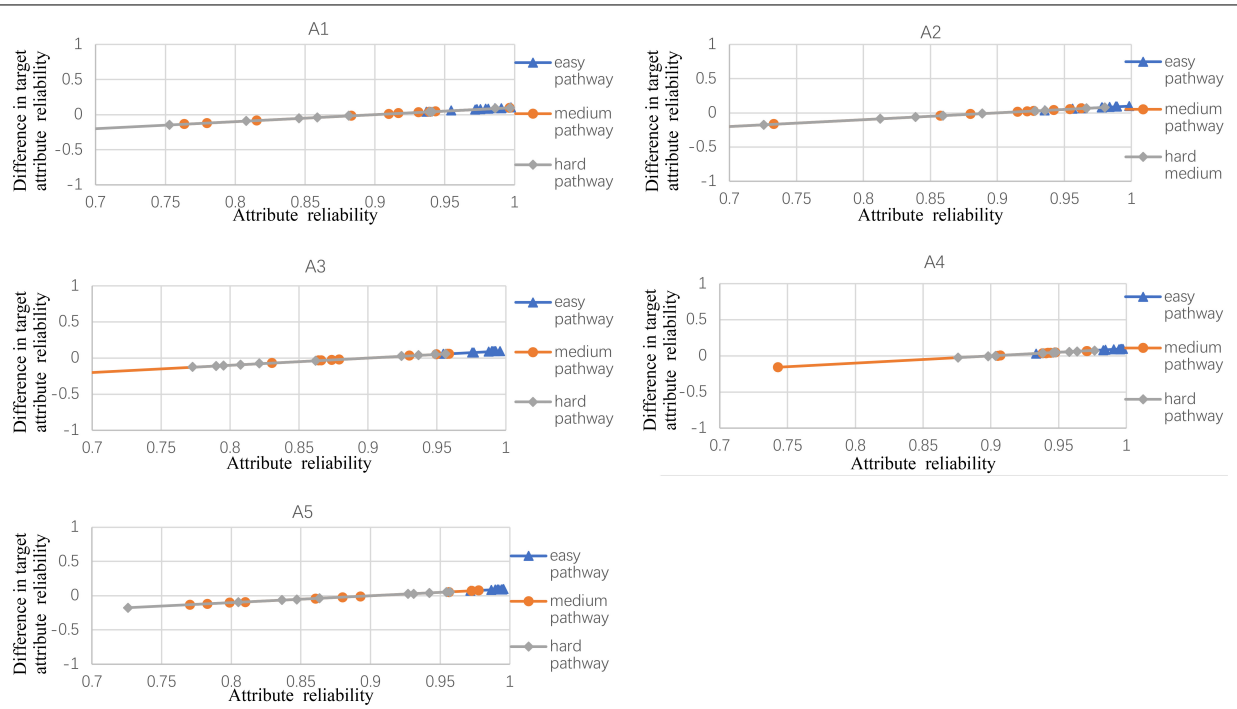


FIGURE 5 | The difference between observed and target reliability with five-attribute, 25-item, and 10-panel conditions.

pathway is higher than those of the medium and easy pathways. Therefore, these results show that the proposed $Diff_j$ can describe the item difficulty of CDM and can be verified by IRT.

In addition, the standard deviation (SD) of $Diff_j$ in each primary pathway was very small for all experimental conditions, which showed that the items in the same pathway had very similar difficulty levels. We also used the same data to verify the IRT

TABLE 3 | Item difficulties and expected number-correct score statistics for each pathway.

Five panels 25 items	Cognitive diagnosis item difficulties		Rasch item difficulties		Expected number-correct score Based on CD		Cronbach α coefficient
Pathway	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Easy	0.6171	0.0125	–0.3868	0.0930	77.092	43.976	0.960
Medium	0.4593	0.0143	0.5428	0.0895	57.587	28.532	0.924
Hard	0.3794	0.0150	1.0456	0.0924	47.561	27.548	0.916

10 panels 25 items	Cognitive diagnosis item difficulties		Rasch item difficulties		Expected number-correct score Based on CD		Cronbach α coefficient
Pathway	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Easy	0.6148	0.0097	–0.9041	0.1329	153.493	83.948	0.978
Medium	0.4772	0.0213	0.2075	0.1261	119.492	59.728	0.962
Hard	0.3887	0.0143	0.9899	0.1342	97.276	56.009	0.958

Five panels 21 items	Cognitive diagnosis item difficulties		Rasch item difficulties		Expected number-correct score Based on CD		Cronbach α coefficient
Pathway	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Easy	0.6120	0.0139	–0.4374	0.0956	64.162	37.182	0.952
Medium	0.4605	0.0138	0.5019	0.0950	48.533	24.348	0.910
Hard	0.3845	0.0139	1.0091	0.0979	40.411	22.988	0.900

10 panels 21 items	Cognitive diagnosis item difficulties		Rasch item difficulties		Expected number-correct score Based on CD		Cronbach α coefficient
Pathway	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	
Easy	0.6099	0.0101	–0.9456	0.0849	127.912	71.358	0.974
Medium	0.4788	0.0214	–0.2315	0.0823	100.812	51.264	0.955
Hard	0.3932	0.0133	0.3654	0.0849	82.583	48.242	0.950

TABLE 4 | Number of constraints violated in each constraint group for each test pathway.

10 panels, 21 items				10 panels, 25 items				5 panels, 21 items				5 panels, 25 items			
Constraint group	**Easy pathway**	**Medium pathway**	**Hard pathway**	**Constraint group**	**Easy pathway**	**Medium pathway**	**Hard pathway**	**Constraint group**	**Easy pathway**	**Medium pathway**	**Hard pathway**	**Constraint group**	**Easy pathway**	**Medium pathway**	**Hard pathway**
Content category	0	1	0	Content category	0	0	0	Content category	0	0	0	Content category	0	0	0
Item types	0	0	0	Item types	0	0	0	Item types	0	0	0	Item types	0	0	0
Answer keys	0	0	1	Answer keys	0	0	1	Answer keys	0	0	0	Answer keys	0	0	0
Attribute times	0	0	0	Attribute times	0	0	0	Attribute times	0	0	0	Attribute times	0	0	0
Enemy item	0	0	0	Enemy item	0	0	0	Enemy item	0	0	0	Enemy item	0	0	0

difficulty via Rasch model, which results indicated that the two types of difficulty parameters (IRT and CD) were very consistent. From the above results, it is reasonable to use the mean correct probability of all KSs as the item difficulty index for CD-MST.

Table 3 also displayed that the mean expected number-correct scores were calculated under a large sample with 1,000 examinees. It was shown in the sixtth and seventh columns of **Table 3**. First, we calculated each examinee's expected number-correct score in each primary pathway. Then we calculated the mean and SD. As expected, examinees had the highest mean expected number-correct scores in the easy pathway, while they had the lowest mean expected number-correct scores in the hard pathway. It is theoretically reasonable because examinees usually get more scores on easy items.

