Psychometrika (2025), **90**, 1705–1731
doi:10.1017/psy.2025.10044

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

An Extended Two-Parameter Logistic Item Response Model to Handle Continuous Responses and Sparse Polytomous Responses

Seewoo Li[1] and Hyo Jeong Shin[2]

[1]Department of Education, University of California Los Angeles, Los Angeles, CA, USA; [2]Graduate School of Education, Sogang University, Seoul, South Korea

Corresponding author: Seewoo Li; Email: seewooli@ucla.edu

(Received 19 November 2024; revised 12 August 2025; accepted 12 August 2025)

Abstract

The article proposes a novel item response theory model to handle continuous responses and sparse polytomous responses in psychological and educational measurement. The model extends the traditional two-parameter logistic model by incorporating a precision parameter, which, along with a beta distribution, forms an error component that accounts for the response continuity. Furthermore, transforming ordinal responses to a continuous scale enables the fitting of polytomous item responses while consistently applying three parameters per item for model parsimony. The model's accuracy, stability, and computational efficiency in parameter estimation were examined. An empirical application demonstrated the model's effectiveness in representing the characteristics of continuous item responses. Additionally, the model's applicability to sparse polytomous data was supported by cross-validation results from another empirical dataset, which indicates that the model's parsimony can enhance model-data fit compared to existing polytomous models.

Keywords: computerized adaptive testing; continuous-bounded response; item response theory; sparse polytomous response; test linking and equating

1. Introduction

Item response theory (IRT) is being widely used in the field of psychology, education, and behavioral sciences, for many practical applications, such as data analysis, test equating and linking, developments of standard setting, and computerized adaptive testing (CAT). Numerous IRT models have been developed to take into account various features of item response data (van der Linden & Glas, 2010; van der Linden, 2016). In line with the development and expansion of IRT models, this article addresses two psychometric challenges.

Firstly, most of the IRT models used for test equating, standard setting, and CAT struggle with handling continuous response data. For example, in the age of generative artificial intelligence (AI), measuring skills, such as computer programming, often involves items that are continuously scored (e.g., Gerdes et al., 2010; Maiorana et al., 2015; Seo & Cho, 2018). Accurately accrediting test-takers on a reliable and comparable scale requires equating across different test dates, developing a standard setting, or constructing a proficiency scale. In these assessments, item responses can be completion

© The Author(s), 2025. 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.2025.10044 Published online by Cambridge University Press

rates, which are the ratio of completed subtasks to the total subtasks, where the number of subtasks often exceeds 10, or even 100. Moreover, continuous response formats, such as slider or visual analog scale (VAS) items, are increasingly used in computer-based assessments (e.g., Attali et al., 2022; García-Pérez, 2024; Gu, 2018; Open-Source Psychometrics Project, 2020; Toepoel & Funke, 2018; Vall-Llosera et al., 2020). Existing IRT models typically require discretizing continuous responses, leading to a loss of information. Instead, IRT models capable of directly handling continuous responses would provide more appropriate results for test equating and standard setting by preserving the continuous scale of the data.

Secondly, sparse item response data, characterized by a limited number of test-takers for each item or score category, is widely observed in the operational datasets (e.g., Casabianca et al., 2023; Jones et al., 2011; Kallinger et al., 2019; Mitchell et al., 2023). For example, ordinal response categories (e.g., $x = 0, 1, 2, \ldots, N$) with few or no observations in certain categories hinder the application of polytomous IRT models, often necessitating post-hoc adjustments, such as collapsing score categories. This can be further exacerbated in the adaptive testing situation, when easy items dominate item banks to ensure test-taker engagement, resulting in sparse data in lower score categories, as most test-takers may get high scores on these items. Furthermore, a balanced incomplete block design (BIBD), which produces a sparse item response matrix, can be preferred to avoid the effect of test fatigue when many items are added to an item bank at once (e.g., Chen et al., 2023). This problem is expected to become more prevalent as item banks can be rapidly expanded through automated item generation (AIG) using large language models (LLMs) (Attali et al., 2022; Macat International, 2024; Shin et al., 2024; von Davier et al., 2024). Additionally, parsimonious IRT models can be beneficial for assessments with underrepresented groups, such as visually impaired students, where only a limited number of students participate in the assessment. In such cases, accurate and stable parameter estimation is threatened, but a parsimonious IRT model can be considered beneficial (Davey & Pitoniak, 2011; O'Neill et al., 2020).

Several IRT models have been proposed to deal with these challenges (see Section 2.3). However, they may not be suitable for operational applications of IRT in practice. Potential issues include assumption misalignment (Noel & Dauvier, 2007; Samejima, 1973), infeasible parameter estimation (Müller, 1987; Verhelst, 2019), and complex model interpretation (Müller, 1987; Samejima, 1973). Additionally, some models are based on factor analysis (FA) (Ferrando, 2001; Mellenbergh, 1994), and others focus on some special types of item responses (Chen et al., 2019; Kloft et al., 2023; Molenaar et al., 2022; Noel, 2014).

As a novel alternative approach, this article aims to propose a continuous-response IRT model: extended two-parameter logistic (E2PL) item response model, which can handle continuous item responses and sparse item responses. The E2PL extends the original two-parameter logistic model (2PL: Birnbaum, 1968) by incorporating an additional precision parameter that accounts for error, modeled using a beta distribution.

The proposed model offers several advantages. First, by benchmarking the generalized latent variable modeling framework (Skrondal & Rabe-Hesketh, 2004), although it does not strictly belong to the framework as it should be (see Section 3.3), its structure and interpretation are closely aligned with existing models, such as standard IRT and FA. For instance, indices analogous to communality and unique variance in FA can be easily derived, and parameters, such as factor loadings and intercepts, are explicitly specified. Consequently, the item parameters (i.e., item discrimination, difficulty, and precision) are straightforward to interpret. The discrimination and difficulty parameters retain interpretations similar to the 2PL model, while the precision parameter governs the error component. Specifically, the inverse of the precision parameter plays a role analogous to that of the dispersion parameter in the generalized linear model (GLM) framework (Ferrari & Cribari-Neto, 2004; McCullagh & Nelder, 1989; Skrondal & Rabe-Hesketh, 2004). Second, the error term's beta distribution, which is the conjugate prior of binomial distribution, enables the model to accommodate a wide range of response distributions, including skewed or zero-one-inflated data. Third, it can effectively handle sparse score categories of polytomous item response data by transforming ordinal responses to a continuous scale (as demonstrated in Section 6). Being parsimonious, the E2PL can yield better model-data fit than conventional polytomous IRT models, especially when items have many score categories and sparse

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

data limits the accuracy of parameter estimation. Lastly, the bell-shaped item information function of the E2PL is a useful feature that is expected to be used in adaptive testing to administer an item that provides maximum information.

The remainder of this article is organized as follows. Section 2 provides an overview of IRT and reviews existing IRT models for continuous responses. Section 3 presents the mathematical formulation of the E2PL with visual illustrations and through a comparison with the 2PL, discusses its differences from Noel & Dauvier (2007)'s model, and explicates its theoretical item response distribution. A simulation study in Section 4 evaluates the stability of the estimation algorithm, parameter recovery, and the computation time of parameter estimation. Sections 5 and 6 illustrate the application of the E2PL to empirical continuous response data and sparse polytomous data, respectively. Lastly, Section 7 discusses the E2PL's potential and limitations and concludes the article. Detailed parameter estimation procedures are presented in the Appendix.

2. IRT and continuous item responses

2.1. IRT

Unlike classical test theory (CTT), which relies on summed scores, many IRT models use mathematical formulations that allow both item parameters (item characteristics) and ability parameters (test-takers' latent proficiencies) to be calibrated and interpreted on a common scale. Furthermore, with appropriate scale conversions and test designs, scores from different tests can be adjusted and compared on a common scale, a process known as linking, equating, or vertical scaling, depending on the measurement context (Kolen, 2004). These features make IRT favorable for assessment developers and measurement practitioners in developing, administering, analyzing, and reporting tests. Additionally, assuming the item parameters in the item bank are accurate, IRT provides the foundational framework of adaptive testing, which enables test assembly, test selection, test stopping, and proficiency estimation.

2.1.1. 2PL model

As one of the most popular IRT models and for its direct connection to the E2PL, we briefly review the 2PL (Birnbaum, 1968). Assuming one dichotomous item response $x \in \{0,1\}$ from a single test-taker for brevity, the probability of a correct response ($x = 1$) can be expressed as follows:

$$P(x = 1|\theta,a,b) = \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}, \tag{1}$$

where θ, a, and b are ability parameter, item discrimination parameter, and item difficulty parameter, respectively. Depending on varying θ values, the probability of a correct response ranges from 0 to 1, exhibiting an S-shaped curve as in Panel (a) of Figure 1. The item difficulty parameter b determines the inflection point of the symmetric curve at which the probability becomes 0.5, and the item discrimination parameter a determines the steepness of the curve.

