Psychometrika (2026), **91**, 995–1019
doi:10.1017/psy.2026.10097

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

A Joint Model for Graded Responses and Response Times

Xinyu Zhang[1], Xiangbin Meng[1], Wei Gao[1] and Gongjun Xu[2]

[1]School of Mathematics and Statistics, Key Laboratory of Applied Statistics of MOE, Key Laboratory of Big Data Analysis of Jilin Province, Northeast Normal University, Changchun, Jilin, China; [2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA
Corresponding author: Xiangbin Meng; Email: mengxb600@nenu.edu.cn

(Received 25 May 2025; revised 16 December 2025; accepted 12 February 2026)

Abstract
While the joint modeling of item responses and response times (RTs) has received considerable attention, most existing approaches remain limited to dichotomous items and are not applicable to assessments involving polytomous or mixed-format items. To address this limitation, this article proposes a novel joint modeling framework for graded item responses and RTs. Specifically, we develop a conditional RT model given item responses and integrate it with a marginal response model based on Samejima's graded response model, yielding a conditional joint model for graded item responses and RTs. The model is then embedded within a two-level hierarchical framework to account for the relationship between ability and speed at the population level. A key methodological contribution is the development of a stochastic approximation EM (SAEM) algorithm for estimating the proposed model, which efficiently computes its marginal maximum likelihood estimates. Simulation studies demonstrate the accurate parameter recovery of the SAEM algorithm and indicate that the proposed model outperforms the hierarchical model assuming conditional independence across various testing conditions. Finally, an empirical analysis using data from the 2022 Programme for International Student Assessment illustrates the effectiveness of the graded response–response time model in large-scale assessments.

Keywords: graded item response; joint modeling; response times; SAEM algorithm

1. Introduction

In educational and psychological measurement, item response times (RTs), as a type of process data, offer valuable insights into examinees' test-taking behaviors and performance (Lee & Chen, 2011; Ratcliff & Kang, 2021; Schneider et al., 2023; Sinharay, 2020; Ulitzsch et al., 2020a, 2020b, 2022; Wang & Xu, 2015; Wang et al., 2018). With the increasing prevalence of computerized testing, the routine collection of RTs has become more feasible, providing researchers with unprecedented opportunities to explore its psychometric properties and applications across diverse assessment settings.

A central methodological task is the development of appropriate statistical models for analyzing RTs. Early work (e.g., Entink, van der Linden, et al., 2009; Rouder et al., 2003; van der Linden, 2006) focused on modeling RTs independently of item responses, aiming to characterize their statistical properties and infer latent cognitive mechanisms. Several distributional models have been proposed in this context, including the log-normal model (van der Linden, 2006), the Box–Cox normal model (Entink, van der Linden, et al., 2009), the semiparametric model (Wang et al., 2013), and the shifted Weibull distribution

© 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.2026.10097 Published online by Cambridge University Press

(Rouder et al., 2003), among which the log-normal model has gained wide adoption due to its estimation simplicity and robust empirical performance (Sinharay & van Rijn, 2020).

To better leverage RTs for psychological and educational measurement, it is crucial to propose the joint model for RTs and item responses. One of the most widely used frameworks is the hierarchical model (HM) proposed by van der Linden (2007), which specifies separate submodels for RTs and responses at the first level and captures their correlations at the person and item levels through a second-level structure. This framework focuses on the between-person relationship between ability and speed and has been extended in various directions to accommodate different testing contexts (Entink, Fox, et al., 2009; Klein Entink et al., 2009; Ulitzsch et al., 2022; Zhan, Jiao, & Liao, 2018; Zhan, Jiao, Wang, et al., 2018). A key limitation of the HM, however, is its assumption of conditional independence between responses and RTs given the latent traits. Empirical studies have questioned this assumption, demonstrating residual dependence not fully explained by the correlation between latent traits (Bolsinova & Tijmstra, 2019; Bolsinova, De Boeck, et al., 2017; Meng et al., 2015). To address this issue, various conditional joint models for item responses and RTs have been proposed.

One approach is modeling the joint distribution of RTs and responses as a bivariate distribution with a nonzero dependence parameter. For example, Ranger and Ortner (2012) proposed a conditional joint model that incorporates a covariance structure to capture local dependence between speed and accuracy. Extending this approach, Meng et al. (2015) allowed the conditional correlation to vary across both items and individuals, thereby improving the model's flexibility in capturing person- and item-specific dependencies. Another widely used strategy factorizes the joint distribution into the product of a marginal distribution for RTs and a conditional distribution for responses given RTs. This approach provides greater modeling flexibility while maintaining interpretability. For instance, Bolsinova, De Boeck, et al. (2017) introduced a joint model that incorporates residual RTs into the item response model to account for unexplained dependence. Similarly, Bolsinova, Tijmstra, et al. (2017) proposed a joint modeling framework in which the conditional distribution of responses given RTs follows a two-parameter normal ogive (2PNO) model, with the intercept expressed as a linear function of log-transformed RTs. In contrast, Bolsinova and Tijmstra (2019) proposed another strategy by reversing the factorization direction. Specifically, they factorized the joint distribution into the product of a marginal distribution for responses and a conditional distribution for RTs given responses. This formulation allows for modeling RT heterogeneity as a function of response accuracy, explicitly distinguishing between correct and incorrect responses.

In addition, process-based approaches have emerged that seek to model both RTs and responses as outputs of an underlying cognitive process. These models integrate item response theory with decision-making models. For instance, diffusion IRT models embed a diffusion decision process within an IRT framework (Tuerlinckx & De Boeck, 2005) and were implemented through the diffIRT package (Molenaar et al., 2015b). Building on these works, Kang et al. (2022) extended the diffusion IRT framework by allowing person- and item-level variability, thereby allowing the model to capture the conditional dependence between responses and RTs. Further, Kang, Jeon, et al. (2023) introduced a latent space diffusion IRT model in which both persons and items are embedded in a latent space to capture complex dependency between RTs and responses. These models depart from conventional statistical formulations and instead aim to recover the cognitive mechanisms driving observed RT and response patterns.

Despite these methodological advances, most existing joint models have been developed for dichotomous items, and research on polytomous items remains scarce. Molenaar et al. (2015a, 2015c) demonstrated that, for graded response items, the HM can be reformulated as a generalized linear factor model when parameter dependence is ignored, enabling estimation using Mplus. These studies extended the applications of the HM and provided powerful and efficient computational tools for estimating joint models of item responses and RTs. Nevertheless, their work does not propose a dedicated joint modeling framework for graded-response or mixed-format assessments. More recently, Kang, Molenaar, et al.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

(2023) proposed a process-based model for Likert-type items, but their framework is confined to non-cognitive assessments. Consequently, a fully specified joint model for graded responses and RTs in cognitive assessment is still lacking. This gap underscores the need for joint models that can accommodate these item formats and effectively capture the relationship between item responses and RTs.

In this study, drawing on the framework of Bolsinova and Tijmstra (2019), we develop a joint model for graded responses and RTs. The proposed approach develops a conditional RT model given item responses and integrates it with a marginal response model based on Samejima's graded response model (GRM; Samejima, 1969). This integration yields a conditional joint model that captures the within-person relationship between item responses and RTs. We further extend this model into a hierarchical formulation that accommodates between-person relationship between ability and speed. A second major contribution of this work is the development of a stochastic approximation EM (SAEM) algorithm for computing the marginal maximum likelihood estimate (MMLE) of the proposed model. The proposed estimation procedure is capable of handling mixed-format assessments that contain both binary and graded items. Finally, the proposed model and estimation procedure are evaluated through two simulation studies and an empirical analysis, demonstrating their practical applicability and strong performance.

The remainder of the article is organized as follows. Section 2 introduces a joint modeling framework for graded item responses and RTs. Section 3 presents an SAEM algorithm for estimating the proposed model. Section 4 reports simulation studies evaluating parameter recovery and comparing the graded response–response time (GR–RT) model with the HM. Section 5 applies the proposed method to empirical data to demonstrate its practical utility. Finally, we conclude with a discussion of the findings and directions for future research.

2. Model and rationale

In this section, we present a joint modeling framework for graded item responses and RTs, beginning with an empirical example that illustrates the dependence between RTs and response categories. Building on these observations, we construct a conditional joint model and further extend it into a two-level hierarchical structure.

2.1. Empirical investigation of the relationship between RTs and response categories

In this empirical analysis, we examine the differences in RTs across students with varying graded responses, using data from the Programme for International Student Assessment (PISA) scientific literacy assessment. The dataset contains responses from 717 Spanish students to 10 items, including six dichotomous and four polytomous items.

Because RTs are typically modeled on the log scale, we analyze the logarithmic transformation of RTs. Substantial variation in mean log-RTs is observed across response categories. As shown in the left panel of Figure 1, for a four-category item ("DS643Q04R"), the mean log-RTs increases with response category, particularly from category 1 to 2. F tests confirm that these differences are statistically significant at the 0.05 level. This pattern suggests that responses in higher categories tend to be associated with longer RTs. However, part of this relationship may be driven by individual differences in speed. To account for this, we conducted a residual-based analysis.

Specifically, we modeled the RT data using the log-normal model proposed by Klein Entink et al. (2009), which decomposes log-RT into item-level and person-level components. Let T_{ij} denote the RT variable of test taker i on item j, with t_{ij} denoting its realization. This model is given by

$$\log T_{ij} = \beta_j - \varphi_j \nu_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma_j^2),$$

where $\nu_i \in (-\infty, +\infty)$ is the latent speed trait for examinee i, and $\beta_j \in (-\infty, +\infty)$, $\varphi_j \in (0, +\infty)$, and $\sigma_j^2 \in (0, +\infty)$ denote the time intensity, speed discrimination, and the variance parameters for item j,

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

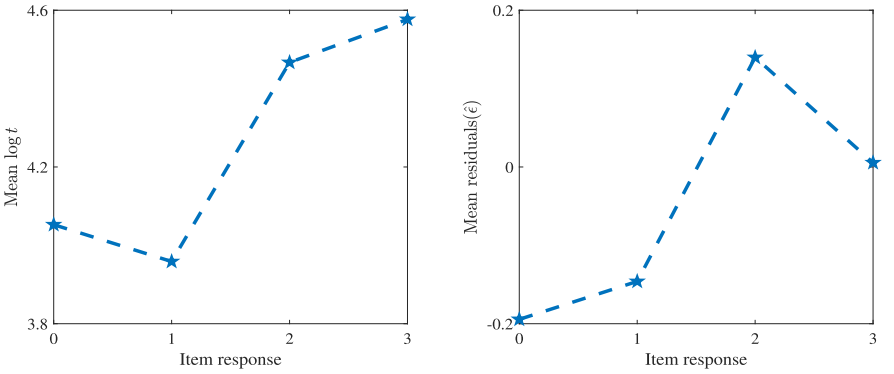


Figure 1. Line charts of mean log–response times ($\log t_{ij}$) and mean residuals ($\hat{\varepsilon}_{ij}$) across different response categories for item "DS643Q04R."

respectively. After estimating the model, we obtained parameter estimates $\hat{\beta}_j, \hat{\varphi}_j, \hat{\sigma}_j$, and $\hat{v}_i$, and computed residuals

$$\hat{\varepsilon}_{ij} = \log t_{ij} - \hat{\beta}_j + \hat{\varphi}_j \hat{v}_i,$$

which represent log-RTs adjusted for individual speed and item-level effects. These residuals allow for a reassessment of RT differences across response categories without confounding due to person or item variability.

The right panel of Figure 1 presents the mean residuals across response categories. Compared to the raw log-RTs, the residual pattern differs slightly, displaying a peak at category 2 followed by a decline. This indicates that part of the trend in raw log-RTs was attributable to individual speed differences. Nevertheless, an overall upward pattern remains, and F tests confirm that the differences in mean residuals across categories are remain statistically significant at the 0.05 level.

Overall, these findings indicate a clear relationship between RTs and response categories. To more adequately capture this relationship, we propose an RT model that explicitly incorporates category-specific effects, as described in the following section.

2.2. A joint model for graded responses and RTs

We begin by defining some notation. Let test takers be indexed by $i = 1, \ldots, N$ and items indexed by $j = 1, \ldots, M$. Let Y_{ij} be the response variable of test taker i to item j, with y_{ij} representing its observed value. Without loss of generality, we assume $y_{ij} = 0, \ldots, K_j - 1$, where $K_j \geq 2$ is the total number of response categories. Finally, let $I_{(\cdot)}$ denote the indicator function.

