Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-04959-w

OPEN

Check for updates

# Two-stage polytomous attribute estimation for cognitive diagnostic models: overcoming computational challenges in large-scale assessments with many polytomous attributes

Yuting Han [1,2,3], Feng Ji [4] & Zhehan Jiang [5,6,7]✉

Cognitive diagnosis models (CDMs) have been advocated as a useful tool in calibrating large-scale assessments, yet the computational challenges are inevitably amplified when the modeling complexity (e.g., the number and the levels of attributes) increases. This study presents a critical scenario, a large-scale national medical certification exam, where CDM with many polytomous attributes (mpCDM) is of great utility, but poses great computational challenges to many popular open-source CDM software packages. We developed a novel two-stage estimation method and assessed its performance through a Monte Carlo simulation study under various conditions of attribute number, item number, item quality, and sample size. Results indicate that the proposed method maintains high accuracy in handling large-scale data while effectively overcoming computational capacity limitations, especially in scenarios with many polytomous attributes, large numbers of items, and substantial sample sizes. Furthermore, we applied the proposed method to a large-scale health examination dataset, demonstrating its effectiveness in practice. This study contributes to the field of psychometrics by offering a simple yet effective solution to the computational challenges inherent in implementing mpCDMs for large-scale assessments, providing a practical tool for diagnostic analyses in educational and professional certification contexts.

[1] Cognitive Science and Allied Health School, Beijing Language and Culture University, Beijing, China. [2] Institute of Life and Health Sciences, Beijing Language and Culture University, Beijing, China. [3] Key Laboratory of Language and Cognitive Science (Ministry of Education), Beijing Language and Culture University, Beijing, China. [4] Department of Applied Psychology and Human Development, University of Toronto, Toronto, Canada. [5] Institute of Medical Education, Peking University, Beijing, China. [6] National Center for Health Professions Education Development, Peking University, Beijing, China. [7] Peking University Health Science Center-Chaoxing Joint Laboratory for Digital and Smart Medical Education, Beijing, China. ✉email: jiangzhehan@gmail.com

## Introduction

Learning is a complex process that relies on various factors for success. Among these factors, receiving feedback plays a crucial role in any learning process. Learners and educators often rely on exam scores to assess mastery over a given knowledge space, facilitating a positive feedback loop in teaching and learning (Haladyna and Kramer, 2004). This aligns well with a growing trend demanding assessment frameworks to offer more detailed diagnostic information to facilitate learning (Eva et al. 2012). A deeper understanding of learning outcomes can help students adjust their strategies and methods in a timely manner for better long-term academic performance. However, most exams, particularly high-stakes ones, typically provide only a composite score by design, which limits the available information to learners and educators (Lee et al. 2011; Park et al. 2018), while a more fine-grained way of reporting can offer diagnostic feedback that can help learners pinpoint areas for improvement (Haladyna and Kramer, 2004). Hence, an increasing number of educational researchers agree that educational assessments should be more diagnostic, thus making teaching and learning more effective (Hattie and Timperley, 2007).

The cognitive diagnosis model (CDM) offers a theoretical and technical solution for the pressing need for personalized feedback in educational measurement. As a probabilistic model integrating cognitive variables, CDM analyzes exam data to reveal individual differences in knowledge structures, cognitive processes, and skills, thereby providing personalized and detailed diagnostic feedback (Leighton and Gierl, 2007). Many scholars in educational measurement, such as Sinharay and colleagues (2011), urge caution in how we report and use sub-scores from educational assessments. They point out that current tools such as item response theory models and CDMs require specific conditions to yield valid and reliable sub-scores (Schoenherr and Hamstra, 2016; Sinharay, 2010). However, Liu et al. (2018) emphasize the unique benefits of using cognitive diagnostic score reports in large-scale assessments. These assessments cover various subdomains and aim primarily to determine whether students pass or fail. CDMs streamline the scoring and categorization process, offering a comprehensive solution. This approach moves away from traditional methods of "standard setting", where cut-off points are subjectively determined (Cizek, 2012), leading to more accurate and less error-prone diagnostic results (Robitzsch et al. 2017).

CDMs involve examinees' response data and the *Q*-matrix, which defines the knowledge, strategies, and skills required to answer specific items, to precisely assess mastery over different cognitive attributes. Many variants of CDMs have been proposed, and they vary in theoretical foundations, model assumptions, and parameter definitions to meet diverse functional and objective requirements (Leighton and Gierl, 2007). Traditional CDMs—such as the saturated model represented by the generalized deterministic input, noisy "and" gate (G-DINA) model (de la Torre, 2011) and the simplified models like the deterministic inputs, noisy "and" gate (DINA) model (de la Torre, 2009; Junker and Sijtsma, 2001) and the deterministic inputs, noisy "or" gate (DINO) model (Templin and Henson, 2006)—assume binary attributes, i.e., mastered ("1") or not mastered ("0").

In education and assessment, cognitive skills are widely recognized to develop across multiple levels, as exemplified by Bloom's taxonomy (Bloom, 1956) and its later revision by Krathwohl (2002), which defines a hierarchical progression from basic skills like remembering to complex abilities such as creating. This multi-level conceptualization of cognitive development is reflected in major international assessments—for instance, the National Assessment of Educational Progress (NAEP) defines mathematics attributes at multiple levels (National Assessment Governing Board, 2008), while the TIMSS 2015 science framework distinguishes between knowing, applying, and reasoning (Jones et al. 2013). These frameworks underscore an important reality in educational measurement: the assessment of knowledge and skills requires moving beyond simple dichotomous classifications of mastery versus non-mastery.

To address this need for more nuanced assessment, researchers have developed various polytomous cognitive diagnosis models (pCDM) to evaluate examinees' attribute mastery at multiple levels. The development of these models began with Templin's (2004) work on extending the reparameterized unified model (RUM; Hartz, 2002) to handle polytomous attributes through the RUM-PA model and its constrained version (*c*RUM-PA). During the same period, Karelitz (2004) introduced the ordered category attribute coding (OCAC) framework with the OCAC-DINA model to define mastery levels as multiple ordered categories. Von Davier (2008, 2014) subsequently proposed the general diagnostic model (GDM) that accommodates polytomous attributes through flexible mapping functions. Chen and de la Torre (2013) made significant contributions with the polytomous generalized DINA (pG-DINA) model, extending the G-DINA framework to account for main effects and interactions in polytomous settings. Sun et al. (2013) further expanded the field by developing a framework for polytomous attributes through the generalized distance discriminating method. The evolution of polytomous CDMs continued with several important developments. Zhan et al. (2016) proposed the reparameterized polytomous attributes DINA (RPa-DINA) model as a restricted version of pG-DINA. Chen and de la Torre (2018) introduced the general polytomous diagnosis model (GPDM) to handle both polytomous responses and attributes. The field saw further advances with Wang and Chen's (2020) response accuracy model (RAM), Zhan and colleagues' (2020) the partial mastery DINA (PM-DINA) model incorporating higher-order latent structures, and Yakar and colleagues' (2021) fully additive model (*f*A-M). Additional contributions include Bao's (2019) polytomous diagnostic classification model (PDCM), Ma's (2022) higher-order general cognitive diagnosis model (HO-PCDM), and Zhan and colleagues' (2023) Ordinal-DINA model for longitudinal diagnosis. Most recently, de la Torre et al. (2025) proposed the saturated polytomous cognitive diagnosis model (sp-CDM), offering a comprehensive framework that subsumes existing polytomous CDMs while allowing differential attribute level contributions. These methodological advances have enabled more refined assessments of cognitive mastery, as demonstrated by Mohsenpour's (2019) successful application in assessing mathematical literacy among adolescent students, providing detailed diagnostic information about varying competency levels.

Large-scale educational assessments often involve multiple attributes that need to be evaluated at different cognitive levels, making polytomous cognitive diagnosis particularly valuable for providing comprehensive diagnostic feedback. However, the application of CDM with many polytomous attributes (mpCDM) to large-scale assessments faces significant practical challenges. The primary concern is model complexity and computational demands. As the number of items and attributes increases, the number of model parameters grows rapidly, making parameter estimation and interpretation increasingly challenging (Bradshaw et al. 2014). This computational burden is evident in commonly used CDM software packages like GDINA (Ma and de la Torre, 2020) and CDM (George et al. 2016), which often struggle to efficiently handle CDMs with more than eight polytomous attributes. While more advanced software solutions employing parallel algorithms exist for estimating high-dimensional diagnostic models with polytomous attributes (Khorramdel et al.

2019; von Davier, 2016), these tools are often developed by major testing organizations and may not be readily accessible to all practitioners. The complexity of implementing and using such software poses additional challenges for widespread adoption. Furthermore, as the number of skills or attributes increases, model identifiability issues may arise (von Davier, 2008).

Beyond computational issues, the calibration process itself presents significant challenges. The detailed calibration required for each attribute in polytomous settings creates a substantial workload and financial burden (Birenbaum et al. 1993), particularly in large-scale assessments that typically measure a range of complex, interrelated skills and knowledge (Tatsuoka, 2009). Resource constraints often limit practitioners to calibrating cognitive levels only at the item level, as conducting multi-level calibrations for each cognitive attribute across all items proves impractical. The effectiveness of this simplified approach in producing satisfactory estimates of examinees' polytomous attributes requires further investigation.

To fully leverage the potential of large-scale educational assessments, this research aims to develop novel polytomous attribute estimation methods. These methods are designed to serve as effective estimation techniques for mpCDMs. The primary objective is to achieve precise estimations of examinees' polytomous attribute mastery under these complex conditions. We will further show that the proposed methods work well for conditions with large sample sizes, high item volumes, and multiple attributes with hierarchical cognitive levels defined at the item level, which is precisely when traditional estimation methods become infeasible in practice.