Scale transformation If a, b, and θ in Equation (1) are replaced with $a^* = a/\alpha$, $b^* = \alpha b + \beta$, and $\theta^* = \alpha\theta + \beta$, it is always satisfied that $P(x = 1|\theta,a,b) = P(x = 1|\theta^*,a^*,b^*)$ for arbitrary α and β. Using this linear transformation, test linking, equating, and vertical scaling can be achieved in a more flexible and interpretable way.

Information function Supposing $P(\theta) = P(x = 1|\theta,a,b)$, the item information function $I(\theta)$ can be written as follows:

$$I(\theta) = a^2 P(\theta)\left(1 - P(\theta)\right), \tag{2}$$

which is Panel (b) of Figure 1. The item information function has its highest value at $\theta = b$, and the peak of the function gets higher with a larger a value. The function tapers to 0 as θ diverges to positive or negative infinity. The bell-shape of the function implies that the amount of information is concentrated

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

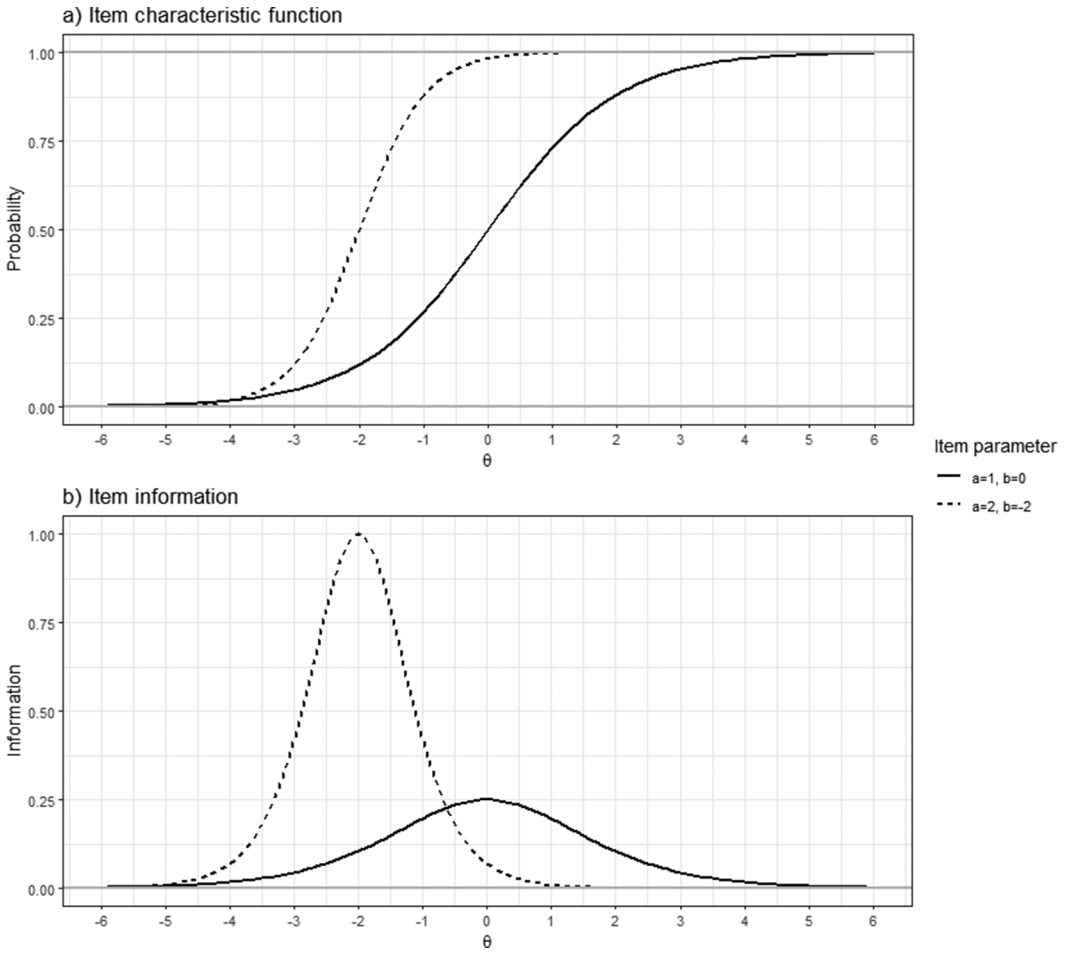


Figure 1. ICFs and item information functions of the 2PL.
Note: For illustrative purposes, values of item parameters are set to $a = 1$ and $b = 0$ for the solid lines and to $a = 2$ and $b = -2$ for the dotted lines.

around $\theta = b$. In particular, the item information function provides useful insight about the range of targeted test-takers' ability levels during the assessment design and CAT.

Asymptotic standard error Since the asymptotic standard error of the ability parameter estimate $\hat{\theta}$ is $[\sum_i I_i(\hat{\theta})]^{-1/2}$, where i denotes the item, the amount of information directly affects the accuracy of the ability parameter estimate.

2.2. Modeling continuous item responses

Along with the advancement of IRT to address a variety of measurement challenges, the introduction of diverse item formats and assessment designs has further driven the development of models capable of handling various types of item responses. These include, but are not limited to, dichotomous, polytomous, and continuous responses.

While numerous IRT models have been developed to address dichotomous and polytomous data, relatively few models have been proposed to flexibly accommodate continuous item responses. One reason for this lag is the traditional compartmentalization of latent variable modeling, which generally categorizes IRT as a framework for analyzing discrete observed variables alongside continuous latent

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

variables (Bartholomew et al., 2011). According to this classification, continuous observed variables (i.e., continuous item responses) are typically handled using FA. Although this distinction is not a strict rule (Cai, 2012), it is often presented in introductory texts on latent variable modeling for its conceptual simplicity. This convention may have led researchers to either apply FA or discrete-response IRT models to continuous-response data.

Another factor contributing to the slower development of continuous-response IRT models is the historical reliance on paper-based assessments, where items are predominantly scored dichotomously or polytomously. It is only more recently, with advancements in information and computer technologies, that continuous item formats have gained broader use in practice (e.g., Attali et al., 2022; García-Pérez, 2024; Gu, 2018; Open-Source Psychometrics Project, 2020; Toepoel & Funke, 2018).

2.3. Existing models for continuous item responses

This section reviews existing IRT models for continuous item responses. Each model has its own distinct characteristics and purposes, but they may present limitations for test development, linking, or equating due to their parameterization, underlying assumptions, or structural features. Certain models are only briefly mentioned here, as they are less relevant to the scope of this article: Mellenbergh (1994) and Ferrando (2001) adopted the identity link function in modeling continuous response, a typical choice within the FA framework, Noel (2014) proposed a model for unfolding responses using the Dirichlet distribution, Chen et al. (2019)'s model assumes a bounded latent space, and Kloft et al. (2023) focused on modeling interval responses on a continuous range using the Dirichlet distribution.

- Samejima (1973)'s model: By taking the number of categories in the graded response model (Samejima, 1969) to infinity, the model postulates an ability parameter and three item parameters: item discrimination, difficulty, and scaling. When a 2PL-type parameterization is applied, the modified difficulty parameter is a combination of the scaling parameter and the original difficulty parameter. The model is based on the normality assumption on the logit-transformed response, and item information is constant across the latent θ scale.
- Müller (1987)'s model: Similar to Samejima (1973)'s model, the model is an extension of Andrich (1982)'s rating scale model by taking the number of categories to infinity. The parameters of interest are ability, item difficulty, and item dispersion. As a Rasch-type model, it holds the specific objectivity property, enabling conditional maximum likelihood estimation. However, parameter estimation may not be practically feasible (Verhelst, 2019) and the absence of the item discrimination parameter can present limitations in practical applications. In addition, item responses are projected on the latent scale using a uniform distribution.
- Noel & Dauvier (2007)'s model: The model utilizes the beta distribution in modeling continuous bounded response, and has ability, item difficulty, and item dispersion parameters. In contrast to the other models that project item responses on the latent space, the model directly handles item responses on their original domain using the beta distribution. The model assumes that the responses are generated from an interpolation mechanism. This model has the most relevant feature to the E2PL, thus, their relationships are discussed in Section 3.4.
- Verhelst (2019)'s model: The model is a direct extension of the Rasch model (Rasch, 1960), having only ability parameter and item difficulty parameter. Despite its simplicity, the model postulates neither probabilistic distribution nor an additional parameter to account for the continuity of responses, such as the scaling or dispersion parameter in the other models discussed above. Furthermore, parameter estimation may not be practically feasible (Verhelst, 2019), and the effectiveness of the practical application has not been examined.
- Molenaar et al. (2022)'s model: Under the assumption that 0s and 1s are inflated, 0s and 1s are separated from the other responses through mixture modeling. Then, responses between 0 and 1 ($0 < x < 1$) can be modeled with any type of model listed above. However, considering that seemingly zero-one-inflated data can be properly modeled by a beta distribution, it is challenging to tell whether 0s and 1s are truly inflated.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

Building on the review of existing methods, the practical need for a new model can be formally motivated as follows. Suppose we are designing an assessment that includes both dichotomous and continuous items, using the 2PL model for dichotomous items to capture item discrimination and difficulty. This structure can be observed in practice, for instance, in programming assessments that combine multiple-choice questions (scored dichotomously) with coding tasks evaluated as percentage-correct (continuous). In this context, the zero-one inflated model by Molenaar et al. (2022) is not suitable for general measurement purposes. Additionally, the models proposed by Müller (1987), Verhelst (2019), and Noel & Dauvier (2007) do not incorporate item discrimination parameters, making them incompatible with the 2PL framework. The interpolation mechanism assumed in Noel & Dauvier (2007) is also unlikely to be appropriate for item responses such as percentage-correct scores. Finally, while Samejima (1973)'s model includes discrimination and difficulty parameters, they are not aligned with those of the 2PL model. In particular, its expected value is given by $\sigma(v(\theta - b))$ (Wang & Zeng, 1998), where v is a scaling parameter and $\sigma(\cdot)$ denotes the sigmoid function, whereas the 2PL uses $\sigma(a(\theta - b))$ as in Equation (1), with a representing item discrimination. Therefore, a new model that is compatible with the 2PL, as well as capable of handling continuous item responses, is required.