In **Table 3**, the Cronbach's α coefficient was used to verify test reliability. The α coefficients varied from 0.900 to 0.978 with an average of 0.945, which indicates that the proposed CD-MST had high reliability. This shows that the assembled test in the study not only satisfies the reliability of CDM but also the reliability of Cronbach's α coefficient.

Table 4 documents the number of constraints violated in each constraint group, and the constraints rae set in **Table 2**. As known in **Table 2**, the constraint group involved 16 categories and 64 constraints. **Table 4** shows that only three of 64 constraints were not satisfied. Specifically, one content balance was not satisfied in the medium pathway with the condition of 21 items and 10 panels, and two answer balances were not satisfied in the hard pathway of the condition of 21 items and 25 items with 10 panels. The overall non-statistical constraint violation rate was about 4.7%, which was an acceptable range. The results indicated that the proposed test assembly had a very good performance in the non-statistical constraints for CD-MST.

CONCLUSION AND DISCUSSION

The MST with the advantages of P&P and CAT is to be applied to many large-scale international examinations. However, the existing MST with the IRT focuses on the examinees' general ability and cannot provide further detailed diagnostic information. Because CD mainly focuses on the multidimensional and discrete cognitive attributes, some test assembly indexes in MST (such as the item difficulty and the Fisher information) are not suitable for CD-MST. There has been no recent research on CD-MST. Although some studies (such as Zheng and Chang, 2015) provided on-the-fly MST (OMST; Zheng and Chang, 2015), which may be a practical method of CD-MST, this may lead to many problems, such as (1) the test developer having difficulty in managing tests before administering, (2) the parallel of the test is difficult to ensure, (3) and the non-statistical constraint also is difficult to satisfy. To address the above issues, a CD-MST framework that not only provides rich diagnostic information about the candidates but also retains the inherent advantages of MST was proposed in this paper. This paper also proposed and employed two statistical indexes, namely, item difficulty and attribute reliability, as the statistical constraints of CD-MST. In this paper, the proposed item difficulty index is a good indicator of the item difficulty based on CD, which has a very significant high correlation with the item difficulty parameter based on IRT (such as the Rasch model). The reliability index also guarantees the reliability and measurement error of tests. These indexes can provide statistical information, which makes it possible to automate test assembly for CD-MST. At the same time, the results showed that the NWADH algorithm under the CD framework successfully satisfied the non-statistical constraints. It showed that the proposed CD-MST framework and statistical indicators are acceptable for CD-MST.

This study employed the NWADH heuristic method to assemble the CD-MST under ATA. The results showed that the statistical and non-statistical constraints were both well satisfied, and the assembled test panels were parallel overall. At the same time, the non-statistical constraints (such as the attribute balance and content balance) were fully considered in CD-MST, which helps improve the content validity and structural validity of CD-MST. Therefore, the proposed CD-MST with NWADH heuristic algorithms not only provides rich diagnostic information but also retains the advantages of MST.

LIMITATIONS AND FURTHER RESEARCH

As an early exploration of CD-MST, despite the promising results, there are still some limitations that need to be studied further. First, even though the CD item difficulty index, the mean correct probability of all KSs, fully represents the item difficulty, it is verified by the IRT model. Further research also can develop other indexes to measure the item difficulty in CDM. For example, Zhan et al. (2018) proposed the probabilistic-input, noisy conjunctive (PINC) model, which defined attribute mastery status as probabilities and reported the probability of knowledge status for examinees from 0 to 1. According to Zhan et al. (2018), classifying an examinee's KSs to 0 or 1 will cause a lot of information of examinees to be lost, so the PINC model can provide more precise and richer information to examinees' KSs than the traditional CDMs. Therefore, researchers can try to use the probability of examinees' KSs to develop a new difficulty index in the future.

Second, attribute reliability was regarded as a quantitative target in this study, which is illustrative but not prescriptive. In future studies, other reliability or information/measurement error indicators may also be considered as quantitative targets. For example, the classification accuracy was proposed by Cui et al. (2012), the classification matches were proposed by Madison and Bradshaw (2015), and the classification consistency was proposed by Matthew and Sinharay (2018). In the future, the comparative analysis of these reliability indexes can be applied to the test assembly in CD-MST.

Third, the NWADH method was used in this study to assemble the panels. Although this method can guarantee the successful completion of the test assembly, there is still a small violation of the constraints. For example, content constraints were slightly violated in this study. Even if this violation

is allowed in the NWADH method, other methods may be considered to ensure that all constraints are met. In fact, the linear programming method and the Monte Carlo method are also widely used in MST. Although these two methods are influenced by the size and quality of the item bank, they can fully meet the test specification. Besides, Luo and Kim (2018) proposed the mixed integer linear programming (MILP) to assemble tests in MST. The result of the MILP method shows that the method had the advantage of the heuristic algorithm and 0–1 linear programming algorithm. Perhaps, the MILP method is also a reasonable ATA method for CD-MST and can resolute the violence of constraints. Therefore, the development of new methods that can fully meet the constraints and successfully assemble tests is also one of the future research directions.

Finally, the test length also needs to be explored in a further study. In the study, the difference between the reliability and the constraints is not significant. The difference between test length levels can be larger (e.g., 21 vs. 42) and be further studied to explore the impact of test length. Researchers can design the different item numbers to explore the best test length that can provide the maximum information and meet the test constraints.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

GL conceptualized the study, developed the methodology, performed the mathematical derivation, conducted the formal analysis, analyzed the data, wrote the original draft of the study, and wrote the article. GL, DT, and YC edited and reviewed the manuscript. GL, XG, and DW performed the simulation studies. DT acquired the funding and resources, performed the investigation, and supervised the study project. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (31960186, 31760288, and 31660278).

REFERENCES

Ackerman, T. (1989). "An alternative methodology for creating parallel test forms using the IRT information function," in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education* (San Francisco, CA).

Belov, D. I., and Armstrong, R. D. (2005). Monte Carlo test assembly for item pool analysis and extension. *Appl. Psychol. Meas.* 29, 239–261. doi: 10.1177/0146621605275413

Cui, Y., Gierl, M., and Chang, H. -H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Educ. Meas.* 49, 19–38. doi: 10.1111/j.1745-3984.2011.00158.x

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

Embretson, S. E., and Yang, X. D. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika* 78, 14–36. doi: 10.1007/s11336-012-9296-y

Gulliksen, H. (1950). *Theory of Mental Tests*. New York, NY: John Wiley, doi: 10.2307/2280760

Haertel, E. H. (1984). An application of latent class models to assessment data. *Appl. Psychol. Meas.* 8, 333–346. doi: 10.1177/014662168400800311

Hartze, M. C. (2002). A bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality. *Am. J. Gastroenterol.* 95, 906–909. doi: 10.1111/j.1572-0241.2000.01927.x

Junker, B., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kim, S., Moses, T., and Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *J. Educ. Meas.* 52, 70–79. doi: 10.1111/jedm.12063

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *J. Educ. Meas.* 14, 117–138. doi: 10.1002/j.2333-8504.1977.tb01128.x

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Appl. Psychol. Meas.* 22, 224–236. doi: 10.1177/01466216980223003

Luecht, R. M., and Burgin, W. (2003). "Test information targeting strategies for adaptive multistage testing designs," in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, (Chicago, IL).