The remainder of the article is structured as follows: we start by introducing the G-DINA model with dichotomous and polytomous attributes, along with the basic idea of the two-stage polytomous attribute estimation method. Then, a Monte Carlo simulation study is conducted to assess the accuracy of the parameter estimates of these two-stage approaches and compare their consistency with the estimates derived from mpCDM. This is followed by an empirical study that validates the practicality of these new methods in real-world settings. The study concludes with some recommendations.

## Methods

**The G-DINA model with dichotomous and polytomous attributes**. This study employs the pG-DINA model for assessing multiple levels of attribute mastery. We begin with its predecessor, the G-DINA model, which provides the essential framework.

CDMs can be classified as simplified or saturated types based on their scope of application (Junker and Sijtsma, 2001; see also Hou, 2013 for a more recent discussion). Simplified CDMs have more restrictive assumptions in item response function construction, offering narrower applicability, simpler structures, and higher parameter estimation precision. Conversely, saturated CDMs, such as the G-DINA model, which considers all potential main effects and interactions, are characterized by fewer limitations and broader applicability but entail more complex structures and demand large sample sizes for accurate and stable parameter estimation (Jiang and Carter, 2019). The G-DINA model incorporates three link functions: identity, logit, and log link functions (de la Torre, 2011). The identity link function within the G-DINA model is formulated as follows:

$$P\left(X_j = 1|\boldsymbol{\alpha}_{lj}^*\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}^* + \sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}^*\alpha_{lk'}^* + \ldots + \delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk}^* \tag{1}$$

In this formula, $\delta_{j0}$ is the intercept for the item $j$, representing the probability of an examinee answering item $j$ correctly without mastering any of its measured attributes. $\delta_{jk}$ is the main effect of attribute $k$, indicating the increase in probability of correctly answering item $j$ as a result of mastering attribute $k$. And $\delta_{jkk'}$ represents the interaction effect between attributes $k$ and $k'$ in item $j$, while $\delta_{j12\ldots K_j^*}$ accounts for the interaction effects among all attributes measured by item $j$. Elements of the $Q$ matrix, $q_{jk}$, are binary variables (0 or 1), where $q_{jk} = 1$ indicates that item $j$ measures attribute $k$, and otherwise not. $K_j^* = \sum_{k=1}^{K} q_{jk}$ denotes the total number of attributes measured by item $j$. $\boldsymbol{\alpha}_{lj}^* = (\alpha_{l1}^*, \ldots, \alpha_{lK_j^*}^*)\prime$ is the reduced attribute vector based on item $j$, where $l = 1, \ldots, K_j^*.\alpha_{lk}^*$ is also a binary variable, where $\alpha_{lk}^* = 1$ indicates mastery of attribute $k$ for pattern $l$. The formula describes the probability of an examinee with an attribute mastery pattern $\alpha_{lj}^*$ correctly answering item $j$. Notably, saturated CDMs can transform into simplified models under certain conditions. By constraining some parameters in Eq. (1) to zero, reparameterizing, and selecting suitable link functions, various submodels within the G-DINA model framework can be obtained. However, these models assume binary attributes, signifying mastery ("1") or non-mastery ("0").

While the G-DINA model provides a comprehensive framework for dichotomous attributes, many educational assessments require evaluating knowledge and skills at multiple levels. Chen and de la Torre (2013) extended the G-DINA framework to accommodate polytomous attributes through the pG-DINA model, which maintains the model's flexibility while enabling a more nuanced assessment of attribute mastery. The item response function (IRF) of this model is formulated as follows:

$$P\left(X_j = 1|\boldsymbol{\alpha}_{lj}^{**}\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}^{**} + \sum_{k'=k+1}^{K_j^*}\sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}^{**}\alpha_{lk'}^{**} + \ldots + \delta_{j12\ldots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lk}^{**} \tag{2}$$

Here, $\alpha_{lj}^{**} = (\alpha_{l1}^{**}, \ldots, \alpha_{lK_j^*}^{**})\prime$ is defined as the collapsed attribute vector, and $\alpha_{lk}^{**} = I(\alpha_{lk} \geq q_{jk})$ is the collapsed dichotomous attribute, where $\alpha_{lk}$ refers to the $P$-level attribute. $K_j^* = \sum_{k=1}^{K} I(q_{jk} > 0)$ is still used to denote the number of required attributes for item $j$, while other parameters follow the same definitions as in the G-DINA model (Chen and de la Torre, 2013). Similar to the G-DINA model, by applying various constraints, the pG-DINA model can be transformed into different polytomous cognitive diagnosis models. In the subsequent sections of this paper, we will primarily focus on the pG-DINA model. This choice of the pG-DINA model as our methodological foundation is motivated by two key considerations: its theoretical comprehensiveness in handling polytomous attributes and the wide availability of software packages implementing the G-DINA framework.

**The two-stage polytomous attribute estimation method**. While the pG-DINA model provides a comprehensive framework for polytomous cognitive diagnosis, its practical implementation becomes particularly challenging in large-scale assessments with many polytomous attributes and hundreds of items, even when adopting a simplified item-level calibration approach where all attributes within an item share the same cognitive level. To address these computational challenges while preserving diagnostic capabilities, we propose a novel two-stage polytomous attribute estimation method, designed specifically for contexts where cognitive levels are uniformly defined at the item level.

The general idea of the proposed two-stage polytomous attribute estimation method involves two steps. Specifically:

**Step 1:** We decompose the polytomous estimation problem into $P$ separate dichotomous estimation problems by grouping items according to their cognitive levels. This decomposition significantly reduces computational complexity since each group only requires estimating binary attribute patterns using the G-DINA model. For each cognitive level $p(p = 1, \ldots, P)$, we treat each level as an independent dichotomous CDM analysis. The Expectation-Maximization algorithm with marginal maximum-likelihood estimation (MMLE/EM) can be used for model parameter estimation (de la Torre, 2009, 2011), and maximum a posteriori estimation (MAP) for attribute mastery patterns.

**Step 2:** To obtain the examinees' polytomous attribute patterns, we propose two methods of merging the dichotomous attribute patterns:

*Maximal merging method (denoted as* P_max*)*. The maximal merging method assigns an examinee's mastery level based solely on the highest cognitive level achieved, regardless of performance at lower levels. For each attribute of the same examinee, the highest cognitive level at which the attribute is estimated to be mastered is taken as the mastery level of that attribute. To illustrate, consider an examinee who masters an attribute at level 3 but not at levels 1 or 2. *P*_max would assign level 3 mastery, focusing exclusively on the highest level of demonstrated capability while disregarding performance at lower levels. Mathematically, the cognitive level of attribute $k$ for examinee $i$ can be estimated using the following formula:

$$\hat{\alpha}_{ik} = \underset{p\in[1,P]}{\mathrm{argmax}}[p\times I(\alpha_{ikp} = 1)], i = 1, \ldots, N; k = 1, \ldots, K; p = 1, \ldots, P \tag{3}$$

where $\alpha_{ikp}$ represents the dichotomous mastery status of attribute $k$ for examinee $i$ in level $p$. The term $I(\alpha_{ikp} = 1)$ is an indicator function, where its value is 1 if $\alpha_{ikp}$ equals 1, and 0 otherwise. Therefore, $p\times I(\alpha_{ikp} = 1)$ is not zero only when $\alpha_{ikp}$ equals 1, and the argmax function will return the index of the maximum value, i.e., the largest $p$ value (attribute level) where $\alpha_{ikp}$ equals 1.

*Linear conditional merging method (denoted as* P_linear*)*. Cognitive development often exhibits hierarchical patterns in learning processes, as reflected in several influential educational theories. Bloom's Taxonomy (1956) systematically organizes cognitive processes from basic to complex levels, progressing from knowledge and comprehension through application to higher-order processes. Anderson and Krathwohl's (2001) revision further emphasized that mastery of higher-level cognitive processes typically requires proficiency in lower-level abilities. This hierarchical progression is reflected in measurement frameworks such as the OCAC (Karelitz, 2004), which explicitly posits that achieving higher mastery levels necessitates the attainment of more fundamental levels. While cognitive diagnostic models like pG-DINA do not explicitly mandate such hierarchical relationships, they can accommodate structural patterns where mastery of certain levels serves as a prerequisite for achieving higher levels (de la Torre and Douglas, 2004).

Drawing from these theoretical frameworks and building upon the *P*_max method, we propose the linear conditional merging method. This approach posits that an examinee can only master a higher-level attribute if all lower-level attributes are mastered. This can be represented by the following equation:

$$\hat{\alpha}_{ik} = \underset{p\in[1,P]}{\mathrm{argmax}}[p\times I(\prod_{v=1}^{v=p}\alpha_{ikv} = 1)],\ i = 1, \ldots, N; \\ k = 1, \ldots, K; p = 1, \ldots, P \tag{4}$$

where $I(\prod_{v=1}^{v=p}\alpha_{ikv} = 1)$ is an indicator function that is equal to 1 only when the mastery status of the attribute $k$ equals 1 for all levels from the first to the $p$th level and 0 otherwise.

*P*_linear enforces hierarchical mastery requirements where higher levels depend on mastery of lower levels, while *P*_max determines mastery based solely on the highest level achieved regardless of performance at lower levels. Consider an example with three attributes where the maximum cognitive level is 2. When an examinee's dichotomous attribute patterns are $\hat{\alpha}_1 = (1, 1, 0)$ and $\hat{\alpha}_2 = (0, 1, 1)$, *P*_max yields a polytomous attribute pattern of $\hat{\alpha} = (1, 2, 2)$, while *P*_linear produces $\hat{\alpha} = (1, 2, 0)$. This difference arises from *P*_linear's enforcement of hierarchical mastery requirements: although the third attribute shows mastery at level 2, the lack of mastery at level 1 results in an overall mastery level of 0 under *P*_linear, whereas *P*_max assigns level 2 based solely on the highest demonstrated capability.