3. E2PL model

3.1. Formulation of the E2PL

3.1.1. Model equations

Moving from the 2PL (see Equation (1)) to the E2PL, we make a transition from binary to continuous response. Below, the response x takes a real number ($0 < x < 1$), thus, the modeled value is no longer a probability for the Bernoulli distribution but an observable quantity. The model equations can be written as follows:

$$\begin{aligned} x &= \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}} + \varepsilon \\ &= \mu + \varepsilon, \end{aligned} \tag{3}$$

$$\mu = E[x|\theta,a,b,v] = \frac{\exp(a(\theta-b))}{1+\exp(a(\theta-b))}, \tag{4}$$

$$Var[x|\theta,a,b,v] = Var[\varepsilon|\theta,a,b,v] = \frac{\mu(1-\mu)}{v+1}, \tag{5}$$

and

$$\begin{aligned} f(\varepsilon \mid \mu,v) = \frac{\Gamma(v)}{\Gamma(v\mu)\Gamma(v(1-\mu))}(\varepsilon+\mu)^{v\mu-1}(1-\varepsilon-\mu)^{v(1-\mu)-1} \\ (\varepsilon+\mu) \sim Beta(v\mu, v(1-\mu)). \end{aligned} \tag{6}$$

The expected response μ is equal to the model equation of the 2PL (Equation (1)). Particularly, the error term ε is introduced to account for the continuous nature of the response using a beta distribution. Through a comparison with the 2PL, Section 3.2.1 illustrates how this error term reflects the continuity. The error term follows a shifted beta distribution where the amount of the shift is $-\mu$ ($-\mu < \varepsilon < 1-\mu$). The precision parameter v can represent the degree to which the density is concentrated around μ. Using Equations (3) and (6), the model can be rewritten in a probabilistic form:

$$P(x|\theta,a,b,v) = \frac{\Gamma(v)}{\Gamma(v\mu)\Gamma(v(1-\mu))}x^{v\mu-1}(1-x)^{v(1-\mu)-1}. \tag{7}$$

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

3.1.2. Information function

The item information function of the E2PL can be expressed using trigamma function $\psi_1(\cdot)$:

$$I(\theta) = (av\mu(1-\mu))^2 \left[\psi_1(v\mu) + \psi_1(v(1-\mu))\right]. \tag{8}$$

In general, the item information function resembles a bell-shaped curve, peaking at $\theta = b$. However, its shape can vary depending on the precision parameter v. When v is close to 3, the function can take a W-shaped form, whereas for values of v less than 2, it tends to look like a bell-shaped curve flipped upside down. Notably, even with small values of v, the item information in the E2PL model is greater than that of the 2PL model, assuming the same a and b parameters are used for both. The greater information of the E2PL can be attributed to its more refined response structure compared with binary responses. Unlike many other dichotomous or polytomous models, the information function in the E2PL does not approach zero as θ approaches $\pm\infty$, instead asymptotically approaching a^2 from below.

3.1.3. Likelihood

We can add subscripts to the equations to express a likelihood function of data. Letting $j = 1,2,\ldots,N$ denote test takers and $i_j \in I_j$ be an ith item among I_j items that the jth test taker responded to, the likelihood can be expressed as follows:

$$\mathcal{L} = \prod_{j=1}^{N} \prod_{i_j \in I_j} P\left(x_{ji_j} | \theta_j, a_{i_j}, b_{i_j}, v_{i_j}\right). \tag{9}$$

The individual item response probabilities (Equation (7)) are multiplied under the assumptions that responses within individuals are independent conditional on the latent trait level θ_j (i.e., the local independence assumption) and that responses are statistically independent across individuals.

3.1.4. Model-data fit and standardized residuals

Following the approach proposed by Ferrari & Cribari-Neto (2004) in the context of beta regression, the standardized residual for test taker j and item i can be computed as:

$$r_{ji} = \frac{x_{ji} - \hat{\mu}_{ji}}{\sqrt{Var\left(x_{ji}\right)}}, \tag{10}$$

where $\hat{\mu}_{ji} = \left(1 + \exp\left(-\hat{a}_i(\hat{\theta}_j - \hat{b}_i)\right)\right)^{-1}$ and $Var\left(x_{ji}\right) = \frac{\hat{\mu}_{ji}(1-\hat{\mu}_{ji})}{\hat{v}_i+1}$ are from Equations (4) and (5). Residual analysis based on this formulation can provide evidence of model misspecification.

The overall model-data fit can be assessed using a pseudo R^2, defined as the squared correlation between the log-odds of $\hat{\mu}_{ji}$ and the log-odds of x_{ji} (Ferrari & Cribari-Neto, 2004). Additionally, K-fold cross-validation provides a more robust estimate of predictive performance. In this case, the log-likelihood under the beta distribution or the root mean squared error (RMSE) of the standardized residuals (i.e., $\sqrt{\frac{1}{T}\sum_j \sum_i r_{ji}^2}$) can serve as a loss function, where T is the total number of item responses evaluated.

Accordingly, for the current version of the E2PL, we advocate the use of standardized residuals, pseudo R^2, and K-fold cross-validation for evaluating model fit, due to their interpretability. More advanced diagnostic and fit assessment methods (e.g., Espinheira et al., 2019; Espinheira et al., 2008; Smithson & Verkuilen, 2006) may be appropriate for future extensions of E2PL.

3.1.5. Parameter estimation

Item parameters can be estimated through marginal maximum likelihood using the expectation–maximization algorithm (MML-EM: Bock & Aitkin, 1981), where the marginal likelihood is obtained by integrating the likelihood $\mathcal{L}$ with respect to θ. Conventionally, the θ distribution is assumed to follow the standard normal distribution. Several types of scores can be used as ability parameter estimates, such

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

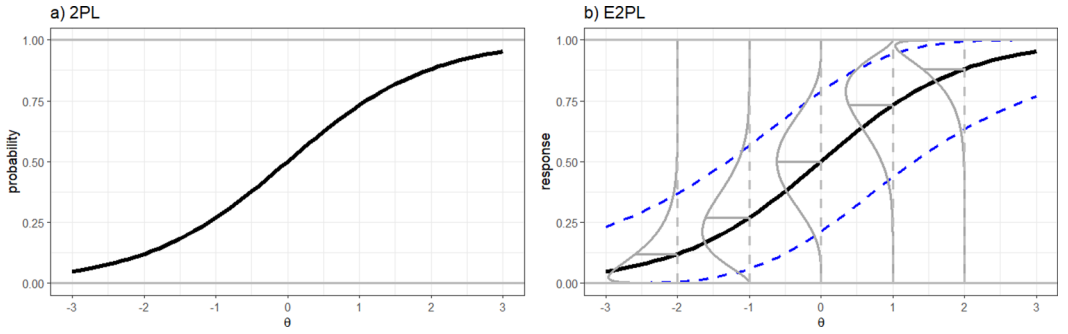


Figure 2. ICFs of the 2PL and the E2PL.
Note: For the illustration, values of item parameters are set to $a = 1$ and $b = 0$ for both functions and $\nu = 10$ for the E2PL. The y-axis is probability in Panel (a) and response in Panel (b). In Panel (b), the blue dashed lines indicate 95% interval conditional on θ. The **gray curves** are probability densities of continuous item responses for the selected θ values of -2, -1, 0, 1, and 2. The **gray line segments** on the μ-curve indicate the means of the densities.

as expected *a posteriori* (EAP), maximum likelihood estimate (MLE), or weighted likelihood estimate (WLE). Details of the parameter estimation are provided in the Appendix.

3.2. Item characteristic function (ICF)

3.2.1. Comparison with the 2PL

Within the generalized latent variable modeling framework (Skrondal & Rabe-Hesketh, 2004), both FA and IRT models can be expressed in the form $\lambda(\mu) = a\theta + c$, where $\lambda(\cdot)$ denotes the link function, typically the identity link for FA and the logit link for IRT. In this formulation, the slope parameter a corresponds to the factor loading, and c represents the intercept. Specifically, in the 2PL model, a is interpreted as item discrimination, while item difficulty is given by $b = -\frac{c}{a}$. The E2PL model adopts this same structural formulation, allowing for analogous interpretations of the a and b parameters. However, the key distinction lies in the modeling of the dispersion structure: the 2PL model assumes a Bernoulli distribution for the observed responses, whereas the E2PL employs a beta distribution. This difference in distributional assumption differentiates the two models.

