Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Summer 2026, Vol. 45, No. 2, e70016

Optimizing Large-Scale Mathematical Assessments: Leveraging Hierarchical Attribute Structures and Diagnostic Classification Models for Enhanced Student Diagnostics

Farshad Effatpanah and Olga Kunina-Habenicht, *Research Unit of Psychological Assessment, Faculty of Rehabilitation Sciences, Technische Universität Dortmund, Dortmund, Germany*
Steve Bernard, Caroline Hornung, and Philipp Sonnleitner, *Luxembourg Centre for Educational Testing (LUCET), Faculty of Humanities, Education, and Social Sciences (FHSE), University of Luxembourg, Luxembourg*

Abstract: *Diagnostic classification models (DCMs) assess students' mastery of cognitive attributes to provide personalized ability profiles. Retrofitting DCMs to large-scale mathematics assessments usually relies on inferred Q-matrices, which can reduce accuracy and diagnostic value. This study evaluated whether constructing items from cognitive models—yielding Q-matrices directly—and incorporating hierarchical relationships among attributes improve diagnostic outcomes. Responses from 5,336 third-grade students to a Luxembourgish image-based, large-scale standardized mathematics exam were analyzed using multiple DCMs and their hierarchical extensions. Items were constructed based on a Q-matrix, derived from the curriculum and cognitive models. The hierarchical A-CDM outperformed other models, classifying students into 60 latent classes with acceptable attribute- and test-level accuracy and more interpretable results than the G-DINA model. Using cognitive model-based item generation and Q-matrices as well as specifying attribute hierarchies enhance the accuracy and interpretability of DCM-based diagnostics in large-scale assessments, complementing traditional psychometric approaches by discerning meaningful within-score differences.*

Keywords: mathematical competence, large-scale standardized assessment, context-embedded, attributes, hierarchical structure, diagnostic classification models

Introduction

Mathematical competency is a critical skill for students' success in professional and academic contexts, especially in the areas of science, technology, engineering, and mathematics (Xu et al., 2023). Many (inter)national large-scale standardized assessments, such as the National Assessment of Educational Progress (NAEP; e.g., Johnson & Carlson, 1994), the Programme for International Student Assessment (PISA; OECD, 2022), and the Trends in International Mathematics and Science Study (TIMSS; e.g., Ferraro & Van de Kerckhove, 2006), have been developed to administer mathematics assessments and measure students' mathematical competency across different regions or countries. These assessments mostly consider mathematical competency as a unitary construct and provide a single overall score to reflect students' mathematical ability (Ribner et al., 2018). However, research in mathematics education and numerical cognition has indicated that mathematical competency is composed of multiple distinct components (e.g., LeFevre et al., 2010). As noted by Ribner et al. (2018), early models of mathematical competency in elementary grades focused on counting and magnitude comparison (e.g., Aunio & Niemivirta, 2010), whereas recent models consist of several dimensions that include, but are not limited to, simple and complex arithmetic operations, symbolic comparison, non-symbolic magnitude comparison, symbolic labeling (e.g., numeral recognition and sequencing), conceptual understanding of counting, and forward and backward counting on the number line (e.g., Cirino, 2011). Difficulty in each of these components can hinder the development of mathematical ability. Given that large-scale assessments aim to ensure educational quality and offer information about "educational contexts for learning" (Martin et al., 2004, p. 3), it is thus indispensable to accurately measure students' mathematical ability where lack of mastery or faulty strategy is identified.

Cognitive diagnostic models (CDMs; Rupp et al., 2010), also known as diagnostic classification models (DCMs), can

provide multidimensional classification-based diagnostic information about examinees' performance (Kunina-Habenicht et al., 2012). DCMs support formative assessment by providing fine-grained diagnostic feedback at both macro (the student cohort) and micro (the individual) levels (Thompson & Clark, 2024). Over the past few decades, several studies have applied DCMs to large-scale mathematics assessments, that were previously developed based on classical test theory (CTT), item response theory (IRT), and Rasch (Rasch, 1960/1980) models (e.g., Choi et al., 2015; Lee et al., 2011; Su et al., 2013; Sun et al., 2023; Wu et al., 2020; Xu et al., 2023; Yamaguchi & Okada, 2018). Previous studies have demonstrated the feasibility of using DCMs to identify examinees' strengths and weaknesses and provide diagnostic feedback, but most have relied on either add-ons to simulation studies or retrofitting to existing tests that can affect diagnostic inferences derived from examinees' cognitive skills. Although the practice of retrofitting DCMs and developing a retrofitted Q-matrix is more convenient and yields greater information about students' latent traits and their skill level, they result in information loss, less accurate modeling of cognitive attributes, and inferior data for the DCM analysis (Liu et al., 2018). The diagnostic power of diagnostic measurement can also be reduced without a principled assessment framework for writing items and constructing a test (Liu et al., 2018). This makes it difficult to specify attributes and casts doubt on the plausibility of latent classes.

Another notable research gap in retrofitting DCMs to mathematical assessments is that they disregarded the potential hierarchical relationships among mathematical attributes and failed to account for how these attributes might be related. Mathematics is hierarchical in nature, where mastering basic attributes is prerequisite for more advanced ones. More specifically, the sequential presentation of materials in schools can cause a dependency among attributes and impact examinees' test performance in large-scale assessments (Effatpanah et al., 2026). A neglected approach in large-scale mathematics assessments is the utility of hierarchical DCMs that can explicitly represent the hierarchy in resultant latent classes, making it easier to diagnose deficiencies in prerequisite attributes.

To address these gaps, this study aims to examine the feasibility of using cognitive models in developing large-scale assessment items that directly provide Q-matrices for DCM analysis, using a Luxembourgish (image-based) context-embedded large-scale standardized mathematics exam. We also apply hierarchical DCMs to investigate whether integrating theoretical assumptions about hierarchical relationships between mathematical attributes can improve its effectiveness and results.

Background

Diagnostic Classification Models (DCMs)

DCMs are confirmatory latent class models that can be used to classify examinees into different latent classes to understand the mastery or non-mastery status of several narrowly defined latent traits or attributes. These attributes include problem-solving strategies, cognitive processes, and (sub)skills required to successfully complete a set of test items or tasks (Birenbaum et al., 1993). In this article, we use the term "attribute" for the categorical latent variables for consistency.

DCMs differ from traditional psychometric models, such as CTT and IRT, in two major ways (Effatpanah et al., 2019). First, in terms of modeling, latent variables in DCMs are discrete or categorical (mastery/non-mastery), whereas ability continuum and estimates in CTT and IRT models are continuous. Second, due to their descriptive and summative nature, the traditional models assign a single total score to students along a continuous unidimensional scale. In contrast, DCMs offer multidimensional attribute profiles by classifying students as non-masters or masters of each attribute measured by the test. Such profiles allow researchers and educators to better identify students' learning status. Students with the same total score are likely to be different in the mastery/non-mastery of several attributes.

Numerous DCMs have been developed, each based on different condensation rules about the way attributes interact and impact item performance. DCMs can be categorized as either specific or general. Specific DCMs merely reflect one type of relationship within a test that can be further classified as compensatory, non-compensatory, or additive. As a non-compensatory model, the Deterministic Inputs, Noisy "And" Gate (DINA; Junker & Sijtsma, 2001) assumes that all required attributes are essential for getting an item right. However, the Deterministic Inputs, Noisy "Or" Gate (DINO; Templin & Henson, 2006) model is a compensatory DCM, assuming that examinees should master at least one required attribute to correctly answer an item. The compensatory reparameterized unified model (C-RUM; Hartz, 2002), linear logistic model (LLM; Maris, 1999), reduced reparameterized unified model (RRUM; Hartz, 2002), and additive CDM (*A*-CDM; de la Torre, 2011) are additive models[1].

On the contrary, general DCMs subsume all specific DCMs as special cases. The examples are the general diagnostic model (GDM; von Davier, 2008), the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), and the generalized DINA (G-DINA; de la Torre, 2011). Based on different link functions, general DCMs are equivalent in their saturated forms, inducing identical model data fit (de la Torre, 2011).

As a most widely general DCM, the saturated G-DINA model takes into account all possible main and interaction effects. It classifies examinees into $2^{K_j^*}$ classes, where K_j^* is the number of required attributes for item j. Each class has its own probability of getting an item right. For the G-DINA model, the probability that an examinee correctly answers item j, which requires two attributes a_1 and a_2 is defined as:

$$P(X_j = 1 | a_1,\ a_2) = \delta_{j0} + \delta_{j1}a_1 + \delta_{j2}a_2 + \delta_{j12}a_1\, a_2 \qquad (1)$$

In Equation (1), δ_{j0} is the intercept for item j (i.e., the probability of a correct response when none of the required attributes has been mastered); $\delta_{j1}a_1$ and $\delta_{j2}a_2$ are two main effects, showing the change in the success probability as a result of the mastery of these two attributes; and $\delta_{j12}a_1\, a_2$ are the interaction effect between the two attributes (a_1 and a_2). For a comprehensive explanation of the response probability formulation in the G-DINA model for items involving several attributes, interested readers are referred to de la Torre (2011).

Several reduced (or specific) DCMs can be derived from the G-DINA model by imposing specific constraints on the pa-

rameters of the model. As shown by de la Torre (2011), by setting both the main and interaction effects to zero, the DINA and DINO models can be derived from the G-DINA model. By removing all interaction effects in the identity, log, and logit link versions of the G-DINA model, the *A*-CDM, RRUM, and C-RUM can be obtained.

Hierarchical DCMs are restricted versions of DCMs. They impose structural assumptions to represent developmental or instructional dependencies among attributes, referred to as attribute hierarchies. In doing so, these models formalize assumptions about how multiple attributes involved in completing a task or an item are interrelated, specifying directional or prerequisite relationships among them. One prominent example is the Hierarchical Diagnostic Classification Model (HDCM; Templin & Bradshaw, 2014) that operationalizes prerequisite relations by limiting the permissible latent classes and imposing corresponding parameter restrictions. The sequential development of attributes, as defined by educational curricula, may establish dependencies between different components of a knowledge domain that may shape how knowledge components are organized cognitively and, consequently, impact examinees' item responses. Hierarchical DCMs also have the potential to identify the inherent latent hierarchical structures among attributes of a cognitive domain. By modeling dependencies among attributes, they provide deeper insight into how knowledge components are organized and interconnected, improving researchers' understanding of the structure of the domain itself (Effatpanah et al., 2026).

Hierarchical Diagnostic Classification Model (HDCM)

In educational contexts, students' item responses may be affected by the sequence in which instructional materials are presented. To capture such possible sequential dependencies among attributes in a particular cognitive domain (e.g., mathematics), researchers have developed several hierarchical DCMs (e.g., Chen & Wang, 2023; Kwon et al., 2024; Ma et al., 2023; Templin & Bradshaw, 2014; Tu et al., 2019). The most commonly used model is the HDCM developed by Templin and Bradshaw (2014). The model extends conventional DCMs by considering attribute dependencies through an attribute hierarchy. In addition to maintaining the flexible framework of general DCMs such as the G-DINA model, it uses inferential statistics to test hypotheses about the postulated hierarchies. For an item involving two attributes within the HDCM, where the mastery of Attribute 2 (a_2) is dependent on the mastery of Attribute 1 (a_1), the reformulation of the G-DINA can represent this hierarchical structure:

$$P(X_j = 1 | a_1,\ a_2) = \delta_{j0} + \delta_{j1} a_1 + \delta_{j2(1)} a_1\, a_2 \qquad (2)$$

In Equation (2), $\delta_{j2(1)} a_1\, a_2$ indicates an interaction for Attribute 2 nested within Attribute 1. As it is evident, there is not any main effect for Attribute 2 due to the constraints imposed by the attribute hierarchy. A main effect for an attribute denotes an increase in the probability of giving a correct answer to a given test item when other attributes have not been mastered. Nonetheless, when the mastery of Attribute 2 requires prior mastery of Attribute 1, the main effect for Attribute 2 must be set to zero (Effatpanah et al., 2026).

The HDCM restricts the attribute space by eliminating inadmissible attribute profiles (latent classes), thereby reducing the number of parameters that must be estimated for each item (Templin & Bradshaw, 2014). This constraint enhances model efficiency because latent classes that are theoretically impossible or absent in the sample are excluded from estimation due to the dependency. For example, consider a test measuring four attributes. A conventional DCM would classify examinees into $2^4 = 16$ latent classes. However, if there is a linear dependency, where the mastery of Attribute 4 relies on the mastery of Attribute 3, Attribute 3 depends on Attribute 2, and Attribute 2 is contingent on Attribute 1, then there will be A+1 (i.e., $4 + 1 = 5$) sound latent profiles. The superiority of HDCMs becomes more important when a large number of attributes are involved in a test, especially in large-scale assessments. For example, in a test with ten attributes, there would be $2^{10} = 1{,}024$ latent classes for conventional DCMs, whereas the HDCM under a linear hierarchy would produce 11 latent classes.

As illustrated in Figure 1, Leighton et al. (2004) detected several types of hierarchical structures: linear, divergent, convergent, independent, mixed, unstructured, and higher-order. In a *linear* structure, all attributes must be mastered sequentially; Attribute 1 is a prerequisite for Attribute 2, and both Attributes 1 and 2 are prerequisites for Attribute 3. When Attribute 1 is not mastered, the other attributes should be absent. A *divergent* structure has several end states coming from a single initial attribute. Although attributes are distinct, there exist correlations in their mastery indicators (Tu et al., 2019). A *convergent* structure indicates that different attributes may develop simultaneously and have equal contribution to other attributes. In this structure, attributes do not require strict sequential mastery. An *independent* structure shows that several attributes are separate from one another. A *mixed* structure reflects the presence of two independent structures, where one set of attributes has a specific structure, the other follows a different one. The other structure in which Attribute 1 is a prerequisite for all other attributes is called *unstructured*. Unlike the divergent structure, in which multiple attributes emerge from a common origin but remain distinct and possibly correlated, the unstructured form shows that without the initial attribute, none of the others can be mastered—emphasizing a strict dependency across all links. And finally, a *higher-order* structure shows that specific attributes (A1 through A6) are related and jointly represent a general competency, denoted as "*G*." The general factor *G* is usually modeled as a continuous latent variable that impacts the probability of mastering the individual binary attributes, thus introducing a dimensional structure above the attribute level (Effatpanah et al., 2026).