Moreover, the item parameters estimated from each cognitive level group using G-DINA can be directly used without additional merging because they are estimated within their respective cognitive levels. The resulting item parameters maintain their interpretability within each cognitive level while significantly reducing computational complexity.

## Simulation study

To evaluate the accuracy of the polytomous attribute mastery patterns obtained using the two-stage estimation methods (*P*_max and *P*_linear) under various conditions and to verify the consistency of these results with those directly derived from the polytomous CDM (pG-DINA), a Monte Carlo simulation study was conducted.

**Simulation design**. A $3 \times 3 \times 3 \times 2$ design was employed, with the following specific factors:

(1) Number of attributes ($K$), set at three levels: 3, 5, and 8 attributes.
(2) Number of items, set based on multiples of the number of attributes, at three levels: 10, 20, and 50 times. For instance, if the number of attributes is 8, then the number of items would be 80, 160, and 400.
(3) Number of examinees ($N$), set at three levels: 500, 1000, and 2000.
(4) Item quality. Two levels were considered: high and low. For high-quality items, both guessing and slipping parameters were set to 0.1 across all items. For low-quality items, these parameters were set to 0.3.

Additionally, the maximum level for all attributes was fixed at 4 ($P = 4$), and a constraint was imposed to ensure that each item measured no more than four attributes.

**Simulation process**. The simulation experiment was conducted according to the following procedure:

(1) *Simulation of true attribute mastery patterns*: A multi-dimensional normal distribution was specified for the attributes. The mean vector was set to zero, and the correlation coefficients between attributes were randomly selected from the range of 0.5–0.8, consistent with typical inter-attribute correlations reported in the literature

**Table 1 The $R_p$ and simplified $Q_p$ matrices for assessing three attributes with the highest attribute level of 4.**

| No. | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | No. | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
|---|---|---|---|---|---|---|---|
| **1** | 1 | 0 | 0 | **15** | 0 | 1 | 1 |
| **2** | 0 | 1 | 0 | **16** | 1 | 1 | 1 |
| **3** | 0 | 0 | 1 | **17** | 2 | 2 | 0 |
| **4** | 2 | 0 | 0 | **18** | 2 | 0 | 2 |
| **5** | 0 | 2 | 0 | **19** | 0 | 2 | 2 |
| **6** | 0 | 0 | 2 | **20** | 2 | 2 | 2 |
| **7** | 3 | 0 | 0 | **21** | 3 | 3 | 0 |
| **8** | 0 | 3 | 0 | **22** | 3 | 0 | 3 |
| **9** | 0 | 0 | 3 | **23** | 0 | 3 | 3 |
| **10** | 4 | 0 | 0 | **24** | 3 | 3 | 3 |
| **11** | 0 | 4 | 0 | **25** | 4 | 4 | 0 |
| **12** | 0 | 0 | 4 | **26** | 4 | 0 | 4 |
| **13** | 1 | 1 | 0 | **27** | 0 | 4 | 4 |
| **14** | 1 | 0 | 1 | **28** | 4 | 4 | 4 |

The entire $Q_p$ matrix for the test is listed in the table, with the $R_p$ matrix highlighted in gray.

(Kunina-Habenicht et al. 2012; Sinharay et al. 2011). The resulting attribute vectors were then discretized into values of 0, 1, 2, 3, and 4.

(2) *Simulation of the Q-matrix*: A polytomous reachability matrix ($R_p$) was first constructed, followed by generating a simplified polytomous $Q$ (denoted as $Q_p$) matrix using an expansion algorithm (Sun et al. 2013). For instance, for three attributes with the highest attribute level of 4, the corresponding $R_p$ matrix and simplified $Q_p$ matrix are shown in Table 1. The reachability matrix is key in cognitive diagnostic analysis for diagnosing each attribute (Sun et al. 2013), and the number of reachability matrices included in the test $Q$-matrix can affect the accuracy rates. In a polytomous attribute circumstance, even if each item only assesses one attribute level, the potential number of attribute assessment patterns $P\times(2^K-1)$ might exceed the number of items to be tested. Therefore, each $Q_p$ matrix in this study included one $R_p$ matrix, with the remaining rows randomly drawn (with replacement) from the simplified $Q_p$ matrix (outside the $R_p$ matrix).

(3) *Simulation of item parameters*: The guessing and slipping parameters were generated according to the specified simulation conditions.

(4) *Simulation of response matrix*: Based on steps (1)–(3), and the pG-DINA model's item response function, the probability $P_{ij}$ of an examinee answering an item correctly was calculated. A random number $r_{ij}$ was generated from $U(0,1)$; if $P_{ij} < r_{ij}$, the examinee $i$ scored 0 on item $j$, otherwise 1. The score matrix for all examinees was simulated.

(5) Parameter estimation using the cognitive diagnosis model (pG-DINA) and the two-stage methods. The MMLE/EM was used for model parameter estimation (de la Torre, 2009, 2011), and MAP for attribute mastery patterns. Additionally, the accuracy of different models, parameter estimation precision, and the consistency between the two-stage methods and the pG-DINA model were computed.

Each condition was repeated 50 times to reduce experimental error. The Monte Carlo simulation process was implemented using custom R scripts, with the data simulation and parameter estimation using the GDINA package (Ma and de la Torre, 2020). The code used for this study is available at https://osf.io/jd9hx/?view_only=51f7d62dee1e4fcdaf039a0fe0000b9d. All simulations and analyses were performed on a computer with an Intel(R) Core(TM) i9-13900K (3.00 GHz) processor and 128 GB RAM.

**Evaluation indices**. The following metrics were used to evaluate the precision of parameter estimates and the consistency between different models:

**Table 2 The average ACCR under various conditions for the pG-DINA model and the two-stage methods.**

| Number of attributes | Test length (multiple of the number of attributes) | $N = 500$ | | | $N = 1000$ | | | $N = 2000$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | pG-DINA | *P*_max | *P*_linear | pG-DINA | *P*_max | *P*_linear | pG-DINA | *P*_max | *P*_linear |
| 3 | 10 | 0.850 | 0.798 | 0.792 | 0.857 | 0.804 | 0.801 | 0.862 | 0.807 | 0.805 |
| | 20 | 0.925 | 0.901 | 0.897 | 0.930 | 0.906 | 0.903 | 0.933 | 0.908 | 0.906 |
| | 50 | 0.990 | 0.986 | 0.985 | 0.991 | 0.987 | 0.986 | 0.992 | 0.988 | 0.987 |
| 5 | 10 | 0.847 | 0.788 | 0.784 | 0.849 | 0.800 | 0.794 | 0.851 | 0.811 | 0.802 |
| | 20 | 0.909 | 0.885 | 0.876 | 0.914 | 0.894 | 0.887 | 0.919 | 0.901 | 0.895 |
| | 50 | 0.980 | 0.976 | 0.973 | 0.983 | 0.979 | 0.976 | 0.985 | 0.981 | 0.979 |
| 8 | 10 | 0.840 | 0.791 | 0.766 | 0.844 | 0.806 | 0.779 | 0.848 | 0.815 | 0.791 |
| | 20 | 0.894 | 0.864 | 0.838 | 0.903 | 0.880 | 0.861 | 0.908 | 0.888 | 0.873 |
| | 50 | 0.951 | 0.947 | 0.935 | 0.967 | 0.962 | 0.954 | 0.973 | 0.968 | 0.962 |

(1) For evaluating the accuracy of estimated attribute profiles, the average attribute correct classification rate (ACCR) and the average pattern correct classification rate (PCCR) were used, calculated as follows:

$$\overline{\mathrm{ACCR}} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{k=1}^{K} W_{ik}}{R \times N \times K} \quad (5)$$

$$\overline{\mathrm{PCCR}} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N}\prod_{k=1}^{K} W_{ik}}{R \times N} \quad (6)$$

Here, $N$ represents the sample size. $K$ is the number of attributes and $R$ is the number of repetitions. And $W_{ik} = I(\hat{\alpha}_{ik} = \alpha_{ik})$, where $\hat{\alpha}_{ik}$ and $\alpha_{ik}$ are the estimated and true attribute mastery levels of the $i$th examinee, respectively. $I$ is an indicator function, where $W_{ik} = 1$ if $\hat{\alpha}_{ik} = \alpha_{ik}$, else $W_{ik} = 0$. ACCR and PCCR reflect the accuracy of individual attributes and overall attribute pattern classification, respectively; higher values indicate better veracity of model parameter estimation.

(2) The mean absolute bias (MAB) is used to evaluate the precision of item parameter estimation, calculated as follows:

$$\mathrm{MAB} = \sum_{r=1}^{R} \frac{\lceil \hat{\tau} - \tau \rceil}{R} \quad (7)$$

Here, $\hat{\tau}$ and $\tau$ are the estimated and true values of the model parameters, respectively. MAB reflects the average deviation between the true and estimated item parameters; lower MAB values indicate higher precision in item parameter estimation.

(3) Attribute-matching rates and pattern-matching rates were used to assess the consistency between pG-DINA and the two-stage estimation methods. The calculations for these rates are the same as ACCR and PCCR, but $\hat{\alpha}_{ik}$ and $\alpha_{ik}$ in Eqs. (5) and (6) were replaced by the attribute mastery levels obtained from the two-stage methods and the pG-DINA model, respectively.

(4) Computational time was measured to evaluate the efficiency of each method. The time was recorded from the start of the estimation process to its completion.