Figure 2 visually illustrates similarities and differences between the 2PL and the E2PL. Due to their common item discrimination and difficulty parameters, the ICF of the 2PL in Panel (a) and the μ curve of the E2PL in Panel (b) are identical. Thus, item interpretations of the two models would be similar or, for convenience, interchangeable. In addition, an identical linear transformation can be carried out for the two models for test linking or equating. Therefore, for mixed-format data of dichotomous and continuous item responses, adopting the pair of the 2PL and the E2PL simplifies the model interpretation and the scale transformation.

Compared with the 2PL's ICF, the ICF of the E2PL in Panel (b) of Figure 2 includes the blue and gray curves to display the beta distribution, the error component of the E2PL. In line with the change of the y-axis from *probability* to *response*, the error term of the E2PL accounts for the continuous scale of the response. The distribution (**gray curves**) and the 95% interval (blue curves) illustrate how the error distribution changes according to θ.

3.2.2. Role of the item parameters

Figure 3 visually illustrates how item parameters influence ICFs in the E2PL. The upper left and right panels show that the effects of the a and b parameters on ICFs are structurally identical to those in the 2PL model: the a parameter alters the steepness of ICFs and the b parameter shifts ICFs along the θ-axis. However, unlike the 2PL, where the Y-axis represents probability, the E2PL models the response directly

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

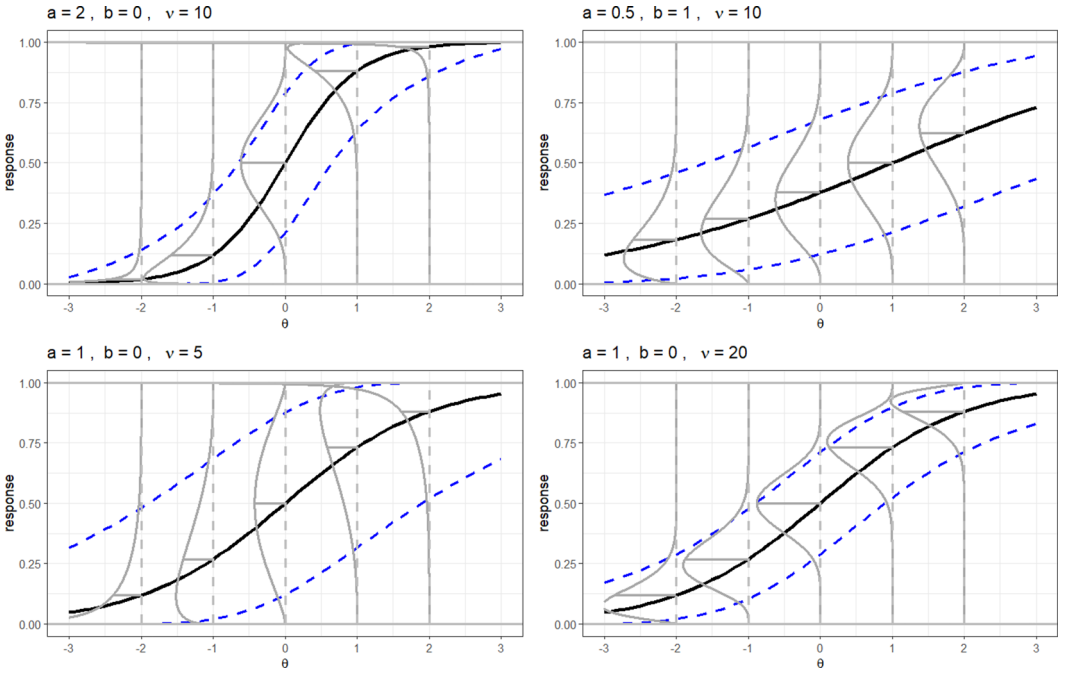


Figure 3. ICFs of the E2PL with varying parameter values.
Note: The blue dashed lines indicate 95% interval conditional on θ. The **gray curves** are probability densities of continuous item responses for selected θ values of -2, -1, 0, 1, and 2. The **gray line segments** on the μ-curve indicate the means of the densities.

on a continuous scale. Accordingly, the a parameter determines the slope of the expected response (i.e., the μ curve), and the expected response reaches 0.5 when $\theta = b$.

Focusing now on the role of the precision parameter ν, when the expected response curves (**solid black lines**) are held constant, the error distributions (**gray curves**) are also identical across panels with the same ν. For example, in the upper-left panel at $\theta = 0$ and the upper-right panel at $\theta = 1$, the distributions are identical because both yield an expected response of $\mu = 0.5$ under the same precision level $\nu = 10$.

To illustrate the effect of the precision parameter, the precision parameter ν is varied between the lower left and right panels. It can be seen that the precision parameter modified only the error variances (**gray curves**) and the width of the interval (blue curves). In brief, after the mean curve (**solid black curves**) is determined by the a and b parameters, the precision parameter ν determines the dispersion, or concentration, of ICFs.

3.3. Communality and unique dispersion

The formulation of the E2PL is well-aligned with the generalized latent variable modeling framework (Skrondal & Rabe-Hesketh, 2004), as it models the mean structure μ using the logit link function and introduces the precision parameter ν to handle the dispersion of data. Here, we use the term *dispersion* instead of *variance* to indicate the role of ν, since statistical variance of the beta distribution depends on μ as in Equation (5). Meanwhile, the beta distribution does not provide a statistically independent structure between μ and ν (Ferrari & Cribari-Neto, 2004), which differentiates the E2PL from the generalized latent variable modeling framework (McCullagh & Nelder, 1989; Skrondal & Rabe-Hesketh, 2004). However, this is a desirable property that allows the model to account for the effects of the item parameters during the calculation of the residual dispersion represented by ν.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

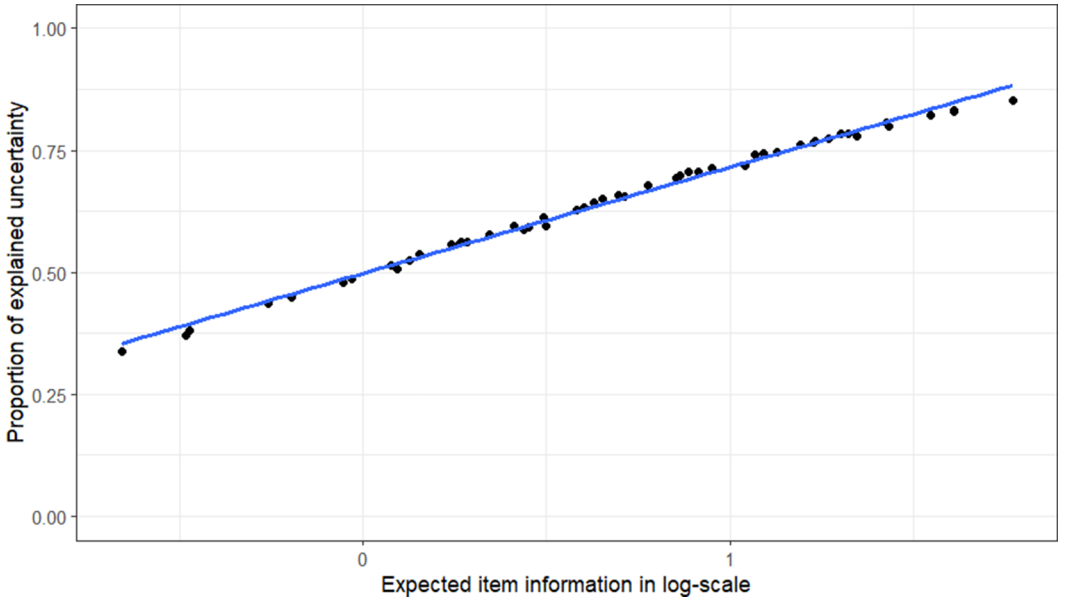


Figure 4. The relationships between the proportion of explained uncertainty $(1 - \frac{\nu_{\text{pre}}}{\nu_{\text{post}}})$ and expected item information $E[I(\theta)]$.
Note: The expected item information values are calculated as $E[I(\theta)] = \int I(\theta)\phi(\theta)\, d\theta$, where $I(\theta)$ is the item information function in Equation (8) and the standard normal distribution $\phi(\theta)$ is used as the latent distribution. The figure is obtained from a randomly generated data for 100,000 test takers and 50 items, where $\theta \sim N(0,1)$, $a \sim Unif(0.5,1.5)$, $b \sim N(0,0.5)$, and $\nu \sim Gamma(10,1)$.