Luecht, R. M., and Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *J. Educ. Meas.* 35, 229–249. doi: 10.2307/1435202

Luecht, R. M., Terry Brumfield, T., and Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Appl. Meas. Educ.* 19, 189–202. doi: 10.1207/s15324818ame1903_2

Luo, X., and Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test. *J. Educ. Meas.* 55, 243–263. doi: 10.1111/jedm.12174

Ma, W., and de la Torre, J. (2017). *GDINA: The Generalized DINA Model Framework*. Available online at: https://github.com/Wenchao-Ma/GDINA

Madison, M. J., and Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educ. Psychol. Meas.* 75, 491–511. doi: 10.1177/0013164414539162

Matthew, S. J., and Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *J. Educ. Meas.* 4, 635–664. doi: 10.1111/jedm.12196

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: The R Foundation for Statistical Computing.

Samejima, F. (1977). A use of the information function for tailored testing. *Appl. Psychol. Meas.* 1, 233–247. doi: 10.1177/014662167700100209

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

Theunissen, T. J. J. M. (1989). Some applications of optimization algorithms in test design and adaptive test. *Appl. Psychol. Meas.* 10, 381–389.

van der Linden, W. J., and Glas, G. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer, 27–52. doi: 10.1007/0-306-47531-6

von Davier, M., and Cheng, Y. (2014). "Multistage testing using diagnostic models," in *Computerized Multistage Testing Theory and Applications*, eds D. L. Yan, A. A. von Davier, and C. Lewis (New York, NY: CRC Press), 219–227.

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

Xiong, X. (2018). A hybrid strategy to construct multistage adaptive tests. *Appl. Psychol. Meas.* 42, 1–14. doi: 10.1177/0146621618762739

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

Yamamoto, K., Shin, H. J., and Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educ. Meas. Issues Pract.* 37, 16–27. doi: 10.1111/emip.12226

Yamamoto, K., Shin, H. J., and Khorramdel, L. (2019). *Introduction of Multistage Adaptive Testing Design in Pisa 2018. OECD Education Working Papers*. Paris: OECD Publishing, doi: 10.1787/b9435d4b-en

Yan, D., von Davier, A., and Lewis, C. (2014). *Computerized Multistage Testing Theory and Applications*. New York, NY: CRC Press.

Yang, L., and Reckase, M. D. (2020). The optimal item pool design in multistage computerized adaptive tests with the p -optimality method. *Educ. Psychol. Meas.* 80, 955–974. doi: 10.1177/0013164419901292

Zhan, P., Wang, W.-C., Jiao, H., and Bian, Y. (2018). Probabilistic-Input, noisy conjunctive models for cognitive diagnosis. *Front. Psychol.* 6:997. doi: 10.3389/fpsyg.2018.00997

Zheng, Y., and Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Appl. Psychol. Meas.* 39, 104–118. doi: 10.1177/0146621614544519

Zheng, Y., Nozawa, Y., Gao, X. H., and Chang, H. H. (2012). *Multistage Adaptive Testing for a Large-Scale Classification test: Design, Heuristic Assembly, and Comparison with other Testing Modes*. Iowa City, IA: ACT Research Report Series.

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fpsyg.2021.517155

Binary Restrictive Threshold Method for Item Exposure Control in Cognitive Diagnostic Computerized Adaptive Testing

Xiaojian Sun[1,2], *Yizhu Gao*[3], *Tao Xin*[4]* *and Naiqing Song*[1,2]

[1] School of Mathematics and Statistics, Southwest University, Chongqing, China, [2] Southwest University Branch, Collaborative Innovation Center of Assessment for Basic Education Quality, Chongqing, China, [3] Faculty of Education, University of Alberta, Edmonton, AB, Canada, [4] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

Although classification accuracy is a critical issue in cognitive diagnostic computerized adaptive testing, attention has increasingly shifted to item exposure control to ensure test security. In this study, we developed the binary restrictive threshold (BRT) method to balance measurement accuracy and item exposure. In addition, a simulation study was conducted to evaluate its performance. The results indicated that the BRT method performed better than the restrictive progressive (RP) and stratified dynamic binary searching (SDBS) approaches but worse than the restrictive threshold (RT) method in terms of classification accuracy. With respect to item exposure control, the BRT method exhibited noticeably stronger performance compared with the RT method, even though its performance was not as high as that of the RP and SDBS methods.

Keywords: cognitive diagnostic computerized adaptive testing, measurement accuracy, item exposure rate, binary searching algorithm, cognitive diagnostic assessment

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, United States

Reviewed by:
Chanjin Zheng,
East China Normal University, China
Miguel A. Sorrel,
Autonomous University of Madrid, Spain

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

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

Received: *03 December 2019*
Accepted: *30 March 2021*
Published: *05 August 2021*

Citation:
Sun X, Gao Y, Xin T and Song N (2021) Binary Restrictive Threshold Method for Item Exposure Control in Cognitive Diagnostic Computerized Adaptive Testing. Front. Psychol. 12:517155. doi: 10.3389/fpsyg.2021.517155

INTRODUCTION

Cognitive diagnostic computerized adaptive testing (CD-CAT; Cheng, 2009, 2010; Chang, 2015) has attracted the attention of numerous researchers and educators over the past few decades (Wang et al., 2012). CD-CAT is a combination of a cognitive diagnostic model (CDM) and computerized adaptive testing (CAT). A key advantage of CD-CAT is that educators can provide remedial instruction for individuals based on the knowledge level of the individuals, which is determined using CDM (e.g., Gierl et al., 2007). In addition, CD-CAT can generate a test tailored to suit an individual's latent trait levels (Mao and Xin, 2013; Chang, 2015; Lin and Chang, 2019). Consequently, the estimation of an individual's latent ability is more accurate when fewer items are used compared with using traditional paper and pencil tests (Weiss, 1982).

One of the major objectives of CD-CAT is to improve classification accuracy. Numerous item selection methods have been developed to achieve this objective. Item selection methods commonly applied including the Kullback–Leibler method (KL; Xu et al., 2003), Shannon entropy method (SHE; Tatsuoka and Ferguson, 2003), posterior weighted KL method (PWKL; Cheng, 2009),

and modified PWKL (MPWKL; Kaplan et al., 2015). Several attempts have been made to develop item selection methods for short-length tests, such as the mutual information (MI; Wang, 2013), posterior weighted CDM discrimination index, and posterior weighted CDM attribute-level CDI (PWACDI) (Zheng and Chang, 2016). All of the aforementioned item selection methods noticeably enhance the classification accuracy of CD-CAT. However, a major attribute of such methods is that they focus largely on maximizing classification accuracy rather than on controlling item exposure, which results in a highly uneven distribution of item bank usage. Although CD-CAT is used mainly for low-risk tests (Leighton and Gierl, 2007; Wang et al., 2011; Mao and Xin, 2013; Lin and Chang, 2019), where item exposure is not a major concern, items may be at risk of overexposure if an individual already knows the items before taking the test (Wang et al., 2011; Mao and Xin, 2013). In addition, it is not appropriate to administer an item bank with a large number of underexposed items because item bank development is a time- and money-consuming process (Wang et al., 2011; Zheng and Wang, 2017). To establish a balance between classification accuracy and item exposure control, several novel item selection methods have been proposed (e.g., Wang et al., 2011; Hsu et al., 2013; Zheng and Wang, 2017).