2.2.1. A conditional model for RTs given item responses

To capture the effect of the ordinal response Y_{ij} on the RTs, we adopt a dummy coding approach. Specifically, we define a dummy variable vector as $\mathbf{Z}_{Y_{ij}}$,

$$\mathbf{Z}_{Y_{ij}} = \left(Z_{ijk}\right)_{1 \leq k \leq K_j - 1}, \quad \text{where} \quad Z_{ijk} = I_{(Y_{ij}=k)}. \tag{1}$$

Let $\mathbf{z}_{y_{ij}} = \left(z_{ijk}\right)_{1 \leq k \leq K_j - 1}$ denote an observation of $\mathbf{Z}_{Y_{ij}}$. Under this encoding scheme, the relationship between y_{ij} and $\mathbf{z}_{y_{ij}}$ is given by

$$\begin{aligned} y_{ij} = 0 &\Leftrightarrow \mathbf{z}_{y_{ij}} = (0, 0, \ldots, 0) \\ y_{ij} = 1 &\Leftrightarrow \mathbf{z}_{y_{ij}} = (1, 0, \ldots, 0) \\ &\vdots \\ y_{ij} = K_j - 1 &\Leftrightarrow \mathbf{z}_{y_{ij}} = (0, 0, \ldots, 1). \end{aligned} \tag{2}$$

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

As noted above, the log-normal model is widely used for modeling RTs because it fits various data well and allows for reliable inference and future extensions. Therefore, we adopt the log-normal model in this study. To incorporate the influence of response categories on RTs, we extend the log-normal model by introducing category-specific parameters. The resulting conditional model for RTs given the observed response $Y_{ij} = y_{ij}$ is

$$\ln T_{ij}|y_{ij} = \beta_j - \varphi_j \nu_i + \mathbf{z}_{y_{ij}}\boldsymbol{\delta}_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma_j^2). \tag{3}$$

Here, ν_i, β_j, φ_j, and σ_j^2 are introduced above; $\boldsymbol{\delta}_j = (\delta_{j1}, \delta_{j2}, \ldots, \delta_{j(K_j-1)})^\top$ represents the item response-specific effect parameter for item j. The parameter $\boldsymbol{\delta}_j$ quantifies the systematic differences in RTs associated with each response category.

2.2.2. Item response model

In this study, we employ the widely-used GRM proposed by Samejima (1969) to model the ordered categorical response variable Y_{ij}. Under the GRM, the probability that test taker i selects response category k $(k = 0, \ldots, K_j - 1)$ for item j is given by

$$p_{ijk} = P(Y_{ij} = k|\theta_i, a_j, b_j, \boldsymbol{\tau}_j) = P(Y_{ij} \geq k|\theta_i, a_j, b_j, \boldsymbol{\tau}_j) - P(Y_{ij} \geq k+1|\theta_i, a_j, b_j, \boldsymbol{\tau}_j),$$

where

$$\begin{aligned} P(Y_{ij} \geq 0 \mid \theta_i, a_j, b_j, \boldsymbol{\tau}_j) &= 1, \\ P(Y_{ij} \geq k \mid \theta_i, a_j, b_j, \boldsymbol{\tau}_j) &= F(a_j\theta_i - b_j - \tau_{jk}), \quad k = 1, 2, \ldots, K_j - 1, \\ P(Y_{ij} \geq K_j \mid \theta_i, a_j, b_j, \boldsymbol{\tau}_j) &= 0. \end{aligned}$$

Here, θ_i denotes the latent ability of test taker i, a_j is the discrimination parameter, b_j is the difficulty parameter, and $\boldsymbol{\tau}_j = (\tau_{j1}, \ldots, \tau_{j(K_j-1)})$ is the thresholds for item j, which correspond to the cut-off points between the response categories. The threshold parameters must satisfy the monotonicity condition $\tau_{j1} < \tau_{j2} < \cdots < \tau_{j(K_j-1)}$. In addition, to ensure identifiability between b_j and $\boldsymbol{\tau}$, we impose the constraint: $\sum_{k=1}^{K_j-1} \tau_{jk} = 0$ for $j = 1, \ldots, M$. The function $F(\cdot)$ refers to the cumulative distribution function (CDF) of either the logistic or the standard normal distribution. In this study, we use the standard normal CDF for convenience in developing the SAEM algorithm for model estimation.

Finally, based on the dummy variable vector $\mathbf{Z}_{Y_{ij}}$ in Equations (1) and (2), the probability mass function of the item response variable Y_{ij} can be expressed as

$$P(Y_{ij} = y_{ij}|\theta_i, a_j, b_j, \boldsymbol{\tau}_j) = p_{ij0}^{1-\sum_{k=1}^{K_j-1} z_{ijk}} \prod_{k=1}^{K_j-1} p_{ijk}^{z_{ijk}}, \quad \text{where } y_{ij} = 0, \ldots, K_j - 1. \tag{4}$$

2.2.3. A joint model for graded responses and RTs

Let $\boldsymbol{\zeta}_i = (\theta_i, \nu_i)^\top$ denote the latent trait vector of examinee i. For item j, let the GRM parameters be $\boldsymbol{\eta}_j = (a_j, b_j, \boldsymbol{\tau}_j)^\top$, and let the RT model parameters be $\boldsymbol{\alpha}_j = (\beta_j, \varphi_j, \boldsymbol{\delta}_j^\top, \sigma_j)^\top$. The complete parameter vector for item j is given by $\boldsymbol{\xi}_j = (\boldsymbol{\eta}_j^\top, \boldsymbol{\alpha}_j^\top)^\top$.

Based on the conditional RT model in Equation (3) and the GRM in Equation (4), the conditional joint model of Y_{ij} and T_{ij} given $\boldsymbol{\zeta}_i$ can be written as

$$f(t_{ij}, y_{ij}|\boldsymbol{\zeta}_i, \boldsymbol{\xi}_j) = \frac{1}{t_{ij}} \phi\left(\frac{\ln t_{ij} - \beta_j + \varphi_j \nu_i - \mathbf{z}_{y_{ij}}\boldsymbol{\delta}_j}{\sigma_j} \right) P(Y_{ij} = y_{ij}|\theta_i, a_j, b_j, \boldsymbol{\tau}_j), \tag{5}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Following the hierarchical modeling framework proposed by van der Linden (2007), the population-level distribution of the latent variables $\boldsymbol{\zeta}_i$ is assumed to follow a bivariate normal distribution

$$\boldsymbol{\zeta}_i = (\theta_i, \nu_i)^\top \sim N_2(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \tag{6}$$

where $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ denote the mean vector and covariance matrix, respectively. For identification purposes, these parameters are typically specified as

$$\boldsymbol{\mu}_p = (0,0)^\top, \quad \boldsymbol{\Sigma}_p = \begin{pmatrix} 1 & \sigma_{\theta\nu} \\ \sigma_{\theta\nu} & 1 \end{pmatrix}.$$

Finally, based on Equations (5) and (6), a two-level hierarchical joint model for Y_{ij} and T_{ij} is given by

$$\begin{cases} f(t_{ij}, y_{ij} | \boldsymbol{\zeta}_i, \boldsymbol{\xi}_j) = \dfrac{1}{t_{ij}} \phi\left(\dfrac{\ln t_{ij} - \beta_j + \varphi_j \nu_i - \mathbf{z}_{y_{ij}} \boldsymbol{\delta}_j}{\sigma_j} \right) P(Y_{ij} = y_{ij} | \theta_i, a_j, b_j, \boldsymbol{\tau}_j), \\ f(\boldsymbol{\zeta}_i | \sigma_{\theta\nu}) = \dfrac{1}{2\pi\sqrt{1-\sigma_{\theta\nu}^2}} \exp\left\{ -\dfrac{\theta_i^2 + \nu_i^2 - 2\sigma_{\theta\nu}\theta_i\nu_i}{2(1-\sigma_{\theta\nu}^2)} \right\}. \end{cases} \tag{7}$$

In this article, the proposed joint model is referred to as the GR–RT model, as it accounts for both graded responses and response times. Notably, when $\boldsymbol{\delta}_j = \mathbf{0}$, the GR–RT model reduces to the basic HM framework proposed by Klein Entink et al. (2009) for graded scoring items, where responses and RTs are conditionally independent. Moreover, since binary scoring is a special case of graded scoring, the GR–RT model provides a unified framework for analyzing test data that include both dichotomous and polytomous items.

Remark 1. The proposed GR–RT model is not a direct extension of the heterogeneity model introduced by Bolsinova and Tijmstra (2019), primarily due to certain practical challenges. In their framework, heterogeneity between correct and incorrect responses is modeled through multiple aspects, including the slope parameters (φ_j), residual variance (σ_j^2), and the joint population distribution of θ and ν. For polytomous items, modeling different slope parameters and residual variances for each response category, along with different distributions of θ and ν, results in a highly complex model structure. This complexity introduces significant challenges: not only does it make parameter estimation more difficult, but it also reduces the overall credibility of the model due to increased uncertainty in the estimation process. Moreover, real-world data often exhibit sparsity in certain response categories, which magnifies the uncertainty and undermines the reliability of statistical inference. Given these practical challenges, the present study adopts a more parsimonious approach by modeling response category-specific heterogeneity solely through intercept terms. While this approach captures only one aspect of heterogeneity, it provides a robust, manageable framework for analyzing polytomous items in the presence of RT data, and it lays the groundwork for future extensions that could incorporate richer heterogeneity structures as the field advances.

The practical applicability of a measurement model depends on the feasibility of accurately estimating its parameters. The complex structure of the GR–RT model presents significant challenges for parameter estimation. Therefore, developing an effective computational method to estimate the parameters of the GR–RT model is a central focus of this study. In the following section, we present an SAEM algorithm designed to compute the MMLE of the GR–RT model.

3. Model estimation

This section presents an SAEM algorithm for estimating the GR–RT model. In Section 3.1, a data augmentation scheme is introduced for the GR–RT model. Building on this data augmentation, Section 3.2 presents an SAEM algorithm for computing the MMLE of the GR–RT model.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Table 1. Relationship between Y_{ij} and $\mathbf{X}_{ij}$ for four-category graded response

		$Y_{ij} \geq 1$	$Y_{ij} \geq 2$	$Y_{ij} \geq 3$	X_{ij1}	X_{ij2}	X_{ij3}
Y_{ij}	0	−	−	−	< 0	< 0	< 0
	1	+	−	−	≥ 0	< 0	< 0
	2	+	+	−	≥ 0	≥ 0	< 0
	3	+	+	+	≥ 0	≥ 0	≥ 0

Note: + indicates the condition is satisfied and − indicates not satisfied.

3.1. A data augmentation scheme for GR–RT model

Data augmentation is fundamental for developing the SAEM algorithm. Following Camilli and Geis (2019), we introduce a latent response vector $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ij(K_j-1)})^\top$, where each X_{ijk} follows a normal distribution:

$$X_{ijk} \overset{\text{i.i.d.}}{\sim} N(a_j\theta_i - b_j - \tau_{jk}, 1^2). \tag{8}$$

Moreover, the relationship between the observed response Y_{ij} and the latent response X_{ijk} is defined as

$$X_{ijk} = \begin{cases} \geq 0, & \text{if } Y_{ij} \geq k, \\ < 0, & \text{otherwise,} \end{cases} \quad \text{for } k = 1, \ldots, K_j - 1. \tag{9}$$

To intuitively illustrate this relationship between $\mathbf{X}_{ij}$ and Y_{ij}, consider a four-category graded response item, where $Y_{ij} \in \{0,1,2,3\}$ and $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, X_{ij3})^\top$. When Y_{ij} takes the values 0, 1, 2, and 3, the corresponding ranges of the elements of $\mathbf{X}_{ij}$ are determined, as shown in Table 1.

We now define some notations. Let $\mathbf{y}_{\cdot j} = (y_{1j}, \ldots, y_{Nj})^\top$, $\mathbf{t}_{\cdot j} = (t_{1j}, \ldots, t_{Nj})^\top$, and $\ln \mathbf{t}_{\cdot j} = (\ln t_{1j}, \ln t_{2j}, \ldots, \ln t_{Nj})^\top$ denote the observed response, RT and log-RT vectors of item j, respectively. The observed response and RT data for all items are denoted by $\mathbf{y} = (\mathbf{y}_{\cdot 1}, \ldots, \mathbf{y}_{\cdot M})$ and $\mathbf{t} = (\mathbf{t}_{\cdot 1}, \ldots, \mathbf{t}_{\cdot M})$. Let $\mathbf{X}_j = (\mathbf{X}_{1j}, \mathbf{X}_{2j}, \ldots, \mathbf{X}_{Nj})$ denote the latent response matrix for item j. Let $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij(K_j-1)})^\top$ denote the realization of $\mathbf{X}_{ij}$, and $\mathbf{x}_j = (\mathbf{x}_{1j}, \mathbf{x}_{2j}, \ldots, \mathbf{x}_{Nj})$ denote the realization of $\mathbf{X}_j$. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^\top$ and $\mathbf{v} = (v_1, \ldots, v_N)^\top$ denote the latent ability and speed vectors, respectively. The combined latent traits are represented as $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \mathbf{v})$. The complete data for the test, including observed responses, RTs, latent responses, and latent traits, are denoted by $\mathbf{u} = (\mathbf{y}, \mathbf{t}, \mathbf{x}^\top, \boldsymbol{\zeta})$, while the complete data for item j are denoted by $\mathbf{u}_{\cdot j} = (\mathbf{y}_{\cdot j}, \mathbf{t}_{\cdot j}, \mathbf{x}_j^\top, \boldsymbol{\zeta})$. Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_M)$ be the collection of GRM parameters and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M)$ be the collection of RT model parameters. The collection of all item parameters in the GR–RT model, combining of $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$, is denoted by $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M)$. Let $\boldsymbol{\Omega} = (\boldsymbol{\xi}, \sigma_{\theta v})$ be the set of parameters to be estimated in the GR–RT model.