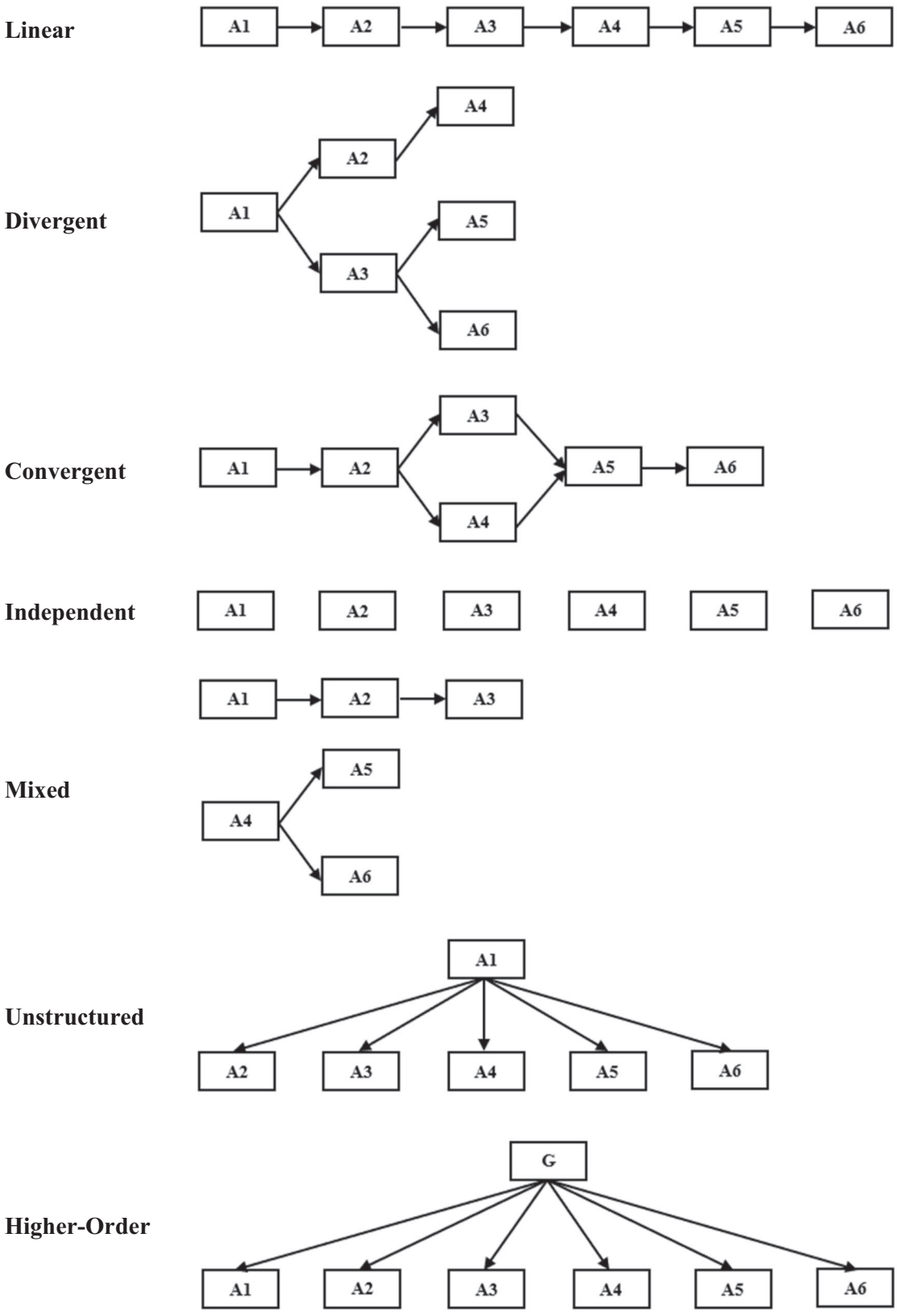
Q-Matrix in DCMs and Large-Scale Assessments

In DCMs, the relationship between attributes and test items is pre-specified in a binary incidence matrix, called Q-matrix (Tatsuoka, 1983). In a Q-matrix, each column corresponds to an attribute, and each row indicates an item. If an item requires its mastery, a value of 1 is inserted; otherwise, a value of 0 is used in the matrix. Using an analogy from confirmatory factor analysis, a Q-matrix is the loading structure of a DCM that hypothesizes item-by-attribute associations, anchored in substantive theory.

It has been widely acknowledged that the Q-matrix is the key factor in determining the quality of cognitive diagnos-

FIGURE 1

Various Types of Attribute Hierarchies



tic information because the diagnostic power of DCMs relies heavily on the theoretical and empirical soundness of the Q-matrix (Lee & Sawaki, 2009). Correct identification of attributes underlying test performance, their numbers, and their associations with test items has been shown to improve the quality of diagnostic information obtained from DCMs (e.g., Kunina-Habenicht et al., 2012; Rupp & Templin, 2008). To ensure the accurate development of a Q-matrix, several methods have been suggested for extracting or identifying relevant attributes. The methods include consulting with domain experts, reviewing existing literature, analyzing dimensionality, utilizing think-aloud protocols, analyzing test content and test specifications, using content domain theories, and conducting eye-tracking research (Leighton et al., 2004). By having a rational cognitive model of task performance, appropriate items can be constructed to measure to what extent students possess expected attributes.

The Q-matrix construction is typically a subjective process. This subjectivity may lead to the presence of misspecifications in the Q-matrix that have negative impacts on the estimation of model parameters, misclassification of examinees, and differential item functioning which may lead to inaccurate inferences (Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016; de la Torre et al., 2022; Kunina-Habenicht et al., 2012). Misspecification refers to adding an unnecessary attribute (overfitting) or omitting a necessary attribute (underfitting) in the Q-matrix (Rupp & Templin, 2008). Consequently, to address this issue, a number of Q-matrix validation methods have been developed to (empirically) identify and modify misspecifications in Q-matrices (e.g., Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016; Li & Chen, 2025; Ma & de la Torre, 2020; just name a few).

These issues are specifically prominent in the context of large-scale assessments, where tests are not typically designed for diagnostic purposes and Q-matrices are constructed usually post hoc. In such settings, developing a theoretically grounded and empirically validated Q-matrix is essential for ensuring that diagnostic interpretations are substantively meaningful and psychometrically sound. Studies have illustrated how this can be achieved in practice. For example, Tjoe and de la Torre (2014) systematically identified and validated cognitive attributes underlying eighth graders' proportional reasoning for cognitive diagnostic assessment (CDA). They first derived attributes from literature and expert input, then validated them through teacher-aligned items and student think-aloud interviews. The study showed how theoretically meaningful and instructionally grounded attributes can be systematically developed. Tatsuoka et al. (2016) argued that defining cognitive attributes for DCMs is an iterative process that must integrate cognitive theory, item design, and psychometric modeling. Using examples from elementary mathematics assessments, they showed how statistical feedback informs revisions to attributes and items.

Previous DCM Studies on Math Assessments

Numerous studies have applied DCMs to mathematics assessments to diagnose students' mathematical abilities. Two lines of research can be identified in the relevant literature. The first line of research involves studies that aimed to develop true diagnostic mathematics tests and checklists. For example, Li et al. (2020) used DCMs to develop a mathematics problem-solving assessment for kindergarteners and found the instrument to be sufficiently reliable in the domain of number and operations. Li and Traynor (2022) created a computational thinking assessment for middle school students. After creating a cognitive model and multiple-choice items, they used the DINA model to classify students' mastery. They suggested that these diagnostic results can help teachers understand typical learning progressions in computational thinking. Recently, Haghayeghi et al. (2024) developed an assessment tool for first-grade students' mathematical abilities using DCMs. After comparing the fit of G-DINA, DINA, and DINO, they found that concepts related to time, multiplication, symmetry, and geometry were the most challenging for students, whereas cardinality, addition, subtraction, weight, and statistics were the easiest.

The second line of research consists of studies applying DCMs to retrofit (inter)national large-scale mathematics exams, such as PISA, TIMSS, and NAEP. These studies share a number of features that distinguish them from the first line and therefore warrant a focused summary. A table summarizing the results of previous studies on the application of DCMs to large-scale mathematics assessments is available in Appendix A. Typical characteristics of retrofit studies are that they: (a) work with publicly available item pools or released items that were not originally designed as diagnostic tests, (b) develop Q-matrices post-hoc using expert judgment, curriculum frameworks, or a combination of expert and data-driven methods, (c) deal with sparse booklet designs and complex sampling (e.g., rotated booklets, country subsamples), and (d) usually emphasize cross-national or large sample comparisons rather than test construction. For example, Lee et al. (2011) used the DINA model to analyze booklets 4 and 5 of TIMSS 2007 Fourth-Grade mathematics assessment. They showed that DCMs can provide fine-grained, attribute-level information useful for classroom instruction. Su et al. (2013) adapted Common Core State Standards attributes to construct Q-matrices for two TIMSS 2003 Eighth-Grade mathematics booklets and developed hierarchical DINO and DINA (i.e., the DINO-H and DINA-H) models. They found that these hierarchical models performed well for items with smaller samples and hierarchically structured attributes. Similarly, Choi et al. (2015) applied the DINA model to the TIMSS 2003 Eighth-Grade math assessment to provide diagnostic insights. Results revealed that the DINA discrimination index highlights comparative differences across countries. Kunina-Habenicht et al. (2017) examined the use of a LCDM for a German diagnostic arithmetic test, comparing its multidimensional scores to those from a unidimensional IRT model. They investigated whether LCDM scores better predicted students' math grades than unidimensional scores from a large-scale assessment. Results showed interpretable item and student parameters and high classification reliability, but the multidimensional scores did not significantly improve outcome prediction over the unidimensional model. In a multi-DCM study (i.e., a study in which several DCMs are compared), Yamaguchi and Okada (2018) compared several DCMs (DINA, DINO, *A*-CDM, LLM, RRUM) and IRT models (i.e., two-parameter logistic (2PL) IRT and three-parameter logistic (3PL) IRT) using TIMSS 2007 Fourth-Grade mathematics data from seven countries. DCMs fit the data better than IRT models, with additive DCMs capturing attribute interactions most effectively.

Along the same lines, several studies have focused on defining and refining cognitive models and Q-matrices. For instance, Terzi and Sen (2019) applied DCMs to TIMSS 2011 Eighth-Grade mathematics assessment. They validated item–attribute links, identified appropriate DCMs for individual items, and examined cross country generalizability. Using the G-DINA model discrimination index (GDI; de la Torre, 2008), they found that validated attribute specifications differed from expert assumptions and that different items fit different models. The study emphasized the importance of item-level Q-matrix validation and model selection. Delafontaine et al. (2022) examined Q-matrix refinement and the need for country-specific Q-matrices using TIMSS 2011 Eighth-Grade mathematics data from five countries. They compared the original expert-designed Q-matrix with two refined versions based on a stepwise validation method (Ma & de la Torre, 2020) and a nonparametric classification method (Chiu, 2013). Both refinements improved model fit, with the stepwise approach performing best and producing meaningful changes in parameters and mastery profiles. The study showed that country-specific refined Q-matrices better captured students' cognitive structures than a single universal Q-matrix. Wu et al. (2020) developed a Q-matrix for 12 PISA 2012 math items and compared student performance across ten countries. They extracted eleven attributes, found LLM provided the best model fit, and revealed different learning trajectories among countries. In another study, Xu et al. (2023) developed a seven-attribute mathematical competency model for TIMSS 2011 Fourth Grade based on the Chinese curriculum and used the DINA model to assess students. Results showed high mastery rate in mathematical cognition and showed DCMs' ability to differentiate students with identical total scores. Using 16 TIMSS 2015 cognitive attributes and the interpretive structural modeling (ISM) method, Sun et al. (2023) further developed a five-level mathematical cognitive model, refined through expert input, ranging from "memorize" to "justify." They argued that it can support adaptive systems and accurately diagnose students' learning and cognitive development. Zhu (2023) also analyzed 21 items from TIMSS 2019 Fourth-Grade assessment across 17 countries using DCMs. Results showed both differences in learning trajectories and commonalities that can inform teaching and cross-country comparisons in mathematics education. Saso et al. (2024, Preprint) examined four approaches for specifying Q-matrices to diagnose students' depth of understanding in Eighth-Grade Japanese mathematics. Using variational Bayesian estimation with G-DINA, the polytomous G-DINA (pG-DINA; Chen & de la Torre, 2013), and the attribute hierarchy method (AHM; Leighton et al., 2004), they found that a linear-hierarchy Q-matrix better captured attribute interactions than polytomous or binary Q-matrices, highlighting the importance of hierarchical relationships among attributes. Zhang et al. (2025) used PISA 2012 math data from Shanghai, Hong Kong, Macau, and Taiwan to identify cognitive attributes and assess students' competencies. DCM results revealed regional differences and offered insights for tailoring curricula and instruction. Recently, Sonnleitner et al. (2026) investigated whether cognitive models originally designed for item generation can be directly used for constructing Q-matrices for DCMs in a large-scale assessment of first graders' math abilities. Using data from Luxembourg's school monitoring program, four developmentally grounded attributes (i.e., counting, addition < 10, addition > 10, and decomposition) were mapped into Q-matrices without post hoc refinement. Items with low fit were dropped entirely. Two DCM approaches were evaluated: a Single-Attribute Hierarchical Model (SAHM), reflecting a strict sequential progression of attributes, and a Multiple-Attribute Hierarchical Model (MAHM), allowing for interaction among attributes. Both models successfully reproduced expected developmental sequences, and decomposition was a pivotal threshold attribute. They argued that cognitive models can effectively shape Q-matrix construction and more precisely reflect learning trajectories.

Although previous research has demonstrated that DCMs can effectively diagnose examinees' mathematical abilities, they have two major limitations. First, except for Sonnleitner et al. (2026), most studies retrofitted onto existing assessments designed under CTT or IRT frameworks. Researchers have argued that retrofitting can be beneficial for determining the diagnostic potential of existing achievement and proficiency tests before investing significant time and resources into developing fully diagnostic tests (Lee & Sawaki, 2009). However, the process of calibrating an existing unidimensional test with a multidimensional DCM may not be efficient and directly contradicts the inferences drawn from the original assessment (Liu et al., 2018). As noted by Liu et al. (2018), in unidimensional IRT, retrofitting requires practitioners to detect and remove multidimensionality from assessments. Several researchers have shown that test construction procedures mainly focus on maximizing total score reliability (e.g., Sinharay, 2014). Consequently, test developers usually choose items with high discrimination values and exclude items that could induce multidimensionality. Additionally, because most tests lack a clear cognitive model during construction, an inferred model is used to develop a Q-matrix for DCMs. However, this inferred model may not fully reflect the cognitive processes underlying performance, leading to inaccurate Q-matrices, implausible latent classes, and compromised diagnostic validity (DiBello et al., 2007). Consequently, retrofitting is severely limited and produces a tenuous fit between the cognitive model and test data (Gierl & Cui, 2008). It is thus important to establish a well-grounded theoretical Q-matrix from the outset when items are constructed.

Second, the studies overlooked the possible hierarchical interactions among mathematical attributes, failing to explain how these attributes may interact or depend on one another. A possible explanation for this oversight can be ascribed to the lack of suitable models that allow testing such hierarchies. Hierarchical DCMs can help researchers rigorously explore and validate these relationships.

The Present Study

The purpose of the present study is, first, to explore whether cognitive models used in item generation can directly inform Q-matrix construction for DCM analyses. Second, it seeks to examine whether incorporating theoretical assumptions about attribute hierarchies into the DCM can enhance its results. Moreover, it investigates to what extent results derived from DCMs provide any additional value beyond traditional approaches (i.e., CTT and IRT). The following research questions were posed:

RQ 1. *Can a Q-matrix resulting from theory-based item generation be effectively applied in DCMs?*

RQ 2. *Does incorporating theoretical assumptions about attribute hierarchies enhance the performance of DCMs?*

RQ 3. *Are results from DCMs in line with traditional types of analyses and do they offer added value?*

To the best of the authors' knowledge, the application of model-based DCMs to context embedded math assessments remains largely underexplored, though recent work by Sonnleitner et al. (2026) has demonstrated this approach using First-Grade data from the Luxembourgish school monitoring system. Therefore, the present study is one of the first examples of empirically applying (non)hierarchical DCMs to an image-based (context embedded) large-scale mathematics exam. Context-embedded tests, which incorporate real world scenarios and practical applications of knowledge, could greatly benefit from the detailed diagnostic feedback that DCMs offer. Results of the study will build on and extend previous research on the application of DCMs to mathematical assessments.

Method

Test Construction, Q-Matrix Development, and Validation

In large-scale assessments, mathematical ability can be conceptualized in two ways: a literacy-oriented or a curriculum-oriented approach (Saß et al., 2017). Literacy-oriented tests (e.g., PISA) assess abilities needed in modern society outside of the educational context, covering areas like quantity, change, space, and uncertainty (OECD, 2022). Curriculum-based tests focus on content and abilities from national curricula and taught in educational contexts, such as numbers and algebra, geometry, measurement, probability, and statistics, that are viewed as basic abilities for mathematical problem-solving and logical reasoning (NCTM, 2000). Despite these differences, the cognitive processes required to solve mathematics items are similar across both approaches (Saß et al., 2017).

The present study is based on item development for the Luxembourgish national school monitoring program (Épreuves Standardisées; ÉpStan) in mathematics, grounded in cognitive models which are empirically validated. For a detailed explanation of cognitive item model development, see Sonnleitner et al. (2025). Specifically, for third grade, eight cognitive item models were developed to define content and contextual features of items assessing specific curricular sub-competencies—for example, "the student is able to read and write numbers from 0 to 100, as well as compare and sort them," "the student can mentally calculate additions and subtractions within the range of 0 to 100," or "the student differentiates between even and odd numbers." These models were implemented in R software (R Core Team, 2025), which allowed the systematic generation of mathematics items with predefined attribute structures, yielding Q-matrices directly applicable in DCM analyses. The R code enabled specification of attributes for each item (e.g., number range, decade crossing), control over the number of item instances sharing identical attributes, and export of items in PDF format for paper-based administration within ÉpStan (cf. Sonnleitner et al., 2025). The models align with the Luxembourgish elementary school curriculum and are informed by established developmental theories of numeracy (e.g., Clements & Sarama, 2007, 2012; Geary, 2006; Ginsburg et al., 2008; MENFP, 2011; von Aster & Shalev, 2007). The schematic flowchart of cognitive item model development and implementation in R (Sonnleitner, 2025) is available in Supplementary Material A.

To ensure theoretical and practical validity, teachers from the relevant grade levels contributed their classroom expertise to refine the models, reflecting instructional realities such as common problem-solving strategies and item difficulty. After item generation, psychologists who were trained and were working at the intersection of mathematical cognition and test development within ÉpStan reviewed the resulting Q-matrices. Their evaluation considered theoretical aspects, including the fit between task characteristics and attribute structure, the number of attributes specified for each item, and potential student problem-solving strategies. Practical considerations were also assessed, such as the distribution of items per attribute and the identification of items exhibiting construct-irrelevant variance in previous analyses that could bias DCM results (cf. Sonnleitner et al., 2025). Notably, due to the highly diverse Luxembourgish setting, math items are mostly presented in image-based real-world problem settings, reducing the impact of students' language background. It must be noted that similar to other large-scale assessment development procedures, the Rasch model (Rasch, 1960/1980) was used during the test development procedure to examine item quality, dimensionality, and discrimination parameters. These analyses ensured that items functioned appropriately across the ability continuum and measured the intended constructs consistently. Items showing poor fit and low discrimination were removed prior to the diagnostic modeling phase. This preliminary IRT-based screening helped establish a psychometrically sound item pool for subsequent hierarchical cognitive diagnostic modeling.

Relevant cognitive attributes were identified based on several sources, including the Luxembourgish mathematics curriculum and prominent developmental models of mathematical ability (Clements & Sarama, 2007, 2012; Geary, 2006; Ginsburg et al., 2008; MENFP, 2011; Sonnleitner et al., 2025; von Aster & Shalev, 2007). As one of the consulted models, Von Aster and Shalev (2007) proposed a hierarchical four-step developmental model of cognitive number representation, predicting different pathways of pathological development. The model starts with an inherited core system representation of numerical magnitude (cardinality; step 1), which underlies understanding number meaning. This supports linking quantities to spoken/written words (step 2) and Arabic symbols (step 3), which in turn enables the development of a mental number line (step 4) where ordinality serves as a second core basis of number. Impairment at any step can hinder subsequent development.

Another influential model consulted for the Q-matrix and test development was Geary's (2006) framework. Geary highlights the role of innate numerical capacities, such as the Approximate Number System (ANS). The system allows humans and children to evaluate and compare quantities autonomously from language and formal education. Geary describes a predictable progression from basic to sophisticated mathematical understanding: children first learn to count by rote, then grasp stable order, one-to-one correspondence,

and cardinality, followed by basic arithmetic, multiplication and division, fractions and decimals, and eventually abstract reasoning such as algebra and geometry. He also emphasizes individual differences influenced by genetic and environmental factors, such as socio-economic background, parental involvement, and instruction quality.

Based on the underlying cognitive models that were used to generate the final item pool administered in the third grade, a set of eight attributes was identified: (1) Addition Simple (ADS) defined as the ability to carry out basic addition operations, that involves small numbers, that is, single-digit or low double-digit (Clements & Sarama, 2007; Geary, 2006); (2) Subtraction Simple (SUS) referring to the ability to carry out subtraction operations, that involves small numbers, that is, single-digit or low double-digit (Clements & Sarama, 2007; Geary, 2006); (3) Addition Complex and Subtraction Complex (ASC) denoting the ability to perform addition and subtraction operations that require carrying over and borrowing as well as multi-digit numbers, where students need to understand place value and adopt regrouping strategies (Artemenko et al., 2018; Clements & Sarama, 2007). Research has indicated that although, from Grades 3 to 4, children generally become more proficient in two-digit subtraction, the borrow operation increases the difficulty in two-digit subtraction, suggesting that it does not develop more rapidly than other aspects of subtraction (Artemenko et al., 2018); (4) Multiplication/Division (MUD) indicating the ability to grasp and perform the concept of multiplication and division, that involves the understanding of proportional relationships (i.e., multiplication as repeated addition and division as repeated subtraction or partitioning) (Clements & Sarama, 2007; Geary, 2006); (5) Modeling (MOD) referring to the ability to apply mathematical knowledge in real-world contexts that requires representing relationships between numbers, abstract thinking, and problem-solving skills (Clements & Sarama, 2007; Geary, 2006); (6) Units of Measurement (UOM) representing the ability to understand and use different measurement units (e.g., time, weight, length) to quantify and compare physical properties (Clements & Sarama, 2007; Geary, 2006); (7) Sequencing and Number Patterns (SNP) showing the ability to recognize and create numerical patterns or sequences; and (8) Dealing with Higher Number Range (HNR) referring to the ability to work with larger numbers, usually beyond two-digit operations (Clements & Sarama, 2007; Geary, 2006). This attribute encompasses understanding place value and using number sense to solve problems. For the test, the scope of numbers ranged from 1 to 20, up to 100, up to 1000, and up to 1 million.

A vector of 0–1 was used to show the relationship between a test item and the required attributes. If an attribute is required to correctly answer an item, it is represented as "1"; otherwise, it is represented as "0." Each attribute pattern can involve one or multiple attribute(s). For instance, the attribute pattern for Item 1 is [10001000], indicating that the item measures Attributes 1 and 5 (i.e., ADS and MOD). To correctly answer the item, students should have mastered the two attributes.

The developed Q-matrix was then empirically validated using the stepwise Wald test method (Ma & de la Torre, 2020) in the *GDINA* R-package version 2.9.12 (Ma et al., 2025). This method provided several modification suggestions. Each suggestion was carefully reviewed by the authors and content experts to ensure theoretical soundness. Most suggested modifications were not conceptually justifiable and were thus discarded. In particular, the proposed deletions of attributes from some items were substantively irrational. By contrast, few suggestions for adding attributes were considered theoretically plausible. For example, it was suggested that the attribute HNR should be added to Item 116. After inspecting the item content and consulting with content experts, this modification was applied to the item and its equivalents across the test booklets. Similarly, MOD was added to Item 110 and its corresponding items. Mesa plots (Ma, 2019) were also examined to evaluate whether refining the q-vectors could improve model fit and the accuracy for items. Overall, the empirical validation procedure supported the adequacy of the Q-matrix, providing evidence for its theoretical soundness. The final Q-matrix is available in Supplementary Material B.

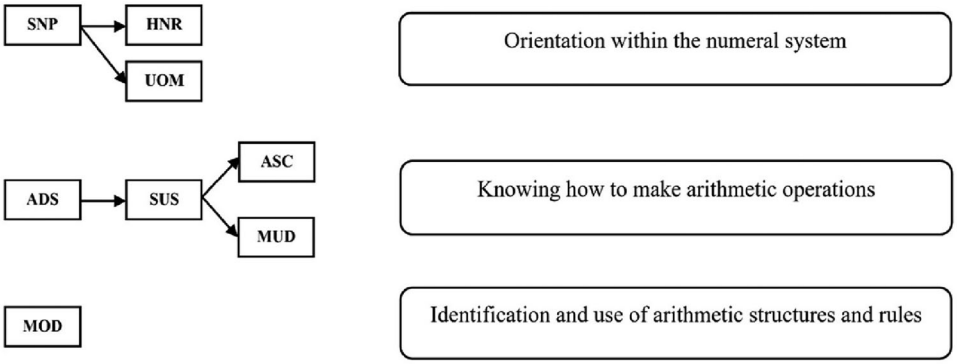
Specifying Attribute Hierarchies

To show how mathematical attributes are interconnected and built upon one another, a hierarchical structure was developed that is aligned with cognitive development models of mathematical competency (e.g., Clements & Sarama, 2007, 2012; Geary, 2006; Ginsburg et al., 2008) and the Luxembourgish mathematics curriculum the test is based on (MENFP, 2011). For example, according to Clements and Sarama (2012), the development of mathematical thinking in children is a gradual process that relies on prior experiences and knowledge. The process begins with understanding basic concepts (e.g., counting, comparing quantities, and recognizing patterns). Then, they start developing an understanding of spatial relationships, measurement, and the concept of classifying and sorting, as well as working with higher number range. Next, they begin to grasp more abstract concepts and thinking enabling them to develop mathematical concepts (e.g., addition, subtraction, multiplication, and division). After that, as children's understanding grows, they will be able to apply their mathematical knowledge in practical or real-life contexts.

Based on developmental models of mathematical ability, experts in mathematical cognition involved in the test development for Luxembourg's school monitoring program reviewed the resulting Q-matrix from the item generation process. Their evaluation, grounded in both theoretical and practical considerations, ensured that the sequential structure of basic arithmetic attributes accurately reflected the cognitive development of third-grade students in mathematics and thus imposed an attribute hierarchy. As can be seen in Figure 2, the model represents three curricular competency strands: (1) orientation within the numeral system, (2) knowing how to make arithmetic operations, and (3) identification and use of arithmetic structures and rules. At the first strand, it is acknowledged that the mastery of SNP is an essential requirement for the mastery of HNR and UOM because students need to identify and understand numerical patterns before learning and working with higher numbers, more complex calculations, and measurement concepts. At the second strand, ADS is a prerequisite for SUS because subtraction is usually taught in connection with addition, and they are conceptually related. The mastery of the two attributes serves as a foundation for the mastery of ASC and MUD, as higher-level attributes. At the third strand, MOD is conceived

FIGURE 2

Hierarchical Structure for the Mathematics Exam



as an independent attribute that may affect different areas, although it is not directly connected to other attributes.

Participants and Materials

The mathematics assessment was administered to the entire Luxembourgish cohort of 5,528 third-grade students using a matrix sampling design. The test consisted of eight distinct booklets. Each student completed two core booklets, which were administered to all participants and served as anchor forms to psychometrically link the entire set of booklets. In addition to these, each student completed one of six supplementary booklets, which were randomly assigned at the class level. Students ($n = 192$) who did not complete both two main booklets were removed from the dataset. In total, a sample of $N = 5{,}336$ was selected for this study. There were 2,613 (49%) boys and 2,716 (50.9%) girls, and 7 (0.1%) unspecified. The individual items were scored dichotomously in that a response was either correct or incorrect. Students also reported their immigration status and socio-economic status (SES). Of the total sample, 816 (15.3%) were first-generation students, 1,705 (32%) second-generation, 2,166 (40.6%) native, and 649 (12.1%) unspecified. SES was also indexed by the International Socio-Economic Index of Occupational Status (ISEI; Ganzeboom, 2010; Ganzeboom et al., 1992) that is based on the level of education, occupation, and income. Its score ranges from 16 to 90. Based on the score range in our dataset, students' SES were categorized into three groups: low (10–29) with 517 students (9.7%), medium (30–45) with 1,657 students (31.1%), and high (46–70) with 2,445 students (45.8%). Also, 717 (13.45%) did not report their SES. Total scores across the booklets ranged from 0 to 60 with a mean of 35.36 and a standard deviation of 10.81. The reliability (Cronbach alpha) of the test was 0.891.

Data Analysis

After developing the Q-matrix and specifying the hierarchical structure, we compared the fit of the G-DINA against several specific DCMs (i.e., DINA, DINO, *A*-CDM, C-RUM, and RRUM) and their hierarchical extensions (i.e., HDCM_DINA, HDCM_DINO, HDCM_*A*-CDM, HDCM_C-RUM, and HDCM_RRUM). The incorporation of the hierarchical extensions of the specific DCMs allows the consideration of hierarchical structure and reduced forms of the measurement models at the same time. The models were compared with regard to relative and absolute fit statistics, including −2log-Likelihood (−2LL), Akaike's Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), the Mean Absolute Difference for the item-pair Correlations (MADcor; DiBello et al., 2007), and the Standardized Root Mean Square Residual (SRMSR; Maydeu-Olivares, 2013). The model with the least information criteria is the best model. The SRMSR is calculated as the square root of the sum of squared differences between observed correlations and model-implied correlations across all item pairs. Hu and Bentler (1999) consider a value < 0.08 as indicative of a substantively negligible amount of misfit. Using the marginal maximum likelihood estimation (MMLE) method and the Expectation Maximization (EM) algorithm, the *CDM* R-package (Robitzsch et al., 2025) was used to fit the models and estimate item parameters. The monotonicity constraint was imposed when fitting the DCMs to ensure that possessing an additional attribute would not decrease the probability of endorsing the item. We also used multiple sets of starting values for all the models to ensure that the −2LL converged to the same value across runs and to minimize the risk of local maxima.

In addition to the above-mentioned indices, several item-level fit statistics were evaluated (Oliveri & von Davier, 2011; von Davier, 2005): (1) Root Mean Square Deviation (RMSD) and its bias-corrected, (2) Mean Deviation (MD), (3) Mean Absolute Deviation (MAD), and (4) the average of Root Mean Square Error of pproximation (RMSEA). The RMSD measures the average squared difference between observed and model-predicted item response probabilities. The bias-corrected RMSD adjusts for sampling error to provide a more accurate estimate of model misfit. The MD shows the average signed difference between observed and expected probabilities, indicating systematic bias in model predictions. The MAD captures the average magnitude of the differences regardless of direction. The mean of RMSEA summarizes average model misfit across items. It is the mean difference between response proportions predicted by the model and those observed for each response category within each latent class weighted by the proportion of the students within the respective latent class. A model with smaller values indicates better fit.

A set of model comparisons was performed to evaluate whether the hypothesized hierarchical structure offered a

more parsimonious but sufficiently accurate fit to the data relative to the fully saturated model. The G-DINA model was the baseline, and a hierarchical DCM showing the best performance represented the restricted alternative. As argued by Templin and Bradshaw (2014), conventional deviance difference tests assume a chi-square distribution with degrees of freedom equal to the difference in model parameters. However, these tests depend on regularity conditions—specifically, that parameters are not on the boundary of the parameter space. In hierarchical DCMs, some parameters are fixed at zero, violating these assumptions (Stoel et al., 2006). As a result, the deviance statistic G follows a mixture of chi-square distributions rather than a standard chi-square, which can lead to under-rejection of the null hypothesis and over-acceptance of hierarchical models. For that reason, Templin and Bradshaw (2014) proposed either deriving the appropriate mixture distribution analytically or estimating correct p-values through simulation. Following their recommendation and similar to Effatpanah et al. (2026), a simulation-based likelihood ratio test was conducted. Both models were first fitted to the observed data ($N = 5{,}336$) to obtain the observed deviance difference. Next, 2,000 datasets were simulated under the null model (the selected HDCM), each with 5,336 cases. For each dataset, both models were re estimated and deviance differences recalculated, producing an empirical distribution of the test statistic. This approach enabled a direct, data-driven comparison of the hierarchical and saturated G-DINA models without relying on asymptotic chi-square assumptions. Absolute (deviance) and relative (AIC, BIC) fit indices were compared across replications.

Classification accuracy and consistency at test- and attribute-level were further examined across the two models. Classification accuracy indicates how well students' classifications agree with their true latent classes. Classification consistency refers to the degree to which a student is consistently classified into the same latent class (or will be indicated as master/non-master of the same attribute) on re-administration of the same or a parallel form of the test. There is not any clear-cut benchmark for classification accuracy and consistency in the DCM literature. However, a test-level accuracy/consistency of 0.7 and an attribute-level accuracy/consistency of 0.8 are generally deemed acceptable (Johnson & Sinharay, 2018). It should be noted that classification accuracies and consistencies should not be used for model comparison, because higher values do not necessarily imply a better-fitting or more valid model. Rather, these indices should be interpreted within the theoretical and statistical framework of each model.

The attribute mastery profiles of students were also compared across the G-DINA and the selected HDCM. A set of demographic variables, such as age, SES, and immigration status, were used to characterize most prevalent latent classes across the selected model.

The attribute prevalence and individual-level mastery profiles were examined. The attribute prevalence shows students' mastery probability of each attribute. Attribute mastery profiles indicate different latent classes into which students are categorized. Probabilities close to 0 or 1 reflects the non-mastery or mastery of an attribute, respectively. As a rule of thumb, students are classified as masters of an attribute when their posterior probability is 0.5 or higher; otherwise, they are classified as non-masters (indicated by 0) if the probability is below 0.5. The mastery probability for the attributes was compared between selected students with the same and different total scores.

Furthermore, we analyzed item discrimination indices to investigate the ability of the test items to differentiate between students based on their mastery of the requisite attributes. A high discrimination index indicates that the item can effectively distinguish between students who have mastered the attributes and those who have not. There are generally no clear-cut criteria for discrimination indices. However, items with discrimination indices ranging between 0.3 and 0.4 could be generally regarded acceptable for adequately distinguishing between students who have mastered the attributes and those who have not (Shi et al., 2024). Finally, tetrachoric correlations among the mathematics attributes were estimated.

Results

Model Fit and Item Parameters

The results of relative and absolute fit statistics across the conventional DCMs and their hierarchical extensions are presented in Table 1. As can be seen, with regard to −2LL, AIC, and BIC, the *A*-CDM and RRUM had the lowest values compared to the other conventional DCMs. The higher BIC value for the G-DINA can be ascribed to its greater number of parameters, which are penalized more by the BIC. With respect to the MADcor and SRMSR, the G-DINA and the additive models (i.e., *A*-CDM, C-RUM, and RRUM) had the smallest values. Similarly, among the hierarchical DCMs, the values of −2LL, AIC, BIC, MADcor, and SRMSR showed the better fit of the HDCM_*A*-CDM and HDCM_RRUM relative to the other rival hierarchical DCMs. Overall, the additive models and their hierarchical extensions showed a better test-level performance compared to the G-DINA model.

Model–data fit at the item level was also compared across the models. For the conventional DCMs, the indices showed that all the models have satisfactory overall model–data fit, with the RMSD values ranging from 0.078 to 0.085 and MAD values between 0.054 and 0.058. Among these, the DINO model revealed the lowest RMSD ($M = 0.078$), lowest MAD ($M = 0.054$), and lowest mean RMSEA (0.111), suggesting the best item-level fit among conventional DCMs. The *A*-CDM, C-RUM, and RRUM showed a comparable performance (RMSD $\approx$ 0.082–0.083; MAD $\approx$ 0.056; RMSEA = 0.117), indicating accurate item-level parameter recovery and unbiased fit (MD values ≈ -0.001). The G-DINA model produced slightly higher RMSD (0.085) and RMSEA (0.120). However, the DINA model showed the poorest item-level fit (RMSD = 0.085; RMSEA = 0.121) and greater variability, suggesting its limited adequacy due to its strict conjunctive assumption. Due to space limitations, the results are available in Appendix B.

For the hierarchical DCMs, the HDCM_*A*-CDM, and HDCM_DINO had the best overall item-level fit. The RMSD values ranged from between 0.065 and 0.122, MAD between 0.047 and 0.090, and RMSEA values from 0.092 to 0.173. The HDCM_*A*-CDM revealed the most optimal performance (RMSD = 0.065; MAD = 0.047; RMSEA = 0.092), indicating highly consistent model–data alignment with minimal bias (MD = −0.001). The HDCM_C-RUM and HDCM_RRUM also had good fit, albeit with slightly higher RMSD (≈0.067–

Table 1

Relative Fit Statistics of (Non)Hierarchical DCMs

	Models	Npar	−2LL	AIC	BIC	MADcor	SRMSR
Conventional DCMs	G-DINA	645	273,468	274,758	279,004	***0.036***	***0.060***
	A-CDM	416	***273,097***	***273,929***	***276,667***	0.037	0.063
	C-RUM	416	273,472	274,304	277,042	0.038	0.063
	RRUM	416	***273,125***	***273,957***	***276,695***	0.037	0.063
	DINA	307	276,480	277,094	279,115	0.045	0.074
	DINO	307	279,478	280,092	282,113	0.042	0.068
Hierarchical DCMs	HDCM_G-DINA	645	286,307	287,597	291,843	0.056	0.087
	HDCM_*A*-CDM	416	***275,386***	***276,218***	***278,957***	***0.035***	***0.060***
	HDCM_C-RUM	416	276,699	277,531	280,269	0.037	0.063
	HDCM_RRUM	416	***275,687***	***276,519***	***279,258***	0.036	0.061
	HDCM_DINA	307	279,469	280,083	282,104	0.041	0.068
	HDCM_DINO	307	281,153	281,767	283,788	0.042	0.069

Note. Npar = Number of Parameters; −2LL = −2 Log-Likelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; MADcor = Mean Absolute Difference Correlation; SRMSR = Standardized Root Mean Square Residual.

Table 2

Classification Accuracy and Consistency at Test and Attribute Level for the G-DINA and HDCM_A-CDM

	G-DINA		HDCM_A-CDM	
Test and Attribute Level	**Accuracy**	**Consistency**	**Accuracy**	**Consistency**
Test	0.605	0.496	0.683	0.577
Sequencing and number patterns (SNP)	0.923	0.879	0.941	0.897
Dealing with higher number range (HNR)	0.913	0.868	0.922	0.873
Units of measurement (UOM)	0.868	0.824	0.909	0.867
Addition simple (ADS)	0.951	0.941	0.987	0.977
Subtraction simple (SUS)	0.963	0.951	0.972	0.952
Addition complex and subtraction complex (ASC)	0.961	0.940	0.958	0.922
Multiplication/division (MUD)	0.912	0.877	0.915	0.867
Modeling (MOD)	0.885	0.852	0.896	0.852

0.068) and RMSEA (≈0.096). These results suggest that introducing a hierarchical structure improves local item-level fit, especially for additive models. However, the HDCM_G-DINA showed poorer item-level fit (RMSD = 0.122; MAD = 0.090; RMSEA = 0.173) and high variability. By taking the results of model–data fit at test- and item-level, the HDCM_*A*-CDM was picked for further analyses based on the purpose of the study, indicating a balance between fit and parsimony.

Across 2,000 replications, the selected hierarchical model (HDCM_*A*-CDM) constantly outperformed the G-DINA model with regard to model–data fit. In every replication, the HDCM_*A*-CDM yielded lower AIC and BIC values. Also, its deviance values were consistently smaller than those of the G-DINA model. The simulated likelihood ratio tests were non-significant across all iterations (mean empirical $p \approx 1.00$), indicating no evidence that the G-DINA model provided a better fit than the hierarchical DCM.

Classification Accuracy

The classification accuracy of the G-DINA and HDCM_*A*-CDM at the attribute and test levels is given in Table 2. The values of attribute-level accuracy and consistency were above 0.85 across the two models. This suggests a high degree of accuracy and consistency in classifying students into different latent classes based on their mastery or non-mastery of each individual attribute. However, for the G-DINA model, the classification accuracy and consistency at the test level were 0.605 and 0.496, respectively. The corresponding values for the HDCM_*A*-CDM were 0.683 and 0.577, indicating higher accuracy and consistency than the G-DINA.

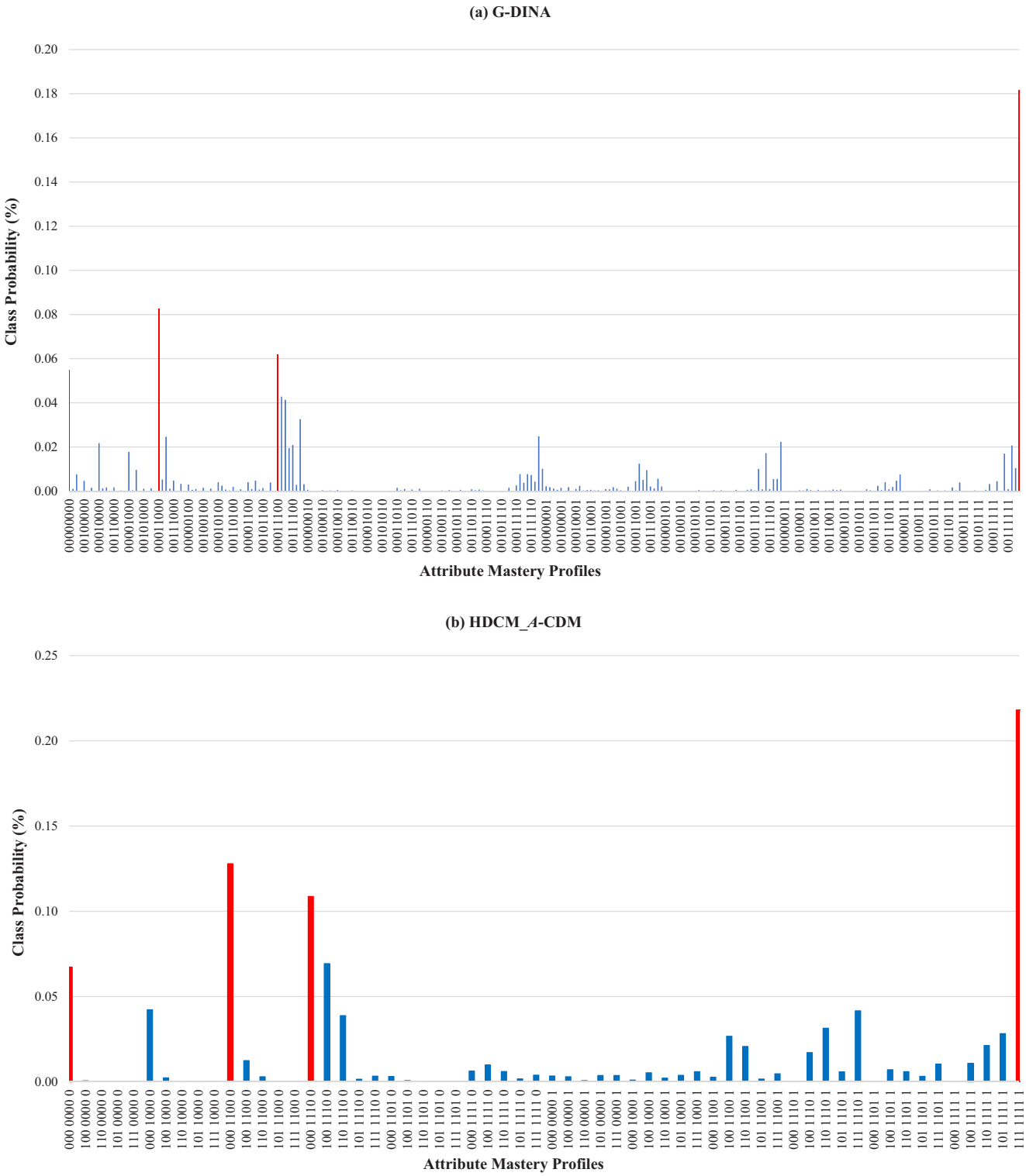
Attribute Mastery Profiles and Their Characteristics

The G-DINA model in this study produced 256 possible latent classes with eight attributes ($2^8 = 256$). The proportion of all the profiles are displayed in Figure 3a, in which 0 indicates non-mastery of the attribute, and 1 indicates mastery of the attribute. A large proportion of students were classified into four mastery profiles of $a_{256} = [11111111]$, $a_{25} = [00011000]$, $a_{57} = [00011100]$, and $a_1 = [00000000]$, with assignment probabilities of almost 18%, 8%, 6%, and 5%, respectively. The profile $a_1 = [00000000]$ indicates that the students have not mastered all the attributes, whereas $a_{256} = [11111111]$ shows that the students have mastered all the required attributes. The profile $a_{25} = [00011000]$ also represents that the students have mastered Attributes 4 and 5 (i.e., ADS and SUS), but they have not mastered the other attributes.

The 60 latent classes derived from the HDCM_*A*-CDM are illustrated in Figure 3b. As can be seen, the hierarchical model considerably reduced the number of latent classes due

FIGURE 3

Attribute Mastery Profile Probabilities across the G-DINA and HDCM_A-CDM



to the constraints imposed by the structure that removed empirically negligible or theoretically implausible profiles. The $a_{60} = [111\ 1111\ 1]$ was the most prevalent latent class, with about 22% students belong to. The second most populated class was $a_{11} = [000\ 1100\ 0]$, indicating that about 13% of students have only mastered ADS and SUS. In the third place, the class $a_{16} = [000\ 1110\ 0]$ showed that almost 11% of students have mastered ADS, SUS, and ASC. The $a_1 = [000\ 0000\ 0]$ was the fourth most populated latent class, showing that the students have not mastered any of the attributes. The attribute mastery probability of the rest of the latent classes varied from 0% to about 7%. The comparison

Table 3

Attribute Mastery Patterns for Six Examinees Based on the HDCM_A-CDM

Student	Total Score	Mastery Profile	Attribute Mastery Probability							
			SNP	HNR	UOM	ADS	SUS	ASC	MUD	MOD
676	12	000 1100 0	0.000	0.000	0.000	0.994	0.975	0.000	0.013	0.001
1682	43	100 1111 0	0.953	0.110	0.303	1.000	0.998	0.997	0.987	0.157
2780	43	111 1110 1	0.998	0.991	0.904	1.000	1.000	0.999	0.297	0.991
2838	43	100 1110 0	1.000	0.184	0.019	1.000	0.999	0.998	0.007	0.072
4458	52	111 1110 1	1.000	0.999	0.987	0.998	1.000	1.000	0.173	0.996
4484	60	111 1111 1	1.000	1.000	0.999	0.999	1.000	0.998	0.998	0.999

Note. SNP = Sequencing and Number Patterns; HNR = Dealing with Higher Number Range; UOM = Units of Measurement; ADS = Addition Simple; SUS = Subtraction Simple; ASC = Addition Complex and Subtraction Complex; MUD = Multiplication/Division; MOD = Modeling.

of the latent class distributions across the two models provided further evidence for the use of the hierarchical model because it significantly eliminated latent classes with zero or very low probability and yielded a more parsimonious and interpretable classification of students.

The seven most prevalent latent classes from the HDCM_*A*-CDM characterized by students' gender, immigration status, and SES are illustrated in Appendix C. As shown, Class 1 (i.e., [000 0000 0]) consisted of 394 students with a larger number of girls (56.6%), more native students (38.3%) from a middle SES (37.8%). Class 256 (i.e., [111 1111 1]) was the largest group, including 1,443 students, of whom 62.7% were boys and 37.3% were girls. This class mostly involved students with higher SES (51.7%) and native ones (61.6%). Overall, higher mastery classes mostly comprised boys and native students with higher SES. Boys were more prevalent in Classes 50 and 60, that is, a_{50} = [111 1111 0] and a_{60} = [111 1111 1], indicating the mastery of more advanced attributes, while there were more girls in mid-level mastery classes, that is, a_{11} = [000 1100 0], a_{16} = [000 1110 0], and a_{17} = [100 1110 0].

Diagnostic Insights and Instructional Implications

The attribute prevalence at group level for the HDCM_*A*-CDM is depicted in Figure 4. The ADS had the highest proportion of students (92%) who have mastered the attribute, followed by the SUS, SNP, and ASC, with mastery probabilities of about 86%, 64%, and 62%, respectively. This shows that about 84% of students have mastered SUS, while approximately 64% and 62% have mastered SNP and ASC, respectively. However, the UOM had the lowest proportion of students (34%) who have mastered it. The MUD was the second most difficult attribute (0.38%), followed by the HNR (42%) and MOD (49%).

DCMs can offer detailed inferences and personalized feedback with regard to each student's mastery or non-mastery of different attributes. Such information will inform teachers about the strengths and weaknesses of students and allow them to design more appropriate materials and activities to improve instruction and dispel students' deficiencies. Table 3 provides the attribute mastery patterns for six students with different total scores, indicating their various mathematical ability levels. For example, the attribute mastery pattern of Student 676 (girl, middle SES, and second generation of immigration) shows that she has only mastered ADS and SUS, so interventions should focus on improving her SNP, HNR, UOM, ASC, MUD, and MOD. Similarly, Student 4458 (girl, high SES, and native) only requires the improvement of MUD because she has mastered the other attributes. Additionally, the attribute mastery patterns of Students 1682, 2780, and 2838 indicated that they have mastered different attributes, although they shared the same total score. This represents the students' diverse range of strengths and weaknesses in their mathematical ability, leading to different profiles with different mastery statuses of the eight attributes.

Item Discrimination

The results of item discrimination indices for the HDCM_*A*-CDM indicated that out of 135 items, 24 items had discrimination values below 0.30, albeit most of them were close to the cut-off value. The average discrimination was 0.467, which was greater than the 0.3 benchmark, showing that the items could generally differentiate between masters and nonmasters of the attributes. Due to space considerations, the results of item discrimination for the HDCM_*A*-CDM are available in Supplementary Material C.

Tetrachoric Correlations among the Attributes

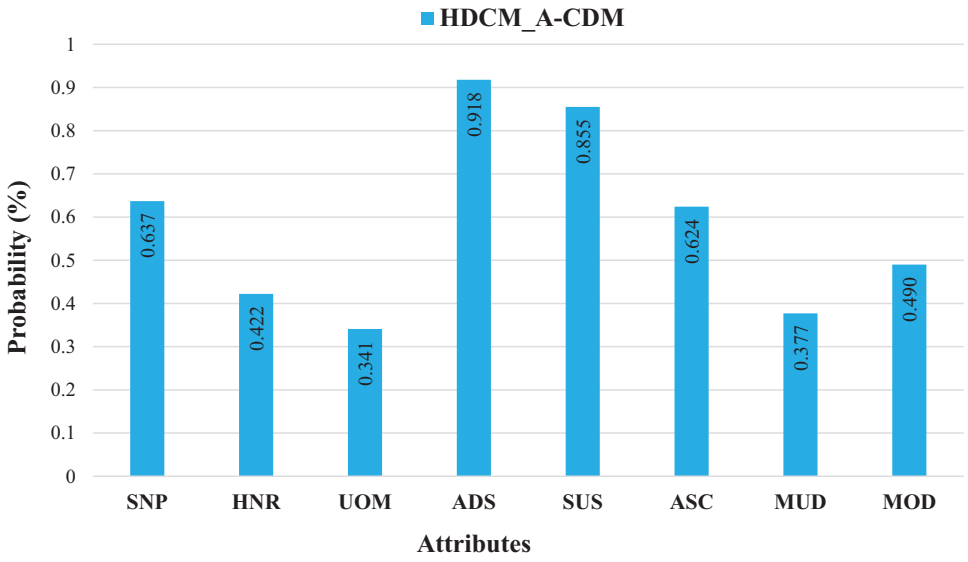
Many attribute pairs generally showed moderate to strong associations, while a few demonstrated weak relationships. The correlation between SNP and HNR ($r = 0.997$), SNP and UOM ($r = 0.994$), ADS and SUS ($r = 0.998$), ADS and ASC ($r = 0.991$), and SUS and ASC ($r = 0.995$) were almost perfect, indicating a strong dependency between the pair of attributes, especially for adjacent attributes. In fact, students who have mastered one of the attributes are more likely to have mastered the other attribute as well. This strong dependency likely occurs due to the proposed hierarchical structure. However, the low association between SUS and MOD ($r = 0.430$) as well as UOM and SUS ($r = 0.432$) showed that the mastery of SUS is not necessarily associated with the mastery of MOD and UOM. The results of tetrachoric correlations are available in Appendix D.

Discussion

This study aimed, first, to investigate whether a Q-matrix directly derived from cognitive model-based item generation can be effectively used for DCM analysis. Second, it investigated whether incorporating theoretical assumptions about attribute hierarchies into DCM can lead to the improvement of diagnostic results. The study also examined whether re-

FIGURE 4

Attribute Prevalence based on the HDCM_A-CDM



sults obtained from DCMs can offer added value to the conventional approaches. In this section, findings are discussed as they relate to the three research questions that were used to guide this study.

Can a Q-Matrix Resulting from Theory-Based Item Generation Be Effectively Applied in DCMs?

Overall, the results of the study provided empirical evidence that cognitive models used for item generation can directly inform a Q-matrix to be effectively used for cognitive diagnostic modeling. More specifically, this approach yields more meaningful and interpretable latent classes for identifying students' mastery status of several attributes that maximizes the practicality of Q-matrices in large-scale standardized assessments.

The comparison of the specific DCMs (i.e., DINA, DINO, C-RUM, *A*-CDM, and RRUM) against the G-DINA at the test level showed that additive models, especially the *A*-CDM and RRUM, had the best performance regarding AIC and BIC, indicating that their modeling structure could better reflect the processing of mathematical ability and the overall pattern of item responses. This finding converges with earlier studies that found better performance of additive (or main effect) models in analyzing mathematical exams, especially when a mathematics test was not originally constructed for diagnostic purposes but retrofitted by DCMs (Wu et al., 2020; Yamaguchi & Okada, 2018). The higher relative fit values for the G-DINA also indicates that its complexity did not translate into better test-level fit. Additionally, the DINA and DINO, as the parsimonious models, had the poorest fit, suggesting that these models are too restrictive to model item responses and explain actual students' knowledge status.

However, at the item level, the DINO model showed the best performance regarding RMSD, bias-corrected RMSD, MD, MAD, and mean RMSEA. It suggests that its simplicity allows it to fit the observed responses for many items, although it was less accurate in modeling the overall test. In fact, a simple model like DINO may reproduce item-level statistics very well but not capture complex interactions across items, leading to lower test-level fit.

The moderate value of the test-level and higher values of attribute-level accuracy and consistency for the G-DINA model further revealed the precision of the model in classifying students into different latent classes based on the mastery/non-mastery of the attributes. This precision is important for ensuring that the resultant latent classes are meaningful and accurately reflect the students' true cognitive profiles. A possible explanation for the moderate test-level accuracy can be the structure of the dataset. There were eight booklets, and students had to complete the first two booklets, causing many missing values for the six remaining booklets. The missing responses supposedly made gaps in attribute estimation, reducing the model's capacity to fully reflect examinees' attribute mastery profiles. DCMs mostly rely on complete response patterns to accurately classify examinees. The classification accuracy of a model can decrease when a considerable portion of the dataset contain missing values (Dai & Svetina Valdivia, 2022).

Furthermore, the results of attribute mastery profiles showed that the G-DINA model could effectively generate different types of attribute mastery profiles, and students were distributed across the latent classes. This indicates the model's high discriminant function. As argued by Lee and Sawaki (2009), a DCM may classify most of the students into profiles in which they have either mastered all attributes or none at all, called flat mastery profiles ([00000000] and [11111111]). It makes individual attribute scores redundant and provides little additional information beyond a total proficiency score. This can happen when a DCM is applied to a non-diagnostic, unidimensional test. In such cases, the utility of profile scoring is questionable, especially in cognitive domains where constructs are highly correlated. The lack of flat mastery profiles in this study supports the discriminant function of the diagnostic model and the effectiveness of the theoretically based Q-matrix.

Does Incorporating Theoretical Assumptions about Attribute Hierarchies Enhance the Performance of DCMs?

Overall, the results provide empirical evidence for supporting the hypothesis that incorporating theoretical assumptions about attribute hierarchy into DCMs enhances the diagnostic results by generating more meaningful, interpretable, and accurate latent classes. The utility of the HDCM is more consistent with the didactic theory of mathematical ability and can better capture the developmental nature of the construct, where the mastery of basic attributes is a prerequisite for the mastery of more advanced attributes.

The model–data fit comparison between the G-DINA and hierarchical DCMs at the test and item level showed the best performance of the HDCM_*A*-CDM. Although the conventional DCMs had lower relative fit statistics values, their hierarchical counterparts, especially the HDCM_*A*-CDM and HDCM_RRUM, showed smaller absolute fit values, except for the HDCM_G-DINA. It indicates that hierarchical DCMs can yield satisfactory model fit while reducing overfitting. The results of model–data fit at the item level also indicated that the HDCM_*A*-CDM had the best overall fit among the rival models. This suggests that introducing a hierarchical structure can enhance item-level fit, especially for additive models. The better performance of the HDCM_*A*-CDM was further supported by a simulation-based likelihood ratio test, with consistently better performance compared to the G-DINA model.

The results of classification accuracy and consistency at test- and attribute-level further revealed that the HDCM_*A*-CDM produced higher values compared to the G-DINA model, especially at the test level. By taking attribute hierarchy into account, this finding suggests that the test items could better distinguish between students with different ability levels. Higher values show the model's improved ability to assign students to correct mastery profiles. Additionally, the higher classification accuracy and consistency of the HDCM_*A*-CDM confirms the results of previous simulation studies that have reported the increased classification accuracy when a hierarchical structure is used (Kwon et al., 2024; Tu et al., 2019). It also corroborates the importance of "restricting the latent class structure within the permissible attribute profile spaces" (Tu et al., 2019, p. 269).

Moreover, unlike the G-DINA model which generated scattered latent classes, the students were assigned to more coherent and meaningful classes in the HDCM_*A*-CDM, that can represent real world differences in mathematical ability. For instance, as indicated in Figure 3b, classes $a_{11} = [000\ 1100\ 0]$ and $a_{16} = [000\ 1110\ 0]$ showed attribute mastery patterns that agree with developmental and didactic theories of mathematical ability regarding the development of cognitive abilities across time. Such latent classes provide a more detailed and feasible understanding of students' performance.

The attribute mastery rates derived from the HDCM_*A*-CDM were consistent with the developmental nature of the mathematical ability. The attribute prevalence showed that a large proportion of students have mastered ADS and SUS, whereas UOM and MUD had lower mastery rates. This finding is in line with expectations for third-grade students, whose cognitive development usually advances from simpler to more complicated mathematical abilities. It also converges with studies that reported the continuous role of math fluency throughout primary school, with ability differing across different arithmetic operations; addition and subtraction are mostly mastered earlier, whereas multiplication and division as well as the concept of measurement units require more time and practice to attain fluency (Gliksman et al., 2022). This finding is also in agreement with Piaget's theory of memory (Piaget & Inhelder, 1968/1973). According to this theory, subtraction is more difficult than addition because students deduce differences from their knowledge of sums rather than being stored and retrieved. This reflects a characteristic of early childhood where positive aspect of actions, perception, and cognition precedes negative aspects (Piaget, 1974/1980).

Additionally, mathematics tests are mostly developed based on a hierarchical framework to reflect the developmental nature of the construct. Therefore, when items are written based on a hierarchical structure, the conventional non-hierarchical DCMs are more likely to generate inaccurate latent classes. As Tu et al. (2019) argued, conventional non-hierarchical DCMs assume that all attribute profiles are present in a population that may induce many equivalence classes with similar response patterns. This could lead to misclassification of attribute profiles due to the non-identifiability problem. However, in this study, the attribute prevalence patterns support the theoretical soundness of the Q-matrix and the hierarchical structure that resulted in generating more sensible diagnostic classifications.

The item discrimination analysis from the HDCM_*A*-CDM provided further evidence for the validity of the diagnostic model, hierarchical structure, and the Q-matrix. Except for 24 items, the test showed acceptable discrimination, indicating that most items effectively distinguished students who had mastered the attributes from those who had not, especially between basic and more advanced mathematical skills. In addition, tetrachoric correlations from the HDCM_*A*-CDM supported the plausibility of the Q-matrix and its hierarchical structure, although some very high correlations likely reflected hierarchical constraints that encouraged co-occurring mastery patterns. Additionally, for attributes arranged in a chain (e.g., ADS → SUS → ASC and ADS → SUS → MUD), this hierarchy may have forced them to behave like nearly the same latent construct, leading to higher correlations. More importantly, the patterns of correlations showed a strong association between attributes within the same strand but moderate associations across them. The presence of moderate correlations between some attributes, such as the HNR and ADS ($r = 0.609$), suggests that the Q-matrix could accurately differentiate between different domains of mathematical ability.

Are Results from DCMs in Line With Traditional Types of Analyses and Do They Offer Added Value?

What is notable is that a distinction should be made between ability parameters of IRT models and the dimensionality of the DCMs attribute because they may be interpreted in different ways. A common belief in the use of IRT models is that a test is considered unidimensional if its items only measure one construct at a time. However, in reality, unidimensional ability of IRT models can involve several abilities, skills, knowledge, and strategies that are relevant to the test (Yamaguchi & Okada, 2018). As argued by Bejar (1983), unidimensionality does not mean that performance on items is affected by a single psychological process. Rather, various

psychological processes are involved in responding to items. "As long as they are involved in unison, that is, performance on each item is affected by the same process and in the same form, unidimensionality will hold" (p. 31). Therefore, the construct represented by the ability parameter of IRT models, usually defined broadly, may consist of several components (Yamaguchi & Okada, 2018). By contrast, in DCMs, constructs are narrowly defined, as categorical latent variables, by breaking down learning concepts into smaller units or "learning quanta" (Fischer, 1973) for optimal teaching and learning. Due to such structural differences among the two frameworks, it is not appropriate to compare the models. Yamaguchi and Okada (2018) argue that although IRT models are not able to fully reflect examinees' actual test-taking behavior, they are more suitable for linking or equating items to create a common scale. On the contrary, DCMs have the potential to more thoroughly capture examinees' cognitive abilities. For that reason, DCMs can complement the results of traditional psychometric models (i.e., CTT and IRT) by providing additional diagnostic insight. Unidimensional IRT models can be used to order examinees on a single latent trait continuum in large-scale standardized tests to make different decisions about examinees, and DCMs can be used to explain examinees' cognitive process of problem solving and provide diagnostic information.

Furthermore, at an individual level, the case analysis of six students' attribute mastery profiles (Table 3) clearly indicated that students with the same total score do not necessarily have the same attribute profiles. Rather, there are different attribute profiles reflecting the various strengths and weaknesses of students in mathematical ability. This suggests that although total scores from traditional models can be used to provide an overall measure of ability, DCMs detect meaningful within-score differences, that are typically ignored by traditional models.

Finally, the study showed an association between latent class membership and several demographic variables, including gender, SES, and immigration status. This finding implies that these variables affect the distribution of examinees across the latent classes. Higher mastery attribute profiles mostly involved male examinees with higher SES and native students, whereas lower and intermediate mastery profiles mainly comprised female examinees with low-to-middle SES and non-native students. This finding is consistent with previous studies highlighting the role of large individual differences in mathematical ability (e.g., Haataja et al., 2024), especially with Sonnleitner et al. (2026) who found almost the same pattern using the DCM.

More specifically, the analysis of the association between gender and class membership revealed that male examinees were more dominant in classes with the mastery of more advanced attributes, whereas female examinees were more prevalent in the intermediate classes. This indicates the effect of gender on the assignment of students to different attribute mastery profiles. This finding is consistent with studies reporting the presence of gender differences between boys and girls at the elementary level (e.g., Jordan et al., 2006; Winkelmann et al., 2008), albeit the reported differences were not high. Studies have also reported gender differences in math performance based on task type. Males generally tend to perform better on mathematical problem solving, math fact retrieval, and items involving tables, while females excel on basic arithmetic, calculus operations, and items involving symbols (O'Neill & McPeek, 1993; Royer et al., 1999).

Furthermore, research has indicated that students from lower SES backgrounds have, on average, lower mathematical ability than students from higher SES backgrounds (Sirin, 2005). Parents with higher SES tend to be more engaged in their children's schooling, and this involvement significantly influences students' mathematics performance, especially in countries with competitive school systems (Niehues et al., 2020). Moreover, the association between attribute mastery profiles and immigration status highlights the importance of considering the differential performance between native and immigrant students. The profile patterns of this study showed that native students exhibited a better performance than their immigrant counterparts, with the mastery of more advanced attributes in higher latent classes. Previous studies have identified several interrelated factors associated with the underachievement of immigrant students, including SES (Giannelli & Rapallini, 2016), parental ethnicity (Kim et al., 2020), language barriers (Toppelberg & Collins, 2010), higher immigrant concentration in schools (Cortes, 2006), educational institutions (Schneeweis, 2011), and the age at which students arrive in the country of immigration (Böhlmark, 2008). Therefore, these factors might have affected the performance of immigrant students and their distribution across the latent classes. It is highly recommended for future studies to include more factors for characterizing latent classes. However, the number of immigrant students in several latent classes with the mastery of more advanced attributes is notably lower. One possible explanation may be that the barriers to learning mathematics in the country of immigration are lower than those for learning language-dependent subjects. As argued by Giannelli and Rapallini (2016), math is a more portable skill, so immigrant students, especially those from countries with high math rankings, may have an advantage over natives. This advantage could be explained by the family influence for second-generation immigrants or by schooling in the country of origin and family influence for first-generation immigrants. Parental influence, especially from highly performing countries in math, can further increase this advantage.

Conclusion

This study has implications for research and practice. The findings of the study highlight the value of cognitive model-based item development for enhancing DCM applications within large-scale assessments. This practice ensures that items are inextricably linked to the underlying cognitive processes or attributes the test is supposed to measure, causing more accurate diagnosis of examinees' strengths and weaknesses and better-informed instructional decisions. The study also offers a framework for considering hierarchical relationships among cognitive skills—especially mathematical ability—into future research. Moreover, hierarchical DCMs help researchers and test developers to adjust tests with instructional objectives and learning standards, inducing more accurate tracking of students' development. They can further provide a more plausible interpretation of test scores and distinguish between students with different proficiency levels; students who have mastered basic attributes from those who are struggling with them.

A notable limitation of the study is that although the items were constructed based on a cognitive model, DCMs were applied to a non-diagnostic test. Future studies can focus on developing a truly diagnostic mathematics test specifically designed for DCMs using cognitive frameworks such as the cognitive design system (CDS; Embretson, 1998), the assessment triangle (Pellegrino et al., 2001), and the evidence-centered design (ECD; Mislevy, 1994).

The main objective of formative assessment is to recurrently assess students' performance and monitor their change in a course of instruction as a result of diagnostic feedback and instruction. It allows all the stakeholders to be accountable for identifying the nature and source of problems and adopting some strategies and approaches to solve these deficiencies. Diagnozing students' weaknesses and resolving them is the ultimate goal of assessment. Therefore, DCMs would be more effective when they are utilized to track the strengths and weaknesses of students over time, rather than being restricted to one-shot assessments. Thanks to the recent development of DCMs, researchers can utilize longitudinal DCMs to measure changes in attribute mastery status over a period of time (Madison & Bradshaw, 2018; Ravand et al., 2025).

Another intriguing direction for future research is the development and application of cognitive diagnostic computerized adaptive testing (CD-CAT) to large-scale standardized assessments. Such models have several advantages (Sorrel et al., 2020): (1) they achieve a more efficient assessment through a carefully designed item bank, (2) they increase test security by using different subsets of items and reducing the probability of item compromise, and (3) they can increase motivation of examinees by administrating items that match their ability level, reducing the test-taking time, and avoiding items that are too easy or too difficult for examinees.

Finally, future studies could also explore the external validity of applying DCMs to mathematical assessments by investigating the extent to which the examinee profiles produced by DCMs match with teachers' perception of their students' mathematical abilities.

Acknowledgments

Open access funding enabled and organized by Projekt DEAL.

Dedication

The first author dedicates this research to the students and innocent souls of the Minab school in Iran—children who could have become the future researchers, scholars, and scientists of our world—and to all Iranians who lost their lives in the war. Their dreams were silenced far too soon, but their memory endures in the hearts of those who remember them. May their souls rest in peace, and may their light continue to guide us toward compassion, justice, and peace.

Funding

This study was funded by the German Research Foundation (DFG; Deutsche Forschungsgemeinschaft) grant KU 3647/2-1, which was awarded to Olga Kunina-Habenicht and the Fonds National de la Recherche Luxembourg grant FAIR-ITEMS (C19/SC/13650128), awarded to Philipp Sonnleitner.

Conflict of Interest Statement

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

Data Availability Statements

The datasets generated and/or analyzed during the current study are available from the last author on reasonable request.

Declaration of Generative AI in Scientific Writing

During the preparation of this work, the authors used DeepL in order to shorten and improve the quality of the written text (language checking and formatting). After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Note

[1]Although all three models are structurally additive, only the *A*-CDM is additive on the probability scale. The C-RUM, LLM, and RRUM are additive on the logit or log scale, not with respect to the probability. Specifically, the *A*-CDM assumes that mastery of each attribute independently and additively increases the probability of a correct response. However, the RRUM is additive on the log scale and describes the multiplicative impact of attribute mastery on the success probability.

ORCID

Farshad Effatpanah https://orcid.org/0000-0003-3970-5588
Olga Kunina-Habenicht https://orcid.org/0000-0002-1646-8260
Steve Bernard https://orcid.org/0009-0007-2784-2098
Caroline Hornung https://orcid.org/0000-0003-0061-200X
Philipp Sonnleitner https://orcid.org/0000-0002-4861-4023

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Artemenko, C., Pixner, S., Moeller, K., & Nuerk, H. C. (2018). Longitudinal development of subtraction performance in elementary school. *The British Journal of Developmental Psychology*, *36*(2), 188–205. https://doi.org/10.1111/bjdp.12215

Aunio, P., & Niemivirta, M. (2010). Predicting children's mathematical performance in grade one by early numeracy. *Learning and Individual Differences*, *20*(5), 427–435. https://doi.org/10.1016/j.lindif.2010.06.003

Bejar, I. I. (1983). *Achievement testing: Recent advances*. Beverly Hills: CA Sage.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach tonumerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in Algebra using the rule-space model. *Journal for Research in Mathematics Education*, *24*(5), 442–459. https://doi.org/10.2307/749153

Böhlmark, A. (2008). Age at immigration and school performance: A siblings analysis using Swedish register data. *Labour Economics*, *15*(6), 1366–1387. https://doi.org/10.1016/j.labeco.2007.12.004

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

Chen, Y., & Wang, S. (2023). Bayesian estimation of attribute hierarchy for cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, *48*(6), 810–841. https://doi.org/10.3102/10769986231174918

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618. https://doi.org/10.1177/0146621613488436

Choi, K. M., Lee, Y.-S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science, and Technology Education*, *11*(6), 1563–1577. https://doi.org/10.12973/eurasia.2015.1421a

Cirino, P. T. (2011). The interrelationships of mathematical precursors in kindergarten. *Journal of Experimental Child Psychology*, *108*(4), 713–733. https://doi.org/10.1016/j.jecp.2010.11.004

Clements, D. H., & Sarama, J. (2007). Early childhood mathematics learning. In F. K. Lester Jr (Ed.), *Second handbook on mathematics teaching and learning* (pp. 461–555). Charlotte, NC: Information Age.

Clements, D. H., & Sarama, J. (2012). Learning and teaching early and elementary mathematics. In J. S. Carlson & J. R. Levin (Eds.), *Instructional strategies for improving students' learning: Focus on early reading and mathematics* (pp. 107–162). IAP Information Age Publishing.

Cortes, K. E. (2006). The effects of age at arrival and enclave schools on the academic performance of immigrant children. *Economics of Education Review*, *25*(2), 121–132 https://doi.org/10.1016/j.econedurev.2004.12.001

Dai, S., & Svetina Valdivia, D. (2022). Dealing with missing responses in cognitive diagnostic modeling. *Psych*, *4*(2), 318–342. https://doi.org/10.3390/psych4020028

Delafontaine, J., Chen, C., Park, J. Y., & Van den Noortgate, W. (2022). Using country specific Q-matrices for cognitive diagnostic assessments with international large-scale data. *Large-scale Assessment in Education*, *10*, 19. https://doi.org/10.1186/s40536-022-00138-4

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. (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., Qiu, X. L., & Santos, K. C. (2022). An empirical Q-matrix validation method for the polytomous G-DINA model. *Psychometrika*, *87*(2), 693–724. https://doi.org/10.1007/s11336-021-09821-x

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol., *26*, pp. 979–1030). Amsterdam: Elsevier. https://doi.org/10.1007/978-3-031-04572-1_16

Doornik J. A. (2007). *Object-oriented matrix programming using Ox*(6th Ed.). London,England: Timberlake Consultants Press.

Effatpanah, F., Ravand, H., Abdi Tabari, M., Chen, Y.-H., & Kunina-Habenicht, O. (2026). How do L2 writing subskills interact hierarchically? Insights from diagnostic classification models. *Assessing Writing*, *68*, 101029. https://doi.org/10.1016/j.asw.2026.101029

Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia*, *9*(12), 1–23. https://doi.org/10.1186/s40468-019-0090-y

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396. https://doi.org/10.1037/1082-989X.3.3.380

Ferraro, D., & Van de Kerckhove, W. (2006). *Trends in international mathematics and science study (TIMSS) 2003 nonresponse bias analysis, technical report*. U.S. Department of Education, Institute of Education Sciences.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Ganzeboom, H. B. (2010, May). A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007. In *Annual Conference of International Social Survey Programme* (Vol., *1*). Lisbon.

Geary, D. C. (2006). Development of mathematical understanding. In D. Kuhn, R. S. Siegler, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Cognition, perception, and language* (6th ed., pp. 777–810). John Wiley & Sons, Inc.

Giannelli, G. C., & Rapallini, C. (2016). Immigrant student performance in Math: Does it matter where you come from? *Economics of Education Review*, *52*, 291–304. https://doi.org/10.1016/j.econedurev.2016.03.006

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives*, *6*(4), 263–268. https://doi.org/10.1080/15366360802497762

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report*, *22*(1), 1–23. https://doi.org/10.1002/j.2379-3988.2008.tb00054.x

Gliksman, Y., Berebbi, S., & Henik, A. (2022). Math fluency during primary school. *Brain Sciences*, *12*(3), 371. https://doi.org/10.3390/brainsci12030371

Haataja, E., Niemivirta, M., Holm, M., Ilomanni, P. K., & Laine, A. (2024). Students' socioeconomic status and teacher beliefs about learning as predictors of students' mathematical competence. *European Journal of Psychology of Education*, *39*, 1615–1636. https://doi.org/10.1007/s10212-023-00791-5

Haghayeghi, M., Moghadamzadeh, A., Ravand, H., Javadipour, M., & Kareshki, H. (2024). Development of a cognitive assessment checklist for first-grade mathematics: Utilizing hierarchical cognitive diagnostic modeling in elementary education. *Journal of Psychoeducational Assessment*, *43*(1), 88–107. https://doi.org/10.1177/07342829241290277

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*[Unpublished doctoral dissertation, University of Illinois at Urbana- Champaign].

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Johnson, E., & Carlson, J. (1994). *The NAEP 1992 technical report*. National Center for Education Statistics. Retrieved from https://files.eric.ed.gov/fulltext/ED376191.pdf

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cog-

nitive diagnostic assessments. *Journal of Educational Measurement*, *55*(4), 635–664. https://doi.org/10.1111/jedm.12196
Jordan, N. C., Kaplan, D., Nabors Oláh, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development*, *77*(1), 153–175. https://doi.org/10.1111/j.1467-8624.2006.00862.x
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064
Kim, Y., Mok, S. Y., & Seidel, T. (2020). Parental influences on immigrant students' achievement-related motivation and achievement: A meta-analysis. *Educational Research Review*, *30*, 100327. https://doi.org/10.1016/j.edurev.2020.100327
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2017). Incremental validity of multidimensional proficiency scores from diagnostic classification models: An illustration for elementary school mathematics. *International Journal of Testing*, *17*(4), 277–301. https://doi.org/10.1080/15305058.2017.1291517
Kwon, T. Y., Huggins-Manley, A. C., Templin, J., & Zheng, M. (2024). Modeling hierarchical attribute structures in diagnostic classification models with multiple attempts. *Journal of Educational Measurement*, *61*(2), 198–218. https://doi.org/10.1111/jedm.12387
Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*(2), 144–177. https://doi.org/10.1080/15305058.2010.534571
Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*(3), 172–189. https://doi.org/10.1080/15434300902985108
LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, *81*(6), 1753–1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x
Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x
Li, J., & Chen, P. (2025). A new Q-matrix validation method based on signal detection theory. *The British Journal of Mathematical and Statistical Psychology*, *78*(2), 522–554. https://doi.org/10.1111/bmsp.12371
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
Li, T., & Traynor, A. (2022). The use of cognitive diagnostic modeling in the assessment of computational thinking. *AERA Open*, *8*. https://doi.org/10.1177/23328584221081256
Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383. https://doi.org/10.1177/0013164416685599
Ma, C., Ouyang, J., & Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, *88*, 175–207. https://doi.org/10.1007/s11336-022-09867-5
Ma, W. (2019). Cognitive diagnosis modeling using the GDINA R package. In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 593–601). Cham: Springer. https://doi.org/10.1007/978-3-030-05584-4_29
Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *The British Journal of Mathematical and Statistical Psychology*, *73*(1), 142–163. https://doi.org/10.1111/bmsp.12156
Ma, W., de la Torre, J., Sorrel, M., & Jiang, Z. (2016-2025). *GDINA: The generalized DINA model framework*. Retrieved from https://CRAN.R-project.org/package=GDINA
Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963–990. https://doi.org/10.1007/s11336-018-9638-5
Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. https://doi.org/10.1007/BF02294535
Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 international science report findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680
MENFP (2011). *Kompetenzraster und Entwicklungsstufen. Grundschule, Zyklen 1 bis 4*. Luxemburg: MENFP.
Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*(4), 439–483. https://doi.org/10.1007/BF02294388
Muthén L. K., Muthén B. O. (1998–2017). *Mplus user's guide*(version 8) [ComputerSoftware and manual]. Los Angeles, CA: Muthén & Muthén.
National Council for Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.
Niehues, W., Kisbu-Sakarya, Y., & Selcuk, B. (2020). Motivation and maths achievement in Turkish students: Are they linked with socio-economic status? *Educational Psychology*, *40*(8), 981–1001. https://doi.org/10.1080/01443410.2020.1724887
Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_3-2018_347-368.pdf
O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In Holland, P. W. & Wainer, H. (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Organization for Economic Cooperation and Development. (2022). *PISA 2022 technical report*. Retrieved from https://www.oecd.org/pisa/
Pellegrino, J. W., Chudowsky, N. J., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy of Sciences.
Piaget, J. (1980). *Experiments in contradiction*. Chicago: University of Chicago Press (D. Coltman, Trans.; original work published 1974).
Piaget, J., & Inhelder, B. (1973). *Memory and intelligence*. New York: Basic Books (A. J. Pomerans, Trans.; original work published 1968).
R Core Team (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (expanded edition). Copenhagen University of Chicago Press. Originally published (1960), Pædagogiske Institut.
Ravand, H., Effatpanah, F., Kunina-Habenicht, O., & Madison, M. J. (2025). A didactic illustration of writing skill growth through a longitudinal diagnostic classification model. *Frontiers in Psychology*, *15*, 1521808. https://doi.org/10.3389/fpsyg.2024.1521808
Ribner, A., Moeller, K., Willougby, M., & Blair, C., & Family Life Project Key Investigators (2018). Cognitive abilities and mathematical competencies at school entry. *Mind, Brain, and Education: The Offi-*

cial Journal of the International Mind, Brain, and Education Society, *12*(4), 175–185. https://doi.org/10.1111/mbe.12160
Robitzsch, A., Kiefer, T., & Wu, M. (2025). *TAM: Test Analysis Modules*. R package version 4.2-21. URL: https://cran.r-project.org/web/packages/TAM
Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2011–2025). *CDM: Cognitive diagnosis modeling*. R package version 8.3–14. Retrieved from https://cran.r-project.org/web/packages/CDM/index.html
Royer, J. M., Tronsky, L. N., Chan, Y., Jackson, S. J., & Marchant, H. (1999). Math-fact retrieval as the cognitive mechanism underlying gender differences in math test performance. *Contemporary Educational Psychology*, *24*(3), 181–266. https://doi.org/10.1006/ceps.1999.1004
Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78–96. https://doi.org/10.1177/0013164407301545
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic* measurement: *Theory*, methods, and applications. New York, NY: Guilford Press.
Saso, S., Oka, M., Uesaka, Y., & Usami, S. (2024, August 19). *Diagnostic assessment of deep understanding using cognitive diagnostic models: A large-scale assessment to promote the use of effective learning strategies*. Preprint. https://doi.org/10.31234/osf.io/hk5y2
Saß, S., Kampa, N., & Köller, O. (2017). The interplay of g and mathematical abilities in large-scale assessments across grades. *Intelligence*, *63*, 33–44. https://doi.org/10.1016/j.intell.2017.05.001
Schneeweis, N. (2011). Educational institutions and the integration of migrants. *Journal of Population Economics*, *24*(4), 1281–1308. https://doi.org/10.1007/s00148-009-0271-6
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136
Shi, X., Ma, X., Du, W., & Gao, X. (2024). Diagnosing Chinese EFL learners' writing ability using polytomous cognitive diagnostic models. *Language Testing*, *41*(1), 109–134. https://doi.org/10.1177/02655322231162840
Sinharay, S. (2014). Analysis of added value of subscores with respect to classification. *Journal of Educational Measurement*, *51*(2), 212–222. https://doi.org/10.1111/jedm.12043
Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417
Sonnleitner, P., Inostroza-Fernandez, P., & Hornung, C. (2026). Mapping first graders's numerical development at scale: Leveraging cognitive models in a large-scale educational assessment. *Educational Assessment*, 1–22. Advance Online Publication. https://doi.org/10.1080/10627197.2026.2633512
Sonnleitner, P., Bernard, S., Michels, M. A., Inostroza-Fernandez, P., Keller, U., Gierl, M. J., Cardoso-Leite, P., & Hornung, C. (2025). Establishing cognitive item models for fair and theory-grounded automatic item generation: A large-scale assessment study with image-based math items. *Applied Measurement in Education*, *38*, 95–117. https://doi.org/10.1080/08957347.2025.2563889
Sorrel, M. A., Barrada, J. R., de la Torre, J., & Abad, F. J. (2020). Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. *PLoS ONE*, *15*(1), e0227196. https://doi.org/10.1371/journal.pone.0227196
Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*(4), 439–455. https://doi.org/10.1037/1082-989X.11.4.439
Su, Y.-L., Choi, K. M., Lee, W. C., Choi, T., & McAninch, M. (2013). Hierarchical cognitive diagnostic analysis for TIMSS 2003 mathematics. *Centre for Advanced Studies in Measurement and Assessment*, *35*, 1–71. https://education.uiowa.edu/sites/education.uiowa.edu/files/2022-10/casma-research-report-35.pdf
Sun, S., Wu, X., & Xu, T. (2023). A theoretical framework for a mathematical cognitive model for adaptive learning systems. *Behavioral Sciences*, *13*(5), 406. https://doi.org/10.3390/bs13050406
Tatsuoka, C., Clements, D. H., Sarama, J., Izsak, A., Orril, C. H., de la Torre, J., & Khasanova, E. (2016). Developing workable attributes for psychometric models based on the Q-matrix. *Journal of Research on Mathematics Education*, *15*, 73–96. https://www.jstor.org/stable/26858791
Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconception based on item response theory. *Journal of Education Measurement*, *20*(4), 345–354 https://doi.org/10.1111/j.1745-3984.1983.tb00212.x
Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317–339. https://doi.org/10.1007/s11336-013-9362-0
Templin, J. L., & 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
Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: Q-matrix validation and item-based model fit evaluation for the TIMSS 2011 assessment. *Sage Open*, *9*(1),. https://doi.org/10.1177/2158244019832684
The MathWorks Inc. (2022). *MATLAB version: 9.13.0* (R2022b), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com
Thompson, W. J., & Clark, A. K. (2024). Improving instructional decision-making using diagnostic classification models. *Educational Measurement: Issues and Practice*, *43*(4), 146–156. https://doi.org/10.1111/emip.12619
Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, *26*, 237–255. https://doi.org/10.1007/s13394-013-0090-7
Toppelberg, C. O., & Collins, B. A. (2010). Language, culture, and adaptation in immigrant children. *Child and Adolescent Psychiatric Clinics of North America*, *19*(4), 697–717. https://doi.org/10.1016/j.chc.2010.07.003
Tu, D., Wang, S., Cai, Y., Douglas, J., & Chang, H.-H. (2019). Cognitive diagnostic models with attribute hierarchies: Model estimation with a restricted Q-matrix design. *Applied Psychological Measurement*, *43*(4), 255–271. https://doi.org/10.1177/0146621618765721
von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine and Child Neurology*, *49*(11), 868–873. https://doi.org/10.1111/j.1469-8749.2007.00868.x
von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2), i–35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307. https://doi.org/10.1348/000711007X193957
Winkelmann, H., van den Heuvel-Panhuizen, M., & Robitzsch, A. (2008). Gender differences in the mathematics achievements of German primary school students: Results from a German large-scale study. *ZDM Mathematics Education*, *40*, 601–616. https://doi.org/10.1007/s11858-008-0124-x
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
Xu, T., Wu, X., Sun, S., & Kong, Q. (2023). Cognitive diagnostic analysis of students' mathematical competency based on the DINA model. *Psychology in the Schools*, *60*(9), 3135–3150. https://doi.org/10.1002/pits.22916

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

Zhang, Y., Zhao, B., Jian, M., & Wu, X. (2025). Cognitive diagnostic analysis of mathematics key competencies based on PISA data. *PLoS ONE*, *20*(2), e0315539. https://doi.org/10.1371/journal.pone.0315539

Zheng, Y., Chiu, C.-Y., & Douglas, J. (2019). *NPCD: Nonparametric methods for cognitive diagnosis*. R Package Version 1.0–11. Retrieved from https://CRAN.R-project.org/package=NPCD

Zhu, Z. (2023). International comparative study of learning trajectories based on TIMSS 2019 G4 data on cognitive diagnostic models. *Frontiers in Psychology*, *14*, 1241656. https://doi.org/10.3389/fpsyg.2023.1241656

Appendix A.

Summary of Studies on the Application of DCMs to Large-Scale Mathematics Assessments

Study	Test	Attributes	Models Used	Software Used
Lee et al. (2011)	TIMSS (4th grade)	Whole numbers, fractions and decimals, number sentences with whole numbers, patterns and relationship, lines and angles, two- and three-dimensional shapes, location and movement, reading and interpreting, and organization and representing	DINA	Ox (Doornik, 2007)
Su et al. (2013)	TIMSS 2003 (8th grade)	Ratios and rates, multi-digit computation, rational numbers, algebraic expressions, equations and inequalities, proportional relationships, fraction operations, problem solving, fraction comparison, and compare two fractions	DINA-H DINO-H	CDM R-package (Robitzsch et al., 2011)
Choi et al. (2015)	TIMSS 2003 (8th grade)	Understand numbers and their representations, know number relationships and systems, understand the meaning and relationships of operations, compute fluently and estimate effectively, understand patterns, relations, and functions, represent and analyze math using algebraic symbols, use mathematical models to analyze relationships, analyze geometric shapes and properties, use coordinate geometry and transformations, apply spatial reasoning and geometric modeling, understand measurable attributes and units, apply techniques and tools for measurement, and understand basic probability concepts	DINA	Ox (Doornik, 2007)
Kunina-Habenicht et al. (2017)	German national large-scale assessment (4th grade)	Addition, subtraction, multiplication, division, modeling skills, skills in using measurement units	LCDM, Unidimensional and multidimensional IRT	Mplus (Muthén & Muthén, 1998–2017).
Yamaguchi and Okada (2018)	TIMSS (4th grade)	Whole numbers, fractions and decimals, number sentences with whole numbers, patterns and relationship, lines and angles, two- and three-dimensional shapes, location and movement, reading and interpreting, and organization and representing	DINA, DINO, *A*-CDM, LLM, RRUM, G-DINA 2PL IRT, and 3PL IRT	CDM (Robitzsch et al., 2017), GDINA (Ma et al., 2017), and TAM (Robitzsch et al., 2018–2025) R-packages

Continued

Study	Test	Attributes	Models Used	Software Used
Terzi and Sen (2019)	TIMSS 2011 (8th grade)	Understands fraction equivalence and ordering; understands decimals; understands ratios and percents; uses arithmetic and algebraic expressions to solve problems; solves one-variable equations and inequalities; solves linear equations and systems of linear equations; uses whole-number operations and identifies arithmetic patterns; draws and describes geometric figures and their relationships; solves problems involving angles, area, surface area, and volume; understands congruence and similarity using models or software; understands perimeter and area and relates area to operations; represents and interprets data; compares two populations; analyzes chance and develops probability models.	G-DINA, DINA, DINO, and *A*-CDM	Ox (Doornik, 2007)
Wu et al. (2020)	PISA	Change and relationships, space and shape, quantity, uncertainty and data, mathematization, mathematical operation, mathematical reality, personal, occupational, societal, and scientific	DINA, DINO, RRUM, *A*-CDM, LLM, LCDM, G-DINA, and mixture models	GDINA R-package (Ma et al., 2020)
Delafontaine et al. (2022)	TIMSS 2011 (8th grade)	Whole numbers and integers; fractions, decimals and proportions; patterns; expressions, equations and functions; lines, angles and shapes; measurement; location and movement; data organization, representation and interpretation; probability	G-DINA and DINA	GDINA (Ma et al., 2020) and NPCD (Zheng et al., 2019) R-packages
Sun et al. (2023)	TIMSS 2015 (4th grade)	Recall, recognize, classify, order, compute, retrieve, measure, determine, represent/model, implement, analyze, integrate/synthesize, evaluate, draw conclusions, generalize, justify	Interpretive structural modeling (ISM)	MATLAB (2022)
Xu et al. (2023)	TIMSS (4th grade)	Recall/recognize, classify/order, compute/measure, retrieve/solve routine problems, represent/model, analyze/integrate, generalize/ justify	DINA	GDINA R-package (Ma et al., 2022)
Zhu (2023)	TIMSS (4th grade)	Whole numbers, expressions, simple equations, and relationships, fractions and decimals, measurement, geometry, reading, interpreting, and representing, using data to solve problems, knowing, applying, and reasoning.	GDM	GDINA R-package (Ma et al., 2022)

Continued

Study	Test	Attributes	Models Used	Software Used
Saso et al. (2024, Preprint)	Japanese standardized achievement test (8th grade)	Processing mathematical terms, understanding the meanings and concrete examples of mathematical terms, formulation and application of equations and formulas, understanding the meanings of procedures and equations, and calculation skills	G-DINA, pG-DINA, and AHM	Julia programming language (Bezanson et al., 2017)
Zhang et al. (2025)	PISA	Mathematical abstraction, logical reasoning, mathematical modeling, intuitive imagination, mathematical operation, and data analysis.	DINA, DINO, RRUM, *A*-CDM, LCDM, LLM and Mixed Model	CDM R-package (Robitzsch et al., 2024)
Sonnleitner et al. (2026)	Luxembourgish school monitoring program (1st grade)	counting, addition <10, addition >10, decomposition	*A*-CDM, Single- and multiple-attribute hierarchical model (SAHM and MAHM)	CDM R-package (Robitzsch et al., 2024)

Appendix B.

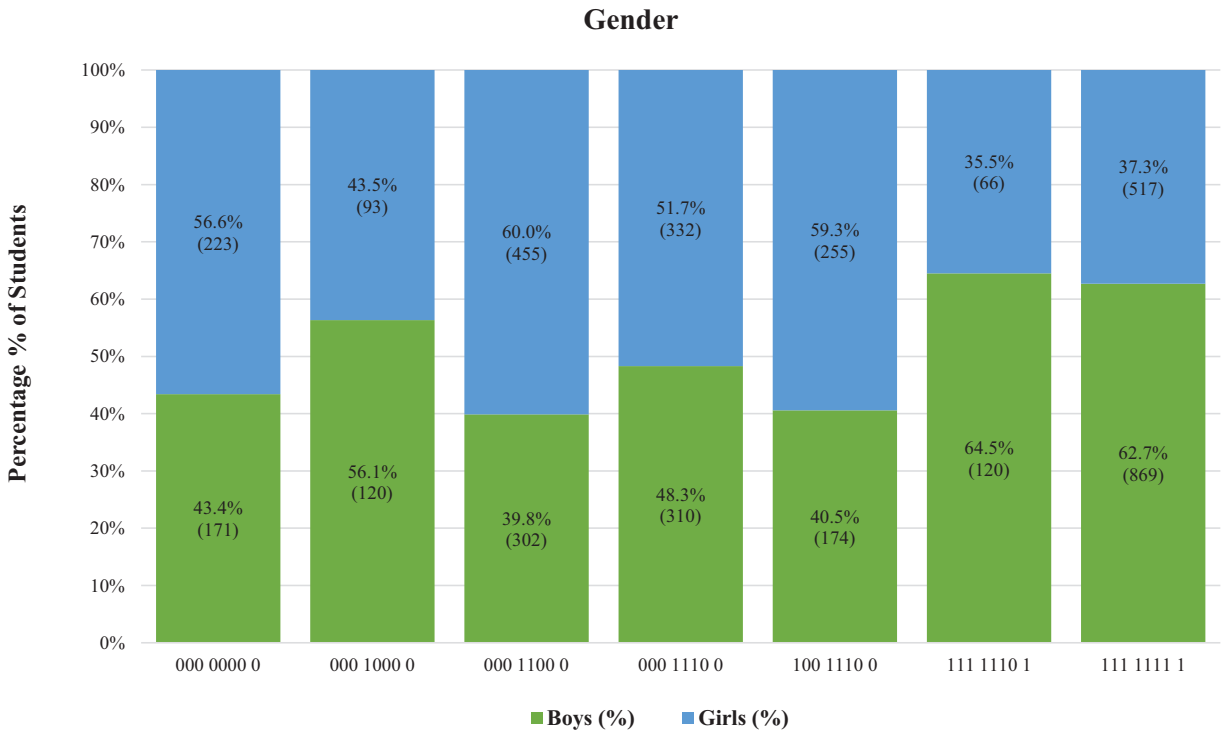
Model–Data Fit Indices at Item Level for (Non)Hierarchical DCMs

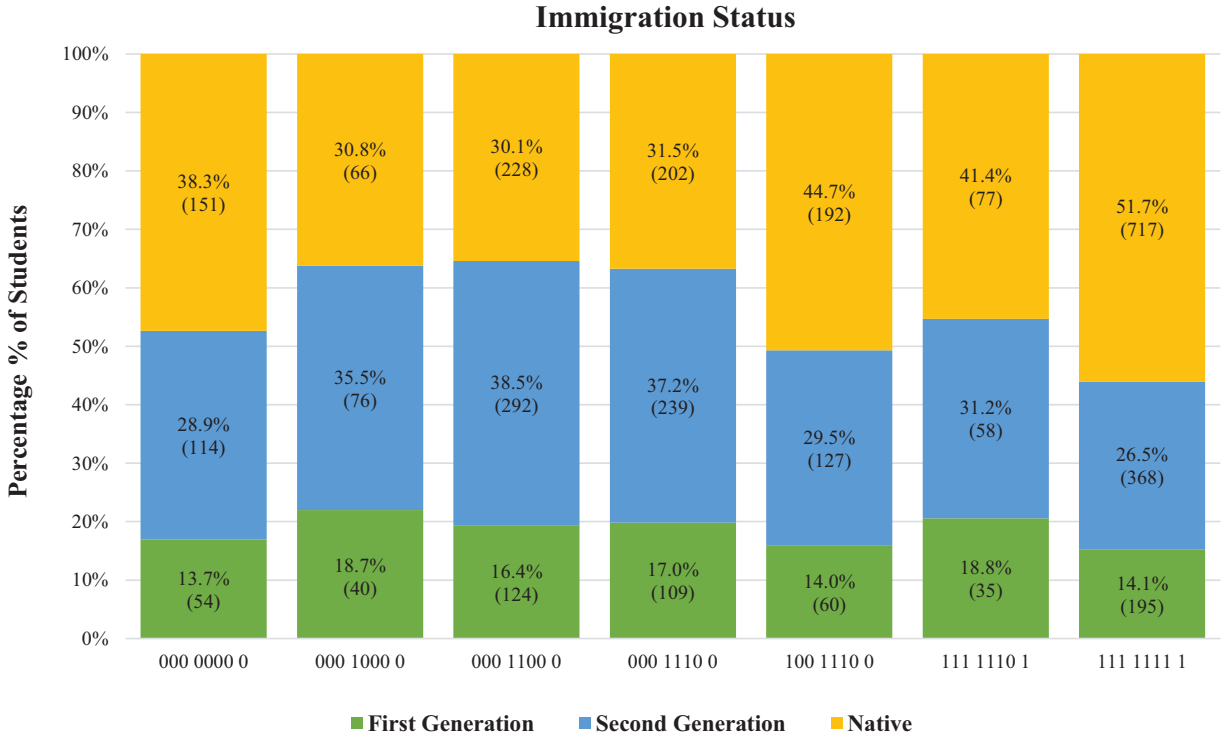
		RMSD				RMSD (Bias-Corrected)				MD				MAD				Mean of RMSEA
	Models	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max	
Conventional DCMs	G-DINA	0.085	0.040	0.028	0.281	0.084	0.039	0.027	0.281	−0.005	0.018	−0.193	0.011	0.058	0.033	0.015	0.199	0.120
	A-CDM	0.082	0.038	0.026	0.201	0.082	0.038	0.025	0.200	−0.001	0.002	−0.013	0.005	0.056	0.033	0.015	0.167	0.117
	C-RUM	0.082	0.039	0.025	0.205	0.081	0.039	0.024	0.204	−0.001	0.004	−0.014	0.025	0.056	0.034	0.015	0.171	0.117
	RRUM	0.083	0.039	0.025	0.202	0.082	0.039	0.024	0.201	−0.001	0.004	−0.014	0.030	0.056	0.034	0.016	0.169	0.117
	DINA	0.085	0.043	0.023	0.204	0.084	0.043	0.022	0.203	−0.002	0.006	−0.057	0.007	0.058	0.040	0.011	0.184	0.121
	DINO	0.078	0.037	0.022	0.166	0.077	0.037	0.022	0.165	−0.001	0.003	−0.012	0.012	0.054	0.030	0.008	0.128	0.111
Hierarchical DCMs	HDCM_G-DINA	0.122	0.066	0.000	0.640	0.121	0.066	0.000	0.640	−0.042	0.065	−0.542	0.110	0.090	0.056	0.000	0.554	0.173
	HDCM_*A*-CDM	0.065	0.034	0.018	0.176	0.064	0.034	0.016	0.175	−0.001	0.002	−0.011	0.006	0.047	0.028	0.011	0.149	0.092
	HDCM_C-RUM	0.067	0.037	0.020	0.179	0.066	0.037	0.020	0.178	0.001	0.006	−0.008	0.037	0.049	0.032	0.012	0.152	0.096
	HDCM_RRUM	0.068	0.036	0.020	0.181	0.066	0.036	0.019	0.180	0.002	0.011	−0.009	0.090	0.049	0.031	0.013	0.155	0.096
	HDCM_DINA	0.070	0.034	0.010	0.175	0.069	0.035	0.009	0.173	−0.001	0.002	−0.008	0.007	0.050	0.029	0.007	0.149	0.099
	HDCM_DINO	0.065	0.035	0.010	0.171	0.064	0.035	0.009	0.170	−0.001	0.003	−0.011	0.009	0.047	0.028	0.003	0.139	0.093

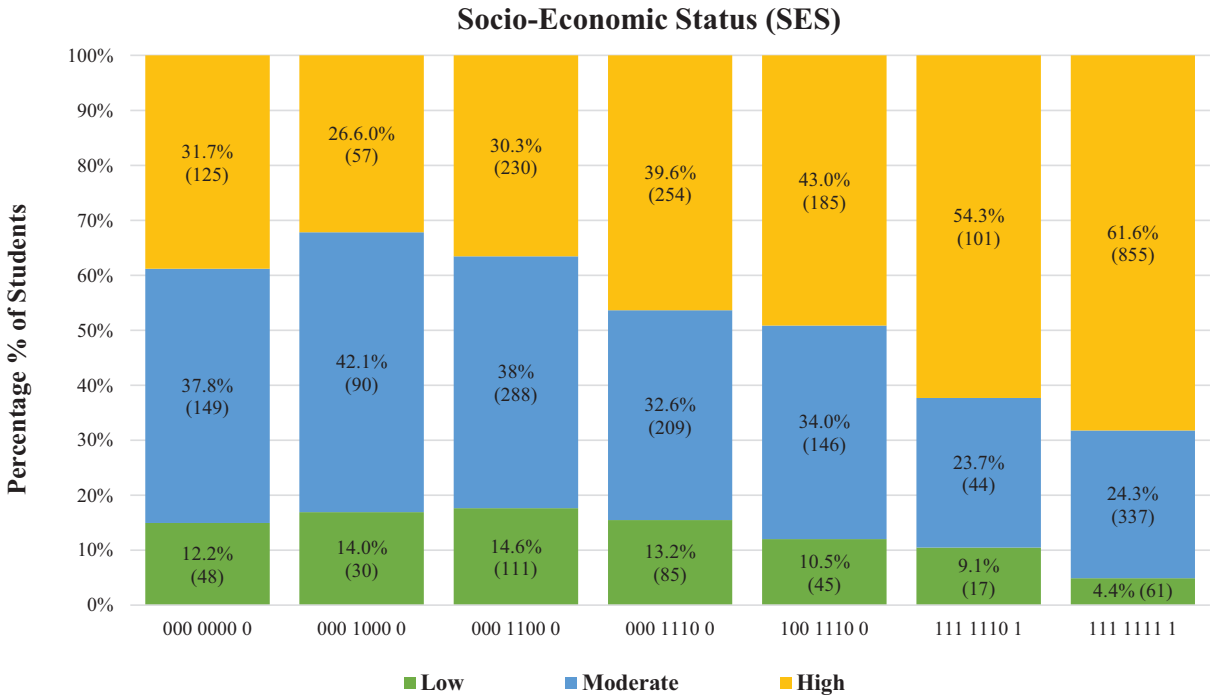
Note. Values are based on one group.
RMSD = Root Mean Square Deviation; MD = Mean Deviation; MAD = Mean Absolute Deviation; RMSEA = Root Mean Square Error of Approximation; *M* = Mean; *SD* = Standard Deviation; Min = Minimum; Max = Maximum.

Appendix C.

Characteristics of the Seven Most Prevalent Latent Classes of the HDCM_A-CDM







Appendix D.

Tetrachoric Correlations among Mathematics Attributes from the HDCM_A-CDM

Model		SNP	HNR	UOM	ADS	SUS	ASC	MUD	MOD
HDCM_ *A*-CDM	**SNP**	1							
	HNR	0.997	1						
	UOM	0.994	0.851	1					
	ADS	0.654	0.609	0.442	1				
	SUS	0.634	0.597	0.432	0.998	1			
	ASC	0.689	0.651	0.653	0.991	0.995	1		
	MUD	0.835	0.753	0.875	0.977	0.982	0.691	1	
	MOD	0.948	0.857	0.934	0.449	0.430	0.440	0.841	1

Note. SNP = Sequencing and Number Patterns; HNR = Dealing with Higher Number Range; UOM = Units of Measurement; ADS = Addition Simple; SUS = Subtraction Simple; ASC = Addition Complex and Subtraction Complex; MUD = Multiplication/Division; MOD = Modeling.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:
Supporting Information