Psychometrika (2026), **91**, 1524–1569
doi:10.1017/psy.2026.10126

PSYCHOMETRIC SOCIETY

APPLICATION AND CASE STUDIES - ORIGINAL

Predicting Multilevel Growth Trajectories Using a Random-Effect Diagnostic Classification Model

Kazuhiro Yamaguchi[1], Haruhiko Mitsunaga[2], Shun Saso[3] and Yuri Uesaka[3]

[1]University of Tsukuba, Japan; [2]Nagoya University, Japan; [3]The University of Tokyo, Japan

Corresponding author: Kazuhiro Yamaguchi; Email: kazz530@gmail.com

(Received 13 December 2024 UTC; revised 26 May 2026 UTC; accepted 1 June 2026 UTC)

Abstract

Learning diagnosis is essential for effective education, with formative assessments shown to significantly enhance academic performance. Diagnostic classification models (DCMs) have been developed to assess students' learning status and provide remedial instruction. However, the impact of mastery or non-mastery of specific attributes on long-term learning development remains uncertain. If certain non-mastered attributes hinder the growth of mathematical ability, early intervention becomes essential. In this study, we developed a random-effects DCM for multilevel growth curves (RDC–MGC) model to identify the specific effects of attribute mastery on individual-level mathematics ability growth. The simulation studies showed that the Bayesian estimation procedure provided appropriate parameter recovery and coverage probabilities, whereas ignoring the multilevel structure resulted in biased parameter estimates. The model was applied to arithmetic test data from second- to sixth-grade elementary school students. Diagnosis was conducted in the second grade, and the effects of mastery on mathematics ability growth from the third to sixth grades were assessed. The results showed that attribute mastery in second grade was associated with both the intercept and slope of individual ability growth, suggesting the potential importance of early-stage diagnostic information for understanding later mathematical development. Potential extensions of the proposed RDC–MGC model are also discussed.

Keywords: Bayesian estimation; diagnostic classification models; growth curve model; mathematical ability; multilevel analysis

1. Introduction

Learning diagnosis, the process of identifying a learner's strengths, weaknesses, and specific gaps in knowledge or skills, is a crucial component of effective learning, as formative assessment significantly improves academic performance (*Effect size* = 0.40, *SE* = 0.06; Hattie, 2023, p. 333). Diagnostic classification models (DCMs; e.g., Rupp et al., 2010; von Davier & Lee, 2019), a family of statistical models, have been applied across various subjects to achieve this purpose. For example, Tatsuoka and Tatsuoka (1997) conducted a computer-adaptive diagnostic test for fraction–subtraction problems, demonstrating that remediation instructions based on diagnostic results improved attribute mastery. Similarly, mathematics data from the Trends in International Mathematics and Science Study (TIMSS) were analyzed using DCMs, revealing differences in attribute mastery status across countries (Lee et al., 2011; Yamaguchi & Okada, 2018). von Davier (2008) applied a general diagnostic model (GDM) to analyze Test of English

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

for International Communication (TOEIC) Internet-based testing (iBT) Reading and Listening data from two test forms, predicting mastery of reading skills (e.g., synthesizing and organizing) and listening skills (e.g., making inferences and connections) based on general abilities. Additionally, Chen and de la Torre (2014) analyzed the Programme for International Student Assessment (PISA) reading data using the generalized deterministic input noisy AND-gate (G-DINA; de la Torre, 2011) model and its sub-models, demonstrating that main effect or saturated models might effectively fit the reading data. These examples highlight the versatility of DCMs in various test applications.

DCMs have been developed to extract information for formative assessment from test data. For example, Ayan and Çıkrıkçı (2021) employed DCMs for math proficiency formative assessment, highlighting the main difference between DCMs and item response theory models (IRT models; e.g., Embretson & Reise, 2000; de Ayala, 2013). IRT models assume a general uni- or multi-dimensional continuous latent trait to order individuals, facilitating the estimation of comparable test scores among test takers, even when different sets of test items are utilized. Horizontal equating or vertical scaling refers to the construction of a common scale from different sets of test items (e.g., Kolen & Brennan, 2014). Therefore, IRT models are appropriate for high-stakes tests whose results have a substantial impact on test takers. DCMs, however, assume discrete latent traits such as the mastery or non-mastery of cognitive elements, termed attributes, to answer a test. These are formally introduced subsequently. DCMs estimate the mastery/non-mastery of latent traits for the following topics that students are required to learn. In other words, DCM results are employed for learning remediation or diagnosis of students' current learning status.

DCMs are primarily used to diagnose whether specific cognitive elements, known as attributes, in a test are mastered by an individual. For example, Tatsuoka and Tatsuoka (1997) developed a remediation selection method based on Mahalanobis distance (Figure 2 in Tatsuoka & Tatsuoka, 1997). In the fraction–subtraction example, students who fail to master attributes such as A5 (reducing the fraction part before getting the common denominator) and A8 (adjusting the whole number part), presented in Table 1 of Tatsuoka and Tatsuoka (1997), were recommended remediation for A5 owing to its proximity to the next attribute mastery pattern. However, this rule-space-based method does not apply to other types of DCMs.

Yamaguchi and Templin (2022c) proposed the use of Hasse diagrams to represent attribute mastery and corresponding observed test scores, demonstrating attributes that provide the largest score changes. Nevertheless, this method is closely related to a specific diagnostic test and lacks generalizability. Moreover, it does not identify the attributes most crucial for future learning. Prior achievement strongly influences subsequent achievement (*Effect size* = 0.73, *SE* = 0.07; Hattie, 2023, p. 79), indicating that the importance of mastering specific attributes may vary for advancing learning. Using feedback based on DCMs, Zhan et al. (2022) provided a framework to combine eye-tracking and problem-solving behavior data to improve feedback and diagnosis about problem-solving skills. They reported empirical applicability and utility of the proposed framework.

The long-term implications of attribute mastery are of interest to teachers or educational psychologists. In particular, identifying cognitive abilities that can predict learning trajectories and determining methods through which these relationships can be effectively modeled are key considerations. Although DCMs can predict the importance of attributes on learning trajectories, a methodology to achieve this has not yet been proposed. To address this issue, we propose an integrated modeling approach for examining how diagnostic attribute mastery predicts subsequent learning trajectories.

Learning trajectories are typically represented by a latent growth curve model, or simply a growth curve (GC) model (Muthén & Muthén, 2000; Oravecz & Muth, 2018; Preacher et al., 2008). The GC model can be conceptualized as a multilevel model where the individual forms level 2, and measurement time constitutes level 1 (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Using the structural equation model (SEM), the GC model can be formulated as a standard latent trait model (Newsom, 2015), which is a single-level model. Furthermore, the SEM can be extended to a multilevel model within educational measurement contexts, referred to as the multilevel SEM (Heck & Reid, 2023; Pritikin et al., 2017; Rabe-Hesketh et al., 2004). Therefore, even the multilevel GC (MGC) model is a three-level

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

model, the multilevel SEM can reduce it as a two-level model. More importantly, this approach enables the representation of group-level (e.g., school-level) trajectories, facilitating the separation of group- and individual-level changes. Consequently, purified versions of individual learning trajectories can be obtained using an MGC model.

Longitudinal growth has also been considered in the context of DCMs, as reported by Yamaguchi and Martinez (2024). Major longitudinal DCMs include the hidden Markov model (HMM; e.g., Bishop, 2006) and latent transition model (LTM; e.g., Collins & Lanza, 2009) type DCMs. HMM- and LTM-type DCMs that model the transition probabilities of attribute masteries between time points are also major approaches. Madison and Bradshaw (2018) proposed a latent transition DCM that is an extension of LTM and explained its estimation based on maximum likelihood with Mplus (Muthén & Muthén, 1998–2017). Wang (2021) developed a penalized expectation–maximization estimation method. Using Bayesian approaches, Wang et al. (2018) developed an HMM-type DCM to attribute-level mastery probability based on covariates and previous attribute mastery status, and its estimation was conducted using a Markov chain Monte Carlo (MCMC) method. Pan et al. (2020) developed multivariate longitudinal DCMs that assumed GC models on attribute mastery probability and multivariate normal distribution on the intercept and slope factors. Zhang and Chang (2020) employed a similar approach to that of Wang et al. (2018) and incorporated random effects on measurement models from learning tools such as exercises, slides, videos, or games that affect attribute mastery.

Previous studies have several limitations. First, all tests are conducted over several time points using DCM as measurement models. In other words, they are interested in mastery status change in attributes over time. Therefore, the model is closed in DCMs. This implies that effective attributes that contribute to long-term improvements in learning cannot be determined using these models. Second, some studies have included random effects in their models. However, these effects are not caused by an individual-level effect, but rather by a group effect, which is naturally introduced by a multilevel data structure. Therefore, attribute masteries are not inherent to individuals because they contain a group-level effect. These models are not suitable for detecting attributes that are essential for long-term general proficiency in individual-level development. To address these issues, we investigate how diagnostic attribute mastery can be incorporated into multilevel growth modeling.

Furthermore, multilevel considerations are equally important for DCMs. Test performance is influenced not only by attribute mastery but also by group-level (e.g., school-level) differences. Construct irrelevant factors, such as group-level variations, may distort the relationship between cognitive elements and item responses. This is similar to the attenuation of correlation coefficients in classical test theory modeling. For example, compared with students at other types of schools, those at preparatory schools may exhibit higher correct item response probabilities even if mastering no attributes is required in the test. However, it may lead to an inappropriate diagnosis of attribute mastery because test answers contain a nuisance element. Therefore, such effects must be accounted for in the application of DCMs.

Multilevel DCMs can incorporate group-level effects as random intercepts within the standard multilevel models. However, previous DCMs have substantially overlooked this straightforward random effect. Higher-order models (de la Torre & Douglas, 2004; Ma, 2022; Zhan, Jiao, Liao, et al., 2019) primarily consider continuous latent traits, similar to IRT models, for attributes. These approaches often utilize multilevel IRT models (Fox, 2010; Fox & Glas, 2001; Wang & Qiu, 2019) for higher-order models in DCMs (Huang, 2017; Wang & Qiu, 2019). Only von Davier (2007) focused on group effects on item response functions. However, major approaches are inconsistent with the categorical nature of latent abilities, which are considered discontinuous or categorical variables in DCMs. To address this issue, the present study directly incorporates group effects as a random effect.

We summarize the limitations of previous studies as follows: (1) previous DCM studies paid limited attention to the use of diagnostic information for predicting subsequent academic growth trajectories and (2) multilevel structures in diagnostic assessments and subsequent tests were not fully incorporated, even though ignoring such structures may lead to biased parameter estimation. To address these limitations, this study proposes an integrated modeling approach that accounts for random group effects for examining individual learning trajectories using a multilevel GC model.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

More specifically, we integrated existing DCM, SEM, and multilevel modeling approaches to formulate the proposed model. We also assessed the performance of the Bayesian estimation procedure for this model and demonstrated the risk of ignoring multilevel structures through Monte Carlo simulation studies. Furthermore, we conducted an empirical analysis using data from second to sixth grade in elementary schools. Importantly, we investigated how second-grade learning states predict subsequent developmental trajectories across later grades. Academic abilities, specifically mathematics/arithmetic skills, were measured using a vertically linked scale. At the second-grade level, students' attribute mastery is estimated through a diagnostic test. From grades three to six, the GC model is applied to assess their progress. The formal specification of this setting is presented below.

Our contributions are as follows. First, the proposed model provides a perspective on how diagnostic information can be used: the use of attribute mastery profiles for examining subsequent learning trajectories. Traditionally, diagnostic information is employed for learning remediation. However, single-time-point diagnostic classification results do not indicate which attributes are important for later learning. The proposed model provides a possible modeling approach for this question. On the other hand, SEM and multilevel modeling approaches have generally not incorporated discrete latent mastery profiles for predictive purposes. Second, the proposed model combines DCMs, SEM, and multilevel models. This model can incorporate group-level variation as a random effect in DCMs and use diagnostic results to examine individual-level development separately from group-level change. Therefore, the proposed model allows examination of the long-term consequences of attribute mastery for subsequent academic development at the individual level while separating group-level effects. Without separating group- and individual-level growth, it is difficult to consider the effect of attribute mastery on later academic proficiency growth. This distinction has received limited attention in the DCM literature. Although the distinction between group-level and individual-level effects has received limited attention, it should be considered in educational testing settings. In addition, because individual developmental trajectories may not always be linear, the SEM-based formulation permits more flexible representations of longitudinal growth. To illustrate this point, we additionally illustrate a partially freely estimated slope-loading specification (Bollen & Curran, 2006, Chapter 4) within the current modeling setting.

The remainder of this article is structured as follows. Section 2 outlines the fundamental concepts of the random-effect DCM for the MGC (RDC–MGC) model. This section introduces the measurement component of the random-effect DCM (RDCM), the MGC model, and the structural model. The posterior distribution of the RDC–MGC is also formulated for Bayesian estimation, which is implemented through an MCMC method. Section 3 shows simulation studies to assess the appropriateness of the estimation method and consequences of ignoring multilevel structure. Section 4 describes the details of the dataset, which comprises a five-year longitudinal mathematical assessment in elementary schools. Moreover, it explains the Q-matrix and attributes defined in the mathematics test, along with a brief discussion of the vertical scaling procedure. Section 5 discusses implications of the proposed model and considers future applications and extensions of the model.

2. Formulation of random-effect diagnostic classification model for multilevel growth curve model

2.1. Overview of the RDC–MGC model

Figure 1 shows the path diagram of the RDC–MGC model, comprising three key components: the RDC measurement model, MGC, and structural model. The first component is the RDC measurement model, which includes two elements: the conventional attribute-related component and group random intercept. Studies have developed various DCMs, including the deterministic input noisy AND-gate model (DINA; de la Torre, 2009; Junker & Sijtsma, 2001; Macready & Dayton, 1977; Maris, 1999), deterministic input noisy OR-gate model (DINO; Templin & Henson, 2006), noisy inputs deterministic AND-gate model (NIDA; Junker & Sijtsma, 2001), additive cognitive diagnostic model (A-CDM; de la Torre, 2011), linear logistic model (LLM; Maris, 1999), reduced reparametrized unified model (R-RUM; Hartz & Roussos, 2008; Roussos et al., 2007), log-linear cognitive diagnostic model (LCDM;

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

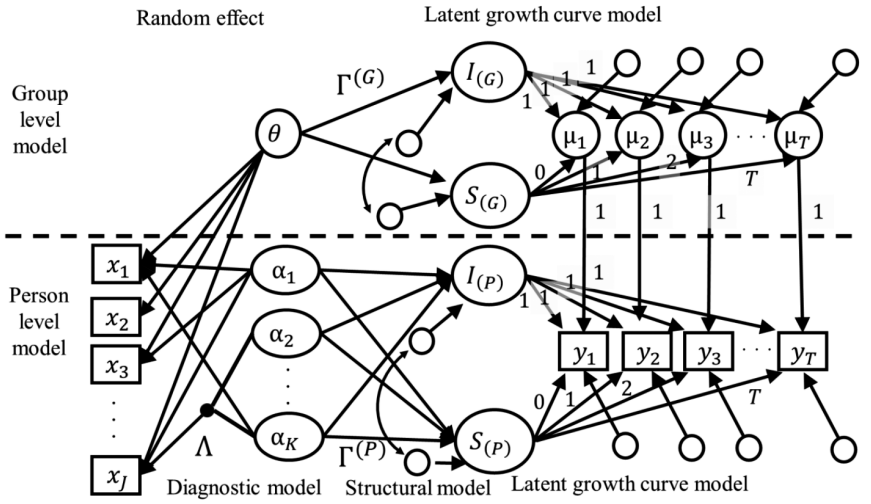


Figure 1. Path diagram of the proposed random-effects diagnostic classification model for multilevel growth curves.

Henson et al., 2009), G-DINA model 2011, and GDM (von Davier, 2008). This study adopts the LCDM-based model because its formulation supports the integration of multilevel effects, similar to multilevel IRT models (Fox, 2010; Huang & Wang, 2014).

The second component is the MGC model, a specialized form of multilevel SEM. It separates observed variables into group-level elements and individual-level components, enabling distinct modeling of these elements. Importantly, the individual growth trajectory derived from the MGC model excludes group-level changes, representing the inherent trajectory of the individual.

The final component is the structural model, which links attribute mastery to the individual intercept and slope of the GC, forming the individual-level model. This model explains how attribute mastery and its interactions account for individual differences in growth trajectories. Additionally, it captures the relationship between group-level differences which are modeled as a random intercept and group-level growth. Although this is a secondary focus, the model can be extended to include group-level covariates to account for variance in the intercept and slope at the group level. The details of these three components of the RDC–MGC model are presented in subsequent sections.

In our model, the multilevel growth model represents academic proficiency, which constitutes our outcome. We aim to construct a model representing the types of attributes that are significant for academic ability growth. Furthermore, we separated growth into two levels and therefore considered the growth between group and individual levels separately. The diagnostic assessment part, which is a single-point measurement, includes independent variables. However, previous DCMs do not consider the group-level baseline effect, which is a random effect. Moreover, they can distort attribute mastery. In other words, we considered the attribute masteries in previous single-level DCMs as the compound of individual attribute masteries and group-level proficiency. We explicitly formulated such random effects and attribute mastery effects that were fixed effects and usual DCM parameters. In summary, the random and fixed effects were assumed in the DCM measurement part. From the above discussion, the most interesting part of this model is the individual structural part that represents the relationship between purified attribute mastery and the growth of academic proficiency. The group-level structural part will be interesting if some group-level covariates can predict the level of growth in academic proficiency. A wider variety of empirical evaluations of group-level structures will be an important topic for future research.

The basic data structure in this model is that the diagnostic measurement is conducted at a single time point, and the attribute mastery patterns predict the growth trajectory. In this case, we assumed

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

the research question to be about which kinds of attributes and their combinations are influential for subsequent learning. This is just an example of research. If we want to use longitudinal DCMs during the first period, we use such a transition as the diagnostic measurement part, and the final mastery pattern can be used as a predictor for the growth trajectory. In addition, we can simultaneously include both a diagnostic model and a usual IRT model in the first measurement phase if we prepare both a diagnostic test and an achievement test at that point. In this case, both continuous latent proficiency and attribute mastery are employed as predictors. Furthermore, the outcome growth model can be assumed for latent proficiency, which is a type of longitudinal IRT model (e.g., Wang & Nydick, 2020) or a second-order factor analysis model (e.g., Wickrama et al., 2016). In this case, tests at different time points should include anchor items to ensure that the same continuous proficiency is being measured. In the current study, for the simplicity of the model, we employed estimated academic proficiency based on vertically scaled tests.

2.2. Measurement component of RDCMs

The RDCM can be expressed in various formulations, as multilevel latent class models can be specified in various ways (e.g., Henry & Muthén, 2010; Vermunt, 2003, 2008). Therefore, RDCM assumes a two-stage sampling procedure for the diagnostic measurement, in which level two units such as schools are randomly sampled first, then the level one units such as students are randomly sampled. Additionally, as previously discussed, numerous DCMs have been developed. In this study, we adopted the LCDM formulation with a random effect. The LCDM defines the probability of a correct item response based on the attribute mastery pattern, model parameters, and Q-matrix.