**Results**. The following section primarily presents the parameter estimation results for high-quality items, while results for low-quality items are provided in Appendix B. This focus on high-quality items allows for a clearer presentation of the methods' performance under optimal conditions, which is typically of greater interest in practical applications. The comprehensive results, including those for low-quality items, are available for readers seeking a more in-depth analysis.

*The accuracy of estimated attribute profiles*. Table 2 presents the average ACCR for both the pG-DINA model and the two-stage polytomous attribute estimation methods under various conditions. The pG-DINA model exhibited high average ACCR values exceeding 0.8 across all conditions, indicating accurate estimation of examinees' attribute mastery levels even when cognitive levels of attributes are marked at the item level. In comparable conditions, the average ACCR of the two-stage polytomous attribute estimation methods slightly trailed behind the pG-DINA model but with negligible differences, all-surpassing 0.75, where the *P*_max method slightly outperformed the *P*_linear method. This suggests a marginal superiority of the *P*_max method in estimating examinees' attribute mastery levels. Notably, the average ACCR increased with a decrease in the number of attributes and an increase in the number of items. In particular, when the number of items was 50 times the number of attributes, even with eight attributes, all three methods achieved average ACCRs above 0.9. Furthermore, the average ACCR improved with larger sample sizes, indicating that with sufficient items (e.g., 50 times the number of attributes) and large sample sizes, all three methods could achieve highly accurate estimations of examinees' attribute mastery levels.

Table 3 presents the average PCCR for the pG-DINA model and the two-stage estimation methods under various conditions. The average PCCR was significantly affected by the number of attributes and the length of items. For instance, with a sample size of 2000, the attribute number set to 8, and item length at 80 (10 times the number of attributes), the average PCCR for the pG-DINA model was only 0.286, indicating that only about a quarter of examinees' attribute profiles could be accurately determined under these conditions. However, extending the item length to 400 (50 times the number of attributes) significantly improved the pG-DINA model's average PCCR to 0.815, which means pG-DINA can accurately determine around 80% of examinees' attribute profiles. The two-stage estimation methods, although close in average PCCR values, generally performed slightly lower than the pG-DINA model. Among them, the *P*_max method showed a slight advantage in estimating attribute patterns over the *P*_linear method. It's noteworthy that in scenarios with large item quantities and ample sample sizes, all methods achieved average PCCR values meeting practical measurement needs, even in the face of high attribute numbers. For instance, in scenarios with a large sample size of 2000, eight attributes, and 400 items, both two-stage estimation methods achieved an average PCCR of around 0.7. This observation underscores the significant impact of item quantity and sample size on estimation accuracy, especially in tests with high attribute numbers.

*Item parameter estimation precision*. Table 4 illustrates the precision of item parameter estimation for both the pG-DINA model

**Table 3 The average PCCR under various conditions for the pG-DINA model and the two-stage methods.**

| Number of attributes | Test length (multiple of the number of attributes) | $N = 500$ | | | $N = 1000$ | | | $N = 2000$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | pG-DINA | P_max | P_linear | pG-DINA | P_max | P_linear | pG-DINA | P_max | P_linear |
| 3 | 10 | 0.623 | 0.516 | 0.505 | 0.637 | 0.527 | 0.524 | 0.649 | 0.534 | 0.531 |
| | 20 | 0.802 | 0.741 | 0.731 | 0.814 | 0.751 | 0.744 | 0.822 | 0.758 | 0.752 |
| | 50 | 0.972 | 0.962 | 0.960 | 0.976 | 0.964 | 0.963 | 0.977 | 0.967 | 0.965 |
| 5 | 10 | 0.449 | 0.317 | 0.314 | 0.455 | 0.344 | 0.338 | 0.463 | 0.368 | 0.355 |
| | 20 | 0.639 | 0.556 | 0.533 | 0.653 | 0.588 | 0.571 | 0.670 | 0.610 | 0.593 |
| | 50 | 0.909 | 0.890 | 0.877 | 0.921 | 0.903 | 0.892 | 0.929 | 0.913 | 0.902 |
| 8 | 10 | 0.261 | 0.168 | 0.160 | 0.274 | 0.201 | 0.174 | 0.286 | 0.220 | 0.192 |
| | 20 | 0.428 | 0.335 | 0.282 | 0.465 | 0.384 | 0.346 | 0.486 | 0.413 | 0.379 |
| | 50 | 0.668 | 0.646 | 0.597 | 0.771 | 0.740 | 0.704 | 0.815 | 0.779 | 0.746 |

**Table 4 The average MAB of item parameter estimation under various conditions for the pG-DINA model and the two-stage methods.**

| Number of attributes | Test length (multiple of the number of attributes) | $N = 500$ | | $N = 1000$ | | $N = 2000$ | |
|---|---|---|---|---|---|---|---|
| | | pG-DINA | Two-stage | pG-DINA | Two-stage | pG-DINA | Two-stage |
| 3 | 10 | 0.089 | 0.115 | 0.061 | 0.085 | 0.040 | 0.058 |
| | 20 | 0.077 | 0.084 | 0.052 | 0.057 | 0.036 | 0.040 |
| | 50 | 0.069 | 0.069 | 0.048 | 0.049 | 0.034 | 0.034 |
| 5 | 10 | 0.162 | 0.212 | 0.118 | 0.155 | 0.084 | 0.102 |
| | 20 | 0.148 | 0.170 | 0.099 | 0.110 | 0.069 | 0.074 |
| | 50 | 0.129 | 0.131 | 0.090 | 0.091 | 0.062 | 0.063 |
| 8 | 10 | 0.198 | 0.255 | 0.148 | 0.189 | 0.114 | 0.133 |
| | 20 | 0.192 | 0.218 | 0.133 | 0.148 | 0.093 | 0.101 |
| | 50 | 0.175 | 0.175 | 0.116 | 0.116 | 0.079 | 0.079 |

**Table 5 The average attribute matching rates between the two-stage methods and the pG-DINA model under various conditions.**

| Number of attributes | Test length (multiple of the number of attributes) | $N = 500$ | | $N = 1000$ | | $N = 2000$ | |
|---|---|---|---|---|---|---|---|
| | | P_max | P_linear | P_max | P_linear | P_max | P_linear |
| 3 | 10 | 0.852 | 0.845 | 0.863 | 0.859 | 0.871 | 0.868 |
| | 20 | 0.933 | 0.929 | 0.940 | 0.937 | 0.944 | 0.941 |
| | 50 | 0.991 | 0.990 | 0.992 | 0.992 | 0.993 | 0.992 |
| 5 | 10 | 0.839 | 0.835 | 0.850 | 0.845 | 0.864 | 0.856 |
| | 20 | 0.910 | 0.901 | 0.921 | 0.915 | 0.928 | 0.922 |
| | 50 | 0.982 | 0.979 | 0.985 | 0.982 | 0.987 | 0.985 |
| 8 | 10 | 0.851 | 0.824 | 0.860 | 0.834 | 0.864 | 0.841 |
| | 20 | 0.888 | 0.860 | 0.902 | 0.884 | 0.910 | 0.897 |
| | 50 | 0.955 | 0.940 | 0.970 | 0.962 | 0.976 | 0.970 |

and the two-stage polytomous attribute estimation methods under various experimental conditions. When the number of items was relatively small, the average MAB values of the two-stage methods were marginally higher than those of the pG-DINA model. However, as the number of items increased, the accuracy of item parameter estimation across different methods converged. Furthermore, the average MAB for item parameter estimation increased with the number of attributes and decreased with longer item lengths and larger sample sizes. The accuracy of item parameter estimation is expected to further improve with increased sample sizes.

*Consistency in estimated profiles between pG-DINA model and the two-stage estimation methods*. Table 5 presents the consistency in estimating examinees' attribute mastery levels between the pG-DINA model and the two proposed two-stage methods. Overall, both two-stage methods demonstrated high consistency with the pG-DINA model, exhibiting attribute matching rates exceeding 0.8. The *P*_max method marginally outperformed the *P*_linear method in this regard. The analysis revealed that the consistency in attribute level estimation improved with larger sample sizes and an increased number of items. However, a declining trend was observed as the number of attributes increased. When the number of attributes was 3 or 5, and the number of items was at least 20 times the number of attributes, both two-stage methods achieved attribute estimation consistency exceeding 0.9 with the pG-DINA model. For scenarios with a higher number of attributes (i.e., 8), a similar level of consistency (>0.9) was achieved when the number of items was at least 50 times the number of attributes. A notable example illustrates the methods' performance under optimal conditions: with 8 attributes, 400 items (50 times the number of attributes), and a sample size of 2000, the consistency in attribute level estimation between the *P*_max method and the pG-DINA model reached 0.976, while the

**Table 6 The average pattern matching rates between the two-stage methods and the pG-DINA model under various conditions.**

| Number of attributes | Test length (multiple of the number of attributes) | *N* = 500 | | *N* = 1000 | | *N* = 2000 | |
|---|---|---|---|---|---|---|---|
| | | *P*_max | *P*_linear | *P*_max | *P*_linear | *P*_max | *P*_linear |
| 3 | 10 | 0.627 | 0.611 | 0.649 | 0.645 | 0.667 | 0.661 |
| | 20 | 0.819 | 0.810 | 0.836 | 0.829 | 0.847 | 0.840 |
| | 50 | 0.975 | 0.973 | 0.979 | 0.977 | 0.980 | 0.978 |
| 5 | 10 | 0.424 | 0.418 | 0.462 | 0.451 | 0.502 | 0.482 |
| | 20 | 0.638 | 0.611 | 0.679 | 0.657 | 0.703 | 0.683 |
| | 50 | 0.919 | 0.903 | 0.931 | 0.919 | 0.939 | 0.928 |
| 8 | 10 | 0.285 | 0.229 | 0.311 | 0.254 | 0.326 | 0.275 |
| | 20 | 0.401 | 0.318 | 0.459 | 0.406 | 0.493 | 0.453 |
| | 50 | 0.694 | 0.62 | 0.787 | 0.744 | 0.828 | 0.792 |