Wang et al. (2011) proposed the combination of the restrictive progressive (RP) and restrictive threshold (RT) methods with the PWKL method to achieve item exposure control for fixed-length tests in CD-CAT. In addition, Hsu et al. (2013) developed the Sympson-Hetter method and considers test overlap control, variable length, online update, and restricted maximum information (SHTVOR) to address item exposure with varied test length in CD-CAT. Recently, Zheng and Wang (2017) applied the binary searching algorithm for item exposure control in CD-CAT. They proposed the dynamic binary searching method for varied-length tests and the stratified dynamic binary searching (SDBS) method for fixed-length tests. However, even though the RP method could generate a more even distribution of item usage for fixed-length tests, the classification accuracy was considerably decreased. In comparison, the RT method achieved higher classification accuracy but a more uneven distribution of item usage. The SDBS method is a promising one-item selection method in terms of the testing efficiency and distribution of item usage, but it has relatively low measurement accuracy and flexibility. In addition, the SDBS method does not take into account item parameters, which potentially resulting in wasted item information and low measurement accuracy. To address the shortcomings of the aforementioned methods for fixed-length CD-CAT, we propose a modified method inspired by Wang et al. (2011) and Zheng and Wang (2017). The new method—the binary restrictive threshold (BRT) method—integrates the binary searching algorithm into the RT method.

The remainder of this paper is organized as follows: First, two commonly used CDMs in CD-CAT, the deterministic input, noisy "and" gate (DINA) model (Junker and Sijtsma, 2001) and the reduced reparameterized unified model (RRUM; Hartz, 2002), are briefly introduced. Subsequently, four item control indices—RP, RT, SDBS, and BRT—are presented to illustrate how such indices balance the trade-off between classification accuracy and item exposure control. Afterward, we perform a simulation study to compare the performance of the novel item exposure index with that of the RP, RT, and SDBS methods. Finally, discussions and conclusions are based on the findings of the simulation study are provided.

CDMs

The DINA Model

The DINA model is one of the most commonly used CDMs in CD-CAT because of its simplicity and ease of explanation (e.g., Cheng, 2010; Chen et al., 2012). It classifies individuals into two classes for each item: those who master all attributes that the item measures and those who lack at least one attribute that the item involves. The DINA model can be expressed as

$$P\left(Y_{ij}=1|\eta_{ij}\right) = \left(1-s_j\right)^{\eta_{ij}} g_j{}^{1-\eta_{ij}},$$
$$\eta_{ij} = \prod_{k=1}^{K} (\alpha_{ik})^{q_{jk}},$$

where Y_{ij} is the response of individual i to item j; η is the ideal response indicating whether an individual master all the required attributes of an item; s is the slip parameter; g is the guess parameter; K is the number of attributes; α_{ik} denotes the deficiency or mastery of the k attribute for individual i; and q_{jk} is the element of the Q matrix.

One limitation of the DINA model is that it cannot distinguish individuals who lack one attribute from those who lack more than one attribute that a specific item measures. By contrast, the RRUM allows the probabilities of different attribute mastery patterns to vary across items.

The RRUM

The RRUM has attracted considerable attention in CD-CAT in recent years (e.g., Dai et al., 2016; Huebner et al., 2018). The item response function of the RRUM can be expressed as follows:

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

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

ITEM EXPOSURE CONTROL INDICES IN FIXED-LENGTH CD-CAT

The RP Method

Wang et al. (2011) developed this item exposure control index. Two components are included in this method: the restrictive component and the progressive component. The former imposes a restriction to make sure that the maximum exposure rate does not exceed a pre-defined value, r. The latter component adds a stochastic element to the item selection methods to

avoid the frequent selection of items with the largest amount of information (Revuelta and Ponsoda, 1998). The RP method can be expressed as follows:

$$RP\_info_j = \left(1 - \exp_j / r\right)\left[(1 - (x/J)) \times R_j + info_j \times \beta x/J\right],$$

where $\exp_j$ is the exposure rate for item j, x is the number of items that have been administered, J is the test length, R_j is a random value that is generated from a uniform distribution $U_{(0,\ \max(infor_j))}$, where $info_j$ refers to the corresponding information of item j, such as PWKL information, and β is the importance parameter that is used to adjust the relative importance of classification accuracy vs. the item exposure control issue. A lower value of β indicates that test security is more important than classification accuracy, and vice versa.

The RT Method

This method is another item exposure control index that was developed by Wang et al. (2011). It also includes two components: a restrictive component and a threshold component that is applied to derive an information interval. Candidate items during each interval can be randomly administered to individuals. The information interval is defined as follows:

$$\begin{aligned} RT\_info_{interval} &= \left[\max\left(info_j\right) - \delta, \max\left(info_j\right)\right], \\ \delta &= \left[\max\left(info_j\right) - \min\left(info_j\right)\right] \times f(x), \\ f(x) &= [1 - (x/J)]^{\beta}, \end{aligned}$$

where δ is the threshold parameter and β is the importance parameter that determines the width of the information interval. The higher the value of β, the narrower the information interval. The rest of the symbols have meanings similar to those in the RP method.

The SDBS Method

Zheng and Wang (2017) developed the SDBS algorithm, which stratifies items on the basis of their discrimination index. This method was inspired by the α-stratification method that is commonly used in IRT-based CAT (Chang and Ying, 1999). The classical testing theory (CTT)–based item discrimination indices for the DINA model and the RRUM are $\left(1 - s_j - g_j\right)$ and $\left(\pi_j^* - \pi_j^* \prod_{k=1}^{K} r_{jk}^{*\,qjk}\right)$, respectively (Rupp et al., 2010). The SDBS can be computed as follows:

$$\begin{aligned} B_j^m &= \left| \sum_{S_{jl}^m = 1} p\left(\boldsymbol{\alpha}_l | Y_{t-1}\right) - 0.5 \right|, \\ S_{jl}^m &= \prod_{k=1}^{K} I\left(q_{jk} \leq \alpha_{lk}\right), \\ p\left(\alpha_l | Y_{t-1}\right) &= \frac{P\left(Y_{j-1} | \boldsymbol{\alpha}_l\right) \pi_0\left(\boldsymbol{\alpha}_l\right)}{\sum_{c=1}^{2^K} P\left(Y_{j-1} | \boldsymbol{\alpha}_c\right) \pi_0\left(\boldsymbol{\alpha}_c\right)}, \end{aligned}$$

where B_j is the binary searching index; S_{jl} is the separation for item j and attribute profile l, where $S_{jl} = 1$ indicates that the attribute profile l possesses all the attributes that item j measures and $S_{jl} = 0$ otherwise; m represents the m^{th} stage; $p\left(\boldsymbol{\alpha}_l | \mathbf{Y}_{t-1}\right)$ is the posterior probability for the l^{th} attribute profile conditional on the first $t - 1$ item responses; $\boldsymbol{Y}_{t-1}$, $P\left(\mathbf{Y}_{t-1} | \boldsymbol{\alpha}_l\right)$ is the joint probability of the first $t - 1$ items conditional on attribute profile α_l; $\pi_0\left(\boldsymbol{\alpha}_l\right)$ is the priori probability; and $I(.)$ is the indicator function, which equals 1 when the expression in the brackets is true and equals 0 otherwise.