Based on Equations (7)–(9), we have

$$\begin{aligned} f(y_{ij}, t_{ij}, \mathbf{x}_{ij}, \boldsymbol{\zeta}_i | \boldsymbol{\xi}_j, \sigma_{\theta v}) &= f(y_{ij}, t_{ij}, \mathbf{x}_{ij} | \boldsymbol{\zeta}_i, \boldsymbol{\xi}_j) f(\boldsymbol{\zeta}_i | \sigma_{\theta v}) \\ &= \prod_{k=1}^{K_j - 1} [I_{(x_{ijk} < 0)} I_{(y_{ij} < k)} + I_{(x_{ijk} \geq 0)} I_{(y_{ij} \geq k)}] \phi(x_{ijk} - a_j\theta_i + b_j + \tau_{jk}) \\ &\quad \times \frac{1}{t_{ij}} \phi\left(\frac{\ln t_{ij} - \beta_j + \varphi_j v_i - \mathbf{z}_{y_{ij}} \boldsymbol{\delta}_j}{\sigma_j} \right) f(\boldsymbol{\zeta}_i | \sigma_{\theta v}). \end{aligned} \tag{10}$$

Further, the complete-data likelihood of $\boldsymbol{\xi}_j$ and $\sigma_{\theta v}$ can be expressed as

$$L(\mathbf{u}_{\cdot j} \mid \boldsymbol{\xi}_j, \sigma_{\theta v}) = \prod_{i=1}^{N} f(y_{ij}, t_{ij}, \mathbf{x}_{ij}, \boldsymbol{\zeta}_i | \boldsymbol{\xi}_j, \sigma_{\theta v}) \propto L_j^1 \times L_j^2 \times L^3,$$

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

where

$$L_j^1 = \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K_j-1}\left(x_{ijk} - a_j\theta_i + b_j + \tau_{jk}\right)^2\right\} I_{\left(\tau_{j1}<\cdots<\tau_{j(K_j-1)},\ \sum_{k=1}^{K_j-1}\tau_{jk}=0\right)}, \tag{11}$$

$$L_j^2 = \sigma_j^{-N}\exp\left\{-\frac{1}{2\sigma_j^2}\sum_{i=1}^{N}\left(\ln t_{ij} - \beta_j + \varphi_j v_i - \mathbf{z}_{y_{ij}}\boldsymbol{\delta}_j\right)^2\right\}, \tag{12}$$

$$L^3 = \left(1-\sigma_{\theta v}^2\right)^{-N/2}\exp\left\{-\frac{1}{2(1-\sigma_{\theta v}^2)}\sum_{i=1}^{N}\left(\theta_i^2 + v_i^2 - 2\sigma_{\theta v}\theta_i v_i\right)\right\}. \tag{13}$$

It can be observed that $L(\mathbf{u}_{\cdot j} \mid \boldsymbol{\xi}_j, \sigma_{\theta v})$ is the product of three normal likelihoods, which places it within the exponential family. When the data likelihood takes the form of an exponential family (Barndorff-Nielsen, 2014), the implementation of the EM algorithm and its extensions is simplified to working with sufficient statistics (Dempster et al., 1977). Therefore, the exponential family structure of the complete data likelihood is particularly advantageous for developing an SAEM algorithm for the GR–RT model.

For brevity, the exponential family form of $L(\mathbf{u}_{\cdot j} \mid \boldsymbol{\xi}_j, \sigma_{\theta v})$ is not given here; instead, we present only the sufficient statistics for $\boldsymbol{\xi}_j$ and $\sigma_{\theta v}$.

Specifically, the sufficient statistics for $\boldsymbol{\eta}_j$ are

$$S_{\boldsymbol{\eta}_j} = \left\{\sum_{i=1}^{N}\sum_{k=1}^{K_j-1}\theta_i x_{ijk}, \sum_{i=1}^{N}\mathbf{x}_{ij}, \sum_{i=1}^{N}\theta_i, \boldsymbol{\theta}^\top\boldsymbol{\theta}\right\}, \tag{14}$$

where $\sum_{i=1}^{N}\mathbf{x}_{ij} = \left(\sum_{i=1}^{N}x_{ij1}, \ldots, \sum_{i=1}^{N}x_{ij(K_j-1)}\right)^\top$.

The sufficient statistics for $\boldsymbol{\alpha}_j$ are

$$S_{\boldsymbol{\alpha}_j} = \left\{\ln\mathbf{t}_j^\top \ln\mathbf{t}_{\cdot j}, \boldsymbol{\Lambda}_j \ln\mathbf{t}_{\cdot j}, \boldsymbol{\Lambda}_j\boldsymbol{\Lambda}_j^\top\right\}, \tag{15}$$

where

$$\boldsymbol{\Lambda}_j = \begin{pmatrix} v_1 & v_2 & \cdots & v_N \\ -1 & -1 & \cdots & -1 \\ -I_{(y_{1j}=1)} & -I_{(y_{2j}=1)} & \cdots & -I_{(y_{Nj}=1)} \\ \vdots & \vdots & \ddots & \vdots \\ -I_{(y_{1j}=K_j-1)} & -I_{(y_{2j}=K_j-1)} & \cdots & -I_{(y_{Nj}=K_j-1)} \end{pmatrix}.$$

Finally, the sufficient statistics for $\sigma_{\theta v}$ are

$$S_{\sigma_{\theta v}} = \left\{\boldsymbol{\theta}^\top\boldsymbol{\theta}, \mathbf{v}^\top\mathbf{v}, \boldsymbol{\theta}^\top\mathbf{v}\right\}. \tag{16}$$

3.2. An SAEM algorithm for the MMLE of GR–RT model

We begin by introducing the EM algorithm for computing the MMLE of the GR–RT model, followed by the proposal of the SAEM algorithm.

3.2.1. EM algorithm

With the introduction of $\mathbf{X}_{ij}$, the E-step of EM algorithm simplifies to working with sufficient statistics, and the M-step directly updates the estimate of $\boldsymbol{\Omega}$ using the updated values of these sufficient statistics. Below, we present the general procedure of the EM algorithm, with a detailed implementation provided in Appendix A of the Supplementary Material.

Let $\boldsymbol{\Omega}^0 = (\boldsymbol{\xi}_1^0, \ldots, \boldsymbol{\xi}_M^0, \sigma_{\theta v}^0)$ be the initial values of $\boldsymbol{\Omega}$, and $\boldsymbol{\Omega}^{r-1} = (\boldsymbol{\xi}_1^{r-1}, \ldots, \boldsymbol{\xi}_M^{r-1}, \sigma_{\theta v}^{r-1})$ denote the parameter values updated at the end of the $(r-1)$th iteration. The rth EM iteration consists of the following steps:

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

1. *E-step:* Compute the conditional expectations of the sufficient statistics $S_{\boldsymbol{\eta}_j}$, $S_{\boldsymbol{\alpha}_j}$, and $S_{\sigma_{\theta v}}$ given $\mathbf{y}$, $\mathbf{t}$, and $\boldsymbol{\Omega}^{r-1}$:

$$s^r_{\boldsymbol{\eta}_j} = E_{(\mathbf{x}_j,\boldsymbol{\theta})}(S_{\boldsymbol{\eta}_j}|\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}), s^r_{\boldsymbol{\alpha}_j} = E_{\mathbf{v}}(S_{\boldsymbol{\alpha}_j}|\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}), \quad \text{for } j = 1,\ldots,M,$$

and

$$s^r_{\sigma_{\theta v}} = E_{\boldsymbol{\zeta}}(S_{\sigma_{\theta v}}|\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}).$$

2. *M-step:* Update the parameter estimate of $\boldsymbol{\Omega}$ by

$$\boldsymbol{\eta}^r_j = \hat{\boldsymbol{\eta}}_j(s^r_{\boldsymbol{\eta}_j}), \boldsymbol{\alpha}^r_j = \hat{\boldsymbol{\alpha}}_j(s^r_{\boldsymbol{\alpha}_j}), \text{for } j = 1,\ldots,M,$$

and

$$\sigma^r_{\theta v} = \hat{\sigma}_{\theta v}(s^r_{\sigma_{\theta v}}),$$

where the expressions of $\hat{\boldsymbol{\eta}}_j(s^r_{\boldsymbol{\eta}_j})$, $\hat{\boldsymbol{\alpha}}_j(s^r_{\boldsymbol{\alpha}_j})$, and $\hat{\sigma}_{\theta v}(s^r_{\sigma_{\theta v}})$ are provided in Equations (A10)–(A13) in the Supplementary Material.

While the M-step in the above EM iteration is straightforward and avoids the need for tedious optimization procedures, the E-step involves complex multidimensional integrals. To address this challenge, we propose an SAEM algorithm, inspired by existing studies (Camilli & Geis, 2019; Lavielle & Mbogning, 2014; Meng & Xu, 2023), which replaces the E-step computation with stochastic approximation, thereby reducing the computational burden.

3.2.2. A dimensionality reduction computation of $E_{(\mathbf{x}_j,\boldsymbol{\theta})}(S_{\boldsymbol{\eta}_j} \mid \mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$

Using the law of total expectation (Kolmogorov, 2018),

$$E_{(\mathbf{x}_j,\boldsymbol{\theta})}(S_{\boldsymbol{\eta}_j} \mid \mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}) = E_{\boldsymbol{\theta}}\left(E_{\mathbf{x}_j}(S_{\boldsymbol{\eta}_j} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}) \mid \mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}\right).$$

From Equation (14), the inner expectation $E_{\mathbf{x}_j}(S_{\boldsymbol{\eta}_j} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$ is given by

$$E_{\mathbf{x}_j}(S_{\boldsymbol{\eta}_j} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}) = \left\{\sum_{i=1}^{N}\sum_{k=1}^{K_j-1}\theta_i E(X_{ijk} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}), \sum_{i=1}^{N}E(\mathbf{X}_{ij} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}), \sum_{i=1}^{N}\theta_i,\boldsymbol{\theta}^\top\boldsymbol{\theta}\right\}, \tag{17}$$

where

$$\sum_{i=1}^{N}E(\mathbf{X}_{ij} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}) = \left(\sum_{i=1}^{N}E(X_{ij1} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}),\ldots,\sum_{i=1}^{N}E(X_{ij(K_j-1)} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})\right)^\top. \tag{18}$$

According to the relationship between X_{ijk} and Y_{ij} defined in Equation (9), the conditional distribution of X_{ijk} given $Y_{ij} = y_{ij}$ is given by

$$X_{ijk}|y_{ij} \sim \begin{cases} N(a_j\theta_i - b_j - \tau_{jk},1)I_{x_{ijk}<0}, & \text{if } y_{ij} < k, \\ N(a_j\theta_i - b_j - \tau_{jk},1)I_{x_{ijk}\geq 0}, & \text{if } y_{ij} \geq k. \end{cases}$$

Further, it can be derived that the conditional expectation of X_{ijk} given $\boldsymbol{\theta}$, $\mathbf{y}$, $\mathbf{t}$, and $\boldsymbol{\Omega}^{r-1}$ is

$$E(X_{ijk} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}) = \begin{cases} a_j^{r-1}\theta_i - b_j^{r-1} - \tau_{jk}^{r-1} - \frac{\phi(-a_j^{r-1}\theta_i + b_j^{r-1} + \tau_{jk}^{r-1})}{\Phi(-a_j^{r-1}\theta_i + b_j^{r-1} + \tau_{jk}^{r-1})}, & \text{if } y_{ij} < k, \\ a_j^{r-1}\theta_i - b_j^{r-1} - \tau_{jk}^{r-1} + \frac{\phi(-a_j^{r-1}\theta_i + b_j^{r-1} + \tau_{jk}^{r-1})}{1-\Phi(-a_j^{r-1}\theta_i + b_j^{r-1} + \tau_{jk}^{r-1})}, & \text{if } y_{ij} \geq k. \end{cases} \tag{19}$$

Finally, by substituting the expression for $E(X_{ijk} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$ in Equation (19) into Equations (17) and (18), the closed form of $E_{\mathbf{x}_j}(S_{\boldsymbol{\eta}_j} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$ is obtained. Let $h^r_j(\boldsymbol{\theta}) = E_{\mathbf{x}_j}(S_{\boldsymbol{\eta}_j} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$, then we

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

have

$$E_{(\mathbf{x}_j,\boldsymbol{\theta})}\left(S_{\boldsymbol{\eta}_j} \mid \mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}\right) = E_{\boldsymbol{\theta}}\left(h_j^r(\boldsymbol{\theta}) \mid \mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1}\right). \tag{20}$$

This reduces the expectation of $S_{\boldsymbol{\eta}_j}$ with respect to $\boldsymbol{\theta}$ and $\mathbf{X}$ to the expectation of $h_j^r(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, thereby achieving dimensionality reduction.