To mathematically illustrate the aforementioned points, let $\nu_{\text{pre}} = (\frac{\bar{X}(1-\bar{X})}{s^2} - 1)$ be the degree of precision for an item before fitting the E2PL with sample mean $\bar{X}$ and sample variance s^2, and $\nu_{\text{post}} = \hat{\nu}$ be the estimate of the precision parameter of the item in Equations (3)–(7). The subscripts for the items are omitted for notational brevity. Then, using the FA notation from MacCallum (2009), $1/\nu_{\text{pre}}$ and $1/\nu_{\text{post}}$ are analogous to the sample variance s^2 and the estimate of the unique variance $\hat{\psi}$ in FA, respectively. As a result, the proportion of the unique dispersion of this item becomes $\frac{1/\nu_{\text{post}}}{1/\nu_{\text{pre}}} = \frac{\nu_{\text{pre}}}{\nu_{\text{post}}}$, and the communality $(1 - \frac{\nu_{\text{pre}}}{\nu_{\text{post}}})$ indicates the proportion of the uncertainty explained by the latent variable. The relationships above are also applicable to multidimensional θ.

Figure 4 shows an almost linear relationship between the explained dispersion and the expected item information in log scale. This trend well reflects the conventional practice in IRT to take item parameters into account, rather than excluding them when explaining the uncertainty of the data. Notably, the formulation of the communality and unique dispersion, as well as their relationships to the item information, are not subject to the scale transformation in Section 2.1.1, as the transformation does not affect μ.

3.4. Differences between the E2PL and Noel and Dauvier's model

It would be worthwhile to clarify the differences between the E2PL and Noel & Dauvier (2007)'s model as they are closely related to one another. Although they introduced a Rasch-type model without item discrimination parameter, it can be easily added to the model. The following discussion assumes Noel and Dauvier's model with the inclusion of the item discrimination parameter to make a fair comparison.

Item responses of Noel and Dauvier's model are assumed to be a manifestation of the interpolation mechanism; an interpolation of one weight (α) pulling a response toward 0 and another weight (β) pulling it toward 1 (i.e., $x = \frac{\alpha}{\alpha+\beta}$), thereby adopting the shape–shape parameterization of the beta distribution. In other words, they modeled the response using $Beta(\alpha,\beta)$, where the expected

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

value of the response x is the interpolation of the two parameters (i.e., $E(x) = \frac{\alpha}{\alpha+\beta}$). In comparison, without assuming a particular mechanism on item responses, the E2PL adopts the mean–precision parameterization, which is a widely used practice in beta regression (Ferrari & Cribari-Neto, 2004).

While the conditional means of the two models (i.e., the μ terms) can be identically expressed using the a and b parameters, the two models differ in their treatment of precision. In the E2PL model, the conditional variance—after accounting for μ—is solely governed by the precision parameter ν, enabling a clear separation of the error component as shown in Equation (3). In contrast, the conditional variance in Noel and Dauvier's model remains a function of all item parameters even after conditioning on μ. Following the parameterization of Ferrari & Cribari-Neto (2004), the precision parameter in Noel's model is given by $2\exp\left(\frac{\tau}{2}\right)\cosh\left(\frac{a(\theta-b)}{2}\right)$, where τ is an additional parameter to account for dispersion. In the E2PL model, by contrast, the precision is directly represented by ν. Because the precision term in Noel and Dauvier's model depends on the latent scale, scale transformations and the derivation of quantities, such as communality and unique dispersion, require additional consideration.

3.5. Distribution of item responses

Each IRT model assumes a particular distribution for item responses. For instance, when the 2PL's item parameters are $a = 1$ and $b = 0$ and $\theta \sim N(0,1)$, the expected probability of observing an item response of $x = 1$ is 0.5.

Similarly, the item response distributions of the E2PL can be derived when the latent distribution is specified. Assuming the standard normal distribution on the latent variable, the response distributions can be mathematically expressed as follows:

$$\begin{aligned} f(x|a,b,\nu) &= \int_{-\infty}^{\infty} f(x|\theta,a,b,\nu)\phi(\theta)\,d\theta \\ &= \int_{-\infty}^{\infty} \frac{\Gamma(\nu)}{\Gamma(\nu\mu)\Gamma(\nu(1-\mu))} x^{\nu\mu-1}(1-x)^{\nu(1-\mu)-1}\phi(\theta)\,d\theta, \end{aligned} \tag{11}$$

where the μ is defined in Equation (4) and $\phi(\theta)$ denotes the standard normal latent distribution. The distribution depends only on item parameters after integrating out the latent variable θ.

Figure 5 shows that skewed or zero-one-inflated response distributions can be generated by the model, as well as bell-shaped distributions. Firstly, when the difficulty parameter b is different from the mean of the latent distribution, a skewed distribution can be formulated. For instance, the mass of the solid line's density ($a = 1$, $b = -1$, $\nu = 10$) is more concentrated near 1 than 0, as the item is relatively easy for the population: $b < E(\theta) = 0$. In comparison, the densities of the dotted and dashed lines ($a = 1$, $b = 0$, $\nu = 10$; $a = 1$, $b = 0$, $\nu = 1$) are both symmetric as their difficulty parameters are equal to the population mean.

Secondly, zero-one-inflated response distributions can be formulated with a small value of ν. For example, the dashed line ($a = 1$, $b = 0$, $\nu = 1$) shows that both 0 and 1 are inflated even when the difficulty parameter b is equal to the population mean. In a strict sense, it may not be an actual zero-one-inflated distribution as the domain of the beta distribution does not contain 0 and 1. However, in practice, responses are almost always rounded to a certain point (e.g., to the nearest hundredth), producing 0s and 1s. Additionally, skewed zero-one-inflated response distributions can be formulated when the precision parameter is small and $b \neq E(\theta)$.

3.6. Preprocessing item responses

To apply the E2PL, a simple transformation of item responses is often necessary, when raw data include minimum and maximum values (e.g., 0% and 100%) that fall outside of the domain of the beta distribution, which excludes 0 and 1. For example, item responses collected as percentages typically include these values as minimum and maximum scores.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

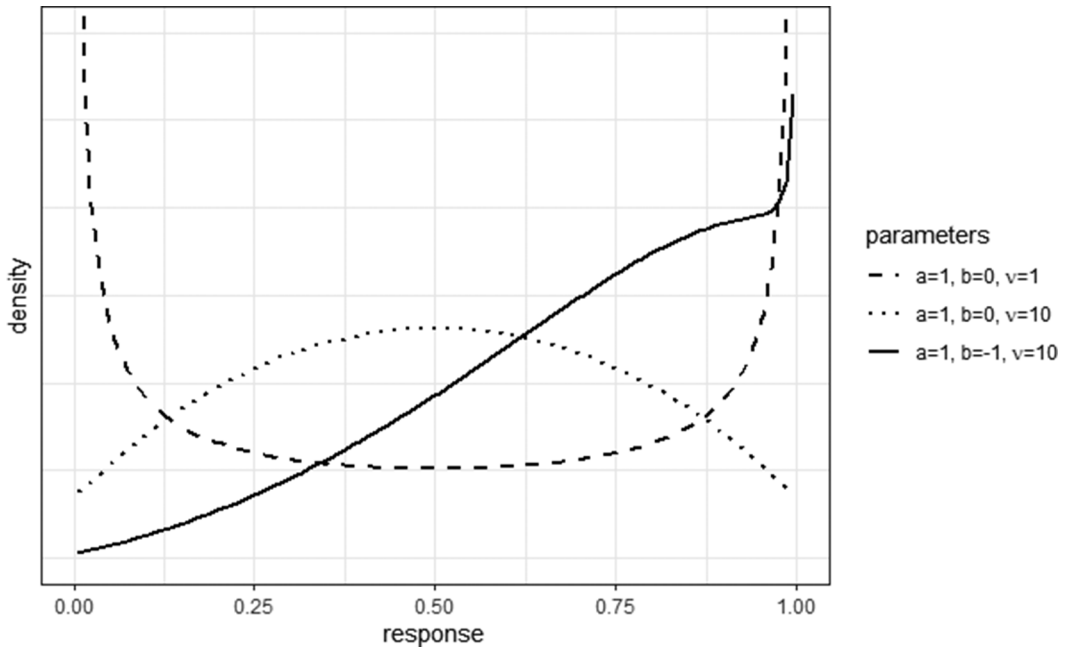


Figure 5. Response distributions of the E2PL.
Note: The distributions are derived from the standard normal latent distribution. The distributions are numerically approximated.

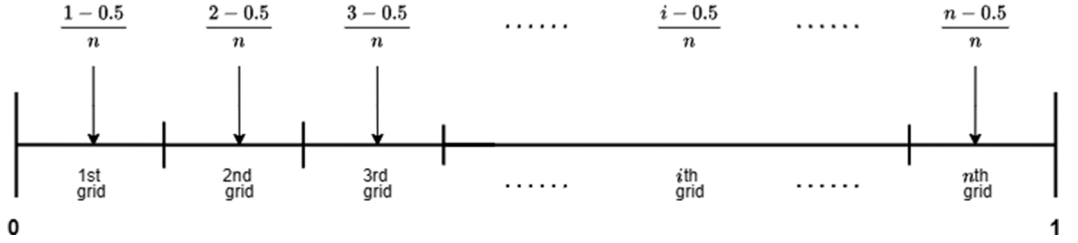


Figure 6. Mapping raw scores on a unit interval.