First, we assume $K \in \mathbb{N}$ attributes for the test, where the subscript $k \in \{1, 2, \ldots, K\}$ represents the attribute index. Subsequently, the $l\left(\in \left\{1, 2, \ldots, L = 2^K\right\}\right)$-th attribute mastery pattern is represented as a latent binary vector of length K: $\boldsymbol{\alpha}_l \in \{0, 1\}^K$. Its k-th element is $\alpha_{lk} \in \{0, 1\}$ representing non-mastery (= 0) and mastery (= 1) of the attribute. For individual $i \in \{1, 2, \ldots, I \in \mathbb{N}\}$, whose mastery status corresponds to $\boldsymbol{\alpha}_l$, we denote this as $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l$. Based on the multilevel structure, individuals (level one unit) are exchangeable given a school (level two unit).

Furthermore, $\boldsymbol{q}_j \in \{0, 1\}^K \backslash 0_K$, where j is an element of the set $\{1, 2, \ldots, J \in \mathbb{N}\}$, and 0_K is the length K of the zero vector, which is a q-vector. The k-th element $q_{jk} \in \{0, 1\}$ is set to 1 if the k-th attribute is measured by item j and 0 otherwise. Using these definitions, the probability of a correct response for individual i belonging to the $m\,(= \{1, 2, \ldots, M \in \mathbb{N}\})$-th group on the test item j is given by

$$\mathrm{P}\left(X_{ijm} = x_{ijm} | \boldsymbol{\alpha}_{im}, \boldsymbol{\lambda}_j, \theta_m, \boldsymbol{q}_j\right) = \frac{\exp\left(f\left(\boldsymbol{\alpha}_{im}, \boldsymbol{\lambda}_j, \theta_m, \boldsymbol{q}_j\right)\right)}{1 + \exp\left(f\left(\boldsymbol{\alpha}_{im}, \boldsymbol{\lambda}_j, \theta_m, \boldsymbol{q}_j\right)\right)}, \tag{1}$$

where $\boldsymbol{\lambda}_j = \{\lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{j12\ldots K}\}$ and

$$\begin{aligned} &f\left(\boldsymbol{\alpha}_{im}, \boldsymbol{\lambda}_j, \theta_m, \boldsymbol{q}_j\right) \\ &\quad = \theta_m + \lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk} q_{jk} \alpha_{imk} + \sum_{k=1}^{K} \sum_{k'<k} \lambda_{jkk'} q_{jk} q_{jk'} \alpha_{imk} \alpha_{imk'} + \cdots + \lambda_{j12\ldots K} \prod_{k=1}^{K} q_{jk} \alpha_{imk}. \end{aligned} \tag{2}$$

The parameter $\boldsymbol{\lambda}_j$ is a vector-fixed effect; therefore, it does not change over school and includes the intercept λ_{j0}, main effects of mastering individual attributes such as λ_{jk}, and interaction effects among attributes, denoted as $\lambda_{j12\ldots K}$. The parameter λ_{j0} corresponds to the item response probability of all non-mastery patterns when the average group (i.e., $\theta_m = 0$). Therefore, it determines the baseline item response probability. If the intercept is negative, set $\theta_m = 0$; the item response probability of the baseline is less than 0.5, which is desirable. The main effect λ_{jk} denotes the increase in the logit of mastering k-th attribute. In DCMs, the main effect is positive because of the monotonicity constraint (e.g., Henson et al., 2009; Yamaguchi & Templin, 2022a). The interaction effect $\lambda_{jkk'}$ determines the probability of an individual simultaneously mastering attributes k and k'. Similarly, the highest interaction term $\lambda_{j12\ldots K}$

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

is the effect on item response probability when an individual masters all attributes required by the j-th item if $\prod_{k=1}^{K} q_{jk} = 1$. A key feature of the RDCM is the random effect θ_m, which is an element of an M-length vector $\boldsymbol{\theta} \in \mathbb{R}^M$, assumed to follow a normal distribution $N\left(0, \sigma_\theta^2\right)$. The random effect modifies the interpretation of the intercept λ_{j0}, which represents the probability of a correct item response for individuals with no attribute mastery in an average group.

Furthermore, in the multilevel IRT model literature, such as Fox and Glas (2001, p. 283), it is indicated that the random intercept is defined as an expected school achievement when controlling other variables. Considering this interpretation, if the RDC–MGC includes covariates of the school or individuals, its random intercept leads to the expected group achievement. The simplest version is the no-covariate case, which is mainly employed in the empirical study, and the random intercept represents school differences in the diagnostic test. Furthermore, the diagnostic measurement models include attribute mastery status, and its intercepts represent the correct response probabilities of the all no-mastering pattern, which is the baseline. Therefore, the random intercept in the simplest version of the RDC–MGC model is the group difference of all the non-mastery patterns. This group difference may explain the guessing probability in the DCM.

Note that the random effect θ_m can formally be a uniform DIF factor. Basically, DIF systematically changes some item response functions caused by group differences, and its factor is represented as a fixed effect. Li and Wang (2015) proposed a unified formulation of DIF in DCMs. Li and Wang (2015, pp. 30–31) reported that using latent proficiencies as grouping factors can cause DIF. Further, Camilli (2006, p. 241) mentioned that DIF can be caused by secondary distribution. For numerous groups, the DIF term can be modified as a random effect in an item (Camilli, 2006, pp. 242–243). Therefore, technically and formally, the group-level proficiency θ_m is a special case of the DIF factor that affects the same difficulty shift on the all-test items.

However, the group-level factor θ_m and usual DIF factor exhibit certain differences. The DIF is assumed for each item, and the group difference of the intercepts can vary across items. Even when the DIF factor is a random effect, it is assumed for items that may be different among groups. In addition, in DIF consideration, the number of assumed groups is relatively small. The group-level factor θ_m is shared for all items and serves as a common factor. Method factors may be similar to the group-level factor θ_m, but they model correlations among sub-items caused by item traits. The random intercept in our study can be considered a group-level difference parameter, as mentioned and can predict the intercept and slope of the growth model part, which will be introduced subsequently, in the real-data analysis. From these points, the random effect may be a group-level proficiency rather than just the uniform DIF.

Note that, although the LCDM formulation is employed, the distribution of attribute mastery should be specified at the attribute mastery level, consistent with latent class formulations (Yamaguchi & Okada, 2020; Yamaguchi & Templin, 2022a, 2022b). This implies that attribute mastery patterns are latent classes rather than a decomposition of each attribute mastery. For example, when the number of attributes is $K = 3$, $L = 2^K = 2^3 = 8$ attribute mastery patterns, $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$, $\ldots$, $(1, 1, 1)$, were assumed. The eight attribute mastery patterns are followed by a categorical distribution. In this formulation, the attributes need not be independent. Specifically, we represent an individual i in group m having a specific attribute mastery pattern $\boldsymbol{\alpha}_l$ as $\mathcal{I}\left(\boldsymbol{\alpha}_{im} = \boldsymbol{\alpha}_l\right)$, which is an indicator function taking value 1 if its argument is true, and 0 otherwise, and its probability of taking value 1 is π_l, satisfying $\sum_l \pi_l = 1$. Therefore, the probability distribution of $\mathcal{I}\left(\boldsymbol{\alpha}_{im} = \boldsymbol{\alpha}_l\right)$ is a categorical distribution:

$$P\left(\boldsymbol{\alpha}_{im} = \boldsymbol{\alpha}_l | \boldsymbol{\pi}\right) = \prod_{l=1}^{L} \pi_l^{\mathcal{I}\left(\boldsymbol{\alpha}_{im} = \boldsymbol{\alpha}_l\right)}. \tag{3}$$

Furthermore, we assume that the exchangeability of individuals results in the joint distribution of $\mathcal{A} = \left\{\boldsymbol{\alpha}_{im}\right\}_{i,\, m=1}^{I_m,\, M}$ can be expressed as

$$P\left(\mathcal{A} | \boldsymbol{\pi}\right) = \prod_{m=1}^{M} \prod_{i=1}^{I_m} \prod_{l=1}^{L} \pi_l^{\mathcal{I}\left(\boldsymbol{\alpha}_{im} = \boldsymbol{\alpha}_l\right)}. \tag{4}$$

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

From these assumptions, the joint distribution of data $\mathrm{X} = \{x_{ijm}\}_{i,j,m=1}^{I_m,J,M}$, latent attribute mastery pattern $\mathcal{A}$, and group random effect $\boldsymbol{\theta}$ is given by

$$\begin{aligned} P\left(\mathrm{X}, \mathcal{A}, \boldsymbol{\theta} | \Lambda_X, \boldsymbol{\pi}, \sigma_\theta^2\right) &= P\left(\mathrm{X} | \Lambda_X, \mathcal{A}, \boldsymbol{\theta}\right) P\left(\mathcal{A} | \boldsymbol{\pi}\right) P\left(\boldsymbol{\theta} | \sigma_\theta^2\right) \\ &= \left(\prod_{m=1}^{M} \prod_{i=1}^{I_m} \left\{ P\left(X_{ijm} = x_{ij} | \boldsymbol{\alpha}_{im}, \boldsymbol{\lambda}_j, \theta_m, \boldsymbol{q}_j\right) \prod_{l=1}^{L} P\left(\boldsymbol{\alpha}_{im} = \boldsymbol{\alpha}_l | \boldsymbol{\pi}\right) \right\}\right) P\left(\boldsymbol{\theta} | \sigma_\theta^2\right), \end{aligned} \quad (5)$$

where $\Lambda_X = \{\boldsymbol{\lambda}_j\}_{j=1}^{J}$. This will be employed subsequently.

2.3. Latent GC model component

We briefly outline the general notation for two-level SEM and then apply it to a two-level GC model. The general multilevel SEM has been extensively investigated in the literature (Depaoli, 2021; Depaoli & Clifton, 2015; Heck & Reid, 2023; Muthén, 1994; Rabe-Hesketh et al., 2004). The two-level GC model decomposes the observed variable as follows:

$$\boldsymbol{y}_{im} = \underbrace{\Lambda_Y^{(P)} \boldsymbol{\eta}_{im}^{(P)} + \boldsymbol{\varepsilon}_{im}^{(P)}}_{\text{Person-level model}} + \underbrace{\boldsymbol{\mu}^{(Ground)} + \Lambda_Y^{(G)} \boldsymbol{\eta}_m^{(G)} + \boldsymbol{\varepsilon}_m^{(G)}}_{\text{Group-level model}}, \quad (6)$$

where $\boldsymbol{y}_{im}, \boldsymbol{\varepsilon}_{im}^{(P)}$, and $\boldsymbol{\varepsilon}_m^{(G)}$ are the outcome vector of T time points and residual vectors for each level: $\boldsymbol{y}_{im} = (y_{1im}, \ldots, y_{tim}, \ldots, y_{Tim})^\top$, $\boldsymbol{\varepsilon}_{im}^{(P)} = \left(\varepsilon_{1im}^{(P)}, \ldots, \varepsilon_{tim}^{(P)}, \ldots, \varepsilon_{Tim}^{(P)}\right)^\top$, $\boldsymbol{\varepsilon}_m^{(G)} = \left(\varepsilon_{1m}^{(G)}, \ldots, \varepsilon_{tm}^{(G)}, \ldots, \varepsilon_{Tm}^{(G)}\right)^\top$. For example, $\boldsymbol{y}_{im}$ contains the academic test scores of an individual i belonging to the m-th school during elementary school period. Equation (6) indicates that such an individual academic score can be decomposed into the school effect part (group-level model) and the person's own part (person-level model). Note that the residual term vectors $\boldsymbol{\varepsilon}_{im}^{(P)}$ and $\boldsymbol{\varepsilon}_m^{(G)}$ are independent from another random variable such as $\boldsymbol{y}_{im}$. However, within the vector, the elements of $\boldsymbol{\varepsilon}_{im}^{(P)}$ or $\boldsymbol{\varepsilon}_m^{(G)}$ can be correlated just the same as usual longitudinal model assumptions. Therefore, for example, we can assume that $\boldsymbol{\varepsilon}_{im}^{(P)}$ or $\boldsymbol{\varepsilon}_m^{(G)}$ is generated from a multivariate normal distribution. However, for the simplicity of the model, we assume that the residuals independently follow different distributions which will be shown later.

Furthermore, consistent with the usual GC model, the latent variables represent the intercept and slope: $\boldsymbol{\eta}_{im}^{(P)} = \left(I_{im}^{(P)}, S_{im}^{(P)}\right)^\top$ and $\boldsymbol{\eta}_m^{(G)} = \left(I_m^{(G)}, S_m^{(G)}\right)^\top$. The upper scripts "$(P)$" and "$(G)$" represent person- and group-level variables, respectively. Therefore, $\boldsymbol{\eta}_{im}^{(P)}$ and $\boldsymbol{\eta}_m^{(G)}$ are the vectors of person-level intercept and slope of individual i in group m and group-level intercept and slope of group m, respectively. More intuitively, $\boldsymbol{\eta}_m^{(G)}$ represents the m-th school's specific growth rate and the initial value of the academic performance, and $\boldsymbol{\eta}_{im}^{(P)}$ corresponds to the individual's specific growth rate and initial status.

The covariance matrix of the intercept and slope can be represented as $\Sigma^{(\mathrm{P})}$ and $\Sigma^{(\mathrm{G})}$, respectively. All the elements of the ground mean vector $\boldsymbol{\mu}^{(Ground)} = \left(\mu_1^{(Ground)}, \ldots, \mu_t^{(Ground)}, \ldots, \mu_T^{(Ground)}\right)^\top$ are fixed to zero for identifiability in this study. Therefore, we set $\boldsymbol{\mu}^{(Ground)} = 0_T$ and omitted this vector for simplicity. Furthermore, we assume a linear GC model for both levels; the time score matrix is $\Lambda_Y^{(P)} = \Lambda_Y^{(G)} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T \end{pmatrix}$. Representing $\boldsymbol{\mu}_m^{(G)} = \Lambda_Y^{(G)} \boldsymbol{\eta}_m^{(G)} + \boldsymbol{\varepsilon}_m^{(G)} = \left(\mu_{1m}^{(G)}, \ldots, \mu_{tm}^{(G)}, \ldots, \mu_{Tm}^{(G)}\right)^\top$, which is the m-th school's average trajectory for the academic performance, the element wise person-level model is given by

$$\begin{aligned} y_{tim} &= \mu_{tm}^{(G)} + I_{im}^{(P)} + (t-1) S_{im}^{(P)} + \varepsilon_{tim}^{(P)}, \\ \varepsilon_{tim}^{(P)} &\sim N\left(0, \sigma_{\varepsilon_{tm}^{(P)}}^2\right), \end{aligned}$$

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

$$\begin{pmatrix} I_{im}^{(P)} \\ S_{im}^{(P)} \end{pmatrix} \sim MVN\left(\boldsymbol{\mu}_{(I,S)_{im}}^{(P)}, \Sigma^{(P)}\right), \tag{7}$$

where the person-level mean vector and covariance matrix of the intercept and slope are $\boldsymbol{\mu}_{(I,S)_{im}}^{(P)} = \left(\mu_{I_{im}^{(P)}}, \mu_{S_{im}^{(P)}}\right)^{\top}$ and $\Sigma^{(P)} = \begin{pmatrix} \sigma_{I^{(P)}}^2 & \sigma_{I^{(P)}S^{(P)}} \\ \sigma_{I^{(P)}S^{(P)}} & \sigma_{S^{(P)}}^2 \end{pmatrix}$, respectively. Here, the μ in the right-hand side of μ_{tm} is fixed to zero to identify the model. The $I_{im}^{(P)} + (t-1)S_{im}^{(P)}$ part in Equation (7) represents the deviation of the individual i from $\mu_{tm}^{(G)}$, and this is the individual-level average score at time point t after controlling group effect $\mu_{tm}^{(G)}$. The mean vector $\boldsymbol{\mu}_{(I,S)_{im}}^{(P)}$ is modeled by attribute mastery later, and representing growth rate and initial academic performance are different among individuals. In addition, $\Sigma^{(P)}$ controls the amount of variation of the initial values and slopes among individuals and relationship between the intercept and slope. If $\sigma_{I^{(P)}S^{(P)}}$ is a positive value, the individual achieved high initial academic performance and tended to extend his/her academic performance faster.

Similarly, the group-level GC model is analogous to the person-level model:

$$\begin{aligned} \mu_{tm}^{(G)} &= I_{tm}^{(G)} + (t-1)S_{tm}^{(G)} + \varepsilon_{tm}^{(G)}, \\ \varepsilon_{tm}^{(G)} &\sim N\left(0, \sigma_{\varepsilon_t^{(G)}}^2\right), \\ \begin{pmatrix} I_{tm}^{(G)} \\ S_{tm}^{(P)} \end{pmatrix} &\sim MVN\left(\boldsymbol{\mu}_{(I,S)_m}^{(G)}, \Sigma^{(G)}\right), \end{aligned} \tag{8}$$

where the group-level mean vector and covariance matrix of the intercept and slope are $\boldsymbol{\mu}_{(I,S)_m}^{(G)} = \left(\mu_{I_m^{(G)}}, \mu_{S_m^{(G)}}\right)^{\top}$ and $\Sigma^{(G)} = \begin{pmatrix} \sigma_{I^{(G)}}^2 & \sigma_{I^{(G)}S^{(G)}} \\ \sigma_{I^{(G)}S^{(G)}} & \sigma_{S^{(G)}}^2 \end{pmatrix}$, respectively. The interpretation of the group-level model is the same as the usual GC model.