3.2.3. General framework of the SAEM algorithm

Based on the E-step in Equation (20), and following Delyon et al. (1999), the rth SAEM iteration consists of the following three steps:

1. *S-step:* Randomly sample $\boldsymbol{\zeta}_i^l = (\theta_i^l, \nu_i^l)(l = 1,\ldots,m_r)$ from $f(\boldsymbol{\zeta}_i|\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$, for $i = 1,\ldots,N$. The detailed sampling procedure is given in Appendix B of the Supplementary Material.
2. *SA-step:* Let $\boldsymbol{\theta}^l = (\theta_1^l,\ldots,\theta_N^l)^\top$ and $\mathbf{v}^l = (\nu_1^l,\ldots,\nu_N^l)^\top$. Using the generated $\boldsymbol{\theta}^l$ and $\mathbf{v}^l$, compute $S^l_{\boldsymbol{\alpha}_j}$, $S^l_{\sigma_{\theta\nu}}$, and $h_j^r(\boldsymbol{\theta}^l)$ according to Equations 15–17, for $l = 1,\ldots,m_r$. The sufficient statistics are then updated as follows:

$$s^r_{\boldsymbol{\eta}_j} = s^{r-1}_{\boldsymbol{\eta}_j} + \gamma_r\left(\frac{\sum_{l=1}^{m_r} h_j^r(\boldsymbol{\theta}^l)}{m_r} - s^{r-1}_{\boldsymbol{\eta}_j}\right),$$

$$s^r_{\boldsymbol{\alpha}_j} = s^{r-1}_{\boldsymbol{\alpha}_j} + \gamma_r\left(\frac{\sum_{l=1}^{m_r} S^l_{\boldsymbol{\alpha}_j}}{m_r} - s^{r-1}_{\boldsymbol{\alpha}_j}\right), \text{ for } j = 1,\ldots,M,$$

and

$$s^r_{\sigma_{\theta\nu}} = s^{r-1}_{\sigma_{\theta\nu}} + \gamma_r\left(\frac{\sum_{l=1}^{m_r} S^l_{\sigma_{\theta\nu}}}{m_r} - s^{r-1}_{\sigma_{\theta\nu}}\right),$$

where $\gamma_r > 0$, $\sum_{r=1}^{+\infty}\gamma_r = +\infty$ and $\sum_{r=1}^{+\infty}\gamma_r^2 < +\infty$.
3. *M-Step:* Compute

$$\boldsymbol{\eta}_j^r = \hat{\boldsymbol{\eta}}_j(s^r_{\boldsymbol{\eta}_j}), \boldsymbol{\alpha}_j^r = \hat{\boldsymbol{\alpha}}_j(s^r_{\boldsymbol{\alpha}_j}), \text{for } j = 1,\ldots,M$$

and

$$\sigma^r_{\theta\nu} = \hat{\sigma}_{\theta\nu}(s^r_{\sigma_{\theta\nu}}),$$

according to Equations (A10)–(A13) in the Supplementary Material.

Remark 2. A crucial element of our SAEM algorithm is the use of $h_j^r(\boldsymbol{\theta}) = E_{\mathbf{x}_j}(S_{\boldsymbol{\eta}_j} \mid \boldsymbol{\theta},\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$, which eliminates the need to sample $\mathbf{x}_j$ during the S-step. This enhances computational efficiency and, more importantly, ensures that the estimates of $\tau_{jk}^r(k = 1,\ldots,K_j-1)$ obtained in the M-step reliably preserve the desired ascending order: $\tau_{j1}^r < \tau_{j2}^r < \cdots < \tau_{j(K_j-1)}^r$. In contrast, if $\mathbf{x}_j$ were randomly sampled during S-step, this order could be violated, as highlighted by Camilli and Geis (2019). To illustrate this advantage, we first define $\tilde{\tau}_{jk,l}^r$ according to Equation (A11) in the Supplementary Material, as follows:

$$\tilde{\tau}_{jk,l}^r = \frac{-\sum_{i=1}^N E(X_{ijk} \mid \boldsymbol{\theta}^l,\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})}{N} + \frac{\left(\sum_{i=1}^N E(\mathbf{X}_{ij} \mid \boldsymbol{\theta}^l,\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})\right)^\top \mathbf{1}_{K_j-1}}{N(K_j-1)},$$

where $\mathbf{1}_{K_j-1}$ denotes a $(K_j-1)\times 1$ column vector with all entries equal to 1. The term $\left(\sum_{i=1}^N E(\mathbf{X}_{ij} \mid \boldsymbol{\theta}^l,\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})\right)^\top \mathbf{1}_{K_j-1}$ represents the summation of the elements within the vector $\sum_{i=1}^N E(\mathbf{X}_{ij} \mid \boldsymbol{\theta}^l,\mathbf{y},\mathbf{t},\boldsymbol{\Omega}^{r-1})$.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Using $\tilde{\tau}^r_{jk,l}$, the estimate τ^r_{jk} in M-step of our SAEM can be equivalently computed as

$$\tau^r_{jk} = \tau^{r-1}_{jk} + \gamma_r \left(\frac{\sum_{l=1}^{m_r} \tilde{\tau}^r_{jk,l}}{m_r} - \tau^{r-1}_{jk} \right).$$

Moreover, it can be easily shown that if the initial values of $\boldsymbol{\tau}_j$ satisfy $\tau^0_{j1} < \cdots < \tau^0_{j(K_j-1)}$, the following order holds:

$$E(X_{ij1} \mid \boldsymbol{\theta}^l, \mathbf{y}, \mathbf{t}, \boldsymbol{\Omega}^{r-1}) > E(X_{ij2} \mid \boldsymbol{\theta}^l, \mathbf{y}, \mathbf{t}, \boldsymbol{\Omega}^{r-1}) > \cdots > E(X_{ij(K_j-1)} \mid \boldsymbol{\theta}^l, \mathbf{y}, \mathbf{t}, \boldsymbol{\Omega}^{r-1}).$$

Thus, the order $\tilde{\tau}_{j1,l} < \cdots < \tilde{\tau}_{j(K_j-1),l}$ is preserved, ensuring that $\tau^r_{j1} < \cdots < \tau^r_{j(K_j-1)}$ holds.

It is important to note that the SAEM procedure does not require the number of response categories K_j to be the same across items, which makes it well-suited for estimating the GR–RT model with mixed-scoring items. In addition, because the HM is nested within the GR–RT model, the SAEM algorithm developed here can be directly applied to estimate the HM.

Large-scale assessments often employ matrix-sampling designs, leading to sparse response matrices with non-random missing data. This missingness is not a flaw in data collection but rather a natural consequence of the test booklet assignment process. To facilitate the flexible application of the GR–RT model in this context, we extended the SAEM algorithm to accommodate non-random missing data by introducing an indicator matrix $\mathbf{D} = \{d_{ij}\}_{N \times M}$, where $d_{ij} = 1$ denotes that item j is administered to respondent i, and $d_{ij} = 0$ otherwise. This extension explicitly accounts for the non-ignorable missingness mechanism, allowing for valid parameter estimation under complex large-scale assessment designs. The detailed procedures of the extended SAEM algorithm are provided in Appendix C of the Supplementary Material.

3.2.4. Some specifications for implementing the SAEM algorithm

Initialization: Several studies, including Delyon et al. (1999), Lavielle and Mbogning (2014), and Meng and Xu (2023), have demonstrated that the SAEM algorithm is robust to the selection of initial values. To evaluate this robustness of our SAEM algorithm, we conducted a simulation study comparing parameter estimation performance with three different sets of initial values: a set of randomly selected values, a set of common values, and the true values of $\boldsymbol{\Omega}$. The parameter recovery results across these different initial values were nearly identical, providing additional evidence of the SAEM algorithm's robustness in terms of initial value selection. For the sake of conciseness, the details of this simulation study are not reported here. In both our simulation and empirical studies, we used the set of common values as the initial values.

The specifications of γ_r: The choice of step size γ_r plays an important role in determining the convergence of the SAEM algorithm. Drawing on existing studies (Camilli & Geis, 2019; Lavielle & Mbogning, 2014; Meng & Xu, 2023), we adopt a two-stage step size sequence selection approach. Specifically, we set $\gamma_r = 1$ during the initial R iterations (the first stage) and $\gamma_r = \frac{1}{(r-R)^\kappa}$ for $r > R$ (the second stage), where $0.5 < \kappa \leq 1$. In this study, we use $\kappa = 1$. The step size sequence is thus expressed as $\gamma_r = I_{(r \leq R)} + I_{(r>R)} \frac{1}{r-R}$. In addition, through simulation studies, we verify that the convergence of the SAEM algorithm is ensured when $R = 500$.

The specifications of m_r: Delyon et al. (1999) have demonstrated that the convergence of the SAEM algorithm is guaranteed with $m_r = 1$. Notably, increasing m_r does not improve computational accuracy but instead adds to the computational burden. Therefore, in practical applications of the SAEM algorithm, m_r is typically set to 1.

Convergence check: The convergence of the SAEM iteration is monitored using $\Delta^r = \left\| \boldsymbol{\Omega}^r - \boldsymbol{\Omega}^{r-1} \right\|_{\max}$. Here, $\| \cdot \|_{\max}$ denotes the maximum of the absolute values within a vector. In this study, the SAEM iteration is recommended to terminate when $\Delta^r < 10^{-4}$ or the maximum number of iterations exceeds 1,000.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

4. Simulation studies

This section presents two simulation studies. Simulation Study 1 evaluated the parameter recovery of the GR–RT model and the HM within the MMLE framework using the SAEM algorithm, and also compared the results obtained from Mplus estimation. Simulation Study 2 compared the performance of the GR–RT model and the HM, and assessed the effectiveness of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selecting the correct model.

4.1. Simulation study 1

4.1.1. Design

This study investigated the performance of the proposed SAEM algorithm under mixed-format testing conditions that included both dichotomous and graded-response items. Without loss of generality, the graded items were specified to have four response categories. To assess the impact of item-type composition, five mixture proportions (MPs) of binary and graded-scoring items were considered, as shown in Table 2.

Since the sample size (N) and test length (M) are two important factors influencing the accuracy of the parameter estimation, this simulation considered three sample sizes (N = 500, 1,000, and 2,000) and two test lengths (M = 20 and 40). The three factors (sample size, test length, and MP) were crossed, resulting in $3 \times 2 \times 5 = 30$ simulation conditions. For each simulation condition, the item response and RT data were generated under two true models: the GR–RT model and the HM. The true parameter values were set as follows.

Let M_g denote the number of graded-scored items, where $j = 1, \ldots, M_g$ indexes the graded items and $j = M_g + 1, \ldots, M$ indexes the binary items. Let $U(l_1, l_2)$ denote the uniform distribution between l_1 and l_2, and $LN(\mu, \sigma^2)$ denote the log-normal distribution with parameters μ and σ^2.

1. *GR–RT model:*
   (a) *Graded-scored items:*
   - *GRM:* For $j = 1, \ldots, M_g$, the discrimination parameter a_j was generated from a uniform distribution, $a_j \sim U(0.5, 2.5)$, allowing for moderate to high discrimination. The location parameter b_j was generated from a standard normal distribution, $b_j \sim N(0,1)$, to center the items on the ability scale. These specifications follow previous studies (e.g., Zhang & Chen, 2024; Wang & Xu, 2015; Meng & Xu, 2023). The threshold parameters τ_{j1}, τ_{j2}, and τ_{j3} were generated as follows: $\tau_{j1} \sim U(-1, -0.3)$, $\tau_{j2} \sim U(-0.25, 0.25)$, and $\tau_{j3} = -\tau_{j1} - \tau_{j2}$, such as $\sum_{k=1}^{3} \tau_{jk} = 0$. These specifications based on empirical results from the data of one of the test booklets of the PISA 2022 science literacy assessment.
   - *Log-normal RT model:* For $j = 1, \ldots, M_g$, $\beta_j \sim U(3,5)$, $\boldsymbol{\delta}_j \sim U(-1,1)$, $\sigma_j \sim U(0.3, 0.7)$, and $\varphi_j \sim LN(0, 0.5)$. These specifications based on empirical results from the data of one of the test booklets of the PISA 2022 science literacy assessment.