To adjust raw responses to fit within the beta distribution while preserving their key characteristics, we recommend the preprocessing scheme shown in Figure 6. This approach divides the open interval (0, 1) into equally spaced grids based on the smallest unit of observation in the raw data and maps the raw responses to the midpoints of these grids. For instance, if percentage responses are measured in 1% increments, the raw scores from 0% to 100% are mapped to the midpoints of 101 equally spaced grids: $\{0\%, 1\%, \ldots, 100\%\} \mapsto \{\frac{0.5}{101}, \frac{1.5}{101}, \ldots, \frac{100.5}{101}\}$. Figure 6 visually illustrates this method.

This transformation is useful in cases where values approach the absolute maximum or minimum, allowing for meaningful representation of such scores. In practical contexts, discrepancies often exist between mathematical maximums, defined as the highest possible score, and practical maximums, which may exhibit slight variations even at the upper boundary. For instance, some maximum scores represent model solutions that are qualitatively distinct from those that merely achieve the highest possible score. A similar rationale applies to minimum scores, such as the differences between zero scores due to blank responses and other minimum scores. This approach supports using the open interval of the beta distribution, rather than the closed interval, to better reflect the nuanced characteristics of item responses near the extremes. Additionally, other transformations, such as slightly adjusting

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

the minimum and maximum values (Noel & Dauvier, 2007), could also be applied when they more accurately capture the characteristics of item responses.

Furthermore, this transformation is justifiable when each item comprises multiple subtasks (see Sections 5 and 6). In such cases, a more sophisticated modeling approach, such as the testlet response model (Bradlow et al., 1999; Wainer et al., 2007), is often appropriate for accounting for local item dependencies. However, testlet models can be sensitive to sample size constraints, potentially increasing estimation error due to the bias–variance trade-off and threatening the validity of inferences. While more parsimonious models may be preferable under limited sample conditions, standard polytomous IRT models are often not suitable alternatives, as their response categories typically do not align directly with individual subtasks. Thus, the proposed transformation provides a practical alternative to testlet modeling by enabling the use of the E2PL that maintains fidelity to the item structure while avoiding issues associated with small samples.

4. Simulation study

A simulation study is conducted to assess the recovery of item parameters and the stability of the parameter estimates of the E2PL. Data were generated and the model was fitted using the `IRTest` package in `R` (Li, 2025), with evaluation metrics computed using built-in `R` functions (R Core Team, 2024). The simulation code is publicly available at https://github.com/SeewooLi/E2PL_simulation_study.

4.1. Data generation and model-fitting

The simulation study utilizes a set of 12 items, designed as the factorial combination of the following item parameters: discrimination ($a \in \{0.5, 1\}$), difficulty ($b \in \{-1, 0, 1\}$), and precision ($\nu \in \{5, 10\}$). Sample sizes of 250, 500, and 1000 are used, and for each sample size, 200 sets of item response data are generated based on the specified item parameters.

Model fitting is conducted using the `IRTest` package (Li, 2025). A custom function `IRTest_Cont`, which is developed for this study, implements the MML-EM procedure (Bock & Aitkin, 1981). Convergence of the MML-EM procedure is defined as the point at which the maximum change in parameter estimates falls below 0.0001 within a maximum of 200 EM iterations, ensuring robust model fitting in this simulation. To address the challenge of directly estimating a bounded parameter (i.e., $\nu > 0$), $\log(\nu)$ is estimated and the changes in $\log\hat{(}\nu)$ are tracked. The MML-EM procedure employs 121 equally spaced quadrature points ranging from −6 to 6, assuming the standard normal latent distribution. EAP scores are used to estimate the ability parameter.

4.2. Evaluation criteria

The accuracy of parameter recovery is assessed using bias and RMSE for the item parameter estimates across 200 replications. To evaluate computational efficiency, mean computation time (MCT) is calculated. An AMD Ryzen 7 5700G processor is used for the study. Finally, the stability of the estimation process is confirmed by verifying the convergence of the estimation procedures throughout the simulation study.

4.3. Results

All 600 MML-EM procedures (200 replications × 3 sample sizes) successfully converged within 200 iterations, demonstrating robust stability in the estimation process.

Table 1 summarizes the biases and RMSEs for parameter recovery. The estimates for all three item parameters (discrimination, difficulty, and precision) were nearly unbiased, with the largest observed bias being less than 0.03. Notably, parameter recovery was satisfactory even for the smallest sample size of 250, and the estimation of the precision parameter remained accurate and stable across conditions.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

Table 1. Biases and RMSEs in parameter recovery

	Sample	Parameter			Bias (RMSE)		
Item	size	a	b	ν	a	b	$\log(\nu)$
1		0.5	−1	5	0.01 (0.07)	0.01 (0.18)	0.02 (0.09)
2		0.5	−1	10	−0.00 (0.05)	−0.01 (0.15)	0.00 (0.10)
3		0.5	0	5	−0.01 (0.06)	−0.01 (0.12)	0.00 (0.08)
4		0.5	0	10	0.00 (0.05)	−0.01 (0.10)	0.01 (0.09)
5		0.5	1	5	−0.00 (0.06)	0.02 (0.18)	0.01 (0.10)
6	250	0.5	1	10	0.00 (0.05)	−0.00 (0.14)	0.01 (0.09)
7		1	−1	5	−0.01 (0.08)	−0.02 (0.12)	0.02 (0.10)
8		1	−1	10	−0.00 (0.08)	−0.01 (0.11)	0.01 (0.11)
9		1	0	5	−0.01 (0.08)	−0.00 (0.09)	0.02 (0.10)
10		1	0	10	0.00 (0.06)	−0.00 (0.08)	0.02 (0.10)
11		1	1	5	−0.02 (0.08)	0.01 (0.11)	0.01 (0.10)
12		1	1	10	−0.00 (0.07)	0.00 (0.09)	0.02 (0.10)
1		0.5	−1	5	0.01 (0.05)	0.00 (0.13)	0.00 (0.06)
2		0.5	−1	10	0.00 (0.03)	−0.00 (0.10)	0.00 (0.06)
3		0.5	0	5	0.00 (0.04)	−0.00 (0.08)	0.00 (0.06)
4		0.5	0	10	0.01 (0.03)	−0.01 (0.06)	0.01 (0.07)
5		0.5	1	5	−0.00 (0.04)	0.01 (0.11)	0.01 (0.06)
6	500	0.5	1	10	0.00 (0.04)	−0.01 (0.10)	0.01 (0.06)
7		1	−1	5	−0.03 (0.06)	−0.03 (0.08)	0.00 (0.07)
8		1	−1	10	−0.00 (0.05)	−0.01 (0.06)	0.02 (0.08)
9		1	0	5	−0.01 (0.06)	−0.00 (0.05)	0.01 (0.07)
10		1	0	10	0.00 (0.05)	−0.01 (0.05)	0.01 (0.08)
11		1	1	5	−0.03 (0.06)	0.02 (0.08)	0.00 (0.07)
12		1	1	10	0.00 (0.04)	−0.01 (0.06)	0.01 (0.09)
1		0.5	−1	5	0.00 (0.03)	−0.00 (0.09)	−0.00 (0.04)
2		0.5	−1	10	0.00 (0.03)	0.00 (0.07)	0.00 (0.04)
3		0.5	0	5	−0.00 (0.03)	−0.00 (0.06)	0.01 (0.04)
4		0.5	0	10	−0.00 (0.02)	0.00 (0.05)	0.00 (0.04)
5		0.5	1	5	−0.00 (0.03)	0.01 (0.09)	0.00 (0.04)
6	1000	0.5	1	10	0.00 (0.03)	0.00 (0.07)	0.00 (0.04)
7		1	−1	5	−0.02 (0.04)	−0.02 (0.06)	0.00 (0.05)
8		1	−1	10	−0.00 (0.03)	−0.01 (0.05)	0.01 (0.05)
9		1	0	5	−0.00 (0.04)	−0.00 (0.04)	0.00 (0.05)
10		1	0	10	0.00 (0.03)	−0.00 (0.03)	0.01 (0.05)
11		1	1	5	−0.02 (0.04)	0.01 (0.06)	0.01 (0.05)
12		1	1	10	0.00 (0.03)	−0.01 (0.05)	0.01 (0.05)

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

Table 2. MCT and RMSE($\hat{\theta}$)

Sample size	MCT	RMSE($\hat{\theta}$)
250	8.74	0.27
500	14.50	0.27
1000	26.29	0.27

Note: The MCTs are measured in seconds.

An additional simulation was conducted to examine parameter recovery for items with small a or ν values. Two more items were added to the simulation design: Item 13 with $a = 0.2$, $b = 0$, and $\nu = 10$, and Item 14 with $a = 1$, $b = 0$, and $\nu = 1$. For Item 13, 95% of the item responses fall within the interval $(0.2, 0.8)$, whereas for Item 14, two-thirds of the responses fall outside this interval. The parameter estimates for Item 13 are unbiased with comparatively larger RMSE for $\hat{b}$, which decreased from 0.20 to 0.10 as the sample size increased from 250 to 1000. In contrast, the discrimination parameter a for Item 14 exhibited a negative bias of approximately -0.13 across the three sample sizes, which can be attributed to the rounding of item responses to the nearest value on a discretized scale ranging from 0.00005 to 0.99995 in steps of 0.0001. This suggests that more fine-grained response options are necessary when the precision parameter ν is small to account for the areas near 0 and 1.