**Table 7 The average computation time (in seconds) under various conditions for the pG-DINA model and the two-stage methods.**

| Number of attributes | Test length (multiple of the number of attributes) | *N* = 500 | | *N* = 1000 | | *N* = 2000 | |
|---|---|---|---|---|---|---|---|
| | | pG-DINA | Two-stage | pG-DINA | Two-stage | pG-DINA | Two-stage |
| 3 | 10 | 0.6 | 0.3 | 0.8 | 0.3 | 1.4 | 0.5 |
| | 20 | 1.2 | 0.4 | 1.3 | 0.4 | 2.0 | 0.6 |
| | 50 | 1.5 | 0.5 | 2.1 | 0.7 | 3.4 | 1.1 |
| 5 | 10 | 28.0 | 1.7 | 92.2 | 1.9 | 183.2 | 3.8 |
| | 20 | 64.9 | 2.5 | 118.7 | 2.2 | 267.7 | 4.0 |
| | 50 | 112.2 | 3.2 | 199.8 | 4.2 | 333.9 | 6.8 |
| 8 | 10 | 3353.3 | 5.8 | 11,948.6 | 11.6 | 37,816.5 | 17.6 |
| | 20 | 5445.7 | 6.2 | 23,242.5 | 12.5 | 45,507.4 | 26.5 |
| | 50 | 11,078.3 | 6.8 | 28,535.8 | 14.7 | 81,619.0 | 27.2 |

*P*_linear method achieved 0.970. These results indicate that in scenarios characterized by large sample sizes and high item volumes, both *P*_max and *P*_linear methods can produce attribute mastery level estimations highly consistent with the pG-DINA model, even when dealing with a substantial number of attributes.

Table 6 presents the consistency in estimating examinees' attribute profiles between the two methods and the pG-DINA model. Similarly, the consistency in attribute pattern estimation decreased with more attributes but increased with larger sample sizes and more items. Notably, the consistency in attribute mastery pattern estimation between the *P*_max method and the pG-DINA model was higher than that between the *P*_linear method and the pG-DINA model. With 8 attributes, 400 items (50 times the number of attributes), and a sample size of 2000, the consistency in attribute mastery pattern estimation between the *P*_max method and the pG-DINA model reached 0.828, indicating that over 80% of examinees' attribute profiles estimated by these two methods were completely consistent. This consistency is expected to further improve with larger sample sizes.

*Parameter estimation time*. The computational efficiency of the pG-DINA model and the proposed two-stage methods was evaluated under various conditions, with the results presented in Table 7. The findings reveal a consistent pattern across all experimental conditions: the two-stage methods demonstrate markedly superior computational efficiency compared to the pG-DINA model. This efficiency advantage becomes increasingly pronounced as model complexity increases. For models with three attributes, both methods exhibit relatively low computation times, with the two-stage methods showing a slight edge. However, as the number of attributes increases to five and eight, a substantial disparity in computation time emerges. For instance, in the most complex scenario examined (8 attributes, 400 items (50 times the number of attributes), 2000 examinees), the pG-DINA model required approximately 81,619 s (about 22.7 h), while the two-stage methods completed the task in merely 27.2 s.

The impact of test length and sample size on computation time is evident for both methods, with longer tests and larger samples generally requiring more processing time. However, the rate of increase is considerably steeper for the pG-DINA model.

Notably, the two-stage methods exhibit remarkable scalability. Even as model complexity increases dramatically, their computation time remains relatively manageable. In contrast, the pG-DINA model shows exponential growth in computation time as the number of attributes increases, particularly evident in the transition from 5 to 8 attributes.

Appendix A presents parameter estimation results for various methods under conditions of lower item quality, where both guessing and slipping parameters were fixed at 0.3. Analysis of these results reveals that all methods, including the MMLE/EM estimation under the pG-DINA model and the proposed two-stage approaches, exhibited diminished accuracy in estimating examinees' attribute profiles under these suboptimal conditions. Notably, while computation times increased for all methods under low item quality conditions, the proposed two-stage methods maintained their substantial efficiency advantage, outperforming the traditional MMLE/EM estimation under the pG-DINA model by several orders of magnitude in computation speed. These observations underscore two crucial points: First, the importance of item quality—irrespective of the chosen estimation method, high-quality items are essential for accurate attribute estimation. Second, even under suboptimal item quality conditions, the proposed two-stage methods demonstrate remarkable computational efficiency compared to traditional methods.

## Empirical study

An empirical study was conducted using data from the National Medical Licensing Examination of China, a large-scale health-related certification exam, to validate the feasibility of the new methods in practice. The test comprised approximately 600 multiple-choice items, with each correctly answered item awarded one point, totaling a maximum score of 600. The passing threshold was set at 360 points, with a recent 5-year pass rate fluctuating between 40% and 50%. The test score reports only included the total score, lacking detailed feedback on individual knowledge mastery. The application of cognitive diagnostic assessment (CDA; Leighton and Gierl, 2007) in such large-scale exams could provide more detailed feedback, enhancing educational effectiveness and enabling more accurate individual ability assessment.

### Method

*Determining attributes and cognitive levels*. This study employed a rigorous process for calibrating the polytomous *Q*-matrix for the examination, using a panel of experts assembled by the China National Medical Examination Center. The panel comprised professionals from the nation's leading medical schools and hospitals, each possessing over a decade of experience in teaching and item development. The calibration process involved a collaborative effort between domain experts and psychometricians. They analyzed each test item to identify the examination key points and cognitive levels assessed, ensuring alignment with the test specifications outlined in the examination syllabus. Through this comprehensive review, the expert panel identified 12 distinct examination key points. Each test item was subsequently mapped to at least one of these key points, establishing a clear link between item content and the broader examination objectives. In addition to key point mapping, each item was classified according to its cognitive demand. The panel employed a four-level hierarchical framework to categorize the cognitive processes required by each item, ranging from lower to higher-order thinking skills: memory, understanding, simple application, and comprehensive application. In constructing the *Q*-matrix, we treated examination points as attributes, with cognitive levels serving as multiple levels, forming a complete polytomous *Q*-matrix. The attributes were coded from A1 to A12. Each cognitive level in this study was assessed by over 100 items. Except for attributes A9 and A12, which were not assessed at the comprehensive application level by any item, all other attributes were measured across various cognitive levels. Additionally, the *Q*-matrix for this study exhibits a complex structure, as over a third of the items simultaneously assessed multiple attributes. The item that assessed the most attributes measured five attributes concurrently. For more detailed statistical information about the polytomous *Q*-matrix used in this study, please refer to Appendix B.

*Data description*. The data, provided by the China National Medical Examination Center, included 225,044 samples, with scores ranging from 82 to 531, an average of 346.03, a standard deviation of 82 points, and a median of 357. ~48.55% of examinees passed the 360-point threshold.

*Analysis*. Drawing from the findings of the simulation study, we opted for the two-stage *P*_max estimation method to analyze the test data. The 600 items were divided into four sets according to their cognitive levels, and each set was analyzed using the GDINA model (de la Torre, 2011), with parameter estimation conducted using the GDINA package in R (Ma and de la Torre, 2020). We used the MMLE/EM algorithm to estimate model parameters and the MAP method to determine attribute mastery status. The *P*_max method then integrated the results from the four item sets, yielding the final parameter estimates. The entire analysis process for this empirical study took approximately 143.55 h to complete on a computer with an Intel(R) Core(TM) i9-13900K (3.00 GHz) processor and 128 GB RAM.

**Table 8 The GOF results for the four sets of items.**

| Cognitive levels | M2 | df | *p* | RMSEA |
|---|---|---|---|---|
| Memorizing | 385,031 | 1764 | 0 | 0.031 |
| Understanding | 828,224 | 5271 | 0 | 0.026 |
| Simple application | 3,448,290 | 23571 | 0 | 0.035 |
| Comprehensive application | 578,734 | 4563 | 0 | 0.024 |

*Evaluation indicators*

Model Fit: The goodness of fit (GOF) between the empirical data and the chosen CDM was evaluated by the M2 statistic (Hansen et al. 2016). However, with increasing sample sizes, M2 becomes overly sensitive to model misfit (Xu et al. 2017). Therefore, the root mean square error of approximation (RMSEA) was further employed to assess the effect size of model-data mismatch, with a recommended upper limit of 0.04 (Steiger, 1980).

Classification accuracy: The test-level and attribute-level classification accuracy indicators (Iaconangelo, 2017; Wang et al. 2015) were used to assess the reliability of diagnostics at both the test and attribute levels. Higher values of these indicators indicate higher precision in model classifications.

Item discrimination under CDM: In CDM analysis, an item's overall discrimination can be defined as the probability of examinees who have mastered all attributes of an item [$p_j(1)$] minus the probability of those who have not mastered any attributes of the item [$p_j(0)$], expressed as:

$$d_j = p_j(1) - p_j(0) \tag{8}$$

Score application: Based on the estimation of examinees' attribute profiles, we presented a comprehensive view of cognitive mastery situations. This encompassed the overall examinee population, various groups, and individuals with typical scores.

**Results**. Table 8 shows the goodness of fit (GOF) results for the four sets of items. Although the M2 statistics for all four sets were significant, considering the large sample size, RMSEA was chosen as the measure of model-data misfit. The RMSEA for all four sets of items was below 0.04, indicating negligible misfit between the data and the model.