The SDBS method tends to select an item with a lower B_j value as the next item to be administered. Because only the q-vector of item j is used during the calculation of B_j and item parameters (e.g., slip and guess parameter) are not taken into consideration, items measuring similar or even different attribute profiles can obtain consistent estimations of B.

The BRT Method

Inspired by Zheng and Wang (2017), the present study attempts to combine the binary searching algorithm with the RT method to develop a novel item exposure control method. In particular, the binary searching algorithm is applied first to obtain the candidate item set that has the lowest binary searching index, B. The RT method is then used to select items from the candidate item set.

Because the BRT method combines the binary searching algorithm with the RT method, we expected it to achieve lower classification accuracy but superior item exposure control compared with the RT method. In addition, we expected the BRT method to achieve higher classification accuracy compared with the SDBS and BRP methods because the RT method, which is involved in the BRT method, can yield higher classification accuracy than the RP and SBDS methods (Wang et al., 2011; Zheng and Wang, 2017) when applied to select the appropriate item to be administered.

In the BRT method, items with the lowest B_j value can be obtained first, and then the RT method is applied to randomly select the next item from these items with the lowest B_j value. The mathematical expression of the BRT can be defined as follows:

$$\begin{aligned} BRT\_info_{interval} &= \left[\max\left(info_j\right) - \delta, \max\left(info_j\right)\right], \\ \delta &= \left[\max\left(info_j\right) - \min\left(info_j\right)\right] \times f(x), \\ f(x) &= [1 - (x/J)]^{\beta}, \\ j &\in \underset{\min(B_j)}{Q}, \end{aligned}$$

where $\underset{\min(B_j)}{Q}$ is the q-vector of an item set with the lowest B_j value.

The difference between the BRT and RT methods is that an additional component, the calculated B_j value of each item, is considered in the BRT method. According to Zheng and Wang (2017), the additional component can be used to control item exposure.

In summary, the steps of the BRT method are as follows:

Step 1. Randomly select an item from the item pool as the first item to be administered to individuals;

Step 2. Estimate each individual's attribute profile;
Step 3. Calculate the binary researching index B_j on the basis of the estimated attribute profile;
Step 4. Determine the candidate item set that has the lowest binary researching index;
Step 5. Calculate the BRT index and select the appropriate item as the next item to be administered.
Step 6. Repeat steps 2 to 5 until the terminal rule is satisfied.

SIMULATION STUDY

Simulation Design

We performed a simulation study to evaluate the performance of the BRT method and then compared the BRT method with other item exposure control methods that have been proposed in previous studies (Wang et al., 2011; Zheng and Wang, 2017). In the present study, we manipulated factors such as model type, test length, number of attributes, and item selection method. These factors were set as follows:

Model Type

We included two model types in the present study, namely the DINA model and RRUM, both of which are models commonly applied in CD-CAT.

Number of Attributes

We applied four and six attributes in the present study, both of which are number of attributes that are commonly applied in CD-CAT (e.g., Cheng, 2009, 2010; Mao and Xin, 2013; Dai et al., 2016; Kang et al., 2017; Huebner et al., 2018; Lin and Chang, 2019). For instance, Wang et al. (2011) adopted four attributes in a simulation study, and Zheng and Wang (2017) applied four and six attributes in their study.

Test Length

There are two levels with respect to test length: 25 items (short length) and 40 items (long length). This setting is consistent with those applied in related studies (e.g., Wang et al., 2011; Zheng and Wang, 2017).

Item Selection Method

Six item selection methods—the random, original PWKL, RP, RT, SDBS, and BRT methods—were used in the current study.

The number of conditions was 2 (model type) × 2 (test length) × 2 (number of attributes) × 6 (item selection method) = 48 in total, among which only the item selection method was a within-group variable; the rest were between-group variables. The simulation study was implemented in R software (R Core Team, 2019), and the codes are available upon request from the corresponding author.

Item Bank and Examinees Generation

Two different item banks were generated on the basis of the number of attributes. Each item bank had 480 items, which is also a setting that related studies have commonly applied (Wang et al., 2011; Zheng and Wang, 2017). The item bank can be represented using the Q matrix, which describes the relationship between items and attributes. That is, the element of the Q matrix is 1 if the item measures the attribute, and the element is 0 otherwise. In the present study, the Q matrix was generated entry-by-entry conditional on independence among attributes. In addition, we assumed that each item involved at least one attribute and measured 20% of the attributes on average, which is similar to the case in Zheng and Wang (2017) study.

As for the item parameters, both the guessing and the slipping parameters of the DINA model were generated from a uniform distribution, $U(0.05, 0.25)$. The baseline and penalty parameters of the RRUM were generated from $U(0.75, 0.95)$ and $U(0.20, 0.95)$, respectively. Other studies have also adopted such settings (e.g., Cheng, 2010; Wang et al., 2011; Chen et al., 2012; Mao and Xin, 2013; Zheng and Wang, 2017).

Two $\boldsymbol{\alpha}$ matrices were generated to represent examinees' mastery of the attributes. Two groups of examinees were simulated, and each group was composed of 2,000 examinees. Similar to the Q matrix, the element of the $\boldsymbol{\alpha}$ matrix was marked as 1 if examinees mastered the attribute and marked as 0 otherwise. The steps for generating the $\boldsymbol{\alpha}$ matrix followed those proposed by Wang et al. (2011), and both the threshold and covariance among attributes were set as 0. Consistent with other studies (e.g., Cheng, 2010; Wang et al., 2011; Chen et al., 2012; Mao and Xin, 2013; Wang, 2013; Kaplan et al., 2015; Zheng and Wang, 2017), only one replication was used in the present study.

The value of the importance parameter (i.e., β) for the RT and RP methods was set to be 2. This is because Wang and her colleagues found that the value 2 can generate a reasonable trade-off between measurement accuracy and item usage (Wang et al., 2011). In regard to the BRT method, the value of the importance parameter was determined by a pilot study with varying β values. Its result showed that the value 0.5 is sufficient to balance the trade-off between measurement accuracy and item usage. Thus, the value 0.5 was selected for the BRT method in the current study. In addition, a total of five strata with equal number of items were used for the SDBS method, which recommended by Zheng and Wang (2017).