   (b) *Binary-scored items:*
   - For $j = M_g + 1, \ldots, M$, the true values of a_j, b_j, σ_j, φ_j, β_j, and $\boldsymbol{\delta}_j$ were generated in the same way as the parameters for the graded-scored items described in (a). Note that there is no $\boldsymbol{\tau}$-parameter, since the GRM is reduced to the 2PNO model.

Table 2. Five mixture proportion (MP) structures of four-category graded and binary items

	MP1	MP2	MP3	MP4	MP5
Four-category graded items	100%	70%	50%	30%	0%
Binary items	0%	30%	50%	70%	100%

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

2. *Hierarchical model:*

   - The true values for the parameters in the GRM and log-normal RT models were generated in the same manner as for GR–RT model. Note that there is no $\boldsymbol{\delta}$-parameter in the HM.

3. To account for a diverse range of correlations between ability and speed, the correlation coefficient $\sigma_{\theta v}$ was drawn from a uniform distribution, $\sigma_{\theta v} \sim U(-1,1)$, for each simulation condition. Given $\sigma_{\theta v}$, the person parameters were then sampled from the following bivariate normal distribution:

$$(\theta_i, v_i)^T \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \sigma_{\theta v}\\ \sigma_{\theta v} & 1\end{pmatrix}\right),$$

   for $i = 1,\ldots,N$.

As noted by previous studies (Molenaar et al., 2015a, 2015c), the HM can be estimated using Mplus. Following the reviewers' suggestion, we successfully implemented Mplus to estimate the GR–RT model. Thus, in this simulation, Mplus was used as a comparison tool to evaluate the performance of our proposed method. The Mplus code used for this purpose is provided in Appendix D of the Supplementary Material.

A total of 100 test datasets were generated for each of the two true models across all 30 testing scenarios. The GR–RT and HM models were fitted to their respective simulated data using both the SAEM algorithm and Mplus. Parameter estimation accuracy was assessed using the root mean square error (RMSE) and absolute bias (ABias), both computed as averages over 100 replications and all items. Taking the "a-parameter" as an example, we denote $\hat{a}_{jg}$ as the estimate of a_j obtained from the gth replication. The RMSE and ABias for the a-parameter were calculated as follows:

$$\text{RMSE}_{\hat{a}} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{\frac{1}{100}\sum_{g=1}^{100}(\hat{a}_{jg}-a_j)^2}, \quad \text{ABias}_{\hat{a}} = \frac{1}{M}\sum_{j=1}^{M}\left|\frac{1}{100}\sum_{g=1}^{100}(\hat{a}_{jg}-a_j)\right|.$$

The computing time of both the SAEM algorithm and Mplus was also recorded for all simulation conditions to evaluate computational efficiency.

4.1.2. Results

The RMSE and ABias values for the SAEM estimation of the GR–RT model are presented in Tables 3 and 4. From these results, the following trends can be observed. First, both the RMSE and ABias for GR–RT model remain consistently small across all conditions. For example, the largest RMSE of the parameters is less than 0.2, and the ABias values are generally below 0.1. As the sample size increases, the RMSE and ABias values significantly decrease. In the case of $N = 2,000$, the RMSE values are smaller than 0.09, and the ABias values are nearly 0. These results demonstrate that increasing the sample size greatly improves the accuracy of parameter recovery and results in more unbiased estimators.

Second, across different proportions of four-category items and binary items, both the RMSE and ABias show only small differences, with no substantial increases in either under any test condition. This indicates that for estimating the GR–RT model, the performance of the SAEM algorithm is robust to the mixing of items in the test. Additionally, the RMSE and ABias are robust with regard to test length (M), which is expected, as the MMLE is theoretically independent of the test length.

Finally, both the RMSE and ABias for the RT model are smaller than those for the IRT model, indicating superior estimation accuracy. This is likely due to the linear nature of the RT model and the continuous nature of RT data, which allow for more precise parameter estimation compared to the discrete response data modeled by the IRT model.

The RMSE and ABias values for the SAEM estimation of the HM are reported in Tables 5 and 6. It is evident that these results follow similar trends to those for GR–RT model. Furthermore, both the RMSE

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Table 3. RMSE (ABias) of the estimates of GR–RT model, under three sample sizes (N = 500, 1,000, and 2,000) and five mixture proportions (MPs) of items (MP1, MP2, MP3, MP4, and MP5)

N	MPs	Item types	a	b	τ	σ	φ	β	$\boldsymbol{\delta}$	$\sigma_{\theta v}$
500	MP1	graded	0.187(0.110)	0.099(0.045)	0.051(0.006)	0.021(0.003)	0.029(0.005)	0.048(0.006)	0.089(0.007)	0.009(0.000)
	MP2	graded	0.138(0.019)	0.111(0.058)	0.054(0.007)	0.021(0.005)	0.047(0.035)	0.070(0.041)	0.090(0.006)	0.012(0.010)
		binary	0.139(0.013)	0.094(0.038)	—	0.020(0.003)	0.049(0.040)	0.063(0.044)	0.059(0.003)	
	MP3	graded	0.127(0.014)	0.091(0.047)	0.051(0.006)	0.020(0.003)	0.026(0.003)	0.101(0.089)	0.081(0.007)	0.082(0.082)
		binary	0.139(0.024)	0.102(0.041)	—	0.018(0.002)	0.023(0.004)	0.114(0.098)	0.057(0.004)	
	MP4	graded	0.137(0.031)	0.104(0.059)	0.051(0.007)	0.022(0.002)	0.033(0.019)	0.067(0.047)	0.082(0.007)	0.022(0.018)
		binary	0.139(0.021)	0.109(0.042)	—	0.018(0.002)	0.032(0.020)	0.068(0.046)	0.056(0.005)	
	MP5	binary	0.161(0.066)	0.115(0.059)	—	0.018(0.003)	0.042(0.030)	0.093(0.079)	0.062(0.004)	0.009(0.005)
1,000	MP1	graded	0.145(0.102)	0.062(0.008)	0.036(0.004)	0.015(0.001)	0.021(0.011)	0.078(0.070)	0.060(0.004)	0.023(0.022)
	MP2	graded	0.101(0.029)	0.063(0.007)	0.037(0.004)	0.015(0.002)	0.044(0.039)	0.038(0.008)	0.062(0.005)	0.009(0.007)
		binary	0.096(0.025)	0.061(0.007)	—	0.013(0.001)	0.048(0.043)	0.027(0.007)	0.037(0.002)	
	MP3	graded	0.096(0.017)	0.069(0.036)	0.037(0.003)	0.014(0.002)	0.044(0.039)	0.034(0.013)	0.057(0.005)	0.039(0.038)
		binary	0.096(0.014)	0.072(0.024)	—	0.013(0.002)	0.047(0.043)	0.035(0.012)	0.041(0.004)	
	MP4	graded	0.097(0.011)	0.075(0.042)	0.036(0.003)	0.015(0.001)	0.018(0.002)	0.056(0.046)	0.055(0.005)	0.027(0.026)
		binary	0.105(0.016)	0.078(0.033)	—	0.013(0.001)	0.016(0.002)	0.057(0.043)	0.040(0.002)	
	MP5	binary	0.103(0.015)	0.102(0.077)	—	0.013(0.002)	0.031(0.024)	0.054(0.042)	0.041(0.002)	0.008(0.003)
2,000	MP1	graded	0.072(0.029)	0.043(0.009)	0.025(0.002)	0.011(0.001)	0.018(0.011)	0.038(0.028)	0.044(0.003)	0.010(0.008)
	MP2	graded	0.083(0.046)	0.076(0.060)	0.026(0.002)	0.010(0.001)	0.019(0.007)	0.047(0.035)	0.048(0.003)	0.004(0.002)
		binary	0.082(0.037)	0.064(0.045)	—	0.009(0.001)	0.017(0.008)	0.043(0.037)	0.030(0.003)	
	MP3	graded	0.065(0.015)	0.050(0.033)	0.025(0.002)	0.010(0.001)	0.017(0.009)	0.029(0.017)	0.042(0.003)	0.006(0.004)
		binary	0.068(0.011)	0.052(0.021)	—	0.009(0.001)	0.017(0.011)	0.032(0.018)	0.031(0.002)	
	MP4	graded	0.087(0.054)	0.049(0.030)	0.026(0.003)	0.011(0.001)	0.040(0.037)	0.044(0.037)	0.041(0.003)	0.028(0.027)
		binary	0.082(0.041)	0.051(0.021)	—	0.010(0.001)	0.036(0.033)	0.043(0.034)	0.028(0.002)	
	MP5	binary	0.073(0.014)	0.046(0.006)	—	0.009(0.001)	0.013(0.001)	0.023(0.010)	0.029(0.003)	0.006(0.004)

Note: The test length is M = 20. MP1: Four-category graded items (100%), binary items (0%); MP2: Four-category graded items (70%), binary items (30%); MP3: Four-category graded items (50%), binary items (50%); MP4: Four-category graded items (30%), binary items (70%); and MP5: Four-category graded items (0%), binary items (100%).

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Table 4. RMSE (ABias) of the estimates of GR–RT model, under three sample sizes (N = 500, 1,000, and 2,000) and five mixture proportions (MPs) of items (MP1, MP2, MP3, MP4, and MP5)

N	MPs	Item types	a	b	τ	σ	φ	β	$\boldsymbol{\delta}$	$\sigma_{\theta v}$
500	MP1	graded	0.152(0.097)	0.099(0.059)	0.049(0.005)	0.020(0.003)	0.086(0.079)	0.066(0.032)	0.091(0.008)	0.005(0.005)
	MP2	graded	0.139(0.038)	0.107(0.064)	0.051(0.006)	0.021(0.003)	0.032(0.017)	0.054(0.014)	0.088(0.007)	0.030(0.028)
		binary	0.136(0.033)	0.110(0.052)	—	0.017(0.002)	0.032(0.019)	0.051(0.017)	0.060(0.006)	
	MP3	graded	0.134(0.013)	0.121(0.080)	0.051(0.006)	0.021(0.004)	0.057(0.050)	0.050(0.007)	0.087(0.007)	0.085(0.085)
		binary	0.129(0.011)	0.111(0.061)	—	0.018(0.002)	0.056(0.050)	0.042(0.007)	0.055(0.005)	
	MP4	graded	0.134(0.021)	0.169(0.140)	0.052(0.006)	0.019(0.003)	0.027(0.004)	0.100(0.083)	0.086(0.008)	0.039(0.039)
		binary	0.151(0.029)	0.158(0.116)	—	0.017(0.002)	0.025(0.003)	0.103(0.089)	0.058(0.005)	
	MP5	binary	0.149(0.021)	0.119(0.060)	—	0.017(0.002)	0.062(0.056)	0.108(0.098)	0.055(0.005)	0.051(0.050)
1,000	MP1	graded	0.119(0.074)	0.059(0.007)	0.036(0.003)	0.014(0.002)	0.020(0.010)	0.048(0.034)	0.057(0.004)	0.022(0.021)
	MP2	graded	0.099(0.048)	0.084(0.058)	0.035(0.004)	0.014(0.002)	0.061(0.057)	0.044(0.023)	0.062(0.005)	0.026(0.025)
		binary	0.091(0.026)	0.077(0.040)	—	0.012(0.001)	0.070(0.068)	0.045(0.029)	0.041(0.005)	
	MP3	graded	0.101(0.037)	0.101(0.079)	0.035(0.004)	0.015(0.002)	0.027(0.020)	0.037(0.015)	0.060(0.004)	0.013(0.011)
		binary	0.097(0.027)	0.089(0.062)	—	0.012(0.002)	0.026(0.020)	0.033(0.015)	0.038(0.003)	
	MP4	graded	0.093(0.009)	0.076(0.043)	0.036(0.003)	0.013(0.002)	0.023(0.007)	0.039(0.010)	0.063(0.005)	0.005(0.004)
		binary	0.099(0.009)	0.077(0.034)	—	0.012(0.001)	0.020(0.006)	0.034(0.010)	0.045(0.004)	
	MP5	binary	0.104(0.016)	0.072(0.017)	—	0.012(0.002)	0.043(0.038)	0.041(0.027)	0.042(0.003)	0.005(0.001)
2,000	MP1	graded	0.070(0.030)	0.051(0.029)	0.025(0.002)	0.010(0.001)	0.012(0.001)	0.036(0.027)	0.041(0.003)	0.004(0.002)
	MP2	graded	0.070(0.021)	0.044(0.011)	0.025(0.002)	0.010(0.001)	0.016(0.007)	0.028(0.011)	0.044(0.003)	0.009(0.009)
		binary	0.068(0.019)	0.048(0.009)	—	0.009(0.001)	0.015(0.008)	0.027(0.012)	0.030(0.003)	
	MP3	graded	0.064(0.010)	0.058(0.040)	0.025(0.002)	0.010(0.001)	0.020(0.011)	0.039(0.027)	0.046(0.003)	0.005(0.005)
		binary	0.063(0.009)	0.054(0.031)	—	0.009(0.001)	0.017(0.010)	0.037(0.027)	0.032(0.003)	
	MP4	graded	0.071(0.019)	0.060(0.037)	0.027(0.003)	0.010(0.001)	0.013(0.005)	0.060(0.054)	0.041(0.003)	0.005(0.000)
		binary	0.072(0.014)	0.056(0.027)	—	0.008(0.001)	0.013(0.006)	0.064(0.059)	0.027(0.002)	
	MP5	binary	0.072(0.007)	0.058(0.029)	—	0.009(0.001)	0.012(0.002)	0.041(0.034)	0.028(0.002)	0.013(0.011)

Note: The test length is M = 40. MP1: Four-category graded items (100%), binary items (0%); MP2: Four-category graded items (70%), binary items (30%); MP3: Four-category graded items (50%), binary items (50%); MP4: Four-category graded items (30%), binary items (70%); and MP5: Four-category graded items (0%), binary items (100%).