Table 9 displays the classification accuracy indicators at both the test and attribute levels. These indicators are used to evaluate the CDM model's diagnostic precision across the entire test and for individual attributes. Among the cognitive levels, items assessing "memorizing" had the highest test accuracy at 0.77, while those assessing "simple application" had the lowest at 0.56. At the attribute level, classification accuracy indicators for all attributes in the four groups exceeded 0.8, indicating high attribute classification precision of the G-DINA model. It is worth noting that since attributes A9 and A12 were not involved in items assessing the "comprehensive application" level, these attributes had no classification accuracy indicators at this level.

**Table 9 The classification accuracy indicators at test and attribute levels.**

| Cognitive levels | Test | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Memorizing | 0.77 | 1.00 | 0.95 | 1.00 | 1.00 | 0.96 | 0.90 | 0.99 | 0.91 | 0.92 | 0.91 | 1.00 | 0.99 |
| Understanding | 0.65 | 0.92 | 1.00 | 0.92 | 1.00 | 1.00 | 1.00 | 0.94 | 0.97 | 0.99 | 0.89 | 0.95 | 0.91 |
| Simple application | 0.56 | 0.91 | 0.91 | 0.94 | 1.00 | 0.98 | 1.00 | 0.93 | 0.91 | 0.92 | 1.00 | 1.00 | 0.83 |
| Comprehensive application | 0.64 | 0.88 | 0.85 | 0.88 | 1.00 | 0.92 | 1.00 | 0.91 | 0.94 | - | 0.94 | 1.00 | - |

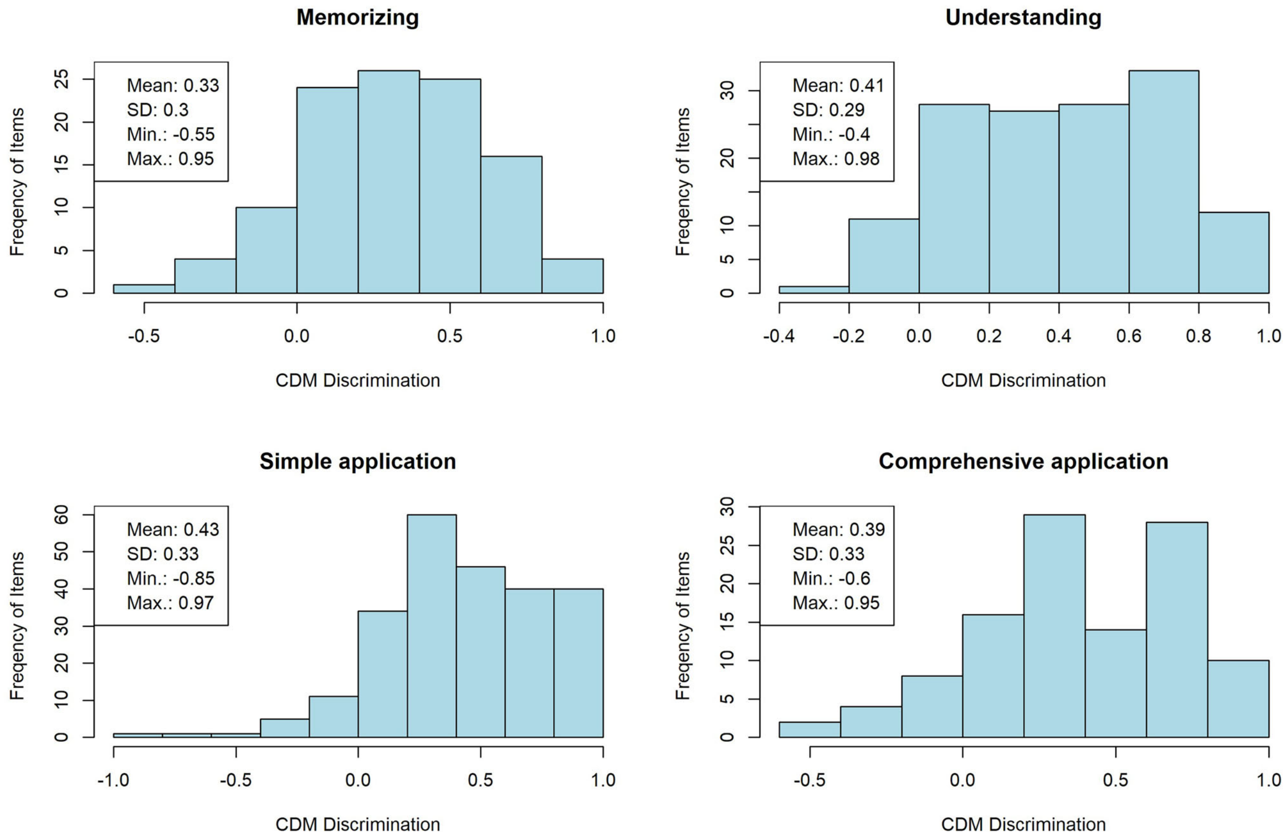

**Fig. 1 Discrimination distribution of item groups by cognitive levels.** The histograms display the cognitive diagnostic discrimination values for test items across four cognitive levels: memorizing (top left), understanding (top right), simple application (bottom left), and comprehensive application (bottom right). Each panel shows the frequency distribution of discrimination indices, with the *x*-axis representing CDM discrimination values and the *y*-axis representing the frequency of items. Mean, standard deviation (SD), minimum, and maximum values are provided in the top-left corner of each panel.

Figure 1 presents the cognitive diagnostic discrimination distribution of all items. Most items had discrimination above 0, effectively distinguishing examinees who had fully mastered the cognitive attributes assessed by the items from those who had not. However, a few items, particularly those assessing "simple application", had negative discrimination, with the lowest reaching −0.85, suggesting a need to focus on improving the quality of these items.

Table 10 displays the proportion of mastery of each cognitive attribute at different levels among all examinees, those who passed the exam, and those who did not. Candidates who passed the exam demonstrated more mastery of attributes at higher cognitive levels ("simple application" and "comprehensive application") compared to those who did not pass, confirming the reasonableness of the attribute mastery estimation results.

In total, 5606 different attribute mastery profiles emerged among all candidates. Table 11 shows the frequency distribution of attribute mastery profiles that occurred in more than 1% of the examinees. The most frequent pattern, occurring in over 10% of candidates, was 444444443443, indicating mastery at the highest level for all attributes except A9 and A12.

Examinees' attribute mastery profiles can be visualized using radar charts. Figure 2 shows the polytomous attribute mastery profiles, represented in different colors, of two examinees who both scored 360. Although they achieved the same total score, their attribute mastery profiles were not identical, demonstrating that cognitive diagnostic assessment can provide richer, personalized feedback compared to traditional test scores.

## Conclusion and discussion

In large-scale examinations, the use of CDMs with many polytomous attributes (mpCDM) addresses a critical need for a more nuanced and detailed assessment of examinees' knowledge and

**Table 10 The mastery percentages of each attribute at each cognitive level for different groups of examinees (%).**

| Group of examinees | Cognitive levels | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| All | Not mastered | 6 | 6 | 7 | 6 | 6 | 6 | 11 | 6 | 7 | 6 | 6 | 3 |
| | Memorizing | 19 | 0 | 23 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 39 |
| | Understanding | 6 | 9 | 5 | 0 | 2 | 0 | 9 | 2 | 16 | 0 | 0 | 20 |
| | Simple application | 5 | 15 | 22 | 0 | 52 | 1 | 8 | 3 | 77 | 24 | 0 | 37 |
| | Comprehensive application | 63 | 70 | 44 | 93 | 39 | 93 | 58 | 89 | 0 | 70 | 94 | 0 |
| Passed | Not mastered | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | Memorizing | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 35 |
| | Understanding | 1 | 4 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 20 |
| | Simple application | 3 | 22 | 25 | 0 | 37 | 0 | 7 | 1 | 99 | 2 | 0 | 45 |
| | Comprehensive application | 96 | 75 | 72 | 100 | 63 | 100 | 92 | 99 | 0 | 98 | 100 | 0 |
| Failed | Not mastered | 12 | 12 | 13 | 12 | 12 | 12 | 21 | 12 | 13 | 12 | 12 | 6 |
| | Memorizing | 37 | 0 | 45 | 0 | 0 | 0 | 28 | 0 | 0 | 0 | 0 | 43 |
| | Understanding | 11 | 14 | 7 | 0 | 4 | 0 | 16 | 3 | 31 | 0 | 0 | 21 |
| | Simple application | 6 | 8 | 18 | 1 | 66 | 1 | 10 | 5 | 56 | 44 | 0 | 30 |
| | Comprehensive application | 33 | 66 | 17 | 87 | 17 | 87 | 26 | 79 | 0 | 43 | 87 | 0 |

**Table 11 Frequency distribution of attribute mastery profiles (frequency > 2250).**

| Profile | Frequency | Percentage |
|---|---|---|
| 444444443443 | 30120 | 13.4 |
| 444444443441 | 17215 | 7.6 |
| 000000000003 | 12138 | 5.4 |
| 444444443442 | 10372 | 4.6 |
| 444434443443 | 8789 | 3.9 |
| 433434443441 | 7467 | 3.3 |
| 444434443441 | 7367 | 3.3 |
| 433444443441 | 4853 | 2.2 |
| 433434443443 | 4349 | 1.9 |
| 444434443442 | 4119 | 1.8 |
| 423434443441 | 3610 | 1.6 |
| 433434443442 | 3533 | 1.6 |
| 141434142341 | 3390 | 1.5 |
| 433444443443 | 3316 | 1.5 |
| 141434143341 | 2657 | 1.2 |
| 433444443442 | 2317 | 1.0 |
| 141434142343 | 2270 | 1.0 |

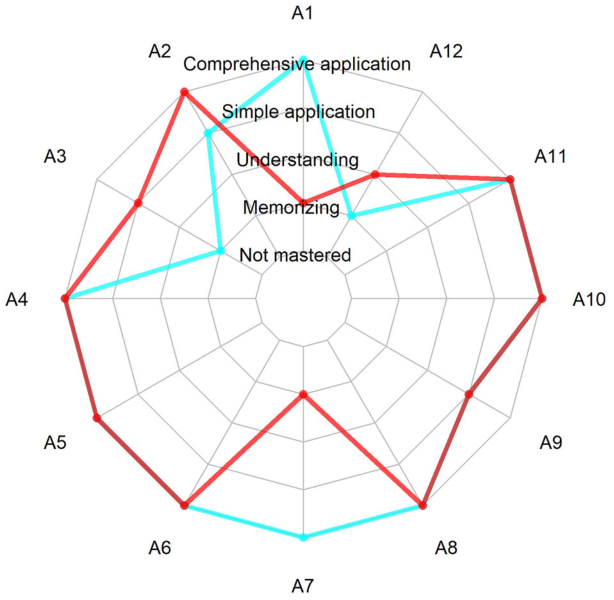

**Fig. 2 The radar chart of polytomous attribute mastery profiles for two examinees scoring 360.** This visualization compares cognitive attribute mastery across 12 attributes (A1–A12) for two examinees with identical total scores. Different colors represent each examinee, with concentric circles indicating mastery levels from "not mastered" (center) to "comprehensive application" (outermost). Despite equal scores, the examinees exhibit distinct mastery profiles, demonstrating how cognitive diagnostic assessment provides more detailed information than traditional scoring methods.

skills. This granular approach enables educators and policymakers to move beyond mere pass/fail determinations or overall score categorizations, diving deeper into the complex fabric of learners' cognitive profiles. It is especially valuable in contexts like national educational assessments, professional certification exams, or large-scale academic surveys, where understanding specific strengths and weaknesses is crucial for tailored educational strategies, targeted interventions, and informed decision-making.

Our study introduces two novel two-stage polytomous attribute estimation methods ($P$_max and $P$_linear) that address several critical challenges in implementing mpCDM for large-scale assessments. Through extensive Monte Carlo simulations and empirical application, we validated the effectiveness and scalability of these methods across various scenarios with high numbers of attributes, extensive item pools, and large sample sizes. The results demonstrate unprecedented capability in handling numerous polytomous attributes, with successful application to 12 four-level attributes in our empirical study. Most notably, these methods achieve remarkable computational efficiency, reducing processing time from hours to seconds compared to traditional approaches, while maintaining accurate and reliable cognitive calibration.

Our methods are particularly well-suited for large-scale assessment contexts where multiple complex skills need to be evaluated simultaneously, such as national certification examinations, professional licensure tests, and comprehensive educational assessments. The approach is most effective when implemented with high-quality items, sufficient test length relative to the number of attributes being measured, and adequate sample sizes. These conditions are typically met in large-scale testing programs where rigorous item development processes are standard practice, comprehensive content coverage is required, and large candidate populations are available. To ensure optimal performance in high-stakes contexts, we recommend supplementing our methods with traditional psychometric indices and conducting thorough item quality control. The empirical success of the national medical licensing examination, which involved multiple polytomous attributes, hundreds of test items, and over 200,000 examinees, exemplifies the ideal application scenario for these methods.

By implementing these methods in R and providing open-source code, we have enhanced the accessibility and reproducibility of complex cognitive diagnostic modeling for a wider range of researchers and practitioners. This approach facilitates further development and application of mpCDM in diverse educational and professional assessment contexts.

Our work paves the way for more precise and comprehensive cognitive diagnostic assessments in large-scale national exams and other high-stakes testing environments, potentially transforming the landscape of educational and professional assessment. The computational efficiency and robust performance of these methods across various conditions significantly enhance the feasibility and effectiveness of implementing complex cognitive diagnostic models in large-scale assessments.

Further discussions from this study include:

Firstly, while our simulation study found high attribute-level classification rates, the pattern-level classification rates were relatively lower, aligning with previous research (Chen and de la Torre, 2013). This reflects an inherent challenge: the expansion of attribute levels from binary to polytomous significantly increases the number of possible attribute mastery patterns (from $2^K$ to $(P+1)^K$), making pattern-level classification more demanding. Our simulation results demonstrated that with high-quality items, sufficient test length, and adequate sample sizes, the two-stage method can achieve excellent pattern-level classification rates comparable to traditional approaches. To ensure optimal performance in assessment contexts, we recommend maintaining high item quality through thorough quality control, ensuring adequate test length relative to the number of attributes, and working with sufficient sample sizes. These conditions, combined with the use of $P$_max rather than $P$_linear, can help maintain high assessment quality while leveraging the substantial computational advantages of the two-stage approach.

Secondly, our results from the two-stage estimation methods suggest that the $P$_linear method, which constructs polytomous attributes based on a strictly linear relationship, is less accurate and consistent with the pG-DINA model compared to the $P$_max method. This strict linear relationship assumption may oversimplify the actual cognitive development process. In reality, cognitive development could be more complex, with different levels of cognitive abilities potentially having more intricate interactions with each other. Therefore, we recommend using the $P$_max method in actual data analysis.

Thirdly, although we only set guessing and slipping parameters in our study, the GDINA package can automatically convert these to the $\delta$ coefficients used in the G-DINA model. This conversion process ensures compatibility with the broader G-DINA framework while maintaining the interpretability of the more intuitive guessing and slipping parameters. Furthermore, our research found that the pG-DINA model and the proposed two-stage methods demonstrated suboptimal accuracy in estimating examinees' attribute mastery levels under conditions of low item quality. This observation underscores the important role that item quality plays in ensuring accurate diagnostic feedback. While advanced modeling techniques such as pG-DINA and the proposed two-stage methods offer powerful tools for attribute estimation, their efficacy is fundamentally dependent on the quality of the input data. The pursuit of methodological refinement must be balanced with a continued focus on the foundational aspects of test design and development to ensure high item quality.

Lastly, although we cannot directly determine the actual classification rate of the $P$_max method for attribute mastery in the empirical study, the simulation study results indicate that relatively accurate estimates can still be achieved even with a large number of attributes, provided that there are a large number of items and samples. It is important to note that due to computational constraints, specifically limitations in the GDINA package, the conditions of the simulation study and empirical study are not entirely consistent. The empirical study incorporated 12 attributes, whereas the simulation study was limited to a maximum of 8 attributes. This limitation stems from the GDINA package's inability to generate simulation data for 12 four-level attributes, not from any inherent limitation in our proposed method. Our proposed method is capable of handling 12 polytomous attributes, as demonstrated in the empirical study. However, for the simulation study, we were constrained by the capabilities of existing software used for comparison. Notwithstanding this disparity, it is noteworthy that both the number of items and the sample size in the empirical study significantly exceeded those in the simulation study. This augmentation in both test length and sample size may potentially mitigate the increased complexity introduced by the additional attributes, allowing our method to perform well even with 12 attributes in the empirical study.

Future research could focus on the following ideas:

Adapt the two-stage polytomous attribute estimation method for attribute-level calibration. While the two-stage polytomous attribute estimation method proposed in this study primarily applies to cognitive level calibration at the item level, it could be adapted for attribute-level calibration, where different attributes assessed in the same item may have different cognitive levels. The concept of dividing attributes into groups, estimating them independently, and then combining them can be further refined and applied to more complex calibration scenarios.

Handle imbalance in item coverage of attributes. In the simulation study, each attribute was measured by a relatively balanced number of items. However, in the empirical study, there was significant variance in the number of items measuring each attribute, with some attributes having as few as four related items. Future research should consider scenarios where the distribution of items measuring each attribute is unequal. This might involve adjusting the simulation conditions or even excluding attributes with insufficient item coverage to ensure more robust and realistic results.

Use mixed models in test data analysis. Instead of relying on a single model, future studies might consider employing mixed models for test data analysis, where different CDMs are selected for different test items (de la Torre and Lee, 2013; de la Torre et al. 2018; Ma et al. 2016; Tu et al. 2017). When the fit of saturated and simplified models is similar, researchers should prefer simplified models based on the Principle of Parsimony (Ma et al. 2016). However, using a single simplified cognitive diagnostic model for the entire test may lead to misfit, affecting the accuracy and reliability of the assessment results (Hou, 2013). While this study exclusively utilized the saturated G-DINA model, considering the model parsimony principle, future research could opt for more simplified models for different items where model fit allows. Additionally, the feasibility of directly combining attribute mastery patterns estimated using different models at different attribute levels within the two-stage method remains an area for further exploration.

Verify polytomous $Q$-matrix based on data-driven methods. In our empirical study, we directly used expert-calibrated polytomous $Q$-matrix. However, accurately calibrating cognitive attributes and levels becomes increasingly challenging as the granularity of cognitive attributes increases and the test structure becomes more complex. Existing research suggests using data-driven methods for *Q*-matrix calibration and validation (Ma et al. 2016). Future studies could explore developing polytomous *Q*-matrix verification methods based on the two-stage polytomous attribute estimation method to enhance the precision of *Q*-matrix calibration.

## Data availability

The datasets generated during the simulation study are accessible at https://osf.io/jd9hx/?view_only=51f7d62dee1e4fcdaf039a0fe0000b9d. The data analyzed during the empirical study are available from the China National Medical Examination Center, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. However, we have provided the R code used in the empirical study, along with a simulated dataset generated based on the settings of the empirical data for reader reference. These resources can be accessed at https://osf.io/jd9hx/?view_only=51f7d62dee1e4fcdaf039a0fe0000b9d.

Received: 25 January 2024; Accepted: 28 April 2025;
Published online: 26 May 2025

## References

Anderson LW, Krathwohl DR (eds) (2001) A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman

Bao Y (2019) A diagnostic classification model for polytomous attributes. Doctoral dissertation, University of Georgia

Birenbaum M, Kelly AE, Tatsuoka KK (1993) Diagnosing knowledge states in algebra using the rule-space model. J Res Math Educ 24(5):442–459. https://doi.org/10.2307/749153

Bloom BS (1956) Taxonomy of educational objectives, vol 1: cognitive domain. McKay, New York, pp. 20–24

Bradshaw L, Izsák A, Templin J, Jacobson E (2014) Diagnosing teachers' understandings of rational number: building a multidimensional test within the diagnostic classification framework. Educ Meas: Issues Pr 33(1):2–14

Chen J, de la Torre J (2013) A general cognitive diagnosis model for expert-defined polytomous attributes. Appl Psychol Meas 37(6):419–437. https://doi.org/10.1177/0146621613479818

Chen J, de la Torre J (2018) Introducing the general polytomous diagnosis modeling framework. Front Psychol 9:1474. https://doi.org/10.3389/fpsyg.2018.01474

Cizek GJ (2012) Setting performance standards: foundations, methods, and innovations. Routledge

de la Torre J (2009) DINA model and parameter estimation: a didactic. J Educ Behav Stat 34(1):115–130

de la Torre J (2011) The generalized DINA model framework. Psychometrika 76(2):179–199

de la Torre J, Douglas JA (2004) Higher-order latent trait models for cognitive diagnosis. Psychometrika 69(3):333–353. https://doi.org/10.1007/BF02295640

de la Torre J, Lee YS (2013) Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. J Educ Meas 50(4):355–373

de la Torre J, Qiu X, Santos KCP (2025) The generalized cognitive diagnosis model framework for polytomous attributes. Psychometrika. https://doi.org/10.1017/psy.2024.16

de la Torre J, van der Ark LA, Rossi G (2018) Analysis of clinical data from a cognitive diagnosis modeling framework. Meas Eval Couns Dev 51(4):281–296

Eva KW, Armson H, Holmboe E, Lockyer J, Loney E, Mann K, Sargeant J (2012) Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. Adv Health Sci Educ 17:15–26

George AC, Robitzsch A, Kiefer T, Groß J, Ünlü A (2016) The R package CDM for cognitive diagnosis models. J Stat Softw 74(2):1–24. https://doi.org/10.18637/jss.v074.i02

Haladyna TM, Kramer GA (2004) The validity of subscores for a credentialing test. Appl Meas Educ 17(4):343–368

Hansen M, Cai L, Monroe S, Li Z (2016) Limited-information goodness-of-fit testing of diagnostic classification item response models. Br J Math Stat Psychol 69(3):225–252

Hartz SM (2002) A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality (Doctoral dissertation), University of Illinois at Urbana-Champaign

Hattie J, Timperley H (2007) The power of feedback. Rev Educ Res 77(1):81–112. https://doi.org/10.3102/003465430298487

Hou L(2013) Differential item functioning assessment in cognitive diagnostic modeling. Unpublished manuscript, University of Delaware

Iaconangelo C (2017) Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models. Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ

Jones LR, Wheeler G, Centurino VAS (2013) TIMSS 2015 science framework. In Mullis IVS, Martin MO (eds) TIMSS 2015 assessment frameworks. Boston College, TIMSS & PIRLS International Study Center, pp. 29–58

Junker BW, Sijtsma K (2001) Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Appl Psychol Meas 25(3):258–272

Jiang Z, Carter R (2019) Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. Behav Res Methods 51(2):651–662. https://doi.org/10.3758/s13428-018-1069-9

Karelitz TM (2004) Ordered category attribute coding framework for cognitive assessments. Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign

Khorramdel L, Shin HJ, von Davier M (2019) GDM software mdltm including parallel EM algorithm. In: von Davier M, Lee Y-S (eds) Handbook of diagnostic classification models. Springer, Cham, Switzerland, pp 603–628

Krathwohl DR (2002) A revision of Bloom's taxonomy: an overview. Theory into Pract 41(4):212–218. https://doi.org/10.1207/s15430421tip4104_2

Kunina-Habenicht O, Rupp AA, 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

Lee YH, de la Torre J, Park YS (2011) A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US National sample using the TIMSS 2007. Int J Test 11(2):144–177

Leighton JP, Gierl MJ (eds) (2007) Cognitive diagnostic assessment for education: theory and applications. Cambridge University Press

Liu R, Qian H, Luo X, Woo A (2018) Relative diagnostic profile: a subscore reporting framework. Educ Psychol Meas 78(6):1072–1088

Ma W (2022) A higher-order cognitive diagnosis model with ordinal attributes for dichotomous response data. Multivar Behav Res 57(2–3):408–421. https://doi.org/10.1080/00273171.2020.1860731

Ma W, de la Torre J (2020) GDINA: an R package for cognitive diagnosis modeling. J Stat Softw 93(14):1–26

Ma W, Iaconangelo C, de la Torre J (2016) Model similarity, model selection, and attribute classification. Appl Psychol Meas 40(3):200–217

Mohsenpour M (2019) Assessing polytomous cognitive attributes of mathematics literacy of 9th grade students: Supplying PGDINA model. J Educ Meas Eval Stud 9(26):109–134

National Assessment Governing Board (2008) Mathematics framework for the 2009 National Assessment of Educational Progress. National Assessment Governing Board

Park YS, Xing K, Lee YS (2018) Explanatory cognitive diagnostic models: incorporating latent and observed predictors. Appl Psychol Meas 42:376–392

Robitzsch A, Kiefer T, Wu M, Yan D (2017) TAM: test analysis modules R package version 2. pp. 12–18

Schoenherr JR, Hamstra SJ (2016) Psychometrics and its discontents: an historical perspective on the discourse of the measurement tradition. Adv Health Sci Educ 21:719–729

Sinharay S (2010) How often do subscores have added value? Results from operational and simulated data. J Educ Meas 47(2):150–174

Sinharay S, Puhan G, Haberman SJ (2011) An NCME instructional module on subscores. Educ Meas: Issues Pract 29(3):29–40

Steiger JH (1980) Statistically based tests for the number of common factors. Paper presented at the Annual Meeting of the Psychometric Society, Iowa City

Sun JN, Xin T, Zhang SM, de la Torre J (2013) A polytomous extension of the generalized distance discriminating method. Appl Psychol Meas 37(7):503–521. https://doi.org/10.1177/0146621613487254

Tatsuoka KK (2009) Cognitive assessment: an introduction to the Rule Space Method. Routledge

Templin JL (2004) Generalized linear mixed proficiency models. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign

Templin J, Henson RA (2006) Measurement of psychological disorders using cognitive diagnosis models. Psychol Methods 11(3):287–305

Tu D, Gao X, Wang D, Cai Y (2017) A new measurement of internet addiction using diagnostic classification models. Front Psychol 8:1768

von Davier M (2008) A general diagnostic model applied to language testing data. Br J Math Stat Psychol 61(2):287–307. https://doi.org/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 Ser 2014:1–13. https://doi.org/10.1002/ets2.12043

von Davier M (2016) High-performance psychometrics: the parallel-E parallel-M algorithm for generalized latent variable models. ETS Res Rep Ser 2016:1–11. https://doi.org/10.1002/ets2.12120

Wang S, Chen Y (2020) Using response times and response accuracy to measure fluency within cognitive diagnosis models. Psychometrika 85(3):600–629. https://doi.org/10.1007/s11336-020-09717-2

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

Xu J, Paek I, Xia Y (2017) Investigating the behaviors of M2 and RMSEA2 in fitting a unidimensional model to multidimensional data. Appl Psychol Meas 41(8):632–644

Yakar L, Dogan N, de la Torre J (2021) Retrofitting of polytomous cognitive diagnosis and multidimensional item response theory models. J Meas Eval Educ Psychol 12(2):97–111. https://doi.org/10.21031/epod.778861

Zhan P, Bian Y, Wang L (2016) Factors affecting the classification accuracy of reparameterized diagnostic classification models for expert-defined polytomous attributes. Acta Psychol Sin 48:318–330. https://doi.org/10.3724/SP.J.1041.2016.00318

Zhan P, Liu Y, Yu Z, Pan Y (2023) Tracking ordinal development of skills with a longitudinal DINA model with polytomous attributes. Appl Meas Educ 36(2):99–114. https://doi.org/10.1080/08957347.2023.2201702

Zhan P, Wang W-C, Li X (2020) A partial mastery, higher-order latent structural model for polytomous attributes in cognitive diagnostic assessments. J Classif 37(2):328–351. https://doi.org/10.1007/s00357-019-09323-7

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 72474004 and No. 32400934).

## Author contributions

Yuting Han: Conceptualization and study design, writing—original draft, data analysis, writing—review and editing. Feng Ji: Writing—original draft, writing—review and editing. Zhehan Jiang: Study design, data collection, formal analysis, writing—review and editing.

## Competing interests

The authors declare no competing interests.

## Ethical approval

This study utilized anonymized, de-identified educational and administrative data from the National Medical Licensing Examination of China. As the study involves only secondary analysis of de-identified data, it was determined to not require formal ethical review according to the institutional policies on research ethics. This determination is in line with national guidelines on the use of de-identified educational data for research purposes. The study was conducted in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

## Informed consent

The requirement for informed consent was waived due to the retrospective nature of the study and the use of de-identified data. This study did not involve any interventions or interactions with human subjects, and the data used cannot be traced back to individuals, thus maintaining privacy and confidentiality.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-04959-w.

**Correspondence** and requests for materials should be addressed to Zhehan Jiang.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.