Evaluation Criteria

Two types of evaluation criteria were used in the current study. The first one was correct classification rate, which includes pattern correct classification rate (PCCR) and attribute correct classification rate (ACCR). Higher values indicate better PCCR and ACCR. The second criteria were item exposure control, which includes the scaled χ^2 (Chang and Ying, 1999), number of items <2% (underused item rate; UIR) and more than 20% (overused item rate; OIR), and the test overlap rate (TOR; Mao and Xin, 2013). Lower values indicated favorable four item exposure control indices. The calculation of such evaluation criteria was performed as follows:

$$
\begin{aligned}
PCCR &= \sum_{i=1}^{N} I\left(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\right)/N, \\
ACCR_k &= \sum_{i=1}^{N} I\left(\hat{\alpha}_{ik} = \alpha_{ik}\right)/N, \\
\chi^2 &= \sum\nolimits_{j=1}^{N_{item}} \left(er_j - J/N_{item}\right)^2/(J/N_{item}),
\end{aligned}
$$

$$er_j = \frac{N_j^{administered}}{N},$$

$$TOR = \frac{\sum_{j=1}^{N_{item}} N_j^{administered} \times \left(N_j^{administered} - 1\right)}{J \times N \times (N-1)},$$

where $\hat{\boldsymbol{\alpha}}_i$ and $\boldsymbol{\alpha}_i$ denote the estimated and true attribute profiles of examinee i, N is the number of individuals, J denotes the test length, N_{item} is the number of items in the item bank, $N_j^{administered}$ is the number of times the j^{th} item is administered (i.e., the number of individuals who answer item j), and er_j is the exposure rate of item j.

RESULTS

Correct Classification Rate

Table 1 presents the correct classification rates for four attributes. The original PWKL and the random methods yielded the highest and lowest PCCRs, respectively, regardless of model type and test length. The RT method yielded the same or slightly lower PCCRs compared with the PWKL method and much higher PCCRs than the RP, SDBS, and BRT methods. In particular, the differences between the RT method and the three item exposure control methods (RP, SDBS, and BRT) were relatively small with respect to the DINA model, which ranged from 0.003 to 0.007 for 25 items, and no difference existed among the methods for 40 items. However, the differences were slightly greater for the RRUM, which ranged from 0.055 to 0.214 and from 0.008 to 0.052 for 25 and 40 items, respectively. The novel BRT method yielded slightly lower PCCRs than did the RT method, while it yielded higher PCCRs than did the RP and SDBS methods. The differences between the BRT method and the other methods (i.e., the RP and SDBS methods) ranged from 0.000 to 0.004 for 25 items, and they shared similar PCCRs for 40 items conditional on the DINA model. The differences ranged from 0.034 to 0.159 and from 0.036 to 0.088, for 25 and 40 items, respectively, conditional on the RRUM. In addition, among the four item exposure control methods (RP, RT, SDBS, and BRT), the SDBS yielded the lowest PCCRs under all conditions except one ($J = 40$, the DINA model). Regarding ACCR, the averaged ACCRs for PWKL and random methods were the greatest and lowest, respectively. The BRT method yielded slightly lower average ACCRs than did the PWKL but slightly higher average ACCRs than the RP and SDBS methods. In addition, the SDBS yielded the lowest average ACCRs among the four item exposure control methods.

The PCCR results for six attributes revealed similar patterns with those for four attributes: the PWKL and random methods yielded the highest and lowest PCCRs, respectively, regardless of model type and test length. As for the remaining four item selection methods, their PCCRs are illustrated in **Table 2**. The RT method yielded the highest PCCRs under all conditions across the four methods. Furthermore, the BRT method yielded lower PCCRs than did the RT method; however, it had higher PCCRs than the RP and SDBS methods. The SDBS method yielded the lowest PCCR under all conditions. In addition, the differences in the PCCRs among the methods were relatively low for the DINA model compared with for the RRUM.

Item Exposure Control

Table 3 presents the item exposure control for four attributes. The PWKL method had the highest scaled χ^2 values, regardless of test length and model type, which indicated that the item exposure rate was quite skewed. In addition, the PWKL method had the highest TOR, UIR, and OIR values. For instance, more than 70% of the items were underused for the PWKL method, irrespective of test length and model type. In addition, the RT method yielded a slightly more even distribution of item usage than did the PWKL method, but it still had higher TORs, UIRs, and OIRs than the other methods, indicating that uneven distribution of item usage occurred. Compared with the RT method, the BRT method produced lower scaled χ^2 values, TORs, UIRs, and OIRs under most conditions. That is, the scaled χ^2 values that the BRT method produced were much lower than those produced by the PWKL and RT methods. The TORs were also lower than those of the RT under all conditions, and the UIRs of the BRT method were lower than those of the RT method under all conditions except one ($J = 40$, the DINA model). As for the RP and SDBS methods, the SDBS method yielded slightly better item exposure control than the RP method when the test length was short ($J = 25$); however, it performed slightly worse than the RP method when the test length was long ($J = 40$). Both the RP and SDBS methods performed better than the BRT method under all four indices (i.e., scaled χ^2 value, TOR, UIR, and OIR). The differences between the BRT method and the RP and SDBS methods were relatively low in terms of the TOR and OIR and higher in terms of the scaled χ^2 value and UIR. In summary, the BRT method yielded relatively poor item usage distribution compared with the RP and SDBS methods but more even distribution of item usage than the original PWKL and RT methods.

Table 4 presents the results of item exposure control for six attributes. Most of the results in the table exhibit a similar pattern to that observed with four attributes. Specifically, the random method had optimal item exposure control under all conditions. As for the scaled χ^2 value and TOR indices, the priority of the rest of the five methods was the RP, SDBS, BRT, RT, and PWKL. With respect to UIR and OIR, the RP method yielded the lowest values under all conditions except one ($J = 25$, the RRUM), in which the SDBS yielded the lowest UIR. The SDBS yielded lower UIRs and OIRs than the BRT and RT methods under all conditions, and the BRT method performed better than the RT method for the two indices under most conditions.

DISCUSSION AND CONCLUSIONS

Inspired by the studies of Wang et al. (2011) and Zheng and Wang (2017), we combined the binary searching algorithm with the RT method to develop the BRT method for CD-CAT. Because the core components of the SDBS method (i.e., a binary searching algorithm) and RT method were integrated into the BRT method, the RT method can be considered a specific case of BRT method, which means that the RT method can be obtained by adding some additional constraints to the BRT method. A simulation study was performed to investigate the performance

TABLE 1 | The correct classification for four attributes.