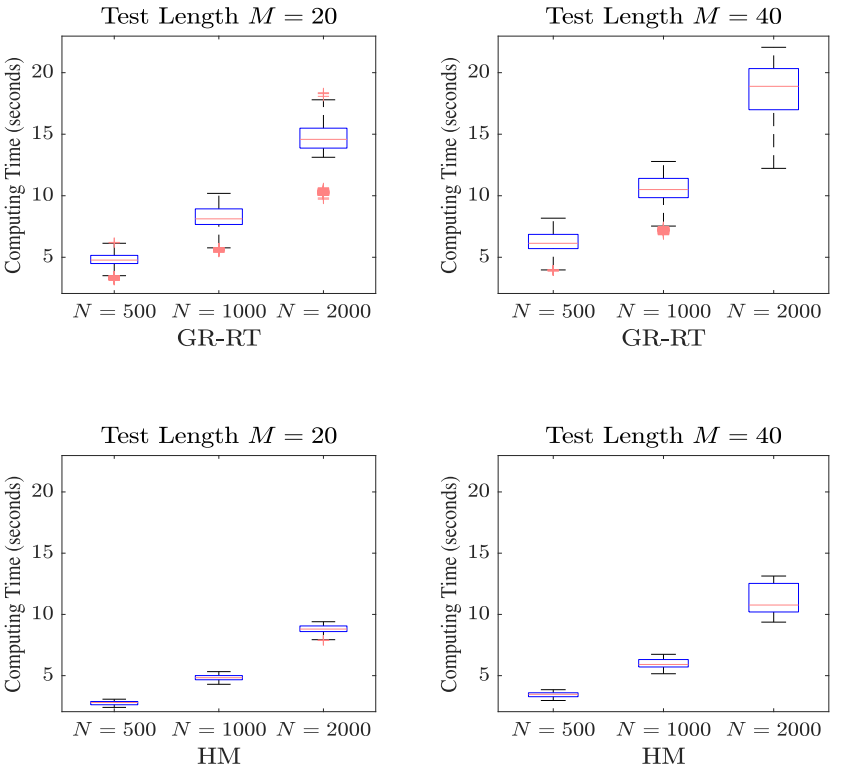


Figure 2. Boxplots of computing time for GR–RT model (top row) and HM (bottom row), across varying sample sizes of N = 500, 1,000, and 2,000 and test lengths of M = 20 and 40 over 100 replications.

and ABias are slightly smaller for HM, which is expected given that HM is a simpler model. Overall, our SAEM algorithm demonstrates effectiveness in estimating the HM. It provides a reliable approach for HM estimation and can be applied to estimate models in the context of graded-score items.

Under both the GR–RT and HM models, the RMSE and ABias values of $\sigma_{\theta v}$ are very small (all below 0.12). However, unlike the other parameters, neither index exhibits a clear decreasing trend as the sample size increases. This is due to the simulation design, in which the true values of $\sigma_{\theta v}$ were randomly drawn from $U(-1,1)$ in each simulation condition. As a result, different sample sizes correspond to different underlying true values, which can obscure the expected improvement in estimation accuracy with increasing sample size.

The computing time of the SAEM algorithm is displayed using boxplots, please see Figure 2. The top two figures present results for GR–RT model, while the bottom figures show results for HM. It can be observed that as the sample size increases, the computing time also increases. However, the maximum time required for a single run does not exceed 22 seconds, indicating that the SAEM algorithm demonstrates high computational efficiency. Additionally, the computing time for M = 40 is slightly greater than that for M = 20, and the computing time for GR–RT model is slightly longer than that for HM under the same simulation conditions. These phenomena are expected, as a larger sample size, longer test length, and the higher model complexity of GR–RT model result in increased computational cost.

Due to space limitations, the results obtained from Mplus are provided in Appendix D of the Supplementary Material. The RMSE and ABias values for the GR–RT and HM parameter estimates using Mplus are summarized in Tables D1–D4 in the Supplementary Material. As shown, the RMSE and ABias values for certain parameters (e.g., a and b) are significantly larger, with some conditions

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Table 5. RMSE (ABias) of the estimates of HM, under three sample sizes (N = 500, 1,000, and 2,000) and five mixture proportions (MPs) of items (MP1, MP2, MP3, MP4, and MP5)

N	MPs	Item types	a	b	τ	σ	φ	β	$\sigma_{\theta v}$
500	MP1	graded	0.151(0.046)	0.118(0.080)	0.056(0.005)	0.019(0.002)	0.039(0.024)	0.024(0.003)	0.013(0.011)
	MP2	graded	0.137(0.015)	0.153(0.117)	0.059(0.007)	0.020(0.003)	0.029(0.015)	0.057(0.051)	0.015(0.012)
		binary	0.127(0.017)	0.124(0.096)	—	0.018(0.002)	0.031(0.021)	0.081(0.077)	
	MP3	graded	0.140(0.023)	0.102(0.040)	0.056(0.005)	0.020(0.002)	0.037(0.026)	0.047(0.040)	0.022(0.020)
		binary	0.152(0.015)	0.111(0.032)	—	0.020(0.002)	0.051(0.044)	0.070(0.066)	
	MP4	graded	0.159(0.051)	0.118(0.041)	0.061(0.007)	0.023(0.001)	0.043(0.032)	0.031(0.014)	0.030(0.027)
		binary	0.165(0.033)	0.129(0.040)	—	0.021(0.002)	0.059(0.053)	0.037(0.027)	
	MP5	binary	0.180(0.041)	0.194(0.130)	—	0.021(0.001)	0.036(0.004)	0.051(0.043)	0.014(0.001)
1,000	MP1	graded	0.129(0.090)	0.068(0.032)	0.039(0.004)	0.014(0.001)	0.032(0.023)	0.026(0.019)	0.036(0.035)
	MP2	graded	0.098(0.015)	0.077(0.046)	0.041(0.005)	0.014(0.002)	0.020(0.009)	0.017(0.003)	0.015(0.013)
		binary	0.088(0.009)	0.064(0.030)	—	0.013(0.001)	0.022(0.015)	0.016(0.006)	
	MP3	graded	0.113(0.059)	0.070(0.028)	0.039(0.004)	0.015(0.002)	0.033(0.027)	0.036(0.030)	0.024(0.022)
		binary	0.120(0.053)	0.078(0.019)	—	0.014(0.002)	0.049(0.046)	0.052(0.049)	
	MP4	graded	0.113(0.034)	0.082(0.026)	0.043(0.003)	0.017(0.002)	0.029(0.021)	0.021(0.009)	0.013(0.010)
		binary	0.110(0.019)	0.091(0.027)	—	0.014(0.001)	0.039(0.034)	0.023(0.015)	
	MP5	binary	0.119(0.029)	0.091(0.014)	—	0.015(0.002)	0.048(0.041)	0.033(0.027)	0.011(0.003)
2,000	MP1	graded	0.096(0.065)	0.043(0.012)	0.028(0.002)	0.010(0.001)	0.015(0.002)	0.038(0.036)	0.012(0.011)
	MP2	graded	0.069(0.020)	0.045(0.005)	0.028(0.003)	0.010(0.001)	0.012(0.002)	0.013(0.005)	0.006(0.004)
		binary	0.064(0.020)	0.039(0.004)	—	0.009(0.000)	0.011(0.003)	0.013(0.007)	
	MP3	graded	0.073(0.014)	0.077(0.060)	0.027(0.003)	0.011(0.001)	0.014(0.004)	0.016(0.010)	0.006(0.000)
		binary	0.080(0.011)	0.073(0.048)	—	0.010(0.000)	0.014(0.007)	0.021(0.018)	
	MP4	graded	0.072(0.005)	0.061(0.026)	0.031(0.004)	0.012(0.001)	0.024(0.020)	0.016(0.008)	0.015(0.013)
		binary	0.074(0.009)	0.061(0.025)	—	0.010(0.001)	0.033(0.031)	0.018(0.013)	
	MP5	binary	0.074(0.013)	0.070(0.035)	—	0.010(0.001)	0.019(0.010)	0.033(0.030)	0.005(0.003)

Note: The test length is M = 20. MP1: Four-category graded items (100%), binary items (0%); MP2: Four-category graded items (70%), binary items (30%); MP3: Four-category graded items (50%), binary items (50%); MP4: Four-category graded items (30%), binary items (70%); and MP5: Four-category graded items (0%), binary items (100%).

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Table 6. RMSE (ABias) of the estimates of HM, under three sample sizes (N = 500, 1,000, and 2,000) and five mixture proportions (MPs) of items (MP1, MP2, MP3, MP4, and MP5)

N	MPs	Item types	a	b	τ	σ	φ	β	$\sigma_{\theta v}$
500	MP1	graded	0.146(0.068)	0.183(0.156)	0.057(0.007)	0.018(0.002)	0.035(0.005)	0.051(0.044)	0.008(0.003)
	MP2	graded	0.125(0.036)	0.085(0.012)	0.056(0.007)	0.019(0.002)	0.025(0.007)	0.028(0.013)	0.116(0.115)
		binary	0.156(0.030)	0.122(0.017)	—	0.019(0.001)	0.025(0.008)	0.028(0.014)	
	MP3	graded	0.133(0.032)	0.127(0.095)	0.054(0.006)	0.019(0.002)	0.028(0.014)	0.047(0.040)	0.049(0.048)
		binary	0.142(0.025)	0.138(0.066)	—	0.020(0.002)	0.032(0.019)	0.059(0.053)	
	MP4	graded	0.135(0.026)	0.149(0.112)	0.056(0.006)	0.018(0.002)	0.025(0.007)	0.072(0.068)	0.008(0.004)
		binary	0.150(0.032)	0.151(0.090)	—	0.021(0.002)	0.027(0.009)	0.087(0.083)	
	MP5	binary	0.154(0.029)	0.180(0.132)	—	0.021(0.002)	0.033(0.005)	0.102(0.098)	0.039(0.038)
1,000	MP1	graded	0.094(0.022)	0.076(0.042)	0.039(0.004)	0.013(0.001)	0.028(0.014)	0.030(0.024)	0.022(0.022)
	MP2	graded	0.109(0.069)	0.114(0.096)	0.039(0.004)	0.013(0.001)	0.045(0.041)	0.055(0.052)	0.003(0.003)
		binary	0.113(0.053)	0.118(0.085)	—	0.013(0.001)	0.052(0.049)	0.065(0.062)	
	MP3	graded	0.109(0.050)	0.085(0.056)	0.039(0.005)	0.014(0.001)	0.018(0.004)	0.022(0.013)	0.022(0.021)
		binary	0.108(0.033)	0.096(0.043)	—	0.015(0.001)	0.018(0.006)	0.024(0.016)	
	MP4	graded	0.102(0.030)	0.096(0.064)	0.039(0.004)	0.013(0.001)	0.020(0.010)	0.021(0.012)	0.025(0.024)
		binary	0.107(0.027)	0.097(0.055)	—	0.015(0.001)	0.022(0.012)	0.023(0.015)	
	MP5	binary	0.117(0.038)	0.085(0.028)	—	0.015(0.001)	0.051(0.045)	0.042(0.038)	0.039(0.038)
2,000	MP1	graded	0.067(0.019)	0.049(0.021)	0.028(0.002)	0.009(0.001)	0.016(0.004)	0.021(0.017)	0.002(0.001)
	MP2	graded	0.090(0.066)	0.075(0.059)	0.028(0.002)	0.009(0.001)	0.021(0.017)	0.044(0.042)	0.021(0.021)
		binary	0.093(0.058)	0.081(0.058)	—	0.009(0.001)	0.024(0.020)	0.051(0.050)	
	MP3	graded	0.062(0.006)	0.108(0.099)	0.027(0.002)	0.009(0.001)	0.012(0.001)	0.050(0.049)	0.004(0.003)
		binary	0.068(0.007)	0.098(0.077)	—	0.010(0.001)	0.012(0.002)	0.064(0.062)	
	MP4	graded	0.095(0.070)	0.079(0.064)	0.028(0.002)	0.009(0.001)	0.031(0.028)	0.025(0.022)	0.012(0.012)
		binary	0.094(0.059)	0.078(0.054)	—	0.010(0.001)	0.038(0.035)	0.030(0.027)	
	MP5	binary	0.077(0.013)	0.058(0.020)	—	0.010(0.000)	0.018(0.005)	0.012(0.001)	0.006(0.002)

Note: The test length is M = 40. MP1: Four-category graded items (100%), binary items (0%); MP2: Four-category graded items (70%), binary items (30%); MP3: Four-category graded items (50%), binary items (50%); MP4: Four-category graded items (30%), binary items (70%); and MP5: Four-category graded items (0%), binary items (100%).

producing values exceeding 10. These inflated values likely reflect instability in Mplus computations, particularly when the test includes mixed-format items, leading to extreme parameter estimates for both the GR–RT and HM models.