For Bayesian analysis, the conditional likelihood, given the group-level score $\boldsymbol{\mu}_Y^{(G)} = \left\{\mu_{tm}^{(G)}\right\}_{t,\,m=1}^{T,\,M}$ and other person-level parameters, is expressed as where

$$P\left(Y|\boldsymbol{\mu}_{(I,S)}^{(P)}, \Sigma^{(P)}, \boldsymbol{\sigma}_{\varepsilon^{(P)}}^2, \boldsymbol{\mu}_Y^{(G)}\right) = \prod_{m=1}^{M}\prod_{i=1}^{I_m}\prod_{t=1}^{T} P\left(y_{tim}|\boldsymbol{\mu}_{(I,S)_{im}}^{(P)}, \Sigma^{(P)}, \sigma_{\varepsilon_{tm}^{(P)}}^2, \mu_{tm}^{(G)}\right), \tag{9}$$

$Y = \left\{y_{tim}\right\}_{t,\,i,\,m=1}^{T,\,I_m,\,M}$ and $\boldsymbol{\sigma}_{\varepsilon^{(P)}}^2 = \left\{\sigma_{\varepsilon_{tm}^{(P)}}^2\right\}_{t,\,m=1}^{T,\,M}$. Similarly, the conditional distribution of the group-level score $\boldsymbol{\mu}_Y^{(G)}$, given the group-level parameters, is

$$P\left(\boldsymbol{\mu}_Y^{(G)}|\boldsymbol{\mu}_{(I,S)}^{(G)}, \Sigma^{(G)}, \boldsymbol{\sigma}_{\varepsilon^{(G)}}^2\right) = \prod_{m=1}^{M}\prod_{t=1}^{T} P\left(\mu_{tm}^{(G)}|\boldsymbol{\mu}_{(I,S)_m}^{(G)}, \Sigma^{(G)}\right), \tag{10}$$

where $\boldsymbol{\sigma}_{\varepsilon^{(G)}}^2 = \left\{\sigma_{\varepsilon_1^{(G)}}^2, \ldots, \sigma_{\varepsilon_T^{(G)}}^2\right\}$. Combining the expressions of the aforementioned two conditional likelihoods, the joint distribution of observed data and latent group-level score is given by

$$\begin{aligned} P\left(Y, \boldsymbol{\mu}_Y^{(G)}|\boldsymbol{\mu}_{(I,S)}^{(P)}, \Sigma^{(P)}, \boldsymbol{\sigma}_{\varepsilon^{(P)}}^2, \boldsymbol{\mu}_{(I,S)}^{(G)}, \Sigma^{(G)}, \boldsymbol{\sigma}_{\varepsilon^{(G)}}^2\right) &= P\left(Y|\boldsymbol{\mu}_{(I,S)}^{(P)}, \Sigma^{(P)}, \boldsymbol{\sigma}_{\varepsilon^{(P)}}^2, \boldsymbol{\mu}\right) P\left(\boldsymbol{\mu}_Y^{(G)}|\boldsymbol{\mu}_{(I,S)}^{(G)}, \Sigma^{(G)}, \boldsymbol{\sigma}_{\varepsilon^{(G)}}^2\right) \\ &= \left\{\prod_{m=1}^{M}\prod_{i=1}^{I_m}\prod_{t=1}^{T} P\left(y_{tim}|\boldsymbol{\mu}_{(I,S)_{im}}^{(P)}, \Sigma^{(P)}, \sigma_{\varepsilon_{tm}^{(P)}}^2, \mu_{tm}^{(G)}\right)\right\} \\ &\quad \left\{\prod_{m=1}^{M}\prod_{t=1}^{T} P\left(\mu_{tm}^{(G)}|\boldsymbol{\mu}_{(I,S)_m}^{(G)}, \Sigma^{(G)}\right)\right\}. \end{aligned} \tag{11}$$

The joint and prior distributions, which will be introduced subsequently, are used to derive the posterior distribution of the model parameters.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

2.4. Structural model component

The structural model represents the relationship between attribute mastery and the intercept and slope parameters in the person-level model. The person-level structural models are expressed as follows:

$$\mu_{I_{im}^{(P)}} = \gamma_{I0}^{(P)} + \sum_{k=1}^{K} \gamma_{Ik}^{(P)} \alpha_{imk} + \sum_{k=1}^{K} \sum_{k'<k} \gamma_{Ikk'}^{(P)} \alpha_{imk} \alpha_{imk'} + \cdots + \gamma_{I12\ldots K}^{(P)} \prod_{k=1}^{K} \alpha_{imk}, \tag{12}$$

$$\mu_{S_{im}^{(P)}} = \gamma_{S0}^{(P)} + \sum_{k=1}^{K} \gamma_{Sk}^{(P)} \alpha_{imk} + \sum_{k=1}^{K} \sum_{k'<k} \gamma_{Skk'}^{(P)} \alpha_{imk} \alpha_{imk'} + \cdots + \gamma_{S12\ldots K}^{(P)} \prod_{k=1}^{K} \alpha_{imk}, \tag{13}$$

where the regression coefficients $\Gamma_I^{(P)} = \left\{\gamma_{I0}^{(P)}, \gamma_{I1}^{(P)}, \ldots, \gamma_{I12\ldots K}^{(P)}\right\}$ and $\Gamma_S^{(P)} = \left\{\gamma_{S0}^{(P)}, \gamma_{S1}^{(P)}, \ldots, \gamma_{S12\ldots K}^{(P)}\right\}$ represent the mastery and interaction effects on the intercept and slope, respectively. Similar to the DCM measurement model, the person-level intercept on the growth model slope is $\gamma_{S0}^{(P)}$, and it is an average slope value of all non-mastering groups. In addition, the person-level main effect of the first attribute on the growth model slope is $\gamma_{I1}^{(P)}$. This implies a difference in the growth speed between the non-mastering group and mastering first group. The person-level interaction on the growth model slope, $\gamma_{I12\ldots K}^{(P)}$, represents the effect of mastering all attributes measured in the test. $\Gamma_I^{(P)}$ represents the effects of person-level attribute mastery on the growth intercept factor. The growth intercept factor is the initial academic proficiency level. Therefore, $\Gamma_I^{(P)}$ represents how attribute masteries predict the initial academic proficiency level.

The random effect θ and group-level growth are represented as follows:

$$\mu_{I_m^{(G)}} = \gamma_I^{(G)} \theta_m, \tag{14}$$

$$\mu_{S_m^{(G)}} = \gamma_S^{(G)} \theta_m. \tag{15}$$

Note that, if some group-level covariates are known, we can use them to explain group-level growth in Equations (14) and (15). The group-level effects are $\gamma_{\mathrm{I}}^{(\mathrm{G})}$ and $\gamma_{\mathrm{S}}^{(\mathrm{G})}$, which represent the average change in the group-level intercept and slope factors by one-unit increase in the group-level random effect (e.g., group-level test answering skills), respectively. Using the above parameters, the joint distribution given in Equation (11) can be modified as

$$\begin{aligned}
&P\left(Y, \boldsymbol{\mu}_{\boldsymbol{Y}}^{(G)} | \boldsymbol{\mu}_{(I,S)}^{(P)}, \Sigma^{(P)}, \boldsymbol{\sigma}_{\varepsilon^{(P)}}^2, \boldsymbol{\mu}_{(I,S)}^{(G)}, \Sigma^{(G)}, \boldsymbol{\sigma}_{\varepsilon^{(G)}}^2\right) \\
&= P\left(Y, \boldsymbol{\mu}_{\boldsymbol{Y}}^{(G)} | \Gamma_I^{(P)}, \Gamma_S^{(P)}, \mathcal{A}, \Sigma^{(P)}, \boldsymbol{\sigma}_{\varepsilon^{(P)}}^2, \gamma_I^{(G)}, \gamma_S^{(G)}, \boldsymbol{\theta}, \Sigma^{(G)}, \boldsymbol{\sigma}_{\varepsilon^{(G)}}^2\right) \\
&= P\left(Y | \Gamma_I^{(P)}, \Gamma_S^{(P)}, \mathcal{A}, \Sigma^{(P)}, \boldsymbol{\sigma}_{\varepsilon^{(P)}}^2, \boldsymbol{\mu}_{\boldsymbol{Y}}^{(G)}\right) P\left(\boldsymbol{\mu}_{\boldsymbol{Y}}^{(G)} | \gamma_I^{(G)}, \gamma_S^{(G)}, \theta_m, \Sigma^{(G)}, \boldsymbol{\sigma}_{\varepsilon^{(G)}}^2\right) \\
&= \left\{\prod_{m=1}^{M} \prod_{i=1}^{I_m} \prod_{t=1}^{T} P\left(y_{tim} | \Gamma_I^{(P)}, \Gamma_S^{(P)}, \boldsymbol{\alpha}_{im}, \Sigma^{(P)}, \sigma_{\varepsilon_{tm}^{(P)}}^2, \mu_{tm}^{(G)}\right)\right\} \\
&\quad \left\{\prod_{m=1}^{M} \prod_{t=1}^{T} P\left(\mu_{tm}^{(G)} | \gamma_I^{(G)}, \gamma_S^{(G)}, \theta_m, \Sigma^{(G)}\right)\right\}.
\end{aligned} \tag{16}$$

2.5. Full conditional distribution and estimation

To employ the Bayesian estimation method, we must specify priors for the model parameters. The mixing parameter $\boldsymbol{\pi}$ follows a Dirichlet distribution with parameter $\boldsymbol{\delta}$, and the LCDM parameters follow the normal distributions: $\boldsymbol{\pi} \sim Dir(\boldsymbol{\delta})$ and $\lambda \sim N\left(\mu_\lambda, \sigma_\lambda^2\right)$. The main effects should be positive; hence, we impose constraints on the prior distribution (Zhan, Jiao, Man, et al., 2019). The prior for the variance parameter of the random effect θ_m is expressed as σ_θ^2, which independently and identically follows an inverse gamma distribution: $\sigma_\theta^2 \sim IG(a, b)$. Combining the above prior settings with the conditional complete likelihood expressed in Equation (5), the joint distribution of the RDCM part can be expressed

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

as follows:

$$P\left(\mathrm{X}, \mathcal{A}, \boldsymbol{\theta}, \Lambda_X, \boldsymbol{\pi}, \sigma^2_\theta\right) = P\left(\mathrm{X}|\Lambda_X, \mathcal{A}, \boldsymbol{\theta}\right) P\left(\mathcal{A}|\boldsymbol{\pi}\right) P\left(\boldsymbol{\pi}\right) P\left(\Lambda_X\right) P\left(\boldsymbol{\theta}|\sigma^2_\theta\right) P\left(\sigma^2_\theta\right)$$

$$= \prod_{m=1}^{M} \prod_{i=1}^{I_m} \left\{ P\left(X_{ijm} = x_{ij}|\boldsymbol{\alpha}_{im}, \boldsymbol{\lambda}_j, \theta_m, \boldsymbol{q}_j\right) \prod_{l=1}^{L} P\left(\boldsymbol{\alpha}_{im} = \boldsymbol{\alpha}_l|\boldsymbol{\pi}\right) \right\}$$

$$\times \left(\prod_{j=1}^{J} P\left(\boldsymbol{\lambda}_j\right) \right) P\left(\boldsymbol{\pi}\right) \left(\prod_{m=1}^{M} P\left(\theta_m|\sigma^2_\theta\right) \right) P\left(\sigma^2_\theta\right), \tag{17}$$

where the priors are expressed by omitting their parameters, such as $P\left(\boldsymbol{\pi}\right)$, for notational simplicity.

Next, we set priors for the MGC model and structural part. The gamma parameters $\boldsymbol{\Gamma} = \left\{ \Gamma_I^{(P)}, \Gamma_S^{(P)}, \gamma_I^{(G)}, \gamma_S^{(G)} \right\}$ are separately assumed to follow normal distributions. The priors for the MGC can be referenced from the Bayesian GC model (Depaoli & Clifton, 2015; Oravecz & Muth, 2018). Measurement error variances, such as $\sigma^2_{\varepsilon_i^{(G)}}$ or $\sigma^2_{\varepsilon_{tm}^{(P)}}$, are again assumed to follow inverse gamma distributions. The covariance matrices for the intercept and slope $\Sigma^{(P)}$ and $\Sigma^{(G)}$ are assumed to be generated from an inverse Wishart distribution: $\Sigma^{(P)} \sim IW\left(\Psi^{(P)}, \nu^{(P)}\right)$ and $\Sigma^{(G)} \sim IW\left(\Psi^{(G)}, \nu^{(G)}\right)$, respectively, where $\Psi^{(P)}$ and $\Psi^{(G)}$ are parameter matrices, and $\nu^{(P)}$ and $\nu^{(G)}$ are the degrees of freedom. Assuming the above priors and the conditional distribution expressed in Equation (11), the conditional distribution, given attribute mastery and random effects, is expressed as

$$\begin{aligned} & P\left(Y, \boldsymbol{\mu}_Y^{(G)}|\Gamma_I^{(P)}, \boldsymbol{\gamma}_S^{(P)}, \mathcal{A}, \Sigma^{(P)}, \boldsymbol{\sigma}^2_{\varepsilon^{(P)}}, \gamma_I^{(G)}, \gamma_S^{(G)}, \boldsymbol{\theta}, \Sigma^{(G)}, \boldsymbol{\sigma}^2_{\varepsilon^{(G)}}\right) \\ & = P\left(Y|\Gamma_I^{(P)}, \Gamma_S^{(P)}, \mathcal{A}, \Sigma^{(P)}, \boldsymbol{\sigma}^2_{\varepsilon^{(P)}}, \boldsymbol{\mu}_Y^{(G)}\right) P\left(\boldsymbol{\mu}_Y^{(G)}|\gamma_I^{(G)}, \gamma_S^{(G)}, \boldsymbol{\theta}, \Sigma^{(G)}, \boldsymbol{\sigma}^2_{\varepsilon^{(G)}}\right) \\ & = \left\{ \prod_{m=1}^{M} \prod_{i=1}^{I_m} \prod_{t=1}^{T} P\left(y_{tim}|\Gamma_I^{(P)}, \Gamma_S^{(P)}, \boldsymbol{\alpha}_{im}, \Sigma^{(P)}, \sigma^2_{\varepsilon_{tm}^{(P)}}, \mu_{tm}^{(G)}\right) \right\} \\ & \quad \left\{ \prod_{m=1}^{M} \prod_{t=1}^{T} P\left(\mu_{tm}^{(G)}|\gamma_I^{(G)}, \gamma_S^{(G)}, \theta_m, \Sigma^{(G)}\right) \right\}. \end{aligned} \tag{18}$$

Finally, from the above joint and conditional distributions, the joint distribution of data and parameters can be expressed as

$$\begin{aligned} & P\left(\mathrm{X}, \mathcal{A}, \boldsymbol{\theta}, \Lambda_X, \boldsymbol{\pi}, \sigma^2_\theta, Y, \boldsymbol{\mu}_Y^{(G)}, \boldsymbol{\Gamma}, \Sigma^{(P)}, \boldsymbol{\sigma}^2_{\varepsilon^{(P)}}, \Sigma^{(G)}, \boldsymbol{\sigma}^2_{\varepsilon^{(G)}}\right) \\ & = P\left(\mathrm{X}|\Lambda_X, \mathcal{A}, \boldsymbol{\theta}\right) P\left(\mathcal{A}|\boldsymbol{\pi}\right) P\left(\boldsymbol{\pi}\right) P\left(\Lambda_X\right) P\left(\boldsymbol{\theta}|\sigma^2_\theta\right) P\left(\sigma^2_\theta\right) \\ & \quad \times P\left(Y|\Gamma_I^{(P)}, \Gamma_S^{(P)}, \mathcal{A}, \Sigma^{(P)}, \boldsymbol{\sigma}^2_{\varepsilon^{(P)}}, \boldsymbol{\mu}_Y^{(G)}\right) P\left(\boldsymbol{\mu}_Y^{(G)}|\gamma_I^{(G)}, \gamma_S^{(G)}, \boldsymbol{\theta}, \Sigma^{(G)}, \boldsymbol{\sigma}^2_{\varepsilon^{(G)}}\right) \\ & \quad \times P\left(\boldsymbol{\Gamma}\right) P\left(\Sigma^{(P)}\right) P\left(\boldsymbol{\sigma}^2_{\varepsilon^{(P)}}\right) P\left(\Sigma^{(G)}\right) P\left(\boldsymbol{\sigma}^2_{\varepsilon^{(G)}}\right). \end{aligned} \tag{19}$$

3. Simulation study 1

We verified parameter recoveries (bias), parameter estimation variability (root mean square error: RMSE), and coverage of equal-tailed posterior credible intervals (95% CI coverage) using the Bayesian estimation procedure for the proposed RDC–MGC model in various settings. Attribute recoveries were also evaluated using attribute-level agreement ratio (AAR) and pattern-level agreement ratio (PAR). Furthermore, we assessed the effects on parameter recovery by ignoring multilevel structures. Applying a single-level version of the DCM with a single growth model under the RDC–MGC model data-generating system can provide biased, unstable estimations, and low coverage probabilities. Instead, considering a multilevel structure (e.g., RDC–MGC model) can recover true parameter values and achieve nominal coverage probabilities.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

3.1. Simulation setting

The RDC–MGC model has various parameters or segments that can be manipulated. We manipulated three crucial factors: sample size, item quality of diagnostic measurement, and strength of the structural model. We assumed conditions of 200 and 1,000 sample sizes. In the 200 sample size condition, 10 schools ($M = 10$) and 20 individuals were sampled from each school. This condition represented tracking a few individuals. Similarly, the 1,000 sample-size condition contained 25 schools ($M = 25$) and 40 individuals, that is, approximately one class was sampled from each school. This condition indicates a relatively large-scale educational study, such as our real-data example.

The second factor was the item quality of the diagnostic measurement. High- and low-quality items were assumed, with the former exhibiting relatively stronger disclination power of attribute mastery patterns than the latter. According to Madison and Bradshaw (2018, p. 917), LCDM item parameters were set to high/low-quality items, and the values were changed for the complexity of the item. The Q-matrix employed in this simulation is presented in Table 1, which has three attributes and nine simple structure items (items 1–9), nine middle complex items requiring two attributes (items 10–18), and two complex items measuring three attributes (items 19 and 20). In the condition of high-quality items, simple items had an intercept $\lambda_0 = -2$ and a main effect $\lambda_1 = 3$. The middle complex items had an intercept $\lambda_0 = -2$, two main effects $\lambda_1 = 2$, and a first-order interaction $\lambda_2 = 1$. Finally, the two complex items had an intercept $\lambda_0 = -3$, three main effects $\lambda_1 = 1$, three first-order interactions $\lambda_2 = 0.5$, and one second-order interaction $\lambda_3 = 1$. Low-quality items had half the parameter values as the high-quality items. For example, the simple items had an intercept $\lambda_0 = -1$ and a main effect $\lambda_1 = 1.5$.

The last manipulation factor was the structural parameter strength, classified as strong or weak. The values were considered based on the results of the following real data. In the condition of the strong structural parameter, the main effects of the intercept were set to $\left(\gamma_{I1}^{(P)}, \gamma_{I2}^{(P)}, \gamma_{I3}^{(P)}\right) = (0.8, 1, 1.2)$. The interaction parameters were generated from the main effects $\left(\gamma_{I12}^{(P)}, \gamma_{I13}^{(P)}, \gamma_{I23}^{(P)}\right) = 0.7 \times \left(\gamma_{I1}^{(P)} \times \gamma_{I2}^{(P)}, \gamma_{I1}^{(P)} \times \gamma_{I3}^{(P)}, \gamma_{I2}^{(P)} \times \gamma_{I3}^{(P)}\right)$. The second-order interaction was determined by $\gamma_{I123}^{(P)} = 0.7 \times \prod_k \gamma_{Ik}^{(P)}$. The effect on the slope parameters $\gamma_S^{(P)}$ was half of the corresponding $\gamma_I^{(P)}$. For example, the main effects on the slope were $\left(\gamma_{P1}^{(P)}, \gamma_{P2}^{(P)}, \gamma_{P3}^{(P)}\right) = 0.5 \times \left(\gamma_{I1}^{(P)}, \gamma_{I2}^{(P)}, \gamma_{I3}^{(P)}\right) = (0.4, 0.5, 0.6)$. The weak structural parameter assumed a quarter of the main effects of the intercept in the strong condition, and the same interaction-generating procedures were applied. These values were slightly arbitrary; therefore, various situations should be considered in future studies. The level two structural parameters $\gamma_I^{(G)}$ and $\gamma_S^{(G)}$ were set to -2 and 1, respectively, in all conditions to maintain the simplicity of the simulation.

We set variance-related parameter values to real-data results. First, the variance of the random effect θ_m was $\sigma_\theta^2 = 1.5$. The variance–covariance matrices were $\Sigma^{(P)} = \begin{pmatrix} 0.4^2 & 0.2 \times 0.4 \times 0.2 \\ 0.2 \times 0.4 \times 0.2 & 0.2^2 \end{pmatrix}$ and $\Sigma^{(G)} = \begin{pmatrix} 0.35^2 & 0.1 \times 0.35^2 \\ 0.1 \times 0.35^2 & 0.35^2 \end{pmatrix}$, with 0.2 and 0.1 correlations for person and group levels, respectively. The person-level residual parameters $\sigma^2_{\varepsilon_t^{(P)}}$ were all set to 0.2, and group-level residual variance $\sigma^2_{\varepsilon_t^{(G)}}$ was set to 0.1. We assumed a four-time point that was the same as the real-data analysis.