The MCT for convergence is shown in Table 2. On average, the MML-EM procedure took 8.74, 14.50, and 26.29 seconds to converge for sample sizes of 250, 500, and 1000, respectively. These times are considered efficient given the data size, a convergence threshold of 0.0001, and the 121 quadrature points used. The RMSE($\hat{\theta}$) is included for reference; as expected, with test length held constant, RMSE($\hat{\theta}$) values were consistent across the simulation conditions.

5. Empirical Illustration 1: Continuous response data

5.1. Data

To illustrate the application and interpretation of the E2PL model, we use empirical data from a company located in South Korea. This company developed and administered an assessment to measure the programming skills of its employees, aiming to establish a reliable and valid item bank through psychometric methods.

For this analysis, we use 12 items, each containing one to four tasks, with each task subdivided into 4 to 60 subtasks depending on the item's purpose. For instance, Item 3 comprises two tasks, with 50 subtasks in the first task and 60 subtasks in the second. The data is unbalanced, with responses from 1,732 participants who answered between one and ten items. Most participants responded to either one item ($n = 333$) or two items ($n = 1253$), and only one participant responded to ten items.

Following the preprocessing procedure outlined in Section 3.6, the binary subtask scores were aggregated and mapped to a unit interval, transforming them into continuous item responses. First, a task-level score is computed. For example, if a participant completes 17 subtasks out of 50 (resulting in the 18th category from 0 to 50), the task-level score is calculated as $\frac{17.5}{51} \approx 0.34$. The item score is then the average of the task-level scores, producing a continuous value that represents the percentage of task completion.

As in Section 4, the `IRTest` (Li, 2025) package is utilized for the model fitting. Given the sparsity of the data, a more lenient convergence threshold of 0.01 is adopted. To evaluate the overall model-data fit, the pseudo R^2 (see Section 3.1.4) was calculated after model estimation, resulting in a value of 0.494.

Figure 7 presents histograms of the item responses, which reflect the continuous scores obtained from the mapping process in Section 3.6. Across all items, the unit interval is densely populated with item responses. In an extreme case, for example, Item 11 yields 172 unique response values, highlighting the challenges of handling such data with a polytomous IRT model.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

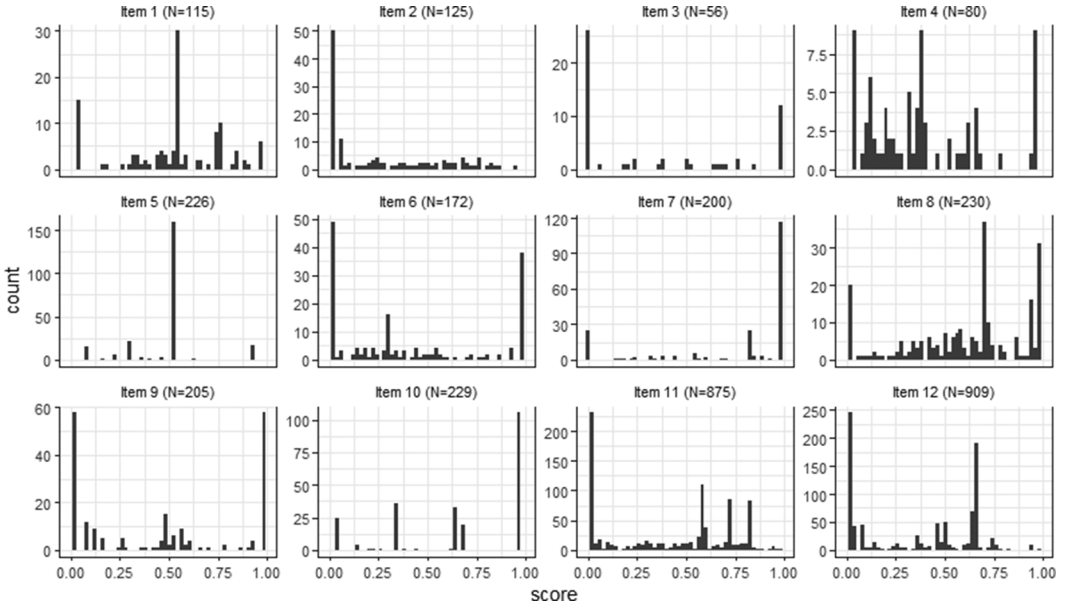


Figure 7. Item response distributions of the programming assessment dataset.

Table 3. Item parameter estimates and communalities of the programming assessment dataset

Item	*a*	*b*	*v*	Communality
1	0.79	−0.03	5.03	0.43
2	0.15	6.85	2.15	0.31
3	0.40	1.07	0.76	0.48
4	0.35	1.02	2.14	0.13
5	0.48	0.17	8.87	0.19
6	0.48	0.47	1.01	0.33
7	0.21	−3.65	0.84	0.46
8	1.46	−0.32	9.53	0.81
9	0.65	0.17	1.20	0.50
10	0.58	−1.00	1.98	0.43
11	0.67	0.75	2.22	0.32
12	0.59	1.17	2.62	0.30

Except for Items 1, 5, and 8, most item response distributions are skewed or exhibit zero-one inflation. For example, Item 2 responses are concentrated near 0, while Item 9 shows inflation near both 0 and 1. Note that the 0s and 1s mentioned above are not exact 0s and 1s, but the closest value to 0 and 1, respectively.

5.2. Item parameter estimates and ICFs

The item parameter estimates are shown in Table 3, and Figure 8 displays the ICFs alongside individual item responses. Figure 8 demonstrates that the E2PL model provides a good fit for the observed item

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

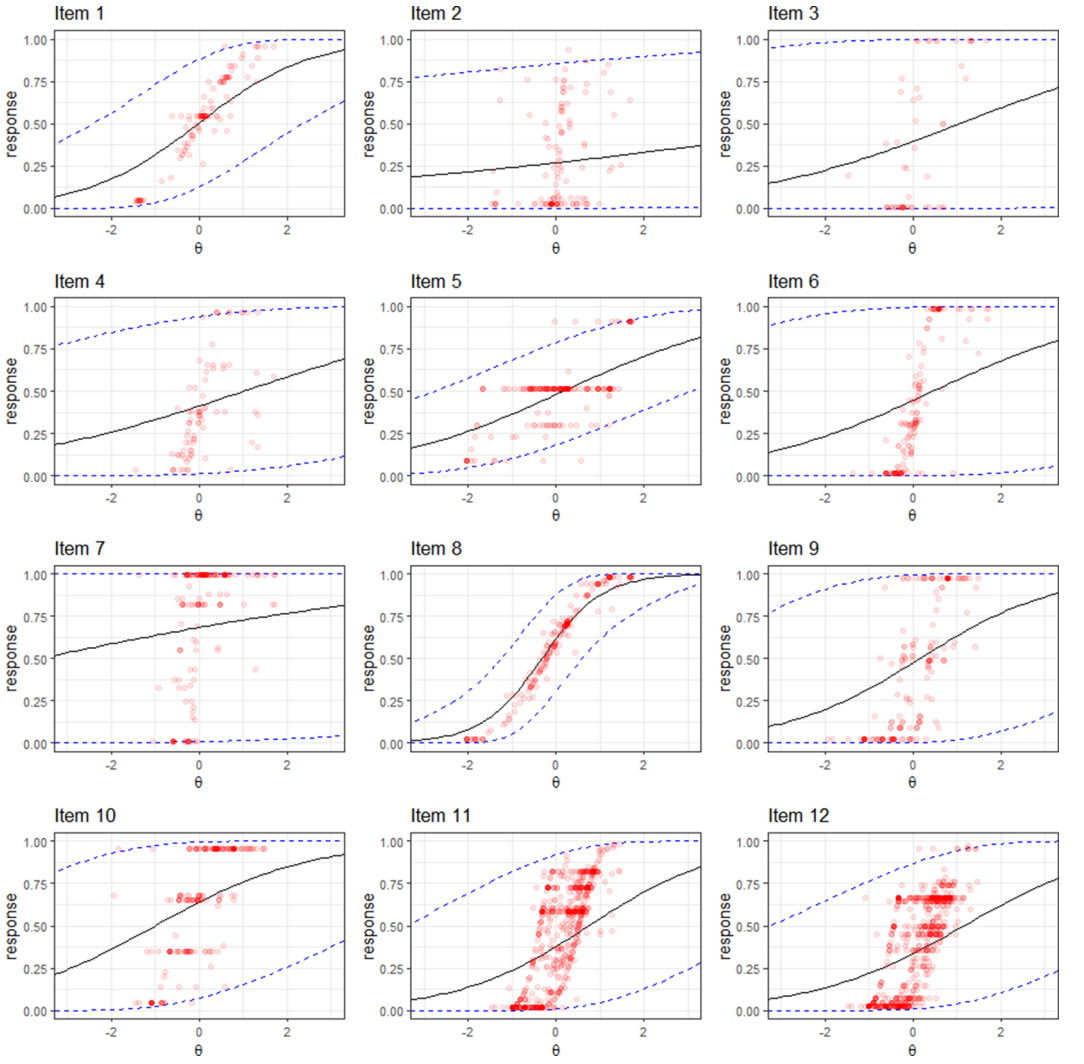


Figure 8. ICFs and item responses of the programming assessment dataset.
Note: The **black lines** are the expected response $\hat{\mu}$, the blue dashed lines indicate the 95% confidence interval, and the red dots are the observed responses.

responses. Notably, when the precision parameter $\hat{\nu}$ is relatively high (e.g., Item 8), responses are tightly clustered around the $\hat{\mu}$ line. Conversely, for items with lower precision parameters (e.g., Items 3, 6, and 7), responses conditional on θ are more dispersed across the response range. For instance, in the interval $-0.5 < \theta < 0.5$ for Item 7, responses span nearly the entire range from 0 to 1.