Item selection method	$J = 25$					$J = 40$				
	PCCR	ACCR				PCCR	ACCR			
		A1	A2	A3	A4		A1	A2	A3	A4
DINA										
PWKL	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
RP	0.997	0.999	1.00	1.000	0.998	1.000	1.000	1.000	1.000	1.000
RT	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SDBS	0.993	0.998	0.998	0.997	0.998	1.000	1.000	1.000	1.000	1.000
BRT	0.997	0.997	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
Random	0.876	0.952	0.978	0.970	0.964	0.962	0.990	0.992	0.994	0.985
RRUM										
PWKL	0.986	0.998	0.994	0.996	0.996	0.999	1.000	1.000	1.000	1.000
RP	0.886	0.972	0.962	0.965	0.968	0.954	0.991	0.986	0.986	0.986
RT	0.975	0.994	0.992	0.994	0.994	0.998	1.000	1.000	0.999	1.000
SDBS	0.761	0.932	0.930	0.930	0.928	0.902	0.975	0.972	0.970	0.969
BRT	0.920	0.978	0.976	0.978	0.976	0.990	0.996	0.998	0.996	0.999
Random	0.572	0.865	0.874	0.865	0.864	0.716	0.913	0.930	0.923	0.904

DINA refers to the deterministic input, noisy "and" gate model; RRUM refers to the reduced reparametrized unified model; PWKL refers to the posterior weighted Kullback-Leibler; RP refers to the restrictive progressive method; RT refers to the restrictive threshold method; SDBS refers to the stratified dynamic binary searching method; BRT refers to the binary RT method; PCCR refers the pattern correct classification rate; and the ACCR refers to the attribute correct classification rate.

TABLE 2 | The correct classification for six attributes.

Item selection method	$J = 25$							$J = 40$						
	PCCR	ACCR						PCCR	ACCR					
		A1	A2	A3	A4	A5	A6		A1	A2	A3	A4	A5	A6
DINA														
PWKL	0.995	0.999	0.998	0.998	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
RP	0.955	0.989	0.990	0.990	0.990	0.992	0.990	0.998	1.000	0.999	0.999	1.000	1.000	1.000
RT	0.992	0.997	0.998	0.999	1.000	0.998	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SDBS	0.958	0.992	0.992	0.993	0.994	0.986	0.993	0.996	1.000	1.000	0.999	1.000	0.998	1.000
BRT	0.980	0.996	0.997	0.997	0.997	0.997	0.995	0.998	1.000	1.000	1.000	1.000	0.999	1.000
Random	0.650	0.942	0.930	0.896	0.931	0.918	0.924	0.829	0.972	0.968	0.966	0.971	0.966	0.960
RRUM														
PWKL	0.904	0.977	0.983	0.982	0.984	0.984	0.980	0.984	0.997	0.996	0.996	0.998	0.998	0.998
RP	0.705	0.944	0.938	0.921	0.940	0.940	0.926	0.847	0.978	0.972	0.964	0.974	0.968	0.958
RT	0.876	0.976	0.971	0.975	0.978	0.982	0.978	0.978	0.998	0.998	0.994	0.995	0.998	0.994
SDBS	0.542	0.908	0.896	0.880	0.896	0.900	0.892	0.772	0.962	0.950	0.956	0.952	0.950	0.948
BRT	0.726	0.946	0.940	0.937	0.939	0.943	0.958	0.907	0.986	0.979	0.978	0.981	0.987	0.985
Random	0.308	0.838	0.828	0.792	0.828	0.811	0.804	0.468	0.910	0.883	0.858	0.892	0.858	0.873

of this novel item exposure control method. According to the results, the BRT method has more discernible merits than the PWKL and RT methods in terms of item exposure control, irrespective of the number of attributes, model type, and test length, although it yields slightly less accurate classification than the PWKL and RT methods under all conditions. The BRT method yields relatively poor item exposure control but more accurate classification under all conditions when compared with the RP and SDBS methods.

The results demonstrate that differences in the PCCRs between the BRT and RT approaches are minor for the DINA model, whereas the BRT method achieves superior item exposure control to the RT method. This is especially true when the scaled χ^2 value and the TOR are examined. These findings indicate that the BRT method, to some degree, is a good candidate for the RT method when the DINA model is applied to CD-CAT with small number of attributes or long length of tests. Compared with the DINA model, the RRUM reveals larger differences in the PCCRs

TABLE 3 | The usage of items for four attributes.

	$J = 25$				$J = 40$			
	χ2	TOR	UIR	OIR	χ2	TOR	UIR	OIR
DINA								
PWKL	213.183	0.496	0.800	0.079	210.750	0.522	0.702	0.127
RP	5.116	0.062	0.000	0.000	2.475	0.088	0.000	0.000
RT	80.505	0.219	0.450	0.052	42.970	0.172	0.000	0.081
SDBS	3.949	0.060	0.010	0.000	4.488	0.092	0.000	0.000
BRT	12.837	0.078	0.208	0.000	9.604	0.103	0.027	0.010
Random	0.233	0.052	0.000	0.000	0.219	0.083	0.000	0.000
RRUM								
PWKL	225.213	0.521	0.825	0.079	228.810	0.560	0.752	0.142
RP	12.477	0.078	0.265	0.000	7.688	0.099	0.000	0.000
RT	145.923	0.356	0.665	0.075	167.067	0.431	0.569	0.135
SDBS	6.844	0.066	0.017	0.002	9.960	0.104	0.002	0.023
BRT	23.454	0.100	0.338	0.012	32.619	0.151	0.258	0.075
Random	0.220	0.052	0.000	0.000	0.215	0.083	0.000	0.000

TABLE 4 | The usage of items for six attributes.

	$J = 25$				$J = 40$			
	χ2	TOR	UIR	OIR	χ2	TOR	UIR	OIR
DINA								
PWKL	205.962	0.481	0.792	0.083	198.364	0.496	0.677	0.127
RP	4.952	0.062	0.000	0.000	2.272	0.088	0.000	0.000
RT	89.060	0.237	0.481	0.050	90.520	0.272	0.196	0.102
SDBS	9.333	0.071	0.190	0.000	10.053	0.104	0.010	0.021
BRT	20.152	0.094	0.298	0.008	21.490	0.128	0.144	0.050
Random	0.196	0.052	0.000	0.000	0.230	0.083	0.000	0.000
RRUM								
PWKL	199.284	0.467	0.783	0.092	200.328	0.500	0.690	0.135
RP	11.023	0.075	0.210	0.000	6.650	0.097	0.000	0.000
RT	122.366	0.307	0.638	0.075	143.225	0.381	0.548	0.125
SDBS	13.047	0.079	0.125	0.008	14.427	0.113	0.029	0.031
BRT	21.810	0.097	0.275	0.015	32.583	0.151	0.183	0.075
Random	0.225	0.052	0.000	0.000	0.205	0.083	0.000	0.000

between the BRT and RT methods. That is, the RT method produces higher PCCRs than the BRT method in all conditions. However, the BRT method performs better than the RT method with regard to item exposure control. These results indicate that there is a trade-off between measure accuracy and item usage when a selection is made from the RT and BRT methods for the RRUM. As for how to choose reference values to interpret the evaluation criteria (e.g., scaled χ^2, overlap rate), there are no definite answer, and reference values can be determined by test purpose. The BRT method can be used to select items to be administered if obtaining an even distribution of item usage is the primary goal, wherein both item exposure control indices (i.e., lower scaled χ^2 and overlap rate) and measurement accuracy (i.e., higher PCCR) are important. In contrast, the RT method can be applied if measurement accuracy is the major consideration, such as classroom settings. In such situations, higher scaled χ^2 and overlap rate are acceptable.