Using the above parameters, we generated 50 simulated data sets for each condition. Individuals' attributes were generated using the modified procedures of Chiu et al. (2009) or Yamaguchi and Martinez (2024). Continuous attributes were first generated using a multivariate normal distribution with zero means and a correlation matrix with 0.4 values, and if the value exceeded 0, the attribute was considered mastered. The mixing parameters $\boldsymbol{\pi}$ of the true attribute mastery pattern were calculated based on the distribution assumption.

For the same dataset, the RDC–MGC model and single-level version of the model, which did not include a random effect θ_m and the group-level growth model, were applied. Therefore, the single-level

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 1. Q-matrix for the first simulation study

Item	Attribute		
	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	1	0
6	0	0	1
7	1	0	0
8	0	1	0
9	0	0	1
10	1	1	0
11	1	0	1
12	0	1	1
13	1	1	0
14	1	0	1
15	0	1	1
16	1	1	0
17	1	0	1
18	0	1	1
19	1	1	1
20	1	1	1

model did not have a group-level structural parameter. We considered the RDC–MGC and single-level models as the two- and single-level models to emphasize the differences between the models.

In the estimation, we assumed 2 chains and 20 thinning intervals. To determine the number of iterations, we conducted preliminary runs of the MCMC procedures. As a result, the MCMC iteration number was 20,000, which was generally enough for the convergence of the MCMC iteration, and the first 10,000 samples were discarded during the burn-in period. Prior settings are shown in the simulation script available on the OSF site: https://osf.io/gpztw.

Posterior means were employed for point estimates. For each parameter, such as λ_{j0}, we calculated $\text{Abs.Bias}_{\lambda_{j0}} = \frac{1}{50}\sum_{n=1}^{50} \mid \widehat{\lambda}_{j0,\,n} - \lambda_{j0}^{\text{True}} \mid$, $\text{RMSE}_{\lambda_{j0}} = \sqrt{\frac{1}{50}\sum_{n=1}^{50}\left(\widehat{\lambda}_{j0,\,n} - \lambda_{j0}^{\text{True}}\right)^2}$, and $95\%\text{CI coverage}_{\lambda_{j0}} = \frac{1}{50}\sum_{n=1}^{50} I\left(\widehat{\lambda}_{j0,\,n}^{L} < \lambda_{j0}^{\text{True}} < \widehat{\lambda}_{j0}^{U}\right)$, where $\widehat{\lambda}_{j0}$ is the estimate based on the n-th data set, $\lambda_{j0}^{\text{True}}$ is the true parameter, and $\widehat{\lambda}_{j0,\,n}^{L}$ and $\widehat{\lambda}_{j0,\,n}^{U}$ are the lower and upper limits of 95% CI with the n-th data set. These values are averaged for the group of parameters. For example, the absolute bias of λ can be expressed as $\text{Abs.Bias}_{\lambda} \times 100 = \frac{100}{\text{No.of } \lambda \text{ parameters}} \sum \text{Abs.Bias}_{\lambda_{j0}}$. Herein, we evaluated only fixed effect parameters because random effects are realized differently for various simulation repetitions. In addition, AAR and PAR were calculated as $\text{AAR}_k = \frac{1}{I\times 50}\sum_{n=1}^{50}\sum_{i=1}^{I} \mathcal{I}\left(\widehat{\alpha}_{ik}^{(n)} = \alpha_{ik}^{\text{True}}\right), \forall k$ and $\text{PAR} = \frac{1}{I\times 50}\sum_{n=1}^{50}\sum_{i=1}^{I} \mathcal{I}\left(\widehat{\boldsymbol{\alpha}}_{i}^{(n)} = \boldsymbol{\alpha}_{i}^{(\text{True})}\right)$ $\widehat{\alpha}_{ik}^{(n)}$ and $\widehat{\boldsymbol{\alpha}}_{i}^{(n)}$ are n-th attribute k and attribute mastery pattern estimates of the individual i. $\widehat{\alpha}_{ik}^{(n)}$ is the

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

expected-a-posteriori estimate of the k-th attribute that takes one if attribute mastery probability is greater than 0.5, and $\widehat{\boldsymbol{\alpha}}_i^{(n)}$ was attribute mastery estimates $\left(\widehat{\alpha}_{i1}^{(n)}, \ldots, \widehat{\alpha}_{iK}^{(n)}\right)^\top$. Because of the computational limitation, the 1,000 MCMC samples were obtained after the first 20,000 iterations for model parameters.

3.2. Results

Table 2 presents 100 times the absolute biases (Abs. Bias) and RMSE, and 95% CI coverage of LCDM parameters (λs) of two- or single-level model results. In conditions of small sample sizes and high-quality items, the two-level model provided slightly better absolute biases, RMSE, and 95% CI coverage than the single model. However, under conditions of small sample sizes and low-quality items, the absolute biases, RMSE, and 95% CI coverage of the two-level model were better than those of the single-level model. For example, the 100 times of absolute bias of the two-level in the strong structural parameter condition was 38.82, and that of the single-level was 69.54. In the condition of 100 times, the RMSE of the two-level model was 44.25, and that of the single-level model was 63.43. The value of the 95% CI coverage for the two-level model was 0.971, and that of the single-level model was 0.790, indicating a notable difference. The absolute bias, the RMSE of the two-level model decreased as the sample size increased, and its 95% CI coverages were close to the theoretically preferable value of 0.95. The absolute biases of the single-level model in the 1,000 sample were smaller than those of the 200 sample size, but the values were worse than those of the two-level models. In the same conditions, the RMSE of the single-level model exceeded that of the two-level model. Moreover, the 95% CI coverages of the single-level model were low even in the large sample size conditions. In summary, for the LCDM item parameter recovery, the two-level model exhibited smaller biases and RMSEs and appropriate 95% CI coverage than the single-level model.

The attribute mixing proportion parameter $\boldsymbol{\pi}$ exhibited the same tendency. The results are presented in Table 3. The notable difference between the two- and single-level models was 95% CI coverage. The two-level model indicated relatively better values of the 95% CI coverages than the single-level models. The high-quality items or strong structural parameter conditions exhibited better absolute biases than the low-quality or weak structural parameter conditions in the two-level models. In addition, the absolute biases of the two-level model were approximately half of those of the single-level model.

The parameter recovery results for structural parameters on the intercept ($\gamma_I^{(P)}$of $\gamma_I^{(G)}$) are presented in Table 4. The person-level structural parameters $\gamma_I^{(P)}$ of the two-level model exhibited smaller biases and RMSEs than those of the single-level model. For example, in the 1,000 sample, high-quality items, and strong structural parameter conditions, the 100 times of the absolute bias and RMSE of the two-level model were 9.21 and 12.69, and the corresponding values of the single-level model were 73.57 and 62.61, respectively. The 95% CI coverages of the two-level model were generally better than those of the single-level model. The decreases in the biases and RMSE of the two-level model because of sample size increase were affected by item quality. The increase in sample size appropriately decreased the biases and RMSEs in the person-level structural parameters in the high-quality items. However, these decreases were slower in the low-quality item condition. The group-level structural parameter $\gamma_I^{(G)}$ observed only in the two-level model was generally a good recovery; its 95% CI coverage was 1 indicating too large 95% CIs. This was, however, better than those under coverage results because of its conservative tendency. The above tendencies were also observed in the recovery of the structural parameters on the slope ($\gamma_S^{(P)}$of $\gamma_S^{(G)}$) presented in Table 5.

The results of the variance-related parameters are presented in Tables 6–8. Table 6 presents the values of the covariance matrix parameter of the intercept $I$and slope S ($\Sigma^{(P)}$ or $\Sigma^{(G)}$). The two-level model exhibited sufficiently small biases and RMSEs of the person-level covariance matrix $\Sigma^{(P)}$, but the single-level models indicated large values. The 95% CIs of $\Sigma^{(P)}$ in the two-level models were relatively better

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 2. One-hundred times the absolute bias (Abs. Bias), root mean square error (RMSE), and 95% credible (95% CI) coverage of log-linear cognitive diagnostic model parameters (λs) of two- and single-level model results in Simulation 1

Sample size	Item condition	Structural parameter condition	Two-level			Single-level		
			Abs. Bias × 100	RMSE × 100	95%CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage
200	High	Strong	41.46	48.22	0.956	49.79	59.00	0.915
		Weak	43.11	49.86	0.961	47.65	56.59	0.939
	Low	Strong	38.82	44.25	0.971	69.54	63.43	0.790
		Weak	46.00	47.87	0.969	71.40	60.73	0.760
1,000	High	Strong	23.06	28.22	0.952	35.03	33.27	0.813
		Weak	24.00	28.65	0.951	33.97	35.30	0.826
	Low	Strong	25.97	31.96	0.925	62.83	57.69	0.571
		Weak	28.86	33.04	0.963	62.40	39.96	0.507

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 3. One-hundred times the Abs. Bias, RMSE, and 95% credible (95% CI) coverage of attribute mastery pattern mixing parameters π of two- and single-level model results in Simulation 1

Sample size	Item quality condition	Structural parameter condition	Two-level			Single-level		
			Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage
200	High	Strong	1.89	2.33	0.958	3.24	3.69	0.815
		Weak	2.25	2.79	0.935	3.69	3.62	0.760
	Low	Strong	2.47	3.03	0.963	5.72	7.07	0.695
		Weak	5.76	6.13	0.900	6.29	6.77	0.605
1,000	High	Strong	0.96	1.33	0.950	2.08	1.78	0.685
		Weak	0.92	1.16	0.953	2.37	1.67	0.633
	Low	Strong	2.16	3.16	0.910	5.50	6.85	0.633
		Weak	2.45	2.89	0.918	5.14	3.96	0.375

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 4. One-hundred times the Abs. Bias, RMSE, and 95% credible (95% CI) coverage of structural parameters on intercept ($\gamma_I^{(P)}$of $\gamma_I^{(G)}$) of two- and single-level model results in Simulation 1

Sample size	Item quality condition	Structural parameter condition	Two-level: $\gamma_I^{(P)}$			Two-level: $\gamma_I^{(G)}$			Single-level: $\gamma_I^{(P)}$		
			Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage
200	High	Strong	16.04	19.60	0.963	15.11	17.72	1.000	62.40	69.77	0.943
		Weak	21.02	25.90	0.957	20.01	24.96	1.000	54.24	57.17	0.931
	Low	Strong	17.73	22.08	0.994	15.22	17.02	1.000	101.51	90.88	0.886
		Weak	20.43	25.07	0.977	18.06	22.26	1.000	74.90	65.47	0.909
1,000	High	Strong	9.21	12.69	0.940	13.89	12.56	1.000	73.57	62.61	0.749
		Weak	9.16	10.99	0.963	14.16	12.46	1.000	67.40	56.82	0.649
	Low	Strong	19.24	29.65	0.911	13.13	11.50	1.000	139.79	100.74	0.594
		Weak	16.69	19.74	0.977	13.85	12.67	1.000	89.52	60.13	0.603

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 5. One-hundred times the Abs. Bias, RMSE, and 95% credible (95% CI) coverage of structural parameters on slope ($\gamma_S^{(P)}$ of $\gamma_S^{(G)}$) of two- and single-level model results

Sample size	Item quality condition	Structural parameter condition	Two-level: $\gamma_S^{(P)}$			Two-level: $\gamma_S^{(G)}$			Single-level: $\gamma_S^{(P)}$		
			Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage
200	High	Strong	10.06	12.53	0.966	10.18	10.63	1.000	28.19	32.72	0.886
		Weak	10.76	13.52	0.977	14.38	14.62	1.000	28.15	31.89	0.943
	Low	Strong	12.73	15.48	0.966	10.67	9.34	1.000	34.52	48.64	0.949
		Weak	10.49	12.70	0.994	14.09	13.89	1.000	32.07	36.43	0.946
1,000	High	Strong	4.83	6.38	0.951	5.73	6.40	1.000	23.15	24.16	0.751
		Weak	4.83	6.01	0.946	6.22	5.69	1.000	28.32	27.56	0.720
	Low	Strong	11.41	16.20	0.923	5.58	5.97	1.000	45.32	54.32	0.780
		Weak	8.67	10.55	0.974	6.48	5.84	1.000	29.60	28.01	0.811

Table 6. One-hundred times the Abs. Bias, RMSE, and 95% credible (95% CI) coverage of covariance matrix parameter of intercept $\boldsymbol{I}$ and slope S ($\Sigma^{(P)}$ or $\Sigma^{(G)}$) of two- and single-level model results in Simulation 1

Sample size	Item quality condition	Structural parameter condition	Two-level: $\Sigma^{(P)}$			Two-level: $\Sigma^{(G)}$			Single-level: $\Sigma^{(P)}$		
			Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage
200	High	Strong	12.16	19.62	0.927	1.67	1.60	0.840	569.50	297.22	0.007
		Weak	11.52	22.18	0.933	1.73	1.75	0.847	455.37	209.04	0.000
	Low	Strong	11.20	18.53	0.927	1.67	1.55	0.820	458.84	241.69	0.127
		Weak	10.61	20.95	0.927	1.73	1.72	0.880	349.74	210.67	0.000
1,000	High	Strong	11.67	13.18	0.853	0.89	1.30	0.947	594.08	162.17	0.000
		Weak	11.04	15.36	0.860	0.70	0.75	0.940	431.07	126.81	0.000
	Low	Strong	11.78	12.91	0.847	2.11	3.11	0.867	432.35	118.18	0.153
		Weak	11.40	15.44	0.860	0.75	0.80	0.967	357.00	133.25	0.000

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 7. One-hundred times the Abs. Bias, RMSE, and 95% credible (95% CI) coverage of residual variance of T time points outcome measures ($\sigma^2_{\varepsilon(P)}$ or $\sigma^2_{\varepsilon(G)}$) of two- and single-level model results in Simulation 1

Sample size	Item quality condition	Structural parameter condition	Two-level: $\sigma^2_{\varepsilon(P)}$			Two-level: $\sigma^2_{\varepsilon(G)}$			Single-level: $\sigma^2_{\varepsilon(P)}$		
			Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage	Abs. Bias × 100	RMSE × 100	95% CI coverage
200	High	Strong	9.71	12.86	0.950	3.21	3.49	0.920	17.79	19.29	0.540
		Weak	9.61	12.62	0.970	3.08	3.28	0.920	17.23	18.82	0.545
	Low	Strong	9.41	12.33	0.975	3.50	3.76	0.925	17.64	18.78	0.560
		Weak	9.04	11.62	0.975	3.38	3.38	0.910	16.86	17.65	0.510
1,000	High	Strong	5.72	6.83	0.980	1.07	1.26	0.950	13.19	11.83	0.335
		Weak	5.69	7.13	0.960	1.28	1.48	0.930	14.83	14.08	0.275
	Low	Strong	9.71	12.86	0.950	3.21	3.49	0.920	17.79	19.29	0.540
		Weak	9.61	12.62	0.970	3.08	3.28	0.920	17.23	18.82	0.545

Table 8. One-hundred times the Abs. Bias, RMSE, and 95% credible (95% CI) coverage of the variance σ^2_θ of random effect θ_m of two- or single-level model results in Simulation 1

Sample size	Item quality condition	Structural parameter condition	Two-level		
			Abs. Bias × 100	RMSE × 100	95% CI coverage
200	High	Strong	121.40	150.85	0.940
		Weak	136.34	165.07	0.880
	Low	Strong	126.72	156.54	0.880
		Weak	158.55	183.48	0.880
1,000	High	Strong	55.35	65.33	0.980
		Weak	53.74	68.74	0.940
	Low	Strong	55.90	65.45	0.980
		Weak	57.97	70.40	0.940

than those of the single-level model. However, even in the 1,000 sample size conditions, the values were approximately 0.85 and were not the best. Increase of sample or group sizes might solve this problem. The group-level covariance results $\Sigma^{(G)}$ were generally good, but the 95% CI coverages in the 200 sample size were in the range of 0.820–0.880 and were not the best. Table 7 presents the results of the residual variances ($\sigma^2_{\varepsilon(P)}$ and $\sigma^2_{\varepsilon(G)}$). Again, the two-level model exhibited better results than the single-level results of the person-level residual variance $\sigma^2_{\varepsilon(P)}$. For example, in the 1,000 sample, high-quality items, and strong structural parameter condition, the 100 times of the absolute bias and RMSE of the two-level model were 5.72 and 6.83, and the corresponding values of the single-level model were 13.19 and 11.83, respectively. The group-level residual variance $\sigma^2_{\varepsilon(G)}$ of the two-level model was sufficiently small. Finally, Table 8 presents the variance parameter σ^2_θ of the random effect θ_m. The 200 sample size conditions indicated approximately 120–160 values for the 100 times of the absolute biases and 150–180 for the 100 times of the RMSEs. However, these values for the 1,000 sample size conditions decreased to approximately 54–58 and 65–70, respectively. The results of the 95% CI coverage were appropriate if the sample size was increased. In our simulation, we set $M = 10$ for the 100 sample size condition and $M = 25$ for the 1,000 sample size condition. Therefore, the group number increased as the sample size increased.

Attribute mastery recovery results are shown in Table 9. High-quality items tended to show higher AARs and PARs. Especially, both AARs and PARs exceeded 0.92 for all conditions with the two-level model in the high-quality items conditions. In addition, strong structural models or larger sample sizes indicated higher AARs and PARs than weak or small sample size conditions. For example, AARs of the 200 sample size, low-quality items, and weak structural condition with the two-level model were 0.763–0.784, but those of the 1,000 sample size, low-quality items, and strong structural condition were 0.884–0.894. More importantly, the single-level model provided drastically lower AARs and PARs than the two-level model. This indicated that ignoring multilevel structure resulted in misdiagnosis.

Summarily, the two-level model provided smaller absolute biases and RMSEs and a preferable 95% CI coverage than the single-level model for all model parameters in almost all conditions. In other words, multilevel structures should be considered when employing a multistage sampling design. In addition, our estimation method was validated using the JAGS language well recovered model parameters at least in the idealized situations.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 9. Attribute-level and pattern-level recovery results of two- and single-level models in Simulation 1

Sample size	Item condition	Gamma parameter	Two-level				Single-level			
			AAR 1	AAR 2	AAR 3	PAR	AAR 1	AAR 2	AAR 3	PAR
200	High	Strong	0.964	0.963	0.971	0.929	0.840	0.854	0.867	0.641
		Weak	0.927	0.931	0.934	0.817	0.822	0.821	0.831	0.582
	Low	Strong	0.879	0.879	0.899	0.779	0.677	0.701	0.714	0.380
		Weak	0.763	0.785	0.784	0.472	0.668	0.681	0.666	0.355
1,000	High	Strong	0.969	0.970	0.976	0.941	0.843	0.850	0.866	0.632
		Weak	0.944	0.946	0.949	0.856	0.833	0.834	0.837	0.607
	Low	Strong	0.884	0.875	0.894	0.742	0.680	0.698	0.693	0.363
		Weak	0.821	0.828	0.842	0.587	0.679	0.681	0.675	0.371

Note: AAR is attribute-level agreement ratio and PAR is pattern-level agreement ratio.

4. Simulation study 2

The first simulation employed three attribute Q-matrix which is relatively simple. To generalize the parameter recovery with the Bayesian estimation method employed in the first simulation, we assumed five attribute Q-matrix that represented a complex situation in the second simulation.

4.1. Simulation setting

Table 10 shows the Q-matrix in the second simulation in which the 1st–10th items were simple needing only one attribute to solve them, and the rest 11th–20th items required two attributes representing complex items. The three attributes loading items were not assumed because the number of combinations of two-way interaction is $\binom{5}{3} = 10$; therefore, the number of item parameters becomes large if we assume items that need three attributes. This is overly complicated, and we did not include these items in the second simulation. In the second simulation, we manipulated the same factors as the first simulation: sample size (200 or 1,000), item quality (high or low), and the structural parameter strength (strong or weak). However, the true parameter values were slightly changed.

First, in the high-quality items condition, simple items had an intercept $\lambda_0 = -2$ and a main effect $\lambda_1 = 4$. The complex items had an intercept $\lambda_0 = -3$, two main effects $\lambda_1 = 2$, and a first-order interaction $\lambda_2 = 2$. Again, low-quality items had half the parameter values as the high-quality items.

In the strong structural parameter condition, the main effects of the intercept were set to $\left(\gamma_{I1}^{(P)}, \gamma_{I2}^{(P)}, \gamma_{I3}^{(P)}, \gamma_{I4}^{(P)}, \gamma_{I5}^{(P)}\right) = (0.6, 0.8, 1, 1.2, 1.4)$. The second interaction was generated in the same way as the first simulation. The two-way interaction was omitted because of the same reason mentioned above. The weak structural parameter assumed half of the main effects of the intercept in the strong condition. Then the interactions were generated the same way. The other parameter settings and MCMC estimation settings were the same as the first simulation. The evaluation perspectives were also the same as the first simulation. The simulation R codes are shown in the OSF file.

4.2. Results

Even with the five attribute Q-matrix condition, we confirm the similar results shown in the first simulation. The two-level model indicated better estimation results than the single-level models: smaller absolute biases and RMSEs, and more appropriate coverages of 95% CI. Therefore, the evidence of the validity of the estimation with the JAGS language was obtained before conducting real-data analysis with the proposed model in the following section.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 10. Q-matrix for the second simulation study

Item	Attribute				
	1	2	3	4	5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	0	0	0	0
7	0	1	0	0	0
8	0	0	1	0	0
9	0	0	0	1	0
10	0	0	0	0	1
11	1	1	0	0	0
12	1	0	1	0	0
13	1	0	0	1	0
14	1	0	0	0	1
15	0	1	1	0	0
16	0	1	0	1	0
17	0	1	0	0	1
18	0	0	1	1	0
19	0	0	1	0	1
20	0	0	0	1	1

Because of the space limitation and to reduce redundancy of the description, the parameter recovery results tables are presented in Tables S1–S6 in the Supplementary Material. We only show attribute mastery recovery results here in Table 11. Most AARs with the two-level in the high-quality items conditions were higher than 0.95, but the ones with the single-level model were at most approximately 0.92. In these conditions, PARs of the two-level model were much better than ones of the single-model. For example, PAR of the two-level model in the condition of the 1,000 sample, high-quality items, and strong structural parameters was 0.931, but corresponding value of the single-level model was 0.638. Differences of AARs or PARs between the two models were evident in low-item conditions. For example, AARs of the two-level in the 200 sample, low-quality items, and strong structural parameter were 0.866–0.895, and those of the single-level model were 0.696–0.725. Furthermore, PAR with the two-level model in the same condition was 0.605, but that with the single-level model was 0.249. In summary, the two-level model relatively better recovered attribute masteries than the single-level model that ignores the multilevel structure.

5. Real-data analysis

In this section, we first describe the following data, item specification, test design, and sampling design and explain the vertical scale construction procedure. The vertical scale was used from the third to the

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 11. Attribute-level and pattern-level recovery results of two- and single-level models in Simulation 2

Sample size	Item condition	Gamma parameter	Two-level						Single-level					
			AAR 1	AAR 2	AAR 3	AAR 4	AAR 5	PAR	AAR 1	AAR 2	AAR 3	AAR 4	AAR 5	PAR
200	High	Strong	0.974	0.974	0.975	0.977	0.982	0.923	0.881	0.891	0.900	0.911	0.916	0.645
		Weak	0.949	0.951	0.957	0.961	0.967	0.828	0.879	0.874	0.881	0.884	0.894	0.586
	Low	Strong	0.870	0.869	0.866	0.879	0.895	0.605	0.696	0.703	0.710	0.728	0.725	0.249
		Weak	0.829	0.832	0.853	0.860	0.876	0.491	0.694	0.687	0.707	0.706	0.722	0.235
1,000	High	Strong	0.978	0.977	0.977	0.979	0.984	0.931	0.883	0.892	0.901	0.911	0.921	0.638
		Weak	0.955	0.959	0.963	0.968	0.973	0.853	0.873	0.877	0.881	0.888	0.898	0.592
	Low	Strong	0.888	0.876	0.868	0.878	0.911	0.615	0.697	0.710	0.715	0.728	0.669	0.229
		Weak	0.843	0.854	0.866	0.883	0.895	0.550	0.693	0.703	0.707	0.716	0.708	0.254

Note: AAR is attribute-level agreement ratio and PAR is pattern-level agreement ratio.

Table 12. Number of items in the anchor and separate tests

	Item difficulty level	No. of items					Total
		Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	
Anchor test	Grade 2	19	9				28
	Grade 3	9	19	10			38
	Grade 4		9	19	9		37
	Grade 5			10	17	10	37
	Grade 6				8	19	27
Separate test	Grade 2	39					39
	Grade 3		37				37
	Grade 4			39			39
	Grade 5				38		38
	Grade 6					38	38

sixth grades in Japanese elementary schools. Next, we explain the attribute definition and Q-matrix construction procedure. These elements were used to define the DCM measurement model at the second-grade level. Additionally, we specify the priors and MCMC settings for the proposed Bayesian data analysis, which are detailed in Section 3.1.3. The parameter estimation results for the linear growth case are presented. We also analyzed a partially freely estimated slope-loading specification as an example of an alternative growth trajectory.

5.1. Data description

5.1.1. Item specification and test design

The difficulty levels of the test items used in this study varied from Grades 2 to 6. All items were produced by Tokyo Shoseki Co., Ltd., Japan, for assessing students' mathematics ability in Japanese elementary schools. The original item response data are available upon request from the company (https://www.tokyo-shoseki.co.jp/company_english/). Data were also obtained from Mitsunaga and Uesaka (2025). These items were organized into five test forms and administered across multiple schools from 2018 to 2022. The detailed sampling procedure is described in the next section.

Several items contained two parts: one required students to "write down the formula," and the other to "calculate the previously answered formula." However, both parts were treated as a single item, considering their local dependence. The combined items were marked as correct only if the examinee answered both parts correctly. All responses were coded as "correct" and "incorrect," with no responses coded as missing.

This study used multiple test forms, which included anchor and separate test forms, as presented in Table 12. Each anchor test form comprised two parts: anchor and unique items. Anchor items were administered to focal grade groups (Grades 2–6) and adjacent grades. For example, the anchor test for Grade 3 contained items designated for Grade 3, as well as items for Grades 2 and 4. In this case, Grade 3 items were referred to as unique items, whereas Grades 2 and 4 items were considered anchor items. The separate test forms contained only items at a single difficulty level, corresponding to Grades 2–6, and the items were identical to those in the anchor test forms. The number of test items administered is presented in Table 12.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

5.1.2. Sampling design

First, we vertically scaled the five tests administered to different grades from 2 to 6. As no anchor items were present in these five tests, termed "original test" forms, we arranged another "anchor tests" forms containing anchor items and administered them to 2,453 students over five grades. Anchor test forms were administered to students in five elementary schools, covering grades 2 through 6. Therefore, 2,453 students completed the anchor tests (Grade 2 = 477, Grade 3 = 504, Grade 5 = 500, and Grade 6 = 488). Eighteen schools were considered for vertical scaling, with an average student number of 64.33 (*SD* = 22.81, Min = 36, and Max = 114). Separate test forms were administered to students attending schools in a particular city. We collected datasets from five consecutive administrations, spanning from 2018 (Grade 2) to 2022 (Grade 6). These datasets contained dichotomous response data from common examinees over five years. For example, the 2018 and 2019 datasets contained data of Grades 2 and 3 students, respectively. In other words, the longitudinal dataset comprised five dichotomous response datasets administered to the same group of students over five years; no common items were included across separate test forms. GCs were drawn by estimating the mathematics proficiency scores of 1,158 samples from 18 schools (average sample size = 64.333, *SD* = 23.470, Min = 36, and Max = 114) who took five original test forms for each academic year. These mathematical proficiency scores for these academic years were represented on the vertical scale using the anchor test. Note that the sample that was different from the students who took anchor tests, and this sample was analyzed for the proposed RDC–MGC model.

5.1.3. Vertical scaling procedure

A combined dataset was used to create a scale that can be interpreted as a common ability scale across the five grades. In this study, IRT analysis with a two-parameter logistic (2PL) IRT model was applied to obtain a common scale that could reflect students' ability levels on the same scale as the item difficulty levels. Before conducting the IRT analysis, we examined local dependencies by comparing Yen's Q_3 statistics (Yen, 1984) for each item; no items had values exceeding 0.3. Scree plots were used to assess unidimensionality for each dataset; the largest eigenvalue was significantly higher than the others in every dataset. Therefore, the dataset was considered to generally support the assumptions of unidimensionality and local independence. Although the vertical scaling included second-grade data, the third through sixth grades were used for longitudinal growth analysis. Second-grade data were employed for DCM measurements.

We discuss the measurement invariance of anchor test items between different grades and samples that answered the original and anchor tests. This is because such invariance can be assessed by calculating DIF or using multiple group IRT models. For either point, the unidimensionality of the anchor tests was verified because we constructed a vertical scale using a 2PL IRT model, which requires meeting the assumption of unidimensionality for each subscale of the class. We calculated polychoric correlation matrices and verified whether the largest eigenvalue marked a peak of scree plots for all grades (Table 4 in Mitsunaga & Uesaka, 2025). Based on the result, we concluded that all subscales had sufficient unidimensionality.

Next, the measurement invariance of the anchor test items between different grades was assessed by Lord's (1980) DIF detection method. We focused on the common items of the adjacent grades (e.g., Grades 3 and 4) test batteries. Then, the item response functions were estimated and assessed DIF between two grades. DIF was detected for almost all cases. However, this was a natural situation for vertical scaling because the different grades had different proficiency levels. The invariance of the anchor test items between the samples that took the original and anchor tests was also verified. Similarly, the DIF between samples was obtained. We analyzed the estimated item response functions of the two groups and found that, unlike discrimination parameters, difficulty parameters were generally similar. The dispersion of the discrimination parameters might be due to differences in the samples. Therefore, only the invariance of item difficulty was acceptable.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

In this study, we used the concurrent calibration method to construct the vertical scale, as this method effectively aligned students' abilities with item difficulty levels across grades (Mitsunaga & Uesaka, 2025). The comparison between separate and concurrent calibrations, two major vertical scaling methods in IRT, was discussed in Lee and Lee (2018) and Hanson and Béguin (2002). Mathematical abilities were estimated using the expected a posteriori (EAP) method, which was used in the GC model. The vertical scaling procedure was performed using the default settings of the "tam.mml.2pl" function in the TAM package (Robitzsch et al., 2022), adopting a multiple group design with the reference group set as Grade 4 students. Mathematical ability was assumed to follow a standard normal distribution.

5.2. Attribute definitions and Q-matrix construction procedure

To define the attributes and Q-matrix for the second-grade test items, we briefly review the background cognitive theory of mathematical tests. Deep understanding involves grasping the interrelations among various types of knowledge, whereas shallow understanding pertains to rote memorization and superficial application of procedures in problem-solving (e.g., Koedinger et al., 2012; Saso et al., 2024). Deep understanding is more conducive to reinforcing long-term memory and flexibly transferring knowledge to other contexts than shallow understanding (e.g., Koedinger et al., 2012). An example of a deep understanding of mathematical procedures is comprehending the principles behind the construction of the procedures or understanding why they are useful. Similarly, a deep understanding of mathematical terms entails knowing their meanings and providing concrete examples. Lachner and Nückles (2016) discussed the former, and Rittle-Johnson and Alibali (1999) the latter.

Saso et al. (2023a, 2023b, 2024) employed DCMs to diagnose attributes, such as shallow and deep understanding of mathematical terms, shallow and deep understanding of mathematical procedures/formulas, and calculation skills. In this study, following Saso et al. (2023a, 2023b, 2024), the third author defined the attributes and constructed an initial Q-matrix. The third author, specializing in educational psychology, holds teaching licenses for elementary school and mathematics at the middle and high school levels in Japan. These attributes were based on the contents of the second-grade level tests. As a result, the following attributes were defined: "A1. Shallow understanding of mathematical procedures," "A2. Deep understanding of mathematical procedures," and "A3. Calculation skills." Note that attributes related to the understanding of mathematical terms were not included in the current study, as the test items did not assess them.

The definitions are as follows: First, the A1 attribute was defined as "understanding how to construct equations from problem statements, reading the hands of a clock, or comparing lengths." Second, the A2 attribute was defined as "understanding the meaning of procedures and connecting them to concrete examples or diagrammatic representations." Third, the A3 attribute was defined as "the ability to correctly perform addition, subtraction, multiplication, and division between integers."

To clarify these attributes, we provided examples of test items. For example, the A1 attribute was measured by Q23, Q24, and Q29. The problem statements for Q23 and Q24 were as follows: "There are 9 apples. There are 3 more oranges than apples. How many oranges are there?" These items required constructing equations to calculate the total number of oranges. Additionally, Q29 asked, "Read the clock when Hiroki got to the park," and included an illustration of a clock. This item required reading the hands of the clock. Q27, which also measured A1, asked, "Compare the lengths. Which one is the longest?" This item assessed the ability to compare lengths.

Examples measuring the A2 attribute included Q33 and Q16. Q33 asked, "Which story matches the math sentence 7 – 4? Choose one and mark the number with an O," requiring the understanding of the equation "7 – 4" and its connection to an actual context. Q16 shown in Figure 2 asked, "Satoko is thinking about how to calculate 7 + 4. Fill in the blank with the correct number," which required understanding the meaning of the equation "7 + 4" and the procedure for calculation with regrouping. Figure 2 illustrates questions related to understanding the principle of calculation. To answer the item, students need to understand the principle of carry-up calculation; they must divide 4 into 3 and 1 against 7 to make 10: (7 + 3) + 1. A2 attribute (deep understanding of mathematical procedures) requires

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

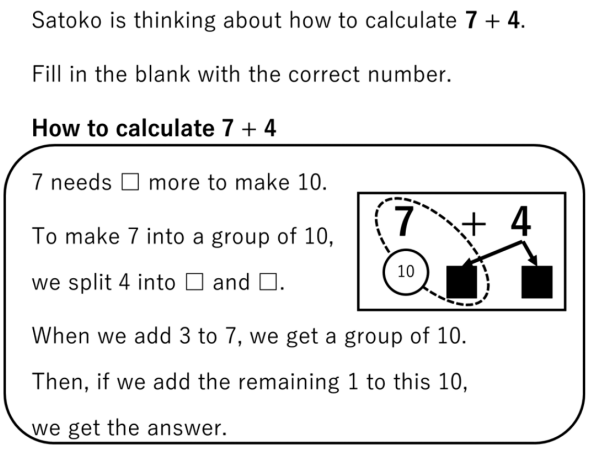


Figure 2. Modified question of Q16.

students to understand and solve problems connecting the principle and daily life situations or charts representing the making-ten strategy. Therefore, this calculation process is not only a simple calculation assumed as the A3 attribute (calculation skills). Finally, Q1 ("5+2") and Q7 ("13-9") measured the A3 attribute by requiring integer calculations.

After the third author constructed the initial Q-matrix, the fourth author, an educational psychologist, and third author discussed and confirmed the structure of the initial Q-matrix from the perspective of the content of the test items and theories of the depth of understanding. Specifically, the third author first explained the definitions of the attributes to the fourth author, and they discussed any discrepancies between the item contents and attribute definitions for each item. As the attributes reflected the nature of the items, they concluded that the items appropriately measured the attributes. The revised Q-matrix was obtained as the final version. Table 13 presents the Q-matrix for the RDCM and the descriptive statistics for each item in the second-grade test form.

Note that shallow and deep understanding could theoretically form a linear attribute hierarchy (Leighton et al., 2004); the Q-matrix reflected this structure. However, this hierarchy was partial, as A3 was not included in the linear structure. The attribute hierarchy is more appropriate for a large number of attributes, as it reduces the number of parameters. This study utilized only three attributes, assuming a structure would result in information loss during parameter estimation. Therefore, no structure was assumed for parameter estimation. Under the final Q-matrix setting, the generic identifiability of the DCM parameter estimation (Chen et al., 2020) was satisfied. The detailed estimation settings will be presented in the next section.

5.3. Data analysis settings

The entire sample size was 1,158. The 35 items from the second grade were used for diagnostic assessment. The mathematics abilities, based on vertically scaled tests from the third to sixth grades, were used for the growth model. These grades will be referred to as the first to fourth time points hereafter. Furthermore, data were collected from 18 schools and used to estimate group-level parameters and random intercepts in the DCM model.

We assumed three types of DCMs for the second grade: the LCDM, C-RUM, and DINA models. The LCDM is a saturated DCM that assumes all possible interaction terms among attributes. The DINA model is the simplest DCM, containing only slip and guessing parameters for each item. In the LCDM notation, the DINA model assumes only the intercept and highest interaction term, which can be considered a reparametrized version of the DINA model (e.g., DeCarlo, 2011). The C-RUM is intermediate between the LCDM and DINA models, as it includes only intercept and main effects. For

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 13. Q-matrix, mean, and standard deviation (SD) of dichotomous item responses for each item

Item	Attribute			Mean	*SD*
	A1: Shallow understanding of mathematical procedures	A2: Deep understanding of mathematical procedures	A3: Calculation skills		
Q1	0	0	1	0.990	0.097
Q2	0	0	1	0.973	0.162
Q3	0	0	1	0.963	0.188
Q4	0	0	1	0.975	0.157
Q5	0	0	1	0.989	0.106
Q6	0	0	1	0.986	0.117
Q7	0	0	1	0.956	0.204
Q8	0	0	1	0.955	0.208
Q9	0	0	1	0.979	0.143
Q10	0	0	1	0.932	0.252
Q11	1	0	1	0.973	0.163
Q12	1	0	1	0.952	0.214
Q13	1	0	1	0.984	0.127
Q14	1	0	1	0.907	0.290
Q15	1	0	1	0.961	0.194
Q16	1	1	1	0.636	0.481
Q17.Q18	1	0	1	0.976	0.154
Q19.Q20	1	0	1	0.869	0.338
Q21.Q22	1	0	1	0.902	0.297
Q23.Q24	1	0	1	0.728	0.445
Q25	1	0	1	0.905	0.293
Q26	1	0	1	0.741	0.438
Q27	1	0	0	0.956	0.204
Q28	1	0	0	0.986	0.117
Q29	1	0	0	0.939	0.240
Q30	1	0	0	0.849	0.358
Q31	1	0	0	0.932	0.252
Q32	1	1	0	0.846	0.361
Q33	1	1	0	0.813	0.390
Q34	1	0	0	0.737	0.440
Q35	1	0	0	0.705	0.456
Q36	1	1	0	0.701	0.458
Q37	1	1	1	0.711	0.454
Q38	1	1	0	0.392	0.488
Q39	1	0	1	0.915	0.279

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

each of these DCMs, we assumed single- and two-level models. In the single-level model, the random intercept, the group-level growth model, and the corresponding structural model were not assumed. The combination of DCMs and model levels results in six possible models: two-level LCDM, single-level LCDM, two-level C-RUM, single-level C-RUM, two-level DINA, and single-level DINA models. These six models were compared first based on the widely applicable information criterion (WAIC; Watanabe, 2010). Then, we interpreted the estimated parameters of the best model based on the WAIC.

Prior settings were as follows: The priors for the DCM parameters followed the approach by Zhan, Jiao, Man, et al. (2019). All priors were modeled with normal distributions, and their variance parameters were fixed at 4. The means of the prior distributions differed among effects: the intercept had a mean of −1.096, whereas the main and interaction effects had means of 0, with the main effects positively constrained. The parameters for the Dirichlet distribution for the attribute pattern mixing parameter were set to $\boldsymbol{\delta} = \mathbf{1}_L$. An inverse gamma distribution with $a = 0.01$ and $b = 0.01$ was used for the variance σ^2_θ of the random effect. The priors for the MGC components were mainly related to variance parameters. The parameters for the inverse gamma distributions for $\sigma^2_{\varepsilon^{(G)}_t}$ and $\sigma^2_{\varepsilon^{(P)}_{tm}}$ were set to $a = 0.01$ and $b = 0.001$, respectively. An inverse Wishart distribution was used for $\Sigma^{(P)}$ and $\Sigma^{(G)}$ with a two-dimensional identity matrix Ψ and 10 degrees of freedom ν. The structural model parameters were controlled by normal distributions. The attribute effects on the intercept, $\gamma^{(P)}_I$s, had a normal distribution with a mean of 0.5 and variance of 1.0, as we assumed that attribute mastery could affect the mathematics ability of the following year. The priors for the attribute effects on the slope $\gamma^{(P)}_S$ and group-level effects $\gamma^{(G)}_I$ and $\gamma^{(G)}_S$were modeled with standard normal distributions.

We employed Bayesian estimation based on the JAGS language (Plummer, 2003). Preliminary MCMC iterations revealed that some structural parameters were not identified, necessitating their fixation. Specifically, the attribute mastery pattern (110) was not estimated in the two-level LCDM; therefore, $\gamma^{(P)}_{I12}$ and $\gamma^{(P)}_{S12}$ were fixed at 0. Additionally, $\gamma^{(P)}_{I13}$ and $\gamma^{(P)}_{S13}$, which were in the single-level LCDM, were fixed at 0 owing to the inverse association observed in the MCMC trace plots for $\gamma^{(P)}_{I1}$ and $\gamma^{(P)}_{I13}$, as well as for $\gamma^{(P)}_{S1}$ and $\gamma^{(P)}_{S13}$. The two- and single-level C-RUM models applied the same constraints as the LCDMs.

In the two- and single-level DINA models, the structural parameters were not well identified; hence, we only assumed the highest interaction parameters, $\gamma^{(P)}_{I123}$ and $\gamma^{(P)}_{S123}$, consistent with their measurement models. The conditions of general identifiability proposed in Chen et al. (2020, p. 128) can be explained as follows: the Q-matrix should measure each attribute, and each attribute should be measured by more than two items. This implies that the Q-matrix in our real-data analysis satisfies generic identifiability. Therefore, the general LCDM measurement model is identifiable. However, a reviewer pointed out that the completeness condition defined in Chiu et al. (2009, p. 643) was not satisfied. In our current data analysis setting, we constrained γ parameters that are the path coefficients of attribute mastery to latent growth intercepts or slope in the DINA model case. However, that was enough based on the comment. We included additional constraints to restrict the attribute mastery pattern and eliminate indistinguishable attribute mastery patterns. Especially, the (010) pattern, which means mastering only a deep understanding of mathematics, could not be estimated with the Q-matrix. Therefore, the pattern was eliminated from the DINA model. Then, we estimated the model parameters.

Furthermore, a reviewer pointed out that a deep understanding of mathematical procedure might not be measured appropriately. Therefore, we conducted the same analysis for the collapsed Q-matrix that has the combined attribute "the understanding of mathematical procedure" which might be more appropriate than the original deep understanding attribute. Because MCMC iterations, except for the two-level LCDM with the collapsed Q-matrix, did not converge well and the model indicated a WAIC value worse than that of the best model with the original Q-matrix, which will be shown subsequently, the results are provided in Supplementary Material B. This indicated that the merged attribute setting was not appropriate for this dataset, and the shallow and deep understandings of mathematical procedures should have been distinguished.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 14. Descriptive statistics of mathematics ability for each grade

Grade	Mean	(*SE*)	*SD*	Median	Minimum	Max	Skewness	Kurtosis
3	−0.856	(0.022)	0.749	−0.802	−3.685	0.521	−0.529	0.209
4	0.135	(0.026)	0.887	0.183	−3.893	1.742	−0.398	0.058
5	1.141	(0.029)	0.997	1.187	−2.579	2.812	−0.383	−0.150
6	2.505	(0.030)	1.032	2.539	−1.845	4.452	−0.224	−0.372

In addition to the above model and the additional comment of the same reviewer, we conducted the same analysis assuming attribute hierarchy between shallow and deep understanding attributes of the original Q-matrix under the two-level LCDM that was the best fit model shown later. This analysis was intended to include intuitive relationship between the two attributes. Note that attribute hierarchy and measurement appropriateness refer to different points. The hierarchy assumption is about the relationship between attributes, and the measurement appropriateness is about between the attribute and items. Therefore, we believe that assuming attribute hierarchy does not solve the problem associated with the deep understanding attribute discussed above. Even under this consideration, for the reader interested in this hierarchy perspective, we conducted the additional analysis including the attribute hierarchy structure. However, MCMC iterations of four parameters did not converge. The WAIC value was 29,431.201(SE = 258.923) and larger than that from the best fitted model shown in Section 5.4. Therefore, it was concluded that the attribute hierarchy assumption was not appropriate in this data set.

Based on the parameter settings and computational memory limitations, different MCMC sampling settings were assumed. For the two-level LCDM, we employed two chains, each with a length of 80,000 iterations. The first three-fourths of the chain length served as the burn-in period; the thinning interval was set to 20. For the single-level LCDM, we used three chains, each with a length of 60,000 iterations. The burn-in period comprised the first half, and the thinning interval was 10. The MCMC settings for the two- and single-level C-RUM and DINA models were similar to those for the single-level LCDM, with the exception of a chain length of 40,000 iterations. These settings resulted in converged MCMC iterations for the model parameters ($\widehat{R} < 1.1$). After convergence, the attribute mastery patterns and log-likelihood for WAIC were sampled. For the LDCMs, 30,000 iterations were used for three chains with a thinning interval of 10, whereas for the other models, 20,000 iterations were used.

The MCMC iterations were conducted using the "jagsUI" package (Kellner, 2024), and the "psych" package (Revelle, 2024) was used to calculate descriptive statistics. Parameter estimates were interpreted using posterior means and 95% CI. Attribute mastery patterns were calculated based on posterior attribute mastery probabilities. If the posterior probability of attribute mastery exceeded 0.5, the attribute was considered mastered; otherwise, it was considered non-mastered. The main data analysis code is available on the Open Science Framework website: https://osf.io/gpztw.

5.4. Results

We first examined the descriptive statistics for the second-grade diagnostic test and the mathematics ability growth from the third to sixth grade. Table 14 presents the means (correct response ratios) and standard deviations (*SD*s) for each item. The test items generally showed a correct response ratio exceeding 0.7, indicating that the test was relatively easy. Item Q38 was the most difficult, with a difficulty index of 0.392.

Table 14 presents descriptive statistics for mathematics abilities from the third to sixth grade. As expected, the ability gradually increased over time. The minimum and maximum values suggested some overlap between grades. For instance, even in the sixth grade, some students scored lower than the average of third-grade students, and some third-grade students outperformed the average fourth-grade student. This implies that certain students may be significantly behind their actual chronological grade

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 15. WAIC values for the estimated models

Measurement model	Two-level		Single-level	
	WAIC	(*SE*)	WAIC	(*SE*)
LCDM	**29,329.382**	(259.045)	29,847.227	(260.411)
C-RUM	29,367.400	(258.766)	29,866.522	(259.709)
DINA	31,638.806	(266.450)	32,469.658	(265.039)

Note: LCDM is log-linear cognitive diagnostic model, C-RUM is compensatory reduced unified model, and DINA is deterministic input noisy AND-gate model.

in mathematics ability. Therefore, early diagnosis for intervention is necessary, and the lack of specific cognitive attributes should be assessed.

Table 12 presents the WAIC values for the six models. The two-level LCDM showed the smallest WAIC value (WAIC = 29,329.382, *SE* = 259.045), followed by the two-level C-RUM (WAIC = 29,367.400, *SE* = 258.766). Generally, the two-level models demonstrated better WAIC values, indicating the necessity of a multilevel structure. Data also indicated that simpler diagnostic measurement models were not appropriate for this case. Therefore, we interpreted the estimated parameters based on the two-level LCDM. After fitting the aforementioned models, we constrained the structural parameters of the two-level LCDM with 95% CI and included zero to refit the model in a stepwise manner. More precisely, we assumed three models that constrained the zero two-way interaction, the zero two-way and one-way interactions, and the main and interaction effects. These stepwise constraint models can help to inform which levels of the structural parameters are important. The WAICs of the constrained two-way interactions model, constrained one- and two-way interaction model, and constrained main and interactions model were 29,324.504 (*SE* = 258.994), 29,366.092 (*SE* = 259.102), and 29,362.482 (*SE* = 259.063), respectively. The constrained two-way interactions model indicated slightly better WAIC than the freely estimated two-level LCDM presented in Table 15, but the difference was small against *SE*s. This difference might be ignorable, but it might be possible that two-way interaction was not more important than the other one-way interactions or main effects whose 95% CIs covered 0. In addition to WAIC values, this was the first application example of the proposed method; it is informative to include parameter estimation results as much as possible. Therefore, we focused on the original two-level LCDM hereafter.

Table 16 presents the LCDM parameter estimates. Because of space limitations, Table 16 only presents posterior means. However, the complete table, including 95% CI (Table S11 in the Supplementary Material), is available on the OSF web page. Notably, all 95% CIs for the interaction parameters included zero. If we set these interactions to zero, the LCDM would reduce to the C-RUM. However, the interactions for Q14, $\lambda_{13} = 1.064$ $(95\%\text{CI}[-0.622, 3.267])$ and Q36, $\lambda_{12} = 0.610$ $(95\%\text{CI}[-0.146, 1.349])$ had lower bounds close to zero, indicating that these interactions were relatively far from the null value. By contrast, other interactions, such as λ_{12} for Q16, showed $\lambda_{12} = 1.064$ $(95\%\text{CI}[-2.537, 3.576])$, indicating that this interaction could be considered negligible.

Table 17 presents the parameter estimates for the mixing parameters of attribute mastery patterns. The pattern (110) was nearly zero: 0.003 $(95\%\text{CI}[0.000, .014])$, as observed in the preliminary analysis. The patterns (111), (101), and (011) exhibited relatively larger values: $\pi_{111} = 0.396$ $(95\%\text{CI}[0.312, .472])$, $\pi_{101} = 0.211$ $(95\%\text{CI}[0.129, 0.296])$, and $\pi_{011} = 0.144$ $(95\%\text{CI}[0.094, 0.193])$. Given that the second-grade diagnostic test was easy, the mixing probabilities reflected the ease of the test. However, although the mixing probability for only one attribute mastery pattern was less than 0.10, the estimates indicate that such a pattern may still exist in the population.

Table 18 presents the parameter estimates for the structural model components. The person-level main effects on the intercept were positive, and their 95% CIs did not include zero: $\gamma_{I1}^{(P)} = 0.852$ $(95\%\text{CI}[0.606, 1.062])$, $\gamma_{I2}^{(P)} = 0.913$ $(95\%\text{CI}[0.439, 1.280])$, and $\gamma_{I3}^{(P)} = 0.358(95\%\text{CI}[0.107, 0.627])$.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 16. Parameter estimate of LCDM parameters

Item	λ_0	λ_1	λ_2	λ_3	λ_{12}	λ_{13}	λ_{23}	λ_{123}
Q1	3.119			0.706				
Q2	1.369			2.016				
Q3	0.820			2.605				
Q4	1.087			3.172				
Q5	1.882			3.648				
Q6	1.779			2.968				
Q7	0.837			2.035				
Q8	0.511			2.817				
Q9	1.560			2.217				
Q10	0.383			1.922				
Q11	0.156	3.370		3.723		−1.841		
Q12	0.174	2.800		1.200		−0.098		
Q13	1.634	1.973		1.220		−0.566		
Q14	−1.427	2.511		2.564		1.064		
Q15	0.239	1.657		2.327		−0.405		
Q16	−3.308	1.915	1.551	0.987	0.421	−0.791	0.596	0.815
Q17.Q18	1.647	0.687		1.485		−0.654		
Q19.Q20	−0.737	0.849		1.539		0.149		
Q21.Q22	−0.784	2.265		1.644		−0.651		
Q23.Q24	−1.094	0.230		0.264		1.482		
Q25	0.046	1.089		0.810		0.067		
Q26	−0.856	1.011		0.728		−0.537		
Q27	1.318	1.533						
Q28	2.666	1.124						
Q29	0.637	2.585						
Q30	−0.548	2.558						
Q31	0.579	2.173						
Q32	−0.193	0.669	0.710		0.761			
Q33	−1.315	1.853	1.324		0.809			
Q34	−1.306	2.139						
Q35	−1.641	2.432						
Q36	−1.306	1.116	0.519		0.610			
Q37	−2.396	2.297	1.137	1.073	−0.333	−1.055	0.461	−0.459
Q38	−2.743	0.950	0.696		0.413			
Q39	−0.456	1.801		1.500		−0.454		

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 17. Estimates of the attribute pattern mixing parameters

Attribute mastery pattern	Mean	95% CI (Lower, Upper)
000	0.077	(0.053, 0.101)
100	0.085	(0.043, 0.136)
010	0.019	(0.006, 0.039)
110	0.003	(0.000, 0.014)
001	0.065	(0.032, 0.104)
101	0.211	(0.129, 0.296)
011	0.144	(0.094, 0.193)
111	0.396	(0.312, 0.472)

Table 18. Parameter estimates of the structural model part

Level	Parameter	Intercept		Slope	
		Mean	95% CI (Lower, Upper)	Mean	95% CI (Lower, Upper)
Person level	Main effect of attribute 1: $\gamma_1^{(P)}$	**0.852**	(0.606, 1.062)	−0.023	(−0.120, 0.076)
	Main effect of attribute 2: $\gamma_2^{(P)}$	**0.913**	(0.439, 1.280)	**−0.237**	(−0.425, −0.016)
	Main effect of attribute 3: $\gamma_3^{(P)}$	**0.358**	(0.107, 0.627)	0.016	(−0.099, 0.126)
	Interaction effect of attributes 1 and 3: $\gamma_{13}^{(P)}$	0.147	(−0.162, 0.458)	0.071	(−0.078, 0.214)
	Interaction effect of attributes 2 and 3: $\gamma_{23}^{(P)}$	−0.253	(−0.720, 0.226)	**0.284**	(0.048, 0.499)
	Interaction effect of attributes 1 to 3: $\gamma_{123}^{(P)}$	−0.247	(−0.493, 0.044)	0.030	(−0.095, 0.149)
Group level	Effect of random intercept: $\gamma^{(G)}$	**−2.107**	(−3.261, −1.220)	**0.936**	(0.397, 1.682)

This indicates that a shallow and deep understanding of mathematical procedures, as well as calculation skills, predict mathematical ability in the third grade. Interestingly, a deep understanding of mathematical procedures negatively affected the slope ($\gamma_{S2}^{(P)} = -0.237, 95\%\text{CI}\,[-0.425, -0.016]$). Furthermore, the interaction between a deep understanding of mathematical procedures and calculation skills positively affected the slope $\left(\gamma_{S23}^{(P)} = 0.284, 95\%\text{CI}\,[0.048, 0.499]\right)$. These effects are discussed subsequently.

Figure 3 illustrates the descriptive growth lines of mathematics ability by attribute mastery patterns. The effects of attribute mastery are evident in the figure. For example, the growth lines of individuals with the (000) attribute mastery pattern (upper-left panel) were generally lower than those of individuals with the (111) pattern (bottom-left panel). In the second upper panels, the slope of the (010) pattern appears slightly flatter than that of the (001) pattern. The estimated effects of attribute mastery on the growth lines were also supported by visual inspection. Notably, some students exhibited atypical growth trajectories. For example, the bottom line of the (001) pattern from third to fifth grade exhibited lower values but caught up in the sixth grade. The line for the (000) pattern, in contrast, did not show improvement throughout the period. These diagnostic test results may highlight severe cases for educational administrators and teachers, and it is crucial to closely monitor such cases. Figure 4 shows

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

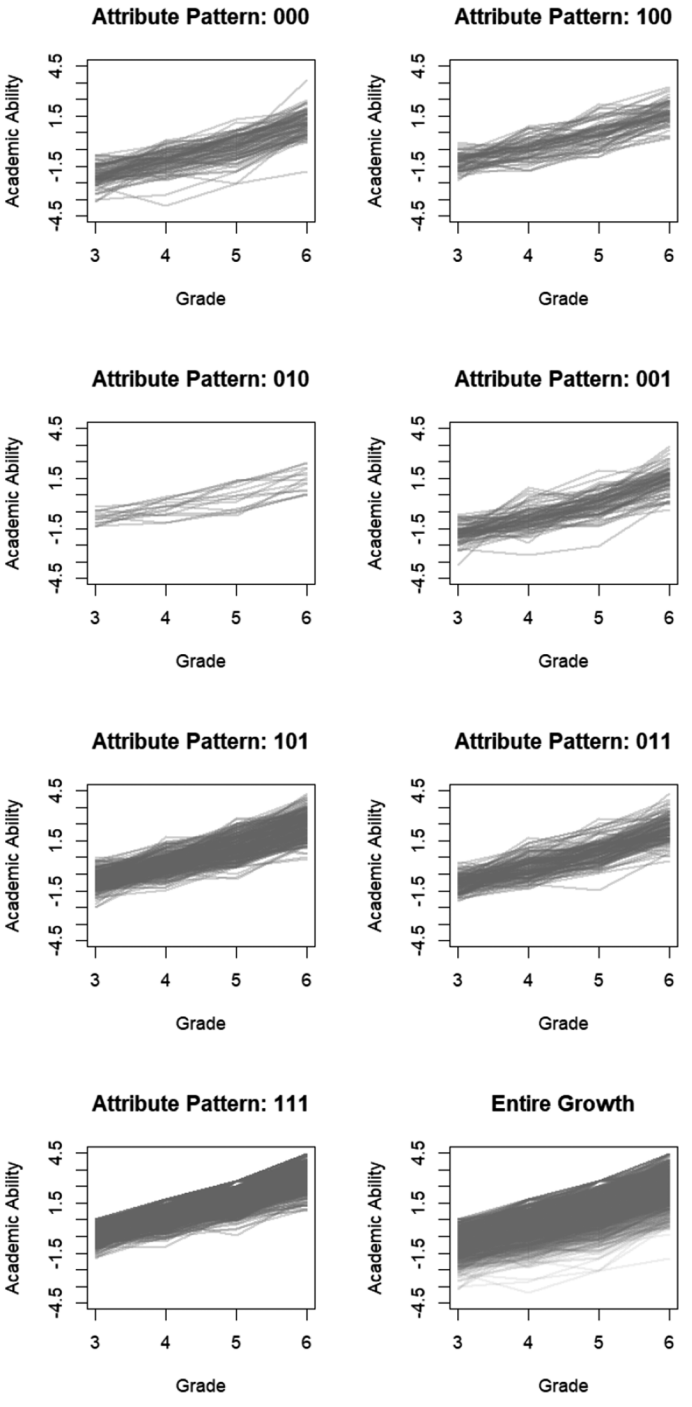


Figure 3. Growth patterns of academic ability for each attribute mastery status.

the GC of the difference in academic ability for each attribute mastery. As seen later, the mastery of each attribute indicated a higher level of academic proficiency at the third grade, intercepted in the proposed model. In these plots, the slopes of the GCs for mastery and non-mastery of attributes were similar. However, mastery of attributes affected the slope parameters, as discussed subsequently.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

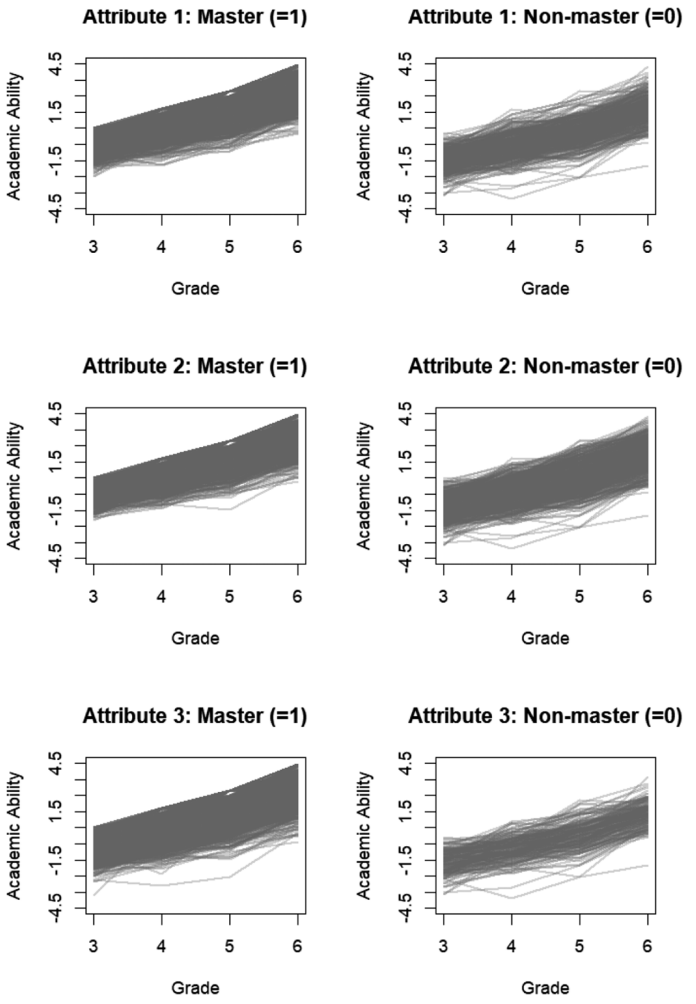


Figure 4. Growth curve difference of academic ability by each attribute mastery.

Table 18 presents the effects of the random intercept on the group-level intercept and slope: $\gamma_I^{(G)} = -2.107$ $(95\%\text{CI}[-3.261, -1.220])$ and $\gamma_S^{(G)} = 0.936$ $(95\%\text{CI}[0.397, 1.682])$, respectively. These results showed that the random intercept could be represented by group-level proficiency rather than just a DIF factor. In our data analysis, the prior for the random intercept was modeled as a normal distribution with a fixed zero mean and variance parameter. The posterior means of θ for schools ranged from 0.613 to 1.244; therefore, the group-level test-answering skill of each school was positive and not centered around zero based on the data. Additionally, the average mathematics ability at the third-grade level was negative. The negative value of $\gamma_I^{(G)}$ predicts a negative group-level mathematics score at the third grade. However, the value of $\gamma_S^{(G)}$ indicates that the increase rate of the group-level is positively affected. These parameters are group-level effects; hence, they should be carefully distinguished from individual-level phenomena.

Table 19 presents the variance, covariance, and correlation parameters. The person-level variances for the intercept and slope, given attribute mastery patterns, are $\sigma^2_{I(P)} = 0.096$ $(95\%\text{CI}[0.072, 0.125])$ and $\sigma^2_{S(P)} = 0.027$ $(95\%\text{CI}[0.022, 0.033])$, respectively. This indicates that individual differences still exist in mathematics ability at the third grade, and the growth rate can be explained by other individual

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 19. Estimates of variance, covariance, and correlation parameters

Parameter	Person level		Group level	
	Mean	95% CI (Lower, Upper)	Mean	95% CI (Lower, Upper)
Variance of intercept: σ_I^2	0.096	(0.072, 0.125)	0.114	(0.049, 0.251)
Variance of slope: σ_S^2	0.027	(0.022, 0.033)	0.108	(0.048, 0.230)
Covariance between intercept and slope: σ_{IS}	0.007	(−0.001, 0.015)	−0.002	(−0.080, 0.066)
Correlation between intercept and slope: ρ_{IS}	0.149	(−0.021, 0.318)	−0.0140	(−0.534, 0.520)
Residual variance at time 1: $\sigma_{\varepsilon_1}^2$	0.182	(0.160, 0.206)	0.008	(0.001, 0.032)
Residual variance at time 2: $\sigma_{\varepsilon_2}^2$	0.227	(0.205, 0.250)	0.003	(0.000, 0.012)
Residual variance at time 3: $\sigma_{\varepsilon_3}^2$	0.229	(0.204, 0.254)	0.008	(0.001, 0.034)
Residual variance at time 4: $\sigma_{\varepsilon_4}^2$	0.181	(0.148, 0.212)	0.179	(0.073, 0.379)
Variance of random intercept: σ_θ^2			1.166	(0.259, 3.385)

factors considered in this study. The group-level variations for the intercept and slope can be expressed as $\sigma_{I^{(G)}}^2 = 0.114$ $(95\%\text{CI}[0.049, 0.251])$ and $\sigma_{S^{(G)}}^2 = 0.108$ $(95\%\text{CI}[0.048, 0.230])$, respectively. The correlation between the intercept and slope at the person-level was slightly positive ($\rho_{I^{(P)}S^{(P)}} = 0.149$); however, the lower bound of the 95% CI was less than zero $(-0.021, 0.318)$. This shows that initial mathematics skills can affect the growth rate at the individual level, but the effect is not significant. The group-level correlation between the intercept and slope was extremely close to zero: $\rho_{I^{(G)}S^{(G)}} = -0.0140$ $(95\%\text{CI}[-0.534, 0.520])$. The person-level residual variances from time 1 to 4 (third to sixth grade) were almost equal and less than the group-level variances. The variance of the random intercept was positive, showing school-level differences in the ability to correctly answer diagnostic tests.

In summary, we demonstrated a multilevel structure for predicting mathematics growth with attribute mastery. Furthermore, the attribute mastery status affected both the intercept and growth of mathematics ability over four years. This effect is evident at the individual level, suggesting the importance of diagnosing individual learning status and providing targeted interventions for those lagging based on this information. The effects on individual intercept results showed that mastery of the three attributes is important to make a basis in the third grade. Especially, understandings of mathematical procedures had a larger effect on the intercept than simple calculation skills. These results indicate that a deep understanding is important even at an individual level. However, the results of the slope representing a change in the mathematical ability score show that a combination of deep understanding and calculation skills is essential for mathematical ability growth. Mastering only deep understanding may suppress learning.

5.5. A modified model

Previously, the model assumed linear growth in academic proficiency. However, Figure 4 suggested that some students might exhibit nonlinear growth trajectories. Therefore, we introduced a simple modification in which some slope-factor loadings were freely estimated (Bollen & Curran, 2006, Chapter 4) to allow alternative representations of growth trajectories. More precisely, we assumed that $\Lambda_Y^{(P)} = \begin{pmatrix} 1 & 0 \\ 1 & \lambda_{Y,2}^{(P)} \\ 1 & \lambda_{Y,3}^{(P)} \\ 1 & 1 \end{pmatrix}$ and $\Lambda_Y^{(G)} = \begin{pmatrix} 1 & 0 \\ 1 & \lambda_{Y,2}^{(G)} \\ 1 & \lambda_{Y,3}^{(G)} \\ 1 & 1 \end{pmatrix}$. We fixed $\lambda_{Y,1}^{(P)} = \lambda_{Y,1}^{(G)} = 0$ and $\lambda_{Y,4}^{(P)} = \lambda_{Y,4}^{(G)} = 1$, while $\lambda_{Y,t}^{(P)}$ and $\lambda_{Y,t}^{(G)}$ for $t = 2$ and 3 were freely estimated (Bollen & Curran, 2006, p. 102). In this specification, $\lambda_{Y,t}^{(P)}$ and

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Table 20. Parameter estimates of the structural model part in freely estimated slope loading model

Level	Parameter	Intercept		Slope	
		Mean	95% CI (Lower, Upper)	Mean	95% CI (Lower, Upper)
Person level	Main effect of attribute 1: $\gamma_1^{(P)}$	**0.861**	(0.644, 1.089)	−0.066	(−0.361, 0.203)
	Main effect of attribute 2: $\gamma_2^{(P)}$	**0.910**	(0.507, 1.242)	**−0.661**	(−1.148, −0.091)
	Main effect of attribute 3: $\gamma_3^{(P)}$	**0.395**	(0.140, 0.645)	−0.007	(−0.311, 0.295)
	Interaction effect of attributes 1 and 3: $\gamma_{13}^{(P)}$	0.094	(−0.193, 0.387)	0.260	(−0.102, 0.626)
	Interaction effect of attributes 2 and 3: $\gamma_{23}^{(P)}$	−0.308	(−0.752, 0.156)	**0.866**	(0.250, 1.457)
	Interaction effect of attributes 1 to 3: $\gamma_{123}^{(P)}$	−0.231	(−0.487, 0.095)	0.036	(−0.301, 0.361)
Group level	Effect of random intercept: $\gamma^{(G)}$	**−1.422**	(−2.127, −0.922)	**2.190**	(1.483, 3.136)

$\lambda_{Y,t}^{(G)}$ represent the proportion of change at each time point relative to the total change from the first to the final time point (Bollen & Curran, 2006, p. 102). This specification allows alternative representations of individual- and group-level nonlinear growth trajectories. Based on this model, we can examine which type of growth trajectory provides a useful representation the development of this academic proficiency development during elementary school.

The prior distribution of $\lambda_{Y,t}^{(P)}$ and $\lambda_{Y,t}^{(G)}$ for $t = 2$ and 3 was specified as $N(1, 1)$. The DCM component was based on the LCDM with a random effect, which showed the best fit in the previous analysis. The structural model specification was the same as that of the previously selected best-fitting model. The MCMC settings were also unchanged from the previous analysis. The same data set as in the previous section was used. Detailed model specifications are provided in the R code available on the OSF webpage.

The WAIC value was $29,405.153$ $(SE = 260.561)$, which was not better than that of the previously selected best-fitting model shown in Table 15. However, the difference was small relative to the standard error. Thus, the freely estimated loading model did not improve model fit, and the results did not provide strong evidence favoring the more flexible growth specification. Nevertheless, we briefly examined the estimated loading values and structural parameters for descriptive purposes.

First, the posterior mean of $\lambda_{Y,2}^{(P)}$ was $0.583\,(95\%\text{CI}\,[0.475, 0.693])$. This result suggests that the growth from third to fourth grade corresponded to 60% of the total person-level growth from third to sixth grade. Similarly, the posterior mean of $\lambda_{Y,3}^{(P)}$ was $0.955\,(95\%\text{CI}\,[0.823, 1.108])$, indicating that a substantial proportion of person-level growth had already occurred by fifth grade. These results suggest that person-level growth may exhibit slight nonlinearity. On the other hand , $\lambda_{Y,2}^{(G)}$ was estimated as 0.275 $(95\%\text{CI}\,[0.249, 0.298])$, and $\lambda_{Y,3}^{(G)}$ was estimated as $0.566\,(95\%\text{CI}\,[0.535, 0.596])$. These estimates suggest that group-level growth was relatively close to linear growth.

Finally, the estimates of the structural model parameters were shown in Table 20. The parameter estimates for the intercept factor were similar to those shown in Table 18. Because the scale of the slope factor changed under the freely estimated loading specification, the magnitudes of the slope-related structural parameters also changed, although the substantive interpretation remained largely similar. Specifically, the main effect of Attribute 2, $\gamma_2^{(P)} = -0.661$ $(95\%\text{CI}\,[-1.148, -0.091])$, and the interaction effect between Attributes 2 and 3, $\gamma_{23}^{(P)} = 0.866$ $(95\%\text{CI}\,[0.250, 1.457])$, were approximately three times larger than those reported in Table 18.

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

In summary, this additional analysis illustrates that the current RDC–MGC formulation can accommodate alternative trajectory specifications within the same longitudinal modeling setting while preserving the substantive interpretation of the relationship between attribute mastery and later growth trajectories.

6. Discussion and future study

This study proposed an RDC–MGC model, which assumes a random intercept for the DCM, a multilevel structure for the GC model, and a structural model linking the DCM and GC model. The primary purpose of the proposed model is to examine how diagnostic information relates to later developmental trajectories while accounting for multilevel data structures. The RDC–MGC model accounts for a group-level difference in the diagnostic test while also distinguishing between individual- and group-level mathematical ability growth. This feature enables the model to reveal the individual-level relationship between attribute mastery and growth, with the structural component representing the pure relationship between attribute mastery and individual growth. Two simulation studies showed that the two-level model with Bayesian estimation method provided smaller absolute biases and RMSEs and more appropriate coverages of 95% CI than single-level models. Real-data analysis demonstrated that the RDC–MGC model is more appropriate than single-level models. The parameter estimates indicated that attribute mastery at the second grade affected both the intercept and slope of mathematics ability from the third to sixth grades. A freely estimated loading specification, which represents a modification of the basic RDC–MGC model, was estimated to illustrate an alternative growth trajectory specification.

This study aimed to develop a model for controlling group-level effects and improving the effect of mastery of attributes on longitudinal growth. The proposed model interpreted the attribute mastery effects as individual effects, a concept that has not been addressed in previous studies. The real-data analysis demonstrated that basic attributes such as understanding of mathematical procedures or calculation skills predicted the individual-level growth during the later period of elementary school. In other words, cognitive abilities that ought to be mastered during the early period of elementary school could be identified. In addition to such empirical findings, the proposed model can be applied to other disciplines, such as language learning. Therefore, the proposed model can facilitate studies on the relationship between attribute mastery and its effects on subsequent ability growth.

The real-data example showed the effects of attribute mastery on both the intercept and slopes. Unlike $\gamma_S^{(P)}$s, $\gamma_I^{(P)}$s can be interpreted easily. In our data analysis, calculation skills had a positive effect, whereas a deep understanding of mathematical procedures had a contradictory effect on the slope.

First, the positive effect of calculation skills may indicate that calculation is a core skill for elementary school-level arithmetic ability. Calculation is essential for achieving high test scores in arithmetic and can also alleviate cognitive burdens when students learn more complex mathematical concepts or procedures. Furthermore, calculation skills may help maintain students' motivation to learn mathematics. In other words, calculation test items can serve as an enjoyable puzzle for students with strong calculation skills. This explains why calculation skills have a positive effect.

Second, the negative effect of a deep understanding of mathematical procedures is more challenging to interpret. In cognitive psychology, a deep understanding of mathematical procedures is generally considered beneficial. However, focusing excessively on understanding the meaning of mathematical procedures may limit the time available for practicing calculation skills. As aforementioned, calculation skills dominate elementary school mathematics, and students who prioritize deep understanding over calculation practice may fail to develop their calculation abilities, negatively impacting their test scores. This does not suggest that calculation practices should be performed without understanding; rather, a balance is necessary.

Several additional explanations might be possible. Mastering only "deep understanding" did not sufficiently represent a deep understanding of mathematical procedures. Although this is confusing,

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

mastering only "deep understanding" indicates that students could talk about the concept or meaning of the mathematical procedures but could not apply the knowledge to solve real mathematical problems. An appropriate "deep understanding" referred to mastering calculation procedures and understanding their meaning. This involved mastering both A2 and A3 attributes. Therefore, the estimation result revealed a problematic mastering pattern. Because of a lack of calculation skills, students who had only mastered "deep understanding" could not appropriately use procedural knowledge to solve mathematical problems, and this slowed the growth of general academic proficiency that was primarily represented as mathematical problem-solving ability, including calculation skills, in this study.

We analyzed the effect of mastering deep understanding on the individual intercept parameter and obtained a value of 0.913, which is the highest among the three attributes (see Table 15). In the third grade, some students might master a deeper understanding higher level than just mastering calculation, and they indicated higher initial scores in the third grade. However, students mastering only "deep understanding" neglected calculation. This meant there might be some heterogeneity in mastering "deep understanding." Using the item measuring the A2 attribute, students can connect procedures to concrete examples or diagrammatic representations; however, calculation is still important during the elementary school period. Therefore, although deep understanding is crucial for learning, the findings of this study suggest some conditions for such deep understanding.

Another consideration is that students might think they deeply understand the procedure, thus ignoring repetitive procedural practice, such as drill-like procedural exercises, in a learning situation. Similarly, an excessive focus on semantic understanding may increase the cognitive load for test items, and the procedures have not become automatic. This can result in overthinking, limiting the student's ability to solve test items and consequently reducing growth rate.

The positive interaction effect of a deep understanding of mathematical procedures and calculation on the slope compensated for the negative effect of a deep understanding of mathematical procedures. The interaction effects of A2 and A3 are understandable because previous studies have emphasized the necessity of both procedural and conceptual understanding (Rittle-Johnson, 2019, for a review). Furthermore, adaptive learners do not persist in a deeper procedure and can use a shallow procedure (Star, 2005; Star & Rittle-Johnson, 2008). This interpretation is consistent with the literature. The estimated positive effect was greater than the negative effect. Therefore, students who learned both attributes tended to grow faster than those who mastered only one of the two attributes. This suggests that the benefits of deep understanding of mathematical procedures may be more likely to emerge when accompanied by sufficient calculation skills. A deep understanding of procedures is crucial for comprehending higher-level mathematical concepts. Our data analysis shows that during elementary school, both understanding and calculation skills must be balanced. They are complementary and both are important to appropriately improve mathematical ability.

This study has several limitations. First, the parameter estimation of the RDC–MGC model is computationally intensive. Our setting required several hours to complete necessary iterations. Furthermore, the large number of model parameters necessitates significant memory capacity on our computer. Therefore, more computationally efficient estimation methods, such as variational inference (Cho et al., 2021; Rijmen & Jeon, 2013; Yamaguchi, 2023; Yamaguchi & Martinez, 2024; Yamaguchi & Okada, 2020) or stochastic optimization techniques (Cai, 2010a, 2010b; Robbins & Monro, 1951; von Davier & Sinharay, 2010; Zhang et al., 2020), are required.

Second, developing a diagnostic test and vertically scaled tests is challenging. Moreover, collecting data for these tests is not straightforward. However, the use of computers in classrooms may facilitate data collection and enhance the extension of this model.

Third, our dataset lacks covariates at both the person and group levels. If covariates were available, they could be used to explain the variance in intercepts and slopes within the proposed model. Furthermore, these covariates could help identify factors that influence individual growth beyond attribute mastery. For instance, teachers' teaching strategies (Carbonneau et al., 2013), students' learning

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

strategies (Aleven & Koedinger, 2002), and motivation (Metallidou & Vlachou, 2007) could all be important factors to assess in future research.

It may be beneficial to consider applicable fields of the RDC–MGC model for future study. Basically, the RDC–MGC model is intended to examine later growth using diagnostic mastery status. Keeping this structure in mind, we can change the diagnostic part. Some DCMs can reveal students' misconceptions (e.g., Kuo et al., 2018; Ozaki et al., 2019); therefore, we can assess what kind of misconceptions and their combination influence the subsequent growth of academic proficiency. For example, Kuo et al. (2018) assumed three misconceptions in the fraction multiplication test in elementary school: "turning the second fraction upside down when multiplying a fraction by a fraction," "solving only the first step of a two-step problem," and "performing incorrect arithmetic operations when confused about the relational terms." These misconceptions may prevent later mathematical learning. Ozaki et al. (2019) analyzed two misconceptions in the Reading Skill Test in Japanese: "in which the nearest word is the subject" and "in which the word beginning a sentence is the subject." These are the basics of reading text and may result in deficiency in reading proficiency. Researchers may be interested in the interaction effects of misconceptions on subsequent learning achievement and overall learning growth.

Furthermore, DCMs have been employed for non-academic contexts such as clinical psychology, which may be an interesting application field of the RDC–MGC model. Templin and Henson (2006) applied DCMs to psychological assessment. Wang et al. (2019) developed a depression diagnostic measure with DCMs, and Bao et al. (2025) applied DCMs for the behavioral and emotional problems of children. Liang et al. (2024) identified mental health symptom profiles of college students, and de la Torre et al. (2017) analyzed data from the Dutch version of Millon Clinical Multiaxial Inventory-III which measures multiple psychological disorders. Those diagnostic results can be predictors of later psychological latent GCs. For example, the GC model has been employed for school or clinical psychology: self-efficacy beliefs in regard to anger and sadness regulation during adolescence (Di Giunta et al., 2023), adolescent depressive symptom changes (Gataviņš et al., 2025), developmental change in mental problems, such as stress or anxiety during college (Liu et al., 2023), and so on. Based on those studies, diagnoses of mental problems can be used to examine how they affect subsequent depression trajectories. Another possible application is examining how emotional problems affect children's self-efficacy beliefs.

Although the current study has several limitations and the main objective of the RDC–MGC model is to predict later latent growth using diagnostic information, several possible extensions can be considered. Here, we discuss the future extensions. First, the measurement part can be changed to other DCMs. This is employed in the real-data analysis section. In addition, for the usual multilevel models (e.g., Raudenbush & Bryk, 2002), the group-level variables can be included in the measurement model to explain group differences. Furthermore, the attribute effects can include random effects that are random slopes in the multilevel models.

Second, the GC model can be extended to the non-linear growth case. For example, quadratic or higher-order slope terms can be added, or some slope paths can be freely estimated (Bollen & Curran, 2006, Chapter 4); the latter specification was shown in the real-data analysis. Piecewise approach with unknown knots and its extensions can be used to capture critical change points in the learning trajectory (e.g., Kohli & Harring, 2013; Kohli et al., 2013). Such growth functions can be assumed at both the individual and group levels. Furthermore, conditional latent curve models (Bollen & Curran, 2006, Chapter 5) enable us to incorporate additional time-invariant covariates. These are some possible extensions. Note that the growth period does not have to be as long as the empirical example in the manuscript. Assuming shorter periods, such as one year or less, may be appropriate for examining development within a school year. For example, if the diagnostic assessment is conducted at the first period of the school year, such information can be used to explain subsequent student growth.

Third, the structural part can also be changed. It is possible to assume nonlinear or polynomial regression functions. Furthermore, as mentioned above, group-level variables can also be introduced into the structural model, and nonlinear functions can be employed. For example, school type (public/private) or teachers' teaching strategies are group-level factors and may improve students' learning and result

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

in higher slopes or intercepts. In addition, interactions among group-level predictors and attributes are cross-level interactions (e.g., Raudenbush & Bryk, 2002, p. 29). Mediation variables can be inserted between the attributes and growth factors. This extension makes it possible to examine the mechanisms through which attributes improve individual learning. Furthermore, if an educational intervention is conducted based on the diagnostic assessment, the treatment variable can be included in the structural model.

As a different perspective, we can set attribute mastery as the final outcome and use growth models as explanatory variables. This is not a minor modification of the RDC–MGC model, but the important part is that the proposed model is based on the combination of DCMs, SEM, and multilevel models. Therefore, various research questions can be examined using the extended model. The extensions of the proposed method are relatively straightforward.

The utility of the above models should ultimately be evaluated by using empirical datasets in future studies. Generally, statistical methods are tailored for specific purposes, so the simplest RDC–MGC model can only address limited educational research questions. However, because the RDC–MGC model is composed of DCMs, multilevel modeling, and SEM, each element can be modified for various study objectives.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/psy.2026.10126.

AI use declaration. During the preparation and revision of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.5) to assist with English language editing, grammar checking, and refinement of academic writing. All AI-generated outputs were reviewed and edited by the authors, who take full responsibility for the content of the manuscript.

Data availability statement. The data analysis code is available on the Open Science Framework page: https://osf.io/gpztw. We appreciate Tokyo Shoseki Co. for providing the real data.

Funding statement. This work was supported by JSPS KAKENHI 20H01720, 22K13810, 23H00985, 23H00065, 23K20759, 24K00485, and 25K17133.

Competing interests. The authors declare no competing interests.

References

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer—Based cognitive tutor. *Cognitive Science*, *26*(2), 147–179. https://doi.org/10.1207/s15516709cog2602_1

Ayan, C., & Çıkrıkçı, R. N. (2021). A study on formative assessment with cognitive diagnostic model application for math proficiency. *The International Journal of Assessment and Evaluation*, *29*(1), 37–55. https://doi.org/10.18848/2327-7920/cgp/v29i01/37-55

Bao, Y., Liu, J., DiStefano, C., & Ding, R. (2025). Identifying children at risk for emotional and behavioral problems: A diagnostic classification model approach. *Psychology in the Schools*, *62*(5), 1355–1368. https://doi.org/10.1002/pits.23394

Bishop, M. (2006). *Pattern recognition and machine learning*. Springer.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. John Wiley & Sons.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). American Council on Education/Praege.

Carbonneau, K. J., Marley, S. C., & Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology*, *105*(2), 380–400. https://doi.org/10.1037/a0031084

Chen, J., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the programme for international student assessment in reading. *Psychology*, *5*(18), 1967–1978. https://doi.org/10.4236/psych.2014.518200

Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, *85*(1), 121–153. https://doi.org/10.1007/s11336-019-09693-2

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665. https://doi.org/10.1007/S11336-009-9125-0

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *74*(S1)52–85. https://doi.org/10.1111/bmsp.12219

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. *718*). John Wiley & Sons.

de Ayala, R. J. (2013). *The theory and practice of item response theory*. (2nd ed.) Guilford Publications.

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

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

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

de la Torre, J., van der Ark, L. A., & Rossi, G. (2017). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281–296. https://doi.org/10.1080/07481756.2017.1327286

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*(1), 8–26. https://doi.org/10.1177/0146621610377081

Depaoli, S. (2021). *Bayesian structural equation modeling*. The Guilford Press.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling*, *22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Di Giunta, L., Lunetti, C., Lansford, J. E., Eisenberg, N., Pastorelli, C., Bacchini, D., Tirado, L. M. U., Iselin, A. R., Basili, E., Gliozzo, G., Favini, A., Cirimele, F., & Remondi, C. (2023). Predictors and outcomes associated with the growth curves of self-efficacy beliefs in regard to anger and sadness regulation during adolescence: A longitudinal cross-cultural study. *Frontiers in Psychology*, *14*, 1010358. https://doi.org/10.3389/fpsyg.2023.1010358

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fox, J.-P. (2010). Multilevel item response theory models. In *Bayesian item response modeling* (pp. 141–191). Springer. https://doi.org/10.1007/978-1-4419-0742-4_6

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(1), 20–23. https://doi.org/10.1007/BF02294839

Gataviņš, M. M., Visoki, E., Hoffman, K. W., Almasy, L., & Barzilay, R. (2025). Adolescent depressive symptom trajectories from before to after the COVID-19 pandemic. *JAMA Network Open*, *8*(12), e2545987. https://doi.org/10.1001/jamanetworkopen.2025.45987

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common item equating design. *Applied Psychological Measurement*, *26*(1), 3–24. https://doi.org/10.1177/0146621602026001001.

Hartz, S., & Roussos, L. (2008). The fusion model for skills diagnosis: Blending theory with practice. *ETS Research Report Series, 08–71*, 1–57. https://www.ets.org/Media/Research/pdf/RR-08-71.pdf

Hattie, J. (2023). *Visible learning: The sequel*. Routledge.

Heck, R. H., & Reid, T. (2023). Multilevel structural equation modeling. In R. H. Hoyle (Ed.), *Hand book of structural equation modeling* (2nd ed., pp. 481–499). The Guilford Press.

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling*, *17*(2), 193–215. https://doi.org/10.1080/10705511003659342

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

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement*, *54*(4), 440–480. https://doi.org/10.1111/jedm.12156

Huang, H. Y., & Wang, W. C. (2014). Multilevel higher-order item response theory models. *Educational and Psychological Measurement*, *74*(3), 495–515. https://doi.org/10.1177/0013164413509628

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

Kellner, K. (2024). *jagsUI: A wrapper around "rjags" to streamline "JAGS" analyses*. (R package version 1.6.2.). https://kenkellner.com/jagsUI/

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research*, *48*(3), 370–397. https://doi.org/10.1080/00273171.2013.778191

Kohli, N., Harring, J. R., & Hancock, G. R. (2013). Piecewise linear–linear latent growth mixture models with unknown knots. *Educational and Psychological Measurement*, *73*(6), 935–955. https://doi.org/10.1177/0013164413496812

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. (3rd ed.) Springer Science + Business Media.

Kuo, B., Chen, C., & de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement*, *42*(3), 179–191. https://doi.org/10.1177/0146621617722791

Lachner, A., & Nückles, M. (2016). Tell me why! Content knowledge predicts process-orientation of math researchers' and math teachers' explanations. *Instructional Science*, *44*(3), 221–242. https://doi.org/10.1007/s11251-015-9365-6

Lee, W., & Lee, G. (2018). IRT linking and equating. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development volume two* (pp. 639–673). John Wiley & Sons.

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

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, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28–54. https://doi.org/10.1111/jedm.12061

Liang, Q., de la Torre, J., Larimer, M. E., & Mun, E.-Y. (2024). Mental health symptom profiles over time: A three-step latent transition cognitive diagnosis modeling analysis with covariates. In M. Stemmler, W. Wiedermann, & F. L. Huang (Eds.), *Dependent data in social sciences research* (pp. 539–562). Springer. https://doi.org/10.1007/978-3-031-56318-8_22.

Liu, X., Zhang, Y., Gao, W., & Cao, X. (2023). Developmental trajectories of depression, anxiety, and stress among college students: A piecewise growth mixture model analysis. *Humanities and Social Sciences Communications*, *10*(1), 1–10. https://doi.org/10.1057/s41599-023-02252-2

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

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

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*(2), 99–120. https://doi.org/10.3102/10769986002002099

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*, 187–212. https://doi.org/10.1007/BF02294535

Metallidou, P., & Vlachou, A. (2007). Motivational beliefs, cognitive engagement, and achievement in language and mathematics in elementary school children. *International Journal of Psychology*, *42*(1), 2–15. https://doi.org/10.1080/00207590500411179

Mitsunaga, H., & Uesaka, Y. (2025). Construction of a vertical scale for mathematics in a Japanese elementary school using a post-hoc anchor test: A comparison of scale aligning methods and classification of longitudinal patterns of ability. *Japanese Journal of Educational Psychology*, *73*(1), 14–27. https://doi.org/10.5926/jjep.73.14 (In Japanese).

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism Clinical and Experimental Research*, *24*(6), 882–891. https://doi.org/10.1111/j.1530-0277.2000.tb02070.x

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide*. (8th ed.) Muthén & Muthén.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. Routledge.

Oravecz, Z., & Muth, C. (2018). Fitting growth curve models in the Bayesian framework. *Psychonomic Bulletin and Review*, *25*(1), 235–255. https://doi.org/10.3758/s13423-017-1281-0

Ozaki, K., Sugawara, S., & Arai, N. (2019). Cognitive diagnosis models for estimation of misconceptions analyzing multiple-choice data. *Behaviormetrika*, *47*(1), 19–41. https://doi.org/10.1007/s41237-019-00100-9

Pan, Q., Qin, L., & Kingston, N. (2020). Growth modeling in a diagnostic classification model (DCM) framework—A multivariate longitudinal diagnostic classification model. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01714

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *The 3rd international workshop on distributed statistical computing* (Vol. *124*, pp. 1–8). http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Preacher, K., Wichman, A., MacCallum, R., & Briggs, N. (2008). *Latent growth curve modeling*. SAGE Publications. https://doi.org/10.4135/9781412984737

Pritikin, J. N., Hunter, M. D., von Oertzen, T., Brick, T. R., & Boker, S. M. (2017). Many-level multilevel structural equation modeling: An efficient evaluation strategy. *Structural Equation Modeling*, *24*(5), 684–698. https://doi.org/10.1080/10705511.2017.1293542

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. https://doi.org/10.1007/BF02295939

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis*. (2nd ed.) Sage Publications.

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research*. (R package version 2.4.6). https://CRAN.R-project.org/package=psych

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research*, *206*(1), 647–662. https://doi.org/10.1007/s10479-012-1181-7

Rittle-Johnson, B. (2019). Iterative development of conceptual and procedural knowledge in mathematics learning and instruction. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 124–147). Cambridge University Press.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, *91*(1), 175–189. https://doi.org/10.1037/0022-0663.91.1.175

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400–407. https://doi.org/10.1214/aoms/1177729586

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules*. (R package version 4.0-16.). https://CRAN.R-project.org/package=TAM

Roussos, L. A., Dibello, L. V, Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnosis system. In J. Leighton, & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Chapter, Cambridge University Press.

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

Saso, S., Oka, M., & Uesaka, Y. (2023a). Development of assessment tools for depth of understanding quantitatively with cognitive diagnostic models. In K. Arai (Ed.), *Advances in information and communication. FICC 2023. Lecture notes in networks and systems* (Vol. *651*, pp. 766–774). Springer. https://doi.org/10.1007/978-3-031-28076-4_55

Saso, S., Oka, M., & Uesaka, Y. (2023b). Diagnosing the depth of understanding using cognitive diagnostic models and their application to a regular test: Examining the applicability of cognitive diagnostic models with a Q-matrix specified through qualitative and quantitative procedures. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *30*(4), 515–530. https://doi.org/10.11225/cs.2023.057

Saso, S., Oka, M., Uesaka, Y., & Usami, S. (2024). Diagnostic assessment of deep understanding using cognitive diagnostic models: A large-scale assessment to promote the use of effective learning strategies. PsyArxiv. https://doi.org/10.31234/osf.io/hk5y2

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (2nd ed.) Sage Publishers.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*(5), 404–411. https://doi.org/10.2307/30034943

Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction*, *18*(6), 565–579. https://doi.org/10.1016/j.learninstruc.2007.09.018

Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: Effect on remedial instruction as empirical validation. *Journal of Educational Measurement*, *34*(1), 3–20. https://doi.org/10.1111/j.1745-3984.1997.tb00504.x

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

Vermunt, J. K. (2003). Multilevel latent class models. In R. M. Stoltzenberg (Ed.), *Sociological methodology* (pp. 213–239). ASA.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*(1), 33–51. https://doi.org/10.1177/0962280207081238

von Davier, M. (2007). Hierarchical general diagnostic models. *ETS Research Reports Provide, RR-07-19*, 1–21. https://doi.org/10.1002/j.2333-8504.2007.tb02061.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

von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Springer Nature.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*(2), 174–193. https://doi.org/10.3102/1076998609346970

Wang, C. (2021). Using penalized EM algorithm to infer learning trajectories in latent transition CDM. *Psychometrika*, *86*, 167–189. https://doi.org/10.1007/s11336-020-09742-1

Wang, C., & Nydick, S. W. (2020). On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics*, *45*(3), 339–368. https://doi.org/10.3102/1076998619882026

Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a new instrument for depression with cognitive diagnosis models. *Frontiers in Psychology*, *10*, 1306. https://doi.org/10.3389/fpsyg.2019.01306

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 57–87. https://doi.org/10.3102/1076998617719727

Wang, W. C., & Qiu, X. L. (2019). Multilevel modeling of cognitive diagnostic assessment: The multilevel DINA example. *Applied Psychological Measurement*, *43*(1), 34–50. https://doi.org/10.1177/0146621618765713

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Wickrama, K. A. S., Lee, T. K., O'Neal, C. W., & Lorenz, F. O. (2016). *Higher-order growth curves and mixture modeling with Mplus: A practical guide*. Routledge/Taylor & Francis Group.

Yamaguchi, K. (2023). Bayesian analysis methods for two-level diagnosis classification models. *Journal of Educational and Behavioral Statistics*, *48*(6), 773–809. https://doi.org/10.3102/10769986231173594

Yamaguchi, K., & Martinez, A. J. (2024). Variational Bayes inference for hidden Markov diagnostic classification models. *British Journal of Mathematical and Statistical Psychology*, *77*(1), 55–79. https://doi.org/10.1111/bmsp.12308

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

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika*, *85*, 973–995. https://doi.org/10.1007/s11336-020-09739-w

Yamaguchi, K., & Templin, J. (2022a). A Gibbs sampling algorithm with monotonicity constraints for diagnostic classification models. *Journal of Classification*, *39*(1), 24–54. https://doi.org/10.1007/s00357-021-09392-7

Yamaguchi, K., & Templin, J. (2022b). Direct estimation of diagnostic classification model attribute mastery profiles via a collapsed Gibbs sampling algorithm. *Psychometrika*, *87*(4), 1390–1421. https://doi.org/10.1007/s11336-022-09857-7

Yamaguchi, K., & Templin, J. (2022c). Observed score reliability indices in diagnostic classification models. *Behaviormetrika*, *49*(1), 47–68. https://doi.org/10.1007/s41237-021-00153-9

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, *44*(3), 251–281. https://doi.org/10.3102/1076998619827593

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*, *44*(4), 473–503. https://doi.org/10.3102/1076998619826040

Zhan, P., Man, K., Wind, S. A., & Malone, J. (2022). Cognitive diagnosis modeling incorporating response times and fixation counts: Providing comprehensive feedback and accurate diagnosis. *Journal of Educational and Behavioral Statistics*, *47*(6), 736–776. https://doi.org/10.3102/107699862211110852

Zhang, S., & Chang, H. (2020). A multilevel logistic hidden Markov model for learning under cognitive diagnosis. *Behavior Research Methods*, *52*(1), 408–421. https://doi.org/10.3758/s13428-019-01238-w

Zhang, S., Chen, Y., & Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology*, *73*(1), 44–71. https://doi.org/10.1111/bmsp.12153

Cite this article: Yamaguchi, K., Mitsunaga, H., Saso, S. and Uesaka, Y. (2026). Predicting Multilevel Growth Trajectories Using a Random-Effect Diagnostic Classification Model. *Psychometrika*, 1524–1569. https://doi.org/10.1017/psy.2026.10126

https://doi.org/10.1017/psy.2026.10126 Published online by Cambridge University Press