As shown in Figure D1 in the Supplementary Material, the computing time of Mplus remains reasonable when both N and M are small, slightly exceeding that of the SAEM algorithm under the same conditions. However, as sample size and test length increase, the computing time grows substantially. For instance, when $N = 2{,}000$ and $M = 40$, a single run of Mplus exceeded 1,000 seconds for the GR–RT model and 700 seconds for the HM, indicating a significant computational burden.

4.2. Simulation study 2

The design and results are presented in Appendix E of the Supplementary Material; here, we briefly summarize the main findings.

The simulation results show that the GR–RT model generally outperforms the HM when the conditional independence assumption is violated ($\boldsymbol{\delta} \neq 0$). Specifically, the GR–RT model produces lower RMSEs with fewer outliers, and its estimates of $\sigma_{\theta v}$ are more closely centered around the true values. In contrast, the HM tends to underestimate $\sigma_{\theta v}$ and yields unstable estimates under these conditions.

When $\boldsymbol{\delta} = 0$, where the HM is the true data-generating model, both models achieve similar accuracy. The HM demonstrates excellent parameter recovery, and the GR–RT model also maintains high accuracy, illustrating its robustness and adaptability across different data structures.

Both models show decreasing RMSEs with increasing sample size, reflecting consistency. However, under $\boldsymbol{\delta} \neq 0$, the HM continues to produce extreme estimates, indicating sensitivity to violations of conditional independence. Estimation stability for both models is also affected by the true value of $\sigma_{\theta v}$, with higher values associated with more stable estimates.

In summary, the GR–RT model provides accurate and stable parameter estimates when residual dependencies exist and remains robust when conditional independence holds. This makes it a flexible and reliable alternative to the HM across a range of testing scenarios.

Finally, we evaluated the performance of AIC and BIC in selecting true data-generating model. Both criteria consistently identified the true data-generating model as the best-fitting model across all 27 simulation conditions. These results highlight the high accuracy of AIC and BIC as model selection criteria in comparing the GR–RT model and the HM.

5. Empirical study

In this section, we present an empirical study conducted to evaluate the performance of the proposed GR–RT model using a mathematics dataset from PISA 2022.

5.1. Data and analysis design

PISA, conducted by the Organization for Economic Co-operation and Development (OECD), assesses the skills and knowledge of 15-year-old students in mathematics, reading, and science every three years across multiple countries. In this study, we focused on the mathematics assessment, as it was the primary focus of PISA 2022 and included a relatively high proportion of polytomous items.

The PISA 2022 mathematics assessment employed three testing formats: computer-based linear testing, computer-based multistage adaptive testing, and paper-based assessment. In this study, we analyzed data from the computer-based linear format, because the GR–RT model assumes a fixed test structure and is therefore not applicable to multistage adaptive testing. To align with the characteristics of large-scale assessments, which involve a large number of examinees and a wide range of items, we selected four test forms from the mathematics assessment dataset.[1,2] These four forms covered all test

[1]In the PISA dataset, forms are identified by the variable BOOKID (labeled "Form Identifier"), see the PISA 2022 Technical Report for detailed information.

[2]The dataset is publicly available on the OECD website: https://www.oecd.org/en/data/datasets/pisa-2022-database.html.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

booklets and contained 233 items (171 dichotomous items and 62 polytomous items with three response categories), yielding responses and RTs from 18,086 students across multiple countries and economies, with each student responding to only one test booklet.

In this empirical study, the data were fitted using two models: the proposed GR–RT model and the HM. Both models were estimated using the proposed SAEM algorithm. To assess estimation precision, bootstrap standard errors (SEs) were computed for the parameter estimates based on 500 bootstrap samples generated by sampling with replacement from the original dataset, while maintaining the original sample size in each replication. For each bootstrap sample, the parameters of both models were estimated, and the bootstrap SEs were calculated as the standard deviations of the resulting estimates across all replications. Further, 95 % bootstrap confidence intervals (CIs) were computed for $\boldsymbol{\delta}$ and $\sigma_{\theta v}$ to evaluate whether they significantly differed from zero at the 0.05 significance level.

Finally, we compared the overall goodness-of-fit of the two models using three approaches: the AIC (Akaike, 1974), the BIC (Schwarz, 1978), and the likelihood ratio test (LRT; Neyman & Pearson, 1928). The AIC and BIC were computed as

$$\begin{aligned} \text{AIC} &= -2\log L(\mathbf{y}, \ln \mathbf{t} \mid \boldsymbol{\Omega}) + 2\omega, \\ \text{BIC} &= -2\log L(\mathbf{y}, \ln \mathbf{t} \mid \boldsymbol{\Omega}) + \omega \log(N), \end{aligned}$$

where

$$L(\mathbf{y}, \ln \mathbf{t} \mid \boldsymbol{\Omega}) = \prod_{i=1}^{N} \iint \prod_{j=1}^{M} f(y_{ij}, \ln t_{ij} \mid \theta_i, v_i, \boldsymbol{\xi}_j, \sigma_{\theta v}) f(\theta_i, v_i \mid \sigma_{\theta v}) d\theta_i dv_i$$

is the marginal likelihood, ω is the total number of estimated parameters, and N is the sample size. The marginal likelihood was approximated via numerical integration; for brevity, the details of the computational procedure are omitted.

For the LRT, we tested the null hypothesis H_0: $\boldsymbol{\delta} = 0$ (there is no difference between the GR–RT model and the HM) against the alternative hypothesis H_1: $\boldsymbol{\delta} \neq 0$. Under H_0 and standard regularity conditions, the LRT statistic G^2 asymptotically follows a chi-square distribution. The 95th percentile of this distribution, denoted by $\chi^2_{0.05}$, was used as the critical value.

5.2. Results

Due to the large number of items ($M = 233$), the bootstrap SEs for the parameters in the GR–RT model and HM were summarized using boxplots, as shown in Figures 3 and 4. Most bootstrap SEs were small, with the majority below 0.1. Although a few SEs exceeded 0.1, all remained within an acceptable range. Further examination indicated that these larger SEs were typically associated with higher parameter estimates and, simultaneously, fewer responding students, suggesting that the observed increase in SEs is statistically reasonable.

Furthermore, we examined the relationship between the parameter estimates from the GR–RT model and the HM. The results showed strong agreement between the two sets of estimates, with all correlations exceeding 0.95. Minor discrepancies were observed for items with larger $\boldsymbol{\delta}_j$ in the GR–RT model, which is consistent with model expectations, as larger $\boldsymbol{\delta}_j$ implies greater divergence between the two models.

The estimates of $\sigma_{\theta v}$ are −0.43 (GR–RT model) and −0.51 (HM), suggesting a negative correlation between θ and v. Notably, the HM appears to underestimate $\sigma_{\theta v}$ compared to the GR–RT model, which is consistent with the findings from Simulation Study 2 (see Appendix E of the Supplementary Material). The corresponding 95% CIs, [−0.45, −0.42] for the GR–RT model and [−0.52, −0.49] for the HM, both excluded zero, confirming the statistical significance of this negative association. These results suggest that higher-ability students tend to respond more slowly, a pattern that aligns with previous research (e.g., Klein Entink et al., 2009; Scherer et al., 2015; Zhan, Jiao, & Liao, 2018). One possible explanation lies in the low-stakes nature of PISA, in which lower-ability respondents may be less motivated (Logan et al., 2011; Wise & Kong, 2005), leading to shorter RTs and more errors than for high-ability respondents. However, the direction of the ability–speed relationship may vary across

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

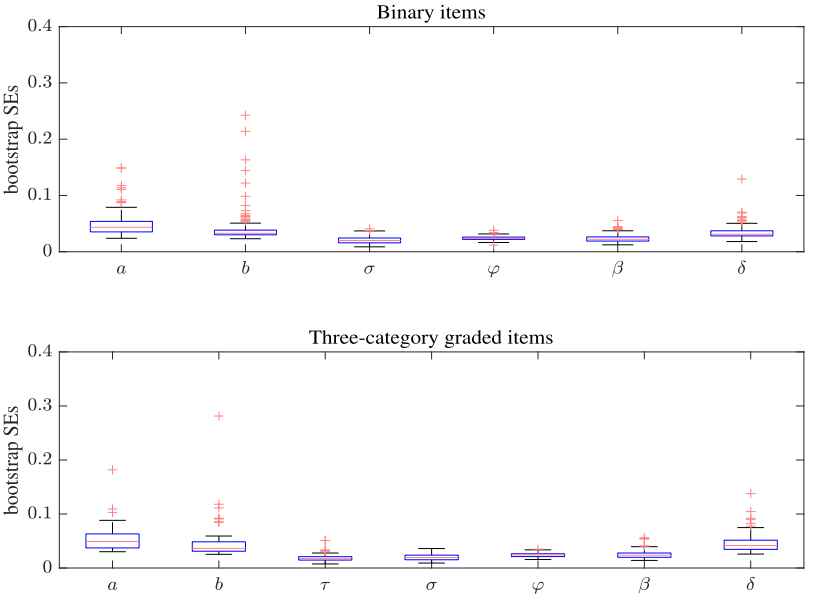


Figure 3. Boxplots of bootstrap standard errors (SEs) for the parameters in the GR–RT model.

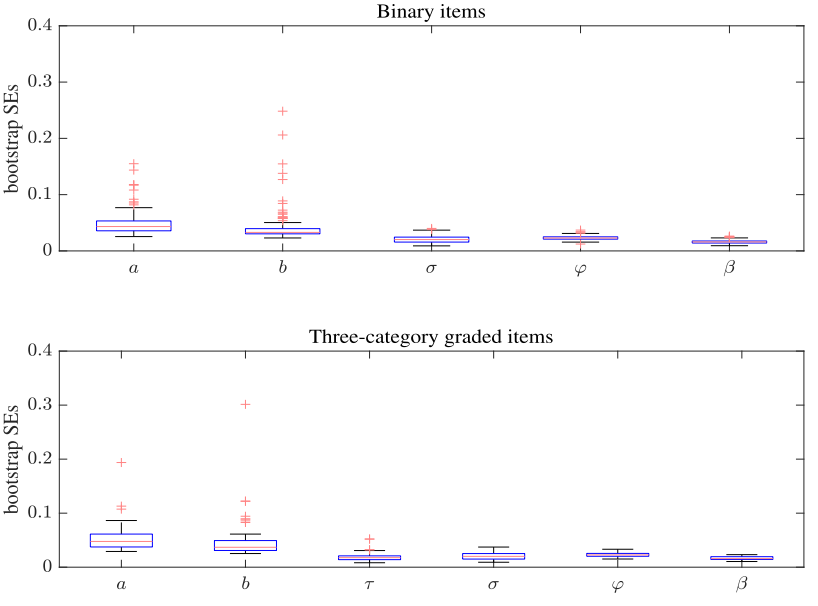


Figure 4. Boxplots of bootstrap standard errors (SEs) for the parameters in the HM.

assessments and populations. Some studies have reported that high-ability examinees respond more quickly (e.g., Fox & Marianti, 2016; Maria et al., 2016; Wang & Xu, 2015), whereas others have observed a curvilinear association between ability and speed (e.g., Molenaar et al., 2015a, 2015c). Therefore, the present findings should be interpreted in the context of this dataset rather than as evidence of a general pattern.