Furthermore, there might be a ceiling effect in measurement accuracy for the DINA model. This is because values of measurement accuracy are close to upper bound (i.e., 1.0) under both 25 and 40 items conditions. We further investigated this effect by conducting a pilot study with varying test length (10, 15, and 20 items) and varying number of attributes (four and six attributes) for the DINA model. Its results showed that the PCCRs are larger than 0.95 for the RT and BRT methods in conditions with 15 items and four attributes and are close to the upper bound in conditions with 20 items. The PCCRs are close to 0.95 under conditions with 20 items and six attributes and close to 1.0 under conditions with 25 items. These results confirmed the ceiling effect of measurement accuracy for the

DINA model. In addition, the pilot study also showed that differences in the PCCRs between the RT and BRT methods are smaller for conditions with four attributes than those with six attributes regardless of which test length is used. In particular, differences in the PCCRs ranged from 0.0 to 0.02 and 0.0 to 0.09 for conditions with four and six attributes, respectively. This result indicated that the number of attributes has a positive effect on the differences of the PCCRs between the RT and BRT methods. The BRT method performs worse in PCCR than the RT method under conditions with large number of attributes and short length of tests.

Overall, the proposed BRT method, to some extent, can better balance the trade-off between correct classification and item exposure compared with prior methods. It yields slightly less accurate classification compared with the original PWKL and RT methods; however, it achieves superior item exposure control. In addition, although the BRT method provides slightly poorer item exposure control than do the RP and SDBS methods, it yields more accurate measurements.

Although the current study presents promising findings, the following potential future directions should be considered. First, the majority of studies that have explored item exposure have been based on the PWKL method. Other flexible methods, such as the SHE, MI, and MPWKL, should be investigated further. Second, both the DINA model and the RRUM are specific CDMs, which assume either conjunctive or disjunctive relationships between items in one tests. By contrast, the general CDMs relax the constraints of the specific CDMs. That is, they allow each item to select the optimal model to achieve optimal results (Ravand, 2016). Whether the new method can be applied to general CDMs is worthy of investigation in the future. It is worth noting that a variety of CDMs have been developed for varying situations in recent years, each of which makes specific assumptions about the relationship between item response and the attributes that item measured. Thus, assumptions that have been made for a situation, to some degree, determines the selection of a CDM. As well as data-driven model selection, for instance, use Akaike's information criterion and Bayesian information criterion to select the CDMs. Third, the application of the new method to the dual-objective CD-CAT (McGlohen and Chang, 2008; Wang et al., 2012, 2014; Dai et al., 2016; Kang et al., 2017; Zheng et al., 2018) could be investigated in the future. The dual-objective CD-CAT combines the IRT model and CDMs; therefore, it may be able to provide both an overall score and specific diagnostic information for individuals. Because the novel item exposure control method is proposed primarily for single-objective CD-CAT, it requires modification before application to dual-objective CD-CAT. Fourth, the new method could be extended to variable-length CD-CAT in the future study. However, it is important to note that the new method needs some modifications before its application to variable-length CD-CAT. This is because the posterior probability of an attribute profile is usually used as termination rule in variable-length CD-CAT. As such, the application of the new method to variable-length CD-CAT would be more complicated than its application to fixed-length CD-CAT. Last, true item parameters, rather than estimated item parameters, are used in the current study. As Huang (2018) and Sun et al. (2020) demonstrated, measurement accuracy is decreased when estimated item parameters are used. In other words, the reliability of item exposure control methods is relatively low with estimated item parameters. Therefore, further studies can consider investigating the reliability of the BRT method when estimated item parameters are used.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the Cultural Experts, and Four Groups of Talented People Foundation of China and National Natural Science Foundation of China (Grant No. 32071093).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.517155/full#supplementary-material

REFERENCES

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

Chang, H. H., and Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Appl. Psychol. Meas.* 23, 211–222. doi: 10.1177/01466219922031338

Chen, P., Xin, T., Wang, C., and Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika* 77, 201–222. doi: 10.1007/s11336-012-9255-7

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

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method. *Educ. Psychol. Meas.* 70, 902–913. doi: 10.1177/0013164410366693

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

Gierl, M. J., Leighton, J. P., and Hunka, S. M. (2007). "Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills," in *Cognitive Diagnostic Assessment for eduCation: Theory and Application*, eds J. Leighton and M. Gierl (Cambridge: Cambridge University Press), 242–247.

Hartz, S. M. C. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). Champaign, IL: University of Illinois at Urbana Champaign.

Hsu, C. L., Wang, W. H., and Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Appl. Psychol. Meas.* 37, 563–582. doi: 10.1177/0146621613488642

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

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

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kang, H. A., Zhang, S., and Chang, H. H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *J. Educ. Meas.* 54, 165–183. doi: 10.1111/jedm.12139

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

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

Lin, C. J., and Chang, H. H. (2019). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educ. Psychol. Meas.* 79, 335–357. doi: 10.1177/0013164418790634

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

McGlohen, M., and Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behav. Res. Methods*. 40, 808–821. doi: 10.3758/BRM.40.3.808

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

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assess.* 34, 782–799. doi: 10.1177/0734282915623053

Revuelta, J., and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *J. Educ. Meas.* 35, 311–327. doi: 10.1111/j.1745-3984.1998.tb00541.x

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

Sun, X., Liu, Y., Xin, T., and Song, N. (2020). The impact of item calibration error on variable-length cognitive diagnostic computerized adaptive testing. *Front. Psychol.* 11:575141. doi: 10.3389/fpsyg.2020.575141

Tatsuoka, C., and Ferguson, T. (2003). Sequential classification on partially ordered sets. *J. R. Stat. Soc.* 65, 143–157. doi: 10.1111/1467-9868.00377

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educ. Psychol. Meas.* 73, 1017–1035. doi: 10.1177/0013164413498256

Wang, C., Chang, H. H., and Douglas, J. (2012). Combining CAT with cognitive diagnosis: a weighted item selection approach. *Behav. Res. Methods*. 44, 95–109. doi: 10.3758/s13428-011-0143-3

Wang, C., Chang, H. H., and Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *J. Educ. Meas.* 48, 255–273. doi: 10.1111/j.1745-3984.2011.00145.x

Wang, C., Zheng, C., and Chang, H. H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *J. Educ. Meas.* 51, 358–380. doi: 10.1111/jedm.12057

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

Xu, X., Chang, H., and Douglas, J. (2003). "A simulation study to compare CAT strategies for cognitive diagnosis," in *Paper Presented at the Annual Meeting of National Council on Measurement in Education* (Chicago, IL).

Zheng, C., and Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Appl. Psychol. Meas.* 40, 608–624. doi: 10.1177/0146621616665196

Zheng, C., He, G., and Gao, C. (2018). The information product methods: a unified approach to dual-purpose computerized adaptive testing. *Appl. Psychol. Meas.* 42, 321–324. doi: 10.1177/0146621617730392

Zheng, C., and Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Appl. Psychol. Meas.* 41, 561–576. doi: 10.1177/0146621617707509

Conflict of Interest: 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership