*Original Research Article*

Asian Journal for Mathematics Education
1-20

DOI: 10.1177/27527263221136301
journals.sagepub.com/home/mea

# Personalized assessment: Applying higher-order cognitive diagnosis models in secondary mathematics

**Ying Zhang[1], Yi Jin[1], Zhenrong Xiong[2], Shing On Leung[3], Gaowei Chen[1], Na Li[2], and Bo Li[2]**

**Abstract**

Personalized assessment is an essential component in education. Although many cognitive diagnosis models (CDMs) have been developed for this purpose, few studies have applied them in secondary mathematical contexts. Using a sample of 391 Grade 11 students from a secondary school in China, the findings indicated that the higher-order generalized deterministic inputs, noisy, "and" gate (higher-order GDINA) model with one-parameter logistic (1PL) best fit the data, and the *Q* matrix validation process achieved acceptable results. At the grade level, most of the participants mastered attributes B1 (i.e., basic concept development of derivatives: simple equations, zero or extreme points, and function range problems), B2 (complex inductive contextualization of derivatives: induction from the known to solve the unknown problems), and B3 (basic routine problem solving of derivatives: graphs that pass through a fixed point or quantitative inequalities). However, less than half of the students mastered attribute B4 (complex transformative contextualization of derivatives: transformation by the combination of numbers and graphs). At the individual level, we selected four representative students with high, medium, and low levels of achievement to examine their individual skill profiles and provide personalized remedial and enhanced feedback. Implications for personalized assessments are discussed.

**Keywords**

cognitive diagnosis, higher-order GDINA model, PISA, *Q* matrix validation, personalized assessment

Date received: 15 July 2022; accepted: 9 October 2022

[1]Faculty of Education, The University of Hong Kong, Hong Kong SAR, China
[2]School of Mathematics and Statistics, Central China Normal University, Wuhan, China
[3]Faculty of Education, The University of Macau, Taipa, Macau SAR, China

**Corresponding Author:**
Bo Li, School of Mathematics and Statistics, Central China Normal University, Wuhan, China.
Email: haoyoulibo@163.com

# 1. Introduction

Personalized assessment is an essential aspect of personalized education (Sadovaya et al., 2016; Tetzlaff et al., 2021; Waldeck, 2006; West, 2011). In educational contexts, assessments are used to observe and ascertain learners' performance and to provide feedback. The feedback, also called diagnostic information, should consider both strengths and weaknesses at the individual (micro) and grade (macro) levels (Leighton & Gierl, 2007; Maghsudi et al., 2021; Waldeck, 2006). Different psychological measurements have been developed to address this issue of providing personalized diagnostic information. Classic test theory (CTT) has been used to measure the reliability and validity of tests (Miller & Lovler, 2018). Baker and Kim (2017) argued that CTT can be used to assess whole tests but does not obtain sufficient data on the unobserved latent traits of students' abilities. This observation stimulated the emergence of item response theory (IRT, see Embretson & Reise, 2000). However, overall test scores or sub-scores have generally been used to rank individual students, which does not satisfy the increasing need for personalized assessments. Rather than assessing students' abilities on a continuous scale, cognitive diagnosis models (CDMs) have been designed to extract students' current learning status and cognitive structures from their test responses (Ravand, 2016; Rupp & Templin, 2008), which can not only provide the feedback at multiple levels but also efficiently contribute to learning and instructional design.

# 2. Literature review

## 2.1 Cognitive diagnosis models

In the past few years, a number of CDMs have been developed (de la Torre, 2009, 2011; Ravand & Robitzsch, 2018; Templin & Henson, 2006) and applied in a variety of contexts (Marszalek et al., 2019; Wu et al., 2020). For example, Marszalek et al. (2019) used the log-linear cognitive diagnostic model to assess the validity of the Social Issues Advocacy Scale (Nilsson et al., 2011). This section provides an overview of the basic notations and terminologies used in CDMs.

Assume that $K$ attributes of $I$ students are being assessed using a test with $J$ items. Accordingly, there are $2^K$ class patterns, denoted by $\alpha = [\alpha_1\ \alpha_2\ \cdots\ \alpha_{2^K}]$. For example, if there are two attributes to be tested, students can be classified into four ($2^2$) class patterns $\alpha_1 = [11]$, $\alpha_2 = [10]$, $\alpha_3 = [01]$, $\alpha_4 = [00]$, where "1" represents mastery of the attribute and "0" represents non-mastery.

The response data of student $i$ to item $j$ is dichotomous, where "1" indicates a correct response and "0" indicates an incorrect response. Thus, the response matrix $X$ of dimension $I \times J$ contains the information of all $I$ students' responses to all $J$ items in this test (George et al., 2016). Associated with a $Q$ matrix of dimension $J \times K$, it represents the relation between all $J$ items and $K$ attributes, where the element of $Q$ matrix $q_{jk} = 1$ indicates that attribute $k$ is tested in item $j$ , and $q_{jk} = 0$ otherwise (Embretson, 1984; Tatsuoka, 1985).

These $X$ and $Q$ matrices are two key input elements of CDMs. However, the correlations among attributes are likely to overlap, which affects the structure of the $Q$ matrix. A clear understanding of the relationships among attributes is necessary for the selection of appropriate models for empirical studies.

## 2.2 Application of CDMs in secondary mathematics

Studies have shown that complicated inter-attribute relationships are common in secondary mathematics contexts. A handful of studies have applied CDMs to mathematical education (Li et al., 2020; Wu, 2019; Yamaguchi & Okada, 2018). For instance, Li et al. (2020) recruited 747 kindergarteners to assess their mathematics problem-solving skills of 11 cognitive attributes. They underscored that CDMs not only provided more accurate information like the mastery patterns of attributes, but

also indicated a practical approach to evaluating the test quality (Li et al., 2020). Similarly, using data from 84 fourth grade elementary school students, Wu (2019) implemented CDMs to evaluate students' mastery of fraction operations. The results showed that remedial instructions based on CDMs were more effective than traditional group courses, and the study also highlighted diagnostic information was effective for all levels of achievement groups, particularly for those with medium and low achievements (Wu, 2019). Likewise, in junior mathematics, Yamaguchi and Okada (2018) analyzed seven countries or regions representing high-, average-, and low-ranked mathematical performances in the Trends in International Mathematics and Science Study 2007 (TIMSS 2007). They illustrated that CDMs performed better than IRT approaches in assessing fourth graders' mathematical literacy (Yamaguchi & Okada, 2018). They also argued that the diagnostic information provided by CDMs could reveal actual students' response behavior and hence the cognitive situations (Yamaguchi & Okada, 2018).

However, the aforementioned literature only addressed the CDMs applied in lower grades and examined relatively simple testing attributes. Higher-grade mathematical contexts have not received sufficient attention. Thus, this study addressed this gap by conducting a detailed investigation of CDMs in secondary mathematical contexts.

## 2.3 Higher-order generalized deterministic, inputs, noisy, "and" gate model

Many CDMs have been developed based on different assumptions. de la Torre (2011) argued that even if the mathematical expressions of diverse CDMs seem similar, it is necessary to carefully select the best models for specific empirical studies. Ravand and Robitzsch (2018) found that when the relationships among attributes are unknown, the generalized deterministic, inputs, noisy, "and" gate (GDINA) model (de la Torre, 2011) is suitable for initial estimates, due to its saturated form with great flexibility. Ma and de la Torre (2020b) also verified that one of the most direct methods of addressing the complex attribute relationship issue was to implement independent models like the GDINA model.

In the GDINA model, as not all attributes are needed for item $j$, the number of latent classes collapses from $2^K$ to $2^{K_j^*}$, where $K_j^* = \sum_{k=1}^{K} q_{jk}$(de la Torre, 2011). The skill classes for student $i$ related to item $j$ are reduced to $\alpha_{j;i}^* = [\alpha_{j;i1}^*, \cdots, \alpha_{j;iK_j^*}^*]$, which only contains the attributes that are examined by item $j$. This gives the total number of attributes that must be mastered to give the correct answer to item $j$, that is, the sum of the $j$-th row of the $Q$ matrix (George et al., 2016). de la Torre (2011) reported the item response function of the GDINA model as follows:

$$P(X_{ij} = 1|\alpha_{j;i}^*, \delta_j) = \delta_{j;0} + \sum_{k=1}^{K_j^*} \delta_{j;k}\alpha_{j;ik}^* + \sum_{k=1}^{K_j^*-1} \sum_{k'=k+1}^{K_j^*} \delta_{j;kk'}\alpha_{j;ik}^*\alpha_{j;ik'}^* + \ldots + \delta_{j;12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{j;ik}^*, \quad (1)$$

where $\delta_j = [\delta_{j;0}, \delta_{j;1}, \ldots, \delta_{K_j^*;1}, \delta_{j;12}, \ldots, \delta_{j;12\ldots K_j^*}]$ are the item parameters that are estimated with regard to item $j$, $\delta_{j;0}$ is the intercept for item $j$, $\delta_{j;k}$ is the main effect due to $\alpha_k^*$, $\delta_{j;kk'}$ is the interaction effect due to $\alpha_k^*\alpha_{k'}^*$, and $\delta_{j;12\ldots K_j^*}$ is the interaction effect due to $\alpha_1^*, \ldots, \alpha_{K_j^*}^*$.

The unclear relationships among attributes make it difficult to determine the structure of the joint attribute distribution (Ma & de la Torre, 2020b). de la Torre and Douglas (2004) explored the capability of higher-order GDINA models to address this problem. In higher-order models, the relationships among attributes are defined within the item response theory framework, which includes models such as the Rasch model (Rasch, 1993), the one-parameter logistic (1PL) model (Thissen, 1982), and the two-parameter logistic (2PL) model (Drasgow, 1989). Under the item response

function of the GDINA model, higher-order GDINA models can be derived for different joint attribute distributions using item response theory models as follows.

Under the two-parameter logistic (2PL) model,

$$P_k(\theta) = P(\alpha_k = 1|\theta) = \frac{\exp(\lambda_{0k} + \lambda_{1k}\theta)}{1 + \exp(\lambda_{0k} + \lambda_{1k}\theta)}, \tag{2}$$

where $\theta$ is the ability parameter of a single dimension, and $\lambda$ is the higher-order structural parameter (Ma & de la Torre, 2020b).

The one-parameter logistic (1PL) model is inferred by setting all $\lambda_{1k} = \lambda_1$. Then,

$$P_k(\theta) = P(\alpha_k = 1|\theta) = \frac{\exp(\lambda_{0k} + \lambda_1\theta)}{1 + \exp(\lambda_{0k} + \lambda_1\theta)}. \tag{3}$$

The Rasch model is described by setting all $\lambda_{1k} = 1$. Then,

$$P_k(\theta) = P(\alpha_k = 1|\theta) = \frac{\exp(\lambda_{0k} + \theta)}{1 + \exp(\lambda_{0k} + \theta)}. \tag{4}$$

In short, many studies have attempted to establish a theoretical framework for higher-order GDINA models. However, there have been few empirical studies in complex mathematical contexts such as secondary schools.

### 2.4 The present study

Using the empirical data collected from a secondary school in China, the present study aimed to investigate the application of the higher-order GDINA models considering the intricacy of interrelationships among attributes in secondary mathematics. Drawing upon the $Q$ matrix validation procedure, this study further examined the diagnostic information from both grade and individual levels. In particular, this study addressed two research questions:

**RQ1**: What kind of higher-order GDINA model best fits the empirical data and how can the $Q$ matrix validation process be used to modify the initial $Q$ matrix?
**RQ2**: What attributes have students mastered at the grade and individual levels and what personalized feedback is most helpful?

## 3. Method

### 3.1 Participants and instruments

The sample in this study consisted of 391 Grade 11 students in a high-performing level secondary school in Luoyang, Henan Province, China. They were from eight different classes and were selected following the school's fixed classes arrangement system. Students had firstly received about five 40-min lecture instructions on the "univariate function and its derivative" module in one week of the spring semester in 2019. The contents of this module were new to all students, but they already had preliminary knowledge in algebra and geometry such as Cartesian coordinates and trigonometric functions. Learning materials were the same and all students completed the instructions. Then a weekly test was designed for the purpose of formative assessment. All students took the test at the same time after accomplishing the instructions, and the test duration was 30 min. Informed consent was obtained from the participants involved in the study.

The test consisted of 13 items (See Appendix A). All the items were multiple choice questions, and the students were required to choose one of the four options coded as A, B, C, or D. A correct answer received one point and an incorrect answer received zero points.

The test was centered on functions and derivatives, which was in alignment with the Chinese mathematical curriculum standards for secondary schools (Ministry of Education, 2018). These basic but essential topics are linked with calculus, which is regarded as the beginning of higher mathematical thinking (Zulnaidi & Zamri, 2017). Vinner (1992) demonstrated that function acts as the core concept in calculus and Selden and Selden (1992) also asserted that functional concept is interconnected with other varieties of sub-topics in mathematics. Secondary school students are therefore supposed to master concrete knowledge about functions (Zulnaidi & Zamri, 2017).

## 3.2 Attribute coding taxonomy

CDMs have been applied in many international assessments (Chen & Chen, 2016; Evran, 2019; Wu et al., 2020; Yamaguchi & Okada, 2018). The global measurement of mathematical literacy has a long history, from the Pilot Twelve-Country Study (Foshay et al., 1962), through the First International Mathematics Study (FIMS64, see Husén, 1967), to more comprehensive international large-scale assessments (see Rutkowski et al., 2014). For example, the Program of International Student Assessment (PISA) is a triennial cyclical international assessment implemented by the Organisation for Economic Co-operation and Development (OECD), and it aims at measuring 15-year-old adolescents' capability to use their reading, mathematics, and science knowledge and skills to deal with life obstacles (OECD, 2004, 2013). The 15-year-old age range in PISA is similar to the secondary school students in China, so it is reasonable to include the PISA assessment framework for comparison.

The PISA test carried out by OECD in 2003, denoted PISA 2003, used an assessment framework with three clusters, the reproduction cluster (RepC), the connection cluster (ConC), and the reflection cluster (RefC). Several sub-domains were incorporated into each cluster (OECD, 2004). The revised framework used to assess mathematical literacy in PISA 2012 had four domains, mathematical content categories (MC), real-world context categories (RWC), mathematical concepts, knowledge, and skills (MCKK), and processes (Pro), with corresponding sub-domains (OECD, 2013). These two cycles of PISA tests were examined in this study as mathematical literacy is one of the three major domains in PISA (OECD, 2004, 2013). For comparison, this study also considered the Chinese mathematical curriculum standards published in 2017 (Ministry of Education, 2018), denoted as Chinese Standards 2017. The Chinese Standards 2017 identified seven areas of core mathematical literacy: computation (Com), data analysis (DA), intuitive imagination (II), mathematical abstraction (MA), mathematical modeling (MM), and logical reasoning (LR). The descriptions of the seven areas and the sub-domains are given in Table 1. Table 1 provides the definitions of mathematical literacy used in PISA 2003, PISA 2012, and the Chinese Standards 2017.

## 3.3 Q matrix construction

To construct the $Q$ Matrix, seven domain content experts, including four professors conducting research in mathematical education and education assessments and three research postgraduate students, participated in this process.

First, the domain content experts assinged attribute codes to the 13 items in the test given to the Grade 11 students based on the PISA 2003, PISA 2012, and Chinese Standards 2017 frameworks. Four attributes were identified, defined, and coded as follows.

**B1.** Basic concept development of derivatives: simple equations, zero or extreme points, and function range problems.

**Table 1.** Mathematical literacy in PISA 2003, PISA 2012, and Chinese standards 2017.

| PISA 2003 | PISA 2012 | Chinese Standards 2017 |
|---|---|---|
| **Reproduction cluster** | **Mathematical content categories** | **Computation** |
| Standard representations and definitions | Quantity | Understanding operation objects |
| Routine computations | Uncertainty and data | Mastering algorithms |
| Routine procedures | Change and relationships | Exploring ideas |
| Routine problem solving | Space and shape | Selecting methods |
| | | Designing programs |
| | | Obtaining results |
| **Connection cluster** | **Real world context categories** | **Data analysis** |
| Modeling | Personal | Collecting data |
| Standard problem | Societal | Organizing data |
| Solving translation and interpretation | Occupational | Extracting information |
| Multiple well-defined | Scientific | Constructing models |
| Methods | | Inferring |
| | | Obtaining conclusions |
| **Reflection cluster** | **Mathematical concepts, knowledge, and skills** | **Intuitive imagination** |
| Complex problem | Fundamental mathematical | Understanding positional relationships, morphological changes, and movement laws of things |
| Solving and posing | Capabilities | |
| Reflection and insight | Communication | Using graphs to describe and analyze mathematical problems |
| Original mathematical | Representation | |
| Approach | Devising strategies | Establishing connections between forms and numbers |
| Multiple complex | Mathematization | |
| Methods | Reasoning and argument | Constructing intuitive models of mathematical problems |
| Generalization | Using symbolic, formal, and technical language | Exploring ideas for solving problems |
| | Operations | |
| | Using mathematical tools | |
| | **Processes** | **Mathematical abstraction** |
| | Formulate | Concepts and concepts |
| | Employ | Quantity and quantity graphics and graphics |
| | Interpret/Evaluate | General laws and structures |
| | | **Mathematical modeling** |
| | | Finding problems |
| | | Asking questions |
| | | Analyzing problems, establishing models, and determining parameters |
| | | Calculating solutions |
| | | Testing results |
| | | Improving models |
| | | Solving practical problems |
| | | **Logical reasoning** |
| | | Inductions |
| | | Analogy |

**B2.** Complex inductive contextualization of derivatives: induction from the known to solve the unknown problems.

**B3.** Basic routine problem solving of derivatives: graphs that pass through a fixed point or quantitative inequalities.

**B4.** Complex transformative contextualization of derivatives: transformation by the combination of numbers and graphs.

The domain experts unanimously agreed that the four defined attributes, all drawn from the "univariate function and its derivative" module, were highly consistent with the taxonomies of PISA 2003, PISA 2012, and Chinese Standards 2017 (See Table 2). For instance, B1 covered all three clusters in PISA 2003, i.e., RepC, ConC, and RefC; three clusters in PISA 2012, i.e., MC, MCKK, and Pro; and five areas in the Chinese Standards 2017, i.e., MA, LR, Com, MM, and II.

After verifying the consistency of the items with these three standards, the domain experts collaboratively constructed an initial $Q$ matrix that matched specific items with attributes (See Table 3).

Here we illustrate the process of constructing the initial $Q$ matrix using Item 4.

***Item 4*** If $x = -2$ is the extreme point of the given function $f(x) = (x^2 + ax - 1)e^{x-1}$, what is the minimum value of $f(x)$ ?

A. $-1$ B. $-2e^{-3}$ C. $5e^{-3}$ D. 1

The domain experts confirmed that Item 4 tested students' understanding of attributes B1 and B2. Therefore, in the initial $Q$ matrix, both attributes were assigned a value of 1 in the Item 4 row, and B3 and B4 were assigned a value of 0, as this item did not test students' mastery of those attributes.

## 3.4 Analysis strategy

After construction of initial $Q$ matrix, model selection was applied to choose the appropriate model for the empirical data in this study. Then a two-step $Q$ matrix validation process was implemented. In the first step, we used the selected model to get the modification elements of initial $Q$ matrix based on the stepwise approach (Ma & de la Torre, 2020a). In the second step, the domain experts made the decisions on acceptance or rejection of these elements to acquire the final $Q$ matrix. After that, we

**Table 2.** Attribute consistency under PISA 2003, PISA 2012, and Chinese standards 2017.

| No. | Attribute Definitions | PISA 2003 | PISA 2012 | Chinese Standards 2017 |
|---|---|---|---|---|
| B1 | Basic concept development of derivatives: simple equations, zero or extreme points, and function range problems. | RepC<br>ConC<br>RefC | MC<br>MCKK<br>Pro | MA<br>LR<br>Com<br>MM<br>II |
| B2 | Complex inductive contextualization of derivatives: induction from the known to solve the unknown problems. | RepC<br>ConC<br>RefC | MC<br>MCKK<br>Pro | MA<br>LR<br>Com<br>MM<br>II |
| B3 | Basic routine problem solving of derivatives: graphs that pass through a fixed point or quantitative inequalities. | RepC<br>ConC<br>RefC | MC<br>MCKK<br>Pro | MA<br>LR<br>Com<br>MM<br>II |
| B4 | Complex transformative contextualization of derivatives: transformation by the combination of numbers and graphs | RepC<br>ConC<br>RefC | MC<br>MCKK<br>Pro | MA<br>LR<br>Com<br>MM<br>II |

*Note.* For PISA 2003, RepC = Reproduction cluster; ConC = Connection cluster; and RefC = Reflection cluster. For PISA 2012, MC = Mathematical content; RWC = Real world context; MCKK = Mathematical concepts, knowledge, and skills; and Pro = Processes. For Chinese Standards 2017, MA = Mathematical abstraction; LR = Logical reasoning; MM = Mathematical modeling; II = Intuitive imagination; Com = Computation; and DA = Data analysis.

**Table 3.** Initial Q matrix.

| | B1 | B2 | B3 | B4 |
|---|---|---|---|---|
| Item 1 | 1 | 0 | 1 | 0 |
| Item 2 | 0 | 0 | 1 | 0 |
| Item 3 | 1 | 0 | 1 | 0 |
| Item 4 | 1 | 1 | 0 | 0 |
| Item 5 | 0 | 1 | 0 | 0 |
| Item 6 | 0 | 1 | 0 | 0 |
| Item 7 | 1 | 0 | 0 | 0 |
| Item 8 | 0 | 0 | 0 | 1 |
| Item 9 | 0 | 0 | 1 | 0 |
| Item 10 | 1 | 0 | 0 | 0 |
| Item 11 | 0 | 0 | 0 | 1 |
| Item 12 | 0 | 0 | 1 | 1 |
| Item 13 | 0 | 1 | 0 | 0 |

conducted the diagnostic information analysis from four aspects: attribute prevalence, latent class and posterior probability percentage, item parameter estimates, and individual skill profiles, using the selected model and the final $Q$ matrix.

We conducted all model selection, $Q$ matrix validation and diagnostic information analysis using *R* software with the *GDINA package version 2.8.8* (Ma et al., 2022). For between-model comparisons, we used six relative fit indices to evaluate the goodness-of-fit of the models and they were: loglikelihood, deviance, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Consistent Akaike Information Criterion (CAIC), and Sample Size-adjusted Bayesian Information Criterion (SABIC). General criteria for the indices indicate that the smaller values of them, the better the goodness-of-fit of the models.

For within-model comparisons, we chose three absolute fit indices and they were: $M_2$, Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMSR). Specifically, the $M_2$ statistic was used to assess the overall model fit for dichotomous responses and the acceptable models should include a nonsignificant value of $M_2$ (Chen et al., 2018; Maydeu-Olivares & Joe, 2006). We also adopted the cut-off RMSEA values of 0.01, 0.05, and 0.08 for excellent, good, and mediocre fit respectively (MacCallum et al., 1996) together with SRMSR smaller than or equal to 0.05 (Erdle et al., 2010).

## 4. Results

### *4.1 Model selection (RQ1)*

To address RQ1, this study investigated which higher-order GDINA models fit the empirical data and how the $Q$ matrix validation process can be used to modify the expert judgment-based initial $Q$ matrix. As discussed in the literature review part about the higher-order GDINA models, the interactions among attributes may affect the joint attribute distribution and thus the structure of the matrix (Ma & de la Torre, 2020b). Research has verified that higher-order GDINA models with Rasch, 1PL, and 2PL joint attribute distributions can be used to select appropriate models for empirical studies (de la Torre, 2011; de la Torre & Douglas, 2004; Ma & de la Torre, 2020b). This study also included the saturated GDINA form for comparison (Table 4).

The saturated GDINA model had the smallest loglikelihood (−2502.771), the lowest deviance (5005.542) and AIC (5103.542). However, Chen et al. (2017) found that BIC, CAIC, and SABIC

**Table 4.** Relative and absolute fit indices for the four models.

| | Saturated GDINA | Higher-order GDINA with Rasch | Higher-order GDINA with 1PL | Higher-order GDINA with 2PL |
|---|---|---|---|---|
| **Relative indices** | | | | |
| Loglikelihood | −2502.771 | −2522.64 | −2516.716 | −2514.32 |
| Deviance | 5005.542 | 5045.28 | 5033.432 | 5028.639 |
| AIC | 5103.542 | 5121.28 | 5111.432 | 5112.639 |
| BIC | 5298.009 | 5272.091 | 5266.211 | 5279.325 |
| CAIC | 5347.009 | 5310.091 | 5305.211 | 5321.325 |
| SABIC | 5142.534 | 5151.519 | 5142.466 | 5146.061 |
| **Absolute indices** | | | | |
| $M_2$ | 29.206 ($df = 42$) $p = .933$ | 46.580 ($df = 53$) $p = .721$ | 39.833 ($df = 52$) $p = .892$ | 33.902 ($df = 49$) $p = .950$ |
| RMSEA | 0 | 0 | 0 | 0 |
| SRMSR | 0.041 | 0.049 | 0.041 | 0.041 |

*Note.* AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent Akaike information criterion; SABIC = sample size-adjusted Bayesian information criterion; RMSEA = root mean square error of approximation; SRMSR = standardized root mean square residual; *df* = degree of freedom.

were more suitable when the model was not specified. Information criteria have been found to be more useful than loglikelihood and deviance (Anderson et al., 1998; Dziak et al., 2020). We found that the higher-order GDINA model with 1PL had the smallest BIC (5266.211), CAIC (5305.211), and SABIC (5142.466) among the four models, indicating that it had the best model fit.

The absolute fit indices also showed that the higher-order GDINA model with 1PL had a good fit. The $M_2$ statistic indicated that the higher-order GDINA model with 1PL was suitable in this setting ($M_2 = 39.833$, $df = 52$, $p = .892$). The zero value of RMSEA for the higher-order GDINA model with 1PL indicated the excellent fit of this model. For SRMSR, the higher-order GDINA model with 1PL had a good fit (SRMSR = 0.041 < 0.05).

Together, the relative and absolute fit indices showed that the higher-order GDINA model with 1PL best fit the data.

## 4.2 Q matrix validation (RQ1)

The initial $Q$ matrix was constructed according to the judgments of the domain experts and thus was subjective. Researchers have proposed multiple methods for partially reducing this subjectivity in $Q$ matrix validation (de la Torre, 2008; de la Torre & Chiu, 2016; Ma & de la Torre, 2020a). Ma and de la Torre (2020a) suggested that a stepwise approach combining the proportion of variance accounted for (PVAF) and the Wald test is suitable when the fixed number of attributes is relatively small, as in this empirical study ($n = 4$). We used the higher-order GDINA model with 1PL to validate the $Q$ matrix and the results suggested that two elements in the initial $Q$ matrix should be modified. Specifically, the results suggested that Items 1 and 3 might not examine B1.

***Item 1*** Given the function $f(x) = 2xf'(e) + \ln x$, where $f'(e)$ is the first order derivative of $f(x)$ at point $e$, then $f(e) = ?$

A. $-e$ B. $e$ C. $-1$ D. 1

***Item 3*** If the function $y = x^3 - 3bx + 1$ is decreasing in the given interval [1, 2], then what is the possible range of $b$, $b \in R$ ?

A. $b \leq 4$ B. $b < 4$ C. $b \geq 4$ D. $b > 4$

These two suggestions about the modifications of two elements in the initial $Q$ matrix were firstly carried out by stepwise statistical methods combing PVAF and the Wald test, and then whether to accept the adoptions was discussed by domain experts (Ma & de la Torre, 2020b). Following consultation, the domain experts accepted the two suggested modifications. That is, they agreed that attribute B1 was not examined in Items 1 and 3. These changes were made to the final $Q$ matrix (See Table 5).

## 4.3 Diagnostic information (RQ2)

To answer RQ2, our analysis of the final $Q$ matrix based on the higher-order GDINA model with 1PL reported four outputs of diagnostic information: (1) attribute prevalence; (2) latent class and posterior probability percentage; (3) item parameter estimates; and (4) individual skill profiles.

Figure 1 summarizes the prevalence of the four attributes. The prevalence of a specific attribute is the probability that students have mastered that attribute regardless of their performance on other attributes. Technically, it is the sum of the probabilities of the latent classes that have mastered that attribute. For example, the attribute prevalence of B1 is the sum of the posterior probabilities of latent classes 1111, 1110, 1100, 1010, 1011, 1101, 1000, and 1001. Overall, more than three-quarters of the students had mastered attributes B1 (77.3%, $n = 302$), B2 (77.0%, $n = 301$), and B3 (81.9%, $n = 320$). In contrast, less than half of the students succeeded in mastering attribute B4 (38.6%, $n = 151$).

To assess the latent class and posterior probabilities, the performance of each student in each of the four attributes was classified as either mastery or non-mastery, resulting in 16 latent classes. All the latent classes are depicted in descending order of posterior probability in Figure 2. More than half of the students (62.18%) were classified into the two dominant latent classes 1111 and 1110. More specifically, the 1111 latent class (mastered B1, B2, B3, B4) had the largest posterior probability percentage (34.50%), indicating that most participants had mastered all four attributes. The latent class 1110 (mastered B1, B2, B3) had the second largest proportion (27.68%).

The item parameter estimates, which were used to measure the quality of the test, are shown in Table 6. Eleven of the items (1, 2, 3, 5, 6, 7, 8, 9, 10, 11, and 13) tested one attribute, and two items (4 and 12) tested two attributes. Moreover, item discrimination has been commonly used to measure the quality of the items and thus the test (Lee et al., 2012; Wang et al., 2018). For the single-attribute items, P(1) and P(0) represent the probabilities of correctly answering items that do or do not test the focal single attribute. For the double-attribute items, P(11) and P(00) represent the probabilities of correctly answering items that do or do not test both attributes, and P(10) and P(01) represent mastery of only the first or only the second attribute. In CDMs, referred to Wu et al.'s (2020) definition, the item discrimination index (IDI) for item $j$ with a single attribute is defined as

$$IDI_j = P_j(1) - P_j(0). \tag{5}$$

Similarly, the IDI for double-attribute item $j$ is denoted as

$$IDI_j = P_j(11) - P_j(00). \tag{6}$$

**Table 5.** Final Q matrix (modified elements only).

| | B1 | B2 | B3 | B4 |
|---|---|---|---|---|
| Item 1 | 0[a] | 0 | 1 | 0 |
| Item 3 | 0[a] | 0 | 1 | 0 |

[a]A change from 1 to 0.

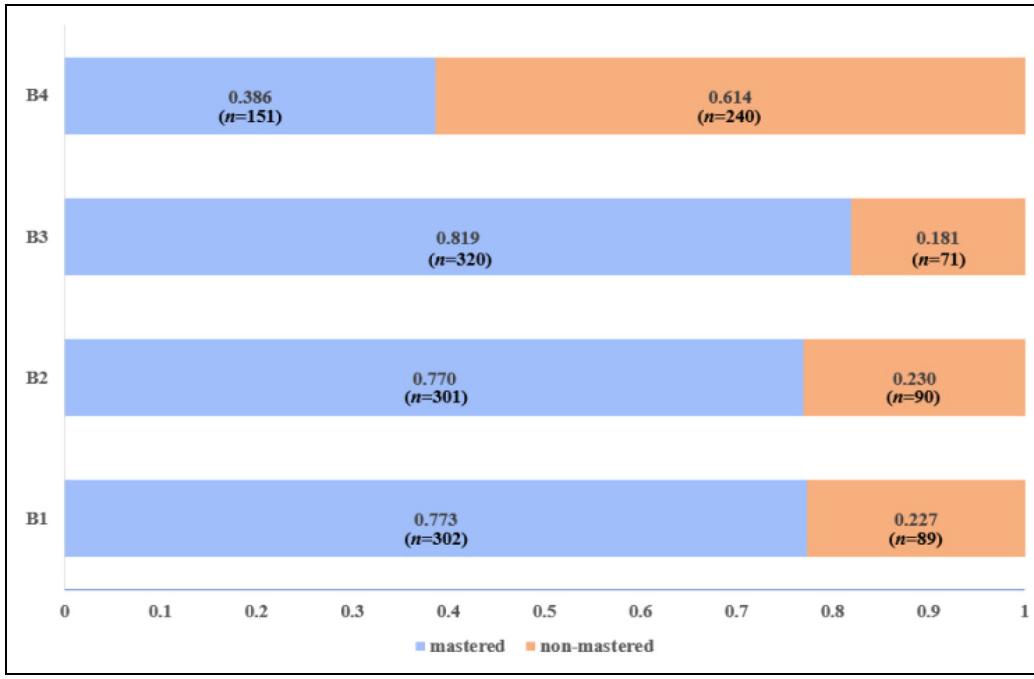



**Figure 1.** Prevalence of attributes B1 to B4.

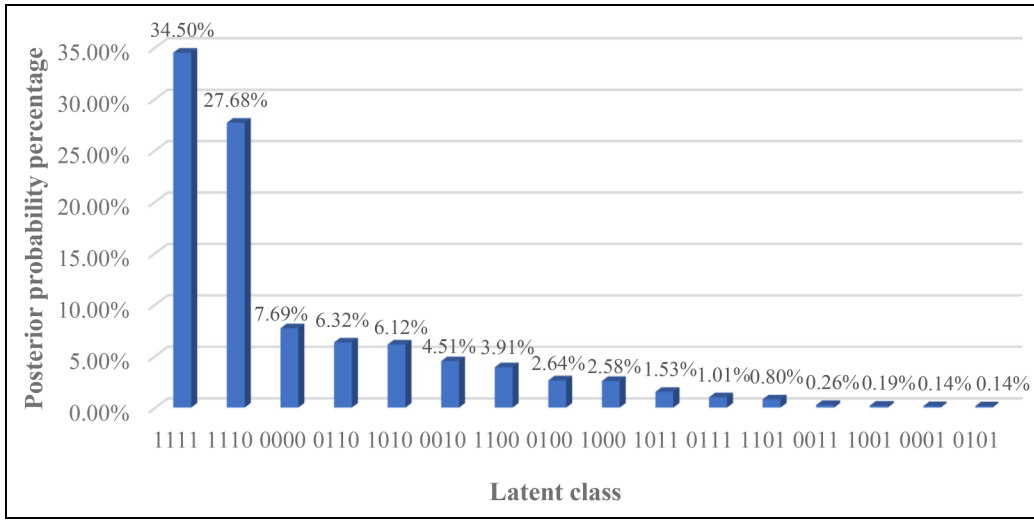



**Figure 2.** Latent class and posterior probability percentage.

Doust et al. (2021) defined 0.2 as the cut-off index value to discriminate capabilities and 0.4 as the ideal value. The IDIs of all 13 items were larger than 0.2, indicating that they could discriminate students' abilities; five items (4, 6, 8, 10, 12) had ideal IDI values of more than 0.4.

Table 7 and Figure 3 show the individual skill profiles of four representative students, S1 to S4. The four selected students represented three different levels of achievement: high (S1), medium (S2

**Table 6.** Item parameter estimates.

| | P(0)<br>P(00) | P(1)<br>P(10) | P(01) | P(11) | IDI |
|---|---|---|---|---|---|
| Item 1 | 0.604 | 0.938 | | | 0.334 |
| Item 2 | 0.606 | 0.859 | | | 0.253 |
| Item 3 | 0.535 | 0.869 | | | 0.334 |
| Item 4 | 0.324 | 0.502 | 0.343 | 0.929 | 0.605 |
| Item 5 | 0.688 | 0.951 | | | 0.263 |
| Item 6 | 0.445 | 0.980 | | | 0.535 |
| Item 7 | 0.576 | 0.946 | | | 0.370 |
| Item 8 | 0.256 | 0.812 | | | 0.555 |
| Item 9 | 0.631 | 0.875 | | | 0.244 |
| Item 10 | 0.238 | 0.711 | | | 0.473 |
| Item 11 | 0.155 | 0.516 | | | 0.361 |
| Item 12 | 0.312 | 0.645 | 0.047 | 1.000 | 0.688 |
| Item 13 | 0.310 | 0.671 | | | 0.361 |

**Table 7.** Individual skill profiles of four representative examinees.

| | B1 | B2 | B3 | B4 |
|---|---|---|---|---|
| S1 | 0.996 | 0.996 | 0.999 | 0.950 |
| S2 | 0.815 | 0.020 | 0.941 | 0.043 |
| S3 | 0.012 | 0.726 | 0.159 | 0.008 |
| S4 | 0.014 | 0.009 | 0.030 | 0.005 |

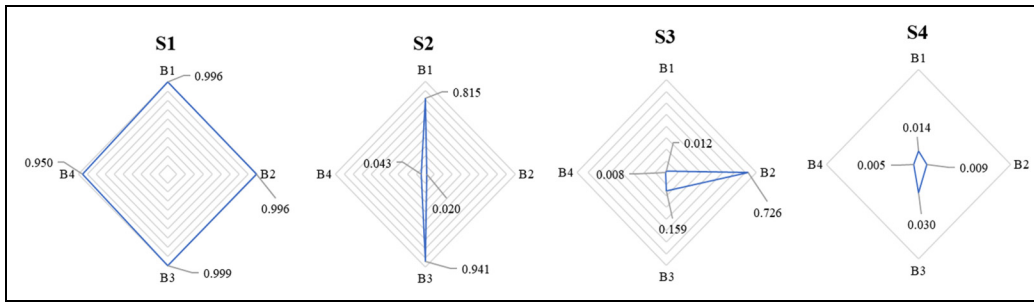



**Figure 3.** Individual skill profiles of four representative examinees.

and S3), and low (S4). Studies have used three measures to estimate individual skill profiles: maximum likelihood estimation (MLE), maximum a posteriori (MAP), and expected a posteriori (EAP) (George et al., 2016; Huebner & Wang, 2011; Ma & de la Torre, 2020b). Huebner and Wang (2011) asserted that EAP is preferred for practical and empirical studies, and George et al. (2016) demonstrated that 0.5 is a suitable EAP cut-off value to distinguish mastery from non-mastery.

Thus, this study adopted the EAP method to estimate mastery probability. Of the four selected students, S1 mastered all four attributes, with mastery probability values all greater than or equal to 0.950. In B1 and B3, S2 had mastery probabilities of 0.815 and 0.941, but in B2 and B4 they

were only 0.020 and 0.043. S3 did well only in B2 with a probability of 0.726 (B1: 0.012; B3: 0.159; B4: 0.008). S4 did not master any of the four attributes and had mastery probabilities less than or equal to 0.030.

## 5. Conclusion and discussion

This empirical study used a sample of 391 Grade 11 students from a secondary school in China. The study examined the use of higher-order CDMs in the $Q$ matrix validation process when there are intricate interrelationships among attributes. Then, this study analyzed students' diagnostic information at both the grade and individual levels.

The results show that the mathematical literacy frameworks used in Chinese secondary mathematics (Chinese Standards 2017) align with the international PISA 2003 and 2012 frameworks for teaching and assessing the "univariate function and its derivative" module. The content in this module is of significance for secondary students, marking the starting point for acquisition of higher mathematical thinking skills in calculus (Zulnaidi & Zamri, 2017). As the study relies on the judgments of domain experts, these results need to be interpreted with caution and carefully applied to other modules. However, the results suggest that our approach is a promising alternative to CDMs based on international large-scale assessments. This finding is consistent with the literature (Chen & Chen, 2016; Evran, 2019; Wu et al., 2020). The study also sheds light on mathematical teaching contexts in Chinese secondary settings. It extends the research on mathematical literacy for kindergarteners and primary school students (Li et al., 2020; Yamaguchi & Okada, 2018).

All 13 items in the assessment implemented in this study test students' mastery of four attributes, which are related to each other in complex ways. Our analysis suggests that the higher-order GDINA model with joint attribute distribution of 1PL is more suitable when there are uncertain relationships among attributes (de la Torre & Douglas, 2004; Ma & de la Torre, 2020b). We also find that the stepwise PVAF modification process developed by Ma and de la Torre (2020b) increases the objectivity of the final $Q$ matrix. The finding that higher-order CDMs are suitable for the $Q$ matrix validation process broadly supports other studies in this area such as Ma and de la Torre (2020a) and de la Torre and Chiu (2016).

The study also shows that the higher-order GDINA model with joint attribute distribution of 1PL provides diagnostic information about the strengths and weaknesses of the students at both the grade and individual levels (Leighton & Gierl, 2007; Maghsudi et al., 2021; Waldeck, 2006). It has four major outputs: attribute prevalence, latent class and posterior probability percentage, individual skill profiles, and item parameter estimates.

Two of the four main outputs, attribute prevalence and latent class and posterior probability percentage, provide diagnostic information at the grade level. In this study, the participants at the grade level had mastery of attributes B1, B2, and B3, as shown by the attribute prevalence values (B1: 77.3%, B2: 77.0%, B3: 81.9%). In contrast, the attribute prevalence of B4 was unsatisfactory (B4: 38.6%), with the majority of the participants failing to understand and internalize the complex transformation of derivatives. The latent class and posterior probability percentages provided more in-depth information. Two latent classes 1111 (mastered B1, B2, B3, B4) and 1110 (mastered B1, B2, B3) dominate the whole proportion and these results are partially consistent with grade-level analysis, suggesting that most participants did not have mastery of B4. And it is likely that mastery of the first three attributes may contribute to the mastery of the fourth attribute B4. Clearly, the diagnostic information provides detailed remedial teaching and guidance for teachers and students (Leighton & Gierl, 2007; Templin & Henson, 2010), allowing instructors to quickly recognize the target population's knowledge status and to make appropriate instructional adjustments, i.e., teachers of this study's participants could enhance students' comprehension of B4-related learning materials.

The individual skill profiles provide information on individuals' cognitive structures. This study examined the cognitive structures of four students with different levels of mastery. S1 had mastered all four attributes, so the instructor could provide more advanced learning materials to encourage self-improvement. For S2 (strengths in B1 and B3 but weaknesses in B2 and B4) and S3 (strengths in B2 but weaknesses in B1, B3, and B4), who had a mixture of strengths and weaknesses, and S4 (weaknesses in all attributes), who had unsatisfactory performance, the instructor might have considered implementing collaborative learning strategies focused on remedial work on the attribute B4, which all three students had failed to master. The instructor could also provide personalized instructions that address specific gaps, such as B2-related instruction for S2, and B1-related instruction for S3. In sum, instructors could devise strategies based on the students' skill profiles and diagnostic information. Together with information about the grade level, these CDMs contribute to the development of personalized assessments, which are crucial to personalized education (Sadovaya et al., 2016; Tetzlaff et al., 2021; Waldeck, 2006; West, 2011).

In general, assessing the mathematical skills of secondary school students is complicated, and this study offers the novel approach of the higher-order GDINA model. Of course, this study has certain limitations.

Test quality is an important condition. Items that have a high guessing rate, namely P(0) for single-attribute items and P(00) for double-attribute items (See Table 6), may impact the cognitive diagnosis even if the item discrimination index value is high (de la Torre et al., 2010). Test quality could be tested and verified using item parameter estimates. In this study, there are six items with relatively high guessing rates i.e., P(0) and P(00) are larger than 0.5 (See Table 6). de la Torre et al. (2010) asserted that this may occur when there is great dependence on the attributes. Thus, future studies should create more items and only use the low guessing rate items in the CDMs. Another issue is the integration of $Q$ matrix validation. The method adopted in this study is consistent with current research (de la Torre & Chiu, 2016; Ma & de la Torre, 2020a) and is based on the assumption that there is a fixed number of attributes, four in this case. Future studies could examine the validation of the number of attributes. Finally, this study examined only one test, which was administered after the derivative module was completed. The effects of personalized assessment for remedy or promotion are unclear and should be considered in future studies. In addition, the students were recruited only from one city in China. The generality can be promoted in future studies by collecting more data from a variety of schools in other cities.

In summary, despite the limitations, this study has several noteworthy theoretical and practical implications. Theoretically, this study contributes to the growing body of literature suggesting that higher-order GDINA models are applicable when confronting the complex structures of attributes, particularly in secondary mathematics. Our study also underpins the consistency of international large-scale assessments like PISA and the national Chinese mathematical curriculum standards in $Q$ matrix construction process. This adds values to a more comprehensive understanding of curriculum schemes from both local and global perspectives. Meanwhile, the issue of subjectivity may exist when domain content experts construct the $Q$ matrix. Our study provides a feasible approach to address this issue by implementing the $Q$ matrix validation process. Practically, the diagnostic information in the findings of our study is valuable for personalized assessment. For one thing, teachers can evaluate test quality by checking item parameter estimates, and design targeted instructional strategies at grade level referring to attribute prevalence and latent class percentages. For another, remedial programs and instructions can also be carried out in accordance with the feedback on individual skill profiles.

## Acknowledgements

We are sincerely grateful to reviewers and editorial team for comments that substantially improved the article. We appreciate Dr. Lei Wang (National Engineering Research Center for E-Learning, NERCEL, Central China Normal University) and Ms. Yan Chen (Luoyang No. 2 Middle School, China) for provision of original test

paper and students' response data in this study. And we also appreciate the suggestions provided by Dr. Yuyao Tong, Mr. Yang Tao, Mr. Pengjin Wang, Mr. Chao Yang, and Mr. Wei Jia (Faculty of Education, The University of Hong Kong).

## Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Contributorship

Ying Zhang initiated the project, conducted the research, and drafted the manuscript. Yi Jin wrote the statistical analysis plan, analyzed the data, and conducted the *Q* Matrix Validation part. Zhenrong Xiong collected the data and completed the theoretical framework of CDMs. Bo Li, Gaowei Chen, Shing On Leung, and Na Li supervised the study, provided important ideas for the research, and revised the draft. All authors read and approved the final manuscript.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Hong Kong Research Grants Council, University Grants Committee, self-determined research funds of CCNU from the colleges' basic research and operation from the Ministry of Education, China, National Natural Science Foundation of China, (grant number 17605221, CCNU19TD006, 61877023)

## ORCID iD

Ying Zhang https://orcid.org/0000-0002-7754-5506

## References

Anderson, D. R., Burnham, K. P., & White, G. C. (1998). Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies. *Journal of Applied Statistics, 25*(2), 263–282. https://doi.org/10.1080/02664769823250

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. New York, NY: Springer.

Chen, F., Liu, Y., Xin, T., & Cui, Y. (2018). Applying the $M_2$ statistic to evaluate the fit of diagnostic classification models in the presence of attribute hierarchies. *Frontiers in Psychology, 9*, 1875. https://doi.org/10.3389/fpsyg.2018.01875

Chen, H., & Chen, J. (2016). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology, 36*(6), 1049–1064. https://doi.org/10.1080/01443410.2015.1076764

Chen, Q., Luo, W., Palardy, G. J., Glaman, R., & McEnturff, A. (2017). The efficacy of common fit indices for enumerating classes in growth mixture models when nested data structure is ignored: A Monte Carlo study. *Sage Open, 7*, 1. https://doi.org/10.1177/2158244017700459

de la Torre, J. (2008). An empirically based method of Q–matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). DINA Model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/S11336-011-9207-7

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8

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

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*(2), 227–249. https://doi.org/10.1111/j.1745-3984.2010.00110.x

Doust, A. R., Khan, W. A., & Al-Ghafri, M. (2021). An item analysis study on TIMSS 2015 mathematics items of Omani and Iranian students comparison IRT and CDM approaches. *International Journal of Mathematics Trends and Technology, 67*(9), 87–95. https://doi.org/10.14445/22315373/IJMTT-V67I9P510

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*(1), 77–90. https://doi.org/10.1177/014662168901300108

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics, 21*(2), 553–565. https://doi.org/10.1093/bib/bbz016

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175–186. https://doi.org/10.1007/BF02294171

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Erdle, S., Irwing, P., Rushton, J. P., & Park, J. (2010). The general factor of personality and its relation to self-esteem in 628,640 internet respondents. *Personality and Individual Differences, 48*(3), 343–346. https://doi.org/10.1016/j.paid.2009.09.004

Evran, D. (2019). An application of cognitive diagnosis modeling in TIMSS: A comparison of intuitive definitions of Q-matrices. *International Journal of Modern Education Studies, 3*(1), 4–17. https://doi.org/10.51383/ijonmes.2019.33

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries*. Hamburg, Germany: UNESCO Institute for Education.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1–24. https://doi.org/10.18637/jss.v074.i02

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407–419. https://doi.org/10.1177/0013164410388832

Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries (Vol. 2)*. Stockholm, Sweden: Almqvist & Wiksell.

Lee, Y.-S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis. *CTT, and IRT Indices: An Empirical Investigation. Asia Pacific Education Review, 13*(2), 333–345. https://doi.org/10.1007/s12564-011-9196-3

Leighton, J. P., & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Li, L., Zhou, X., Huang, J., Tu, D., Gao, X., Yang, Z., & Li, M. (2020). Assessing kindergarteners' mathematics problem solving: The development of a cognitive diagnostic test. *Studies in Educational Evaluation, 66*, 100879. https://doi.org/10.1016/j.stueduc.2020.100879

Ma, W., & de la Torre, J. (2020a). An empirical Q–matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163. https://doi.org/10.1111/bmsp.12156

Ma, W., & de la Torre, J. (2020b). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*, 1–26. https://doi.org/10.18637/jss.v093.i14

Ma, W., de la Torre, J., & Sorrel, M. (2022). *GDINA: The generalized DINA model framework (R package version 2.8.8)*. https://CRAN.R-project.org/package=GDINA

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

Maghsudi, S., Lan, A., Xu, J., & Van Der Schaar, M. (2021). Personalized education in the artificial intelligence era: What to expect next. *IEEE Signal Processing Magazine, 38*(3), 37–50. https://doi.org/10.1109/msp.2021.3055032

Marszalek, J. M., Barber, C., & Nilsson, J. E. (2019). A cognitive diagnostic analysis of the Social Issues Advocacy Scale (SIAS). *Educational Psychology, 39*(6), 839–858. https://doi.org/10.1080/01443410.2019.1585516

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Miller, L. A., & Lovler, R. L. (2018). *Foundations of psychological testing: A practical approach*. Thousand Oaks, CA: Sage.

Nilsson, J. E., Marszalek, J. M., Linnemeyer, R. M., Bahner, A. D., & Misialek, L. H. (2011). Development and assessment of the social issues advocacy scale. *Educational and Psychological Measurement, 71*(1), 258–275. https://doi.org/10.1177/0013164410391581

Ministry of Education. (2018). *Mathematics curriculum standards for general senior high schools*. Beijing, China: People's Education Press.

OECD. (2004). *The PISA 2003 assessment framework*. https://doi.org/10.1787/9789264101739-en

OECD. (2013). *PISA 2012 assessment and analytical framework*. https://doi.org/10.1787/9789264190511-en

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782–799. https://doi.org/10.1177/0734282915623053

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology, 38*(10), 1255–1277. https://doi.org/10.1080/01443410.2018.1489524

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Sadovaya, V. V., Korshunova, O. V., & Nauruzbay, Z. Z. (2016). Personalized education strategies. *International Electronic Journal of Mathematics Education, 11*(1), 199–209. https://doi.org/10.29333/iejme/324

Selden, A., & Selden, J. (1992). *Research perspective on conceptions of functions: summary and overview*. Washington, DC: Mathematical Association of America.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55–73. https://doi.org/10.3102/10769986010001055

Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review, 33*(3), 863–882. https://doi.org/10.1007/s10648-020-09570-w

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*(2), 175–186. https://doi.org/10.1007/bf02296273

Vinner, S. (1992). *The function concept as a prototype for problems in mathematics learning*. Washington, DC: Mathematical Association of America.