Figures 5 and 6 display the 95% CIs for $\boldsymbol{\delta}$ parameter in binary and three-category graded items, respectively. In both figures, CIs that exclude zero are shown in blue, whereas those including zero are shown in red. Most CIs exclude zero (76% for binary items and 84% for three-category items), indicating

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

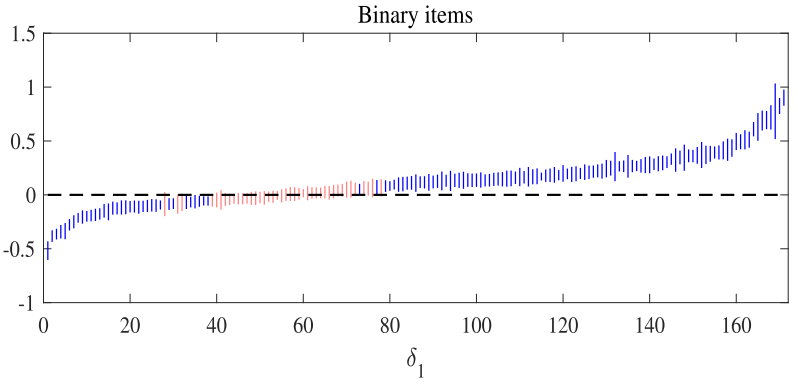


Figure 5. The 95% bootstrap confidence intervals (CIs) for δ_1 of 171 binary items in the GR–RT model. Blue lines indicate items whose CIs do not include zero, and red lines indicate items whose CIs include zero.

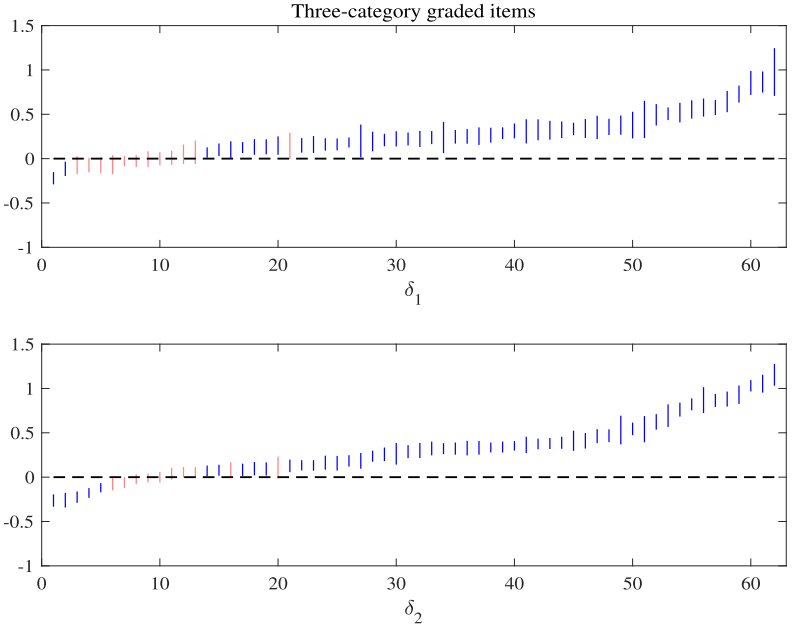


Figure 6. The 95% bootstrap confidence intervals (CIs) for δ_1 and δ_2 of 62 three-category graded items in the GR–RT model. Blue lines indicate items whose CIs do not include zero, and red lines indicate items whose CIs include zero.

that the majority of $\boldsymbol{\delta}$ estimates are significantly different from zero. These results support the inclusion of $\boldsymbol{\delta}$ and confirm the validity of the GR–RT model.

The AIC, BIC, and LRT statistics (G^2) for the GR–RT and HM models are presented in Table 7. Both AIC and BIC suggest that the GR–RT model provides a better fit to the data than the HM. Furthermore, the observed G^2 values exceed the corresponding 0.05 critical values of the chi-square distribution, indicating that the GR–RT model fits the data significantly better than the HM at the 0.05 significance level.

In terms of computational efficiency, the SAEM algorithm took approximately 772 and 673 seconds to estimate the GR–RT and HM models, respectively. Given that the analysis was based on a large-scale dataset with 18,086 respondents and 233 items, these computing times are relatively modest, indicating that the GR–RT model estimated with the SAEM algorithm is both feasible and efficient for large-scale assessment applications.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Table 7. AIC, BIC, and likelihood ratio test (LRT) of the GR–RT and HM (hierarchical model) at the test level

AIC		BIC		LRT		
GR–RT	HM	GR–RT	HM	G^2	$\chi^2_{0.05}$	Decision
1,559,200	1,570,200	1,571,600	1,580,200	11,551	336.06	Reject H_0

Note: G^2 is the likelihood ratio test statistic, and $\chi^2_{0.05}$ is the critical value from the chi-square distribution at the 0.05 significance level. If G^2 exceeds the selected critical value, H_0 is rejected, suggesting that the GR–RT model fits the data significantly better than the HM; otherwise, H_0 is not rejected.

6. Conclusions and further directions

This study develops the GR–RT model, a novel framework for the joint modeling of graded item responses and RTs. A key innovation of the GR–RT model lies in its ability to explicitly capture the influence of response categories on RTs by incorporating response-effect parameters into the RT submodel. In addition, a second-level population structure is introduced to account for the latent association between examinees' speed and ability. A second contribution of this study is the development of an SAEM algorithm for computing the MMLE of the GR–RT model parameters. The proposed SAEM algorithm provides a practical and computationally feasible approach within the frequentist framework and is specifically designed to handle mixed-format tests containing both binary and graded items, thereby enhancing the practical applicability of the GR–RT model in real-world assessment contexts.

To evaluate the performance of the proposed model and estimation method, we conducted two simulation studies and one empirical analysis. The simulation results demonstrated that the SAEM estimation method accurately estimates both the GR–RT model and the HM, and also confirmed the advantages of the GR–RT model over the HM. The empirical analysis further validated the applicability of the GR–RT model in large-scale assessments, confirming that the impact of item responses on RTs is significant for most items. Additionally, AIC, BIC, and likelihood ratio tests consistently supported the GR–RT model's advantages over the HM. In conclusion, the GR–RT model, combined with the SAEM algorithm, provides a robust and effective framework for the joint analysis of graded item responses and RTs.

Despite these contributions, several limitations remain, along with a number of promising avenues for future research. First, as discussed by Remark 1, while the current GR–RT model captures differences in the mean of RTs across response categories, it does not account for potential differences in RT variances or in the joint distribution of θ and ν across these categories. Therefore, developing a more comprehensive heterogeneity model that incorporates these variations represents a valuable and meaningful direction for future research.

Secondly, large-scale assessments are often conducted across diverse populations, such as countries, regions, schools, and other demographic or organizational subgroups, where significant differences in individuals' latent traits and item characteristics may arise between groups. To address this group-level heterogeneity, multiple-group IRT models have been developed. Building on this foundation, proposing a multiple-group GR–RT model that incorporates group-specific parameter variations would further enhance its applicability in large-scale assessments.

Third, with the advancement of computer-based assessments, various types of process data, such as mouse clicks and drags, eye-tracking measures, and neural activation signals, have become easily accessible. These data contain valuable information and provide new opportunities for deeper insights. Therefore, extending the GR–RT framework to model multimodal process data represents a promising direction for future development.

Fourth, multidimensional tests are widely used in educational and psychological measurement. However, there are currently few RT models specifically designed for multidimensional assessments. Therefore, extending the GR–RT model to incorporate multidimensional latent structures and validating its effectiveness and applicability would represent a valuable and meaningful direction for future research.

Finally, the computation of the observed information (OI) matrix remains an important open issue in this study. The OI matrix plays a central role in statistical inference, as its inverse yields

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

the asymptotic SEs of maximum likelihood estimates. However, in this study, we did not derive an algorithmic procedure for computing the OI matrix. Instead, SEs were obtained via nonparametric bootstrap. Given the importance of the OI matrix for uncertainty quantification and hypothesis testing, developing an efficient method for its computation within the SAEM framework is a key direction for future research.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10097.

Funding statement. X.M. is supported by the National Social Science Fund of China (Grant No. 23BTJ055).

Competing interests. The authors declare none.

Disclosure of use of AI tools. In this work, ChatGPT (OpenAI, GPT-4) was used for grammar checking. The authors take full responsibility for the content of the article.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Barndorff-Nielsen, O. (2014). *Information and exponential families: In statistical theory*. John Wiley & Sons.

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modeling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126–1148.

Bolsinova, M., & Tijmstra, J. (2019). Modeling differences between response times of correct and incorrect responses. *Psychometrika*, *84*(4), 1018–1046.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 257–279.

Camilli, G., & Geis, E. (2019). Stochastic approximation EM for large-scale exploratory IRT factor analysis. *Statistics in Medicine*, *38*(21), 3997–4012.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, *27*(1), 94–128.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Entink, R. K., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48.

Entink, R. K., van der Linden, W. J., & Fox, J.-P. (2009). A box–cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*(3), 621–640.

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553.

Kang, I., De Boeck, P., & Ratcliff, R. (2022). Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model. *Psychometrika*, *87*(2), 725–748.

Kang, I., Jeon, M., & Partchev, I. (2023). A latent space diffusion item response theory model to explore conditional dependence between responses and response times. *Psychometrika*, *88*(3), 830–864.

Kang, I., Molenaar, D., & Ratcliff, R. (2023). A modeling framework to examine psychological processes underlying ordinal responses and response times of psychometric data. *Psychometrika*, *88*(3), 940–974.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54.

Kolmogorov, A. N. (2018). *Foundations of the theory of probability: Second English edition*. Courier Dover Publications.

Lavielle, M., & Mbogning, C. (2014). An improved SAEM algorithm for maximum likelihood estimation in mixtures of nonlinear mixed effects models. *Statistics and Computing*, *24*, 693–707.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359.

Logan, S., Medford, E., & Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. Learning & *Individual Differences*, *21*(1), 124–128.

Maria, B., Jesper, T., & Dylan, M. (2016). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 257–279.

Meng, X., & Xu, G. (2023). A mixed stochastic approximation EM (MSAEM) algorithm for the estimation of the four-parameter normal ogive model. *Psychometrika*, *88*(4), 1407–1442.

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*(1), 1–27.

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015a). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74.
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015b). Fitting diffusion item response theory models for responses and response times using the R package diffIRT. *Journal of Statistical Software*, *66*(1), 1–34.
Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015c). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219.
Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference part I. *Biometrika*, *20*(1–2), 175–240.
Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*(2), 128.
Ratcliff, R., & Kang, I. (2021). Qualitative speed-accuracy tradeoff effects can be explained by a diffusion/fast-guess mixture model. *Scientific Reports*, *11*(1), 15169.
Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*(4), 589–606.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97.
Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence*, *48*, 37–50.
Schneider, S., Junghaenel, D. U., Meijer, E., Stone, A. A., Orriens, B., Jin, H., Zelinski, E. M., Lee, P.-J., Hernandez, R., & Kapteyn, A. (2023). Using item response times in online questionnaires to detect mild cognitive impairment. *The Journals of Gerontology: Series B*, *78*(8), 1278–1283.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Sinharay, S. (2020). Detection of item preknowledge using response times. *Applied Psychological Measurement*, *44*(5), 376–392.
Sinharay, S., & van Rijn, P. W. (2020). Assessing fit of the lognormal model for response times. *Journal of Educational and Behavioral Statistics*, *45*(5), 534–568.
Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650.
Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619.
Ulitzsch, E., von Davier, M., & Pohl, S. (2020a). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*, 83–112.
Ulitzsch, E., von Davier, M., & Pohl, S. (2020b). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*, *80*(3), 522–547.
van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.
Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417.
Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.
Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 45–58.
Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183.
Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modeling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 262–286.
Zhan, P., Jiao, H., Wang, W.-C., & Man, K. (2018). *A multidimensional hierarchical framework for modeling speed and ability in computer-based multidimensional tests*. Preprint, arXiv:1807.04003.
Zhang, L., & Chen, P. (2024). A neural network paradigm for modeling psychometric data and estimating IRT model parameters: Cross estimation network. *Behavior Research Methods*, *56*(7), 7026–7058.

Cite this article: Zhang, X., Meng, X., Gao, W. and Xu, G. (2026). A Joint Model for Graded Responses and Response Times. *Psychometrika*, 995–1019. https://doi.org/10.1017/psy.2026.10097

https://doi.org/10.1017/psy.2026.10097 Published online by Cambridge University Press