Waldeck, J. H. (2006). What does "personalized education" mean for faculty, and how should it serve our students? *Communication Education, 55*(3), 345–352. https://doi.org/10.1080/03634520600748649

Wang, W., Song, L., & Ding, S. (2018). The index and application of cognitive diagnostic test from the perspective of classification. *Psychol. Sci, 41*, 475–483. https://doi.org/10.16719/j.cnki.1671-6981.20180234

West, M. (2011). *Using technology to personalized learning and assess students in real-time*. Washington, DC: The Brookings Institution.

Wu, H.-M. (2019). Online individualised tutor for improving mathematics learning: A cognitive diagnostic model approach. *Educational Psychology, 39*(10), 1218–1232. https://doi.org/10.1080/01443410.2018.1494819

Wu, X., Wu, R., Chang, H.-H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology, 11*, 2230. https://doi.org/10.3389/fpsyg.2020.02230

Yamaguchi, K., & Okada, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PLoS ONE, 13*(2), e0188691. https://doi.org/10.1371/journal.pone.0188691

Zulnaidi, H., & Zamri, S. N. A. (2017). The effectiveness of the GeoGebra software: The intermediary role of procedural knowledge on students' conceptual knowledge and their achievement in mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 13*(6), 2155–2180. https://doi.org/10.12973/eurasia.2017.01219a

## Author biographies

**Ying Zhang**, PhD student, Faculty of Education, The University of Hong Kong, Hong Kong SAR. Research focus: mathematics education, learning sciences, teacher professional development, international large-scale assessments, educational measurement.

**Yi Jin**, PhD student, Faculty of Education, The University of Hong Kong, Hong Kong SAR. Research focus: psychometrics, educational assessment.

**Zhenrong Xiong**, MSc graduate, School of Mathematics and Statistics, Central China Normal University, Wuhan, China. Research focus: education big data.

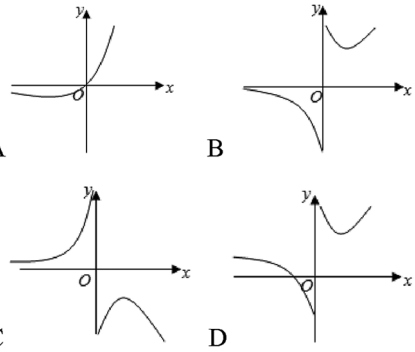
**Shing On Leung**, PhD, Associate Professor, Faculty of Education, The University of Macau, Macau SAR. Research focus: educational measurement, applications of statistics in education and social sciences.

**Gaowei Chen**, PhD, Associate Professor, Faculty of Education, The University of Hong Kong, Hong Kong SAR. Research focus: educational psychology, learning sciences, research methods and methodologies, teacher education and development, and technology-enhanced learning.

**Na Li**, PhD, Lecturer, School of Mathematics and Statistics, Central China Normal University, Wuhan, China. Research focus: comparative education, mathematics education.

**Bo Li**, PhD, Professor, School of Mathematics and Statistics, Central China Normal University, Wuhan, China. Research focus: education big data, applied statistics.

# Appendix A

| Item No. | Item Description |
|---|---|
| 1 | Given the function $f(x) = 2xf'(e) + \ln x$, where $f'(e)$ is the first order derivative of $f(x)$ at point e, then $f(e) = ?$<br>A. $-e$ B. e C. $-1$ D. 1 |
| 2 | Given the function $f(x) = \frac{\cos x}{e^x}$, the tangent line of $f(x)$ at point $(0, f(0))$ is:<br>A. $x+y+1=0$ B. $x+y-1=0$ C. $x-y+1=0$ D. $x-y-1=0$ |
| 3 | If the function$y = x^3 - 3bx + 1$is decreasing in the given interval [1, 2], then what is the possible range of $b$,$b \in R$?<br>A. $b \leq 4$ B. $b<4$ C. $b\geq 4$ D. $b>4$ |
| 4 | If $x = -2$ is the extreme point of the given function $f(x) = (x^2 + ax - 1)e^{x-1}$, what is the minimum value of $f(x)$?<br>A. $-1$ B. $-2e^{-3}$ C. $5e^{-3}$ D. 1 |
| 5 | Knowing that the straight line$y = kx$ is the tangent line of $y = \ln x$, the value of $k$ is:<br>A. e B. $-e$ C. $\frac{1}{e}$ D. $-\frac{1}{e}$ |
| 6 | Assume that $P$ is a point on the curve $C$:$y = x^2 + 2x + 3$ and the range of the tangent inclination angle of curve $C$ at point $P$ is in $\left[\frac{\pi}{4}, \frac{\pi}{2}\right]$, then the value range of the abscissa of point $P$ is:<br>A. $(-\infty, \ -1/2]$ B. $[-1, \ 0]$ C. $[0, \ 1]$ D. $[-1/2, \ +\infty)$ |
| 7 | If the function$f(x) = x^2e^x - a$ has exactly three zero points, the range of the real number $a$ is:<br>A. $(\frac{4}{e^2}, \ +\infty)$ B. $(0, \ \frac{4}{e^2})$ C. $(0, \ 4e^2)$ D. $(0, \ +\infty)$ |
| 8 | If the function $f(x) = x^3 + x^2 - ax - 4$ has exactly one extreme point in the given interval$(-1, 1)$, the range of the real number $a$ is:<br>A. (1,5) B. [1,5) C. (1,5] D. $(-\infty,1)\cup(5,+\infty)$ |
| 9 | The image of the function $f(x) = \frac{e^x}{x}$ is approximately like:<br>A<br>  <br>A B C D |
| 10 | If the function $f(x) = x - \frac{1}{3}\sin 2x + a\sin x$ is monotonously increasing in the interval $(-\infty, \ +\infty)$, the range of number $a$ is:<br>A. $[-1,1]$ B. $[-1,\frac{1}{3}]$ C. $[-\frac{1}{3},\frac{1}{3}]$ D. $[-1,-\frac{1}{3}]$ |
| 11 | If the function $f(x) = \ln x + ax^2 - 2$ has monotonously increasing interval in $\left(\frac{1}{2},2\right)$, the range of the real number $a$ is:<br>A. $(-\infty,-2]$ B. $(-\frac{1}{8},+\infty)$ C. $(-2,-\frac{1}{8})$ D. $(-2,+\infty)$ |
| 12 | Given the function $f(x) = x - 1 - \ln x$,$f(x) \geq kx - 2$ holds for all $x$ in the domain, then the range of the real number $k$ is:<br>A. $\left(-\infty, \ 1-\frac{1}{e^2}\right]$ B. $\left(-\infty, \ -\frac{1}{e^2}\right]$ C. $\left[-\frac{1}{e^2}, \ +\infty\right)$ D. $\left[1-\frac{1}{e^2}, \ +\infty\right)$ |

*(continued)*

(continued)

| Item No. | Item Description |
|---|---|
| 13 | Assume $f'(x)$ is the first order derivative of the odd function $f(x)$ for $x \in R$, given that $f(-1) = 0$, and $xf'(x) - f(x) > 0$ holds when $x > 0$, then what is the range of $x$ when $f(x) > 0$ holds?<br>A. $(-\infty, -1) \cup (-1, 0)$ B. $(0, 1) \cup (1, +\infty)$<br>C. $(-\infty, -1) \cup (0, 1)$ D. $(-1, 0) \cup (1, +\infty)$ |