Given that the estimated $\hat{\nu}$ values for Items 3, 6, and 7 are near or below 1, these items could potentially be simplified into dichotomous items to facilitate scoring. However, any such modifications should not be based solely on the ν parameter, as scoring decisions are also influenced by the overall test design and item format.

5.3. Item information functions

Referring to Equation (8), the magnitude of the item information functions in the E2PL model depends on both a and ν. For illustrative purposes, the item information functions of Items 1, 3, 5, and 12 are

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

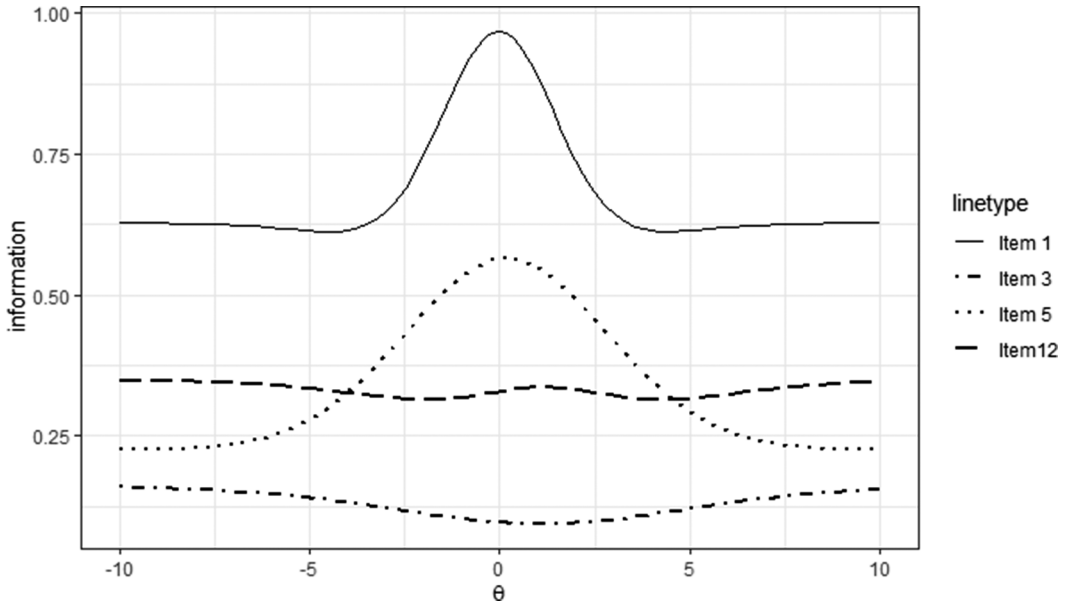


Figure 9. Item information functions of the programming assessment dataset.
Note: Items 1, 3, 5, and 12 are selected for illustrative purposes.

shown in Figure 9. The functions are drawn on a wide range of θ-axis to capture the overall shapes. The information functions for Items 1 and 5 are bell-shaped due to their large $\hat{\nu}$ values. Each function reaches its peak at the corresponding estimated item difficulty parameter $\hat{b}$ (Item 1 at $\theta = -0.03$ and Item 5 at $\theta = 0.17$).

In contrast, due to their small $\hat{\nu}$ values, the item information functions for Items 3 and 12 exhibit local maxima at their respective estimated difficulty parameters $\hat{b}$ (Item 3 at $\theta = 1.07$ and Item 12 at $\theta = 1.17$). As θ approaches $\pm\infty$, all of the information functions asymptotically approach $\hat{a}^2$ from below.

5.4. Discussion

The E2PL model enables the analysis of continuous item response data without requiring discretization, which allows for the extraction of more information compared to discrete data. Specifically, the 2PL has a trade-off between the coverage of the latent space and the amount of item information (see Equation (2)). A high discrimination parameter results in a concentrated area of high item information (a sharp peak in the item information function that quickly diminishes), whereas a low discrimination parameter spreads item information more evenly across a broader latent range, though with a lower overall peak.

In contrast, the E2PL delivers high item information while covering a larger portion of the latent space. For instance, Item 5, with a precision parameter estimate of $\hat{\nu} = 8.87$, maintains high item information, and its relatively low discrimination parameter ($\hat{a} = 0.48$) allows the item information to span almost the entire latent continuum of interest (i.e., $-3 < \theta < 3$).

Furthermore, as an extension of the 2PL, the parameter interpretation of the E2PL remains straightforward. The discrimination and difficulty parameters share the same mathematical interpretation as in the 2PL. The newly introduced precision parameter represents the concentration of item responses around the *S*-shaped mean curve, independent of the latent θ scale. In addition, the scale transformation for the 2PL illustrated in Section 2.1.1 is directly applicable to the E2PL parameter estimates presented in this section, maintaining consistency with the 2PL framework.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

Lastly, the communality indices presented in Table 3 reflect the proportion of uncertainty in each item, represented by $1/\nu$, accounted for by the latent variable. The relatively low communality values for some items (e.g., Items 4 and 5) may indicate potential model misfit, such as multidimensionality in the latent construct or limitations in the design of items intended to assess programming skills.

6. Empirical Illustration 2: Sparse polytomous data

With the growing interest in online learning and increased accessibility to digital devices, more psychometric data are being collected through online platforms. At the same time, controlling item exposure remains a critical concern in the development of computer-based assessments. As a result, adaptive testing is gaining more attention, generally requiring large quantities of items (e.g., Chen et al., 2023; Yan et al., 2016). In such testing situations, item parameters are generally treated as fixed (i.e., pre-calibrated) based on pilot tests. However, during these pilot tests, sparse item response data are often collected, particularly when items are scored polytomously.

If an IRT model is used to analyze the sparse polytomous data, it may pose challenges to the parameter estimation (Davey & Pitoniak, 2011; O'Neill et al., 2020). Moreover, the problem would get worse if the items have many score categories, as the number of item parameters in polytomous IRT models tends to increase with the number of categories (Noel, 2014). In addition to continuous response data, analyses of sparse polytomous data can be enhanced by leveraging the E2PL, which may provide a robust framework for such situations.

6.1. Data

A sparse polytomous dataset has been retrieved from an assessment developed by Macat (Macat International, 2024) for assessing students' critical thinking skills. According to the assessment design, students are assigned different sets of 24 items. Every item has four statements that are scored as either correct or incorrect, with the total score for an item being the sum of correct responses ($x = 0, 1, \ldots, 4$). For this analysis, data from 3,502 participants and 94 items were utilized. Originally, there were 96 items, but two misbehaving items were excluded.

The item response matrix, with dimensions of (3502×94), contains 82,173 valid responses, which constitute approximately 25% of the total possible responses, with the remainder being missing data. This sparseness arises because each student responded to only 24 items. As a result, the number of responses per item varies, ranging from 593 to 1,210. Moreover, among the 470 score categories across the 94 items (94 items × 5 categories), responses were observed fewer than ten times for 35 categories, and fewer than 30 times for 112 categories. For Item 11, the fourth category ($x = 3$) had no observations.

Following the preprocessing procedure outlined in Section 3.6, the item responses were transformed to a unit interval. This transformation divides the unit interval into five equally spaced sections, each with a width of 0.2. The midpoints of these sections were assigned as the continuous item responses for each category: $\{0, 1, 2, 3, 4\} \mapsto \{0.1, 0.3, 0.5, 0.7, 0.9\}$. These converted responses were then treated as continuous item responses in the E2PL.

6.2. Comparison of E2PL and GPCM

6.2.1. Visual comparison

The comparison between the two models begins with a visual examination of their ICFs, as shown in Figure 10. For illustrative purposes, two items (Item 38 and Item 51) out of the 94 total items are selected. The figure displays the mean curve and 95% interval of the E2PL alongside the expected response curve of the generalized partial credit model (GPCM; Muraki, 1992). Additionally, individual item responses are plotted against the ICFs, providing a visual assessment of how each model represents item responses. Based on how the item responses align with the ICFs, the two models demonstrate comparable performance, which supports the use of the E2PL for sparse polytomous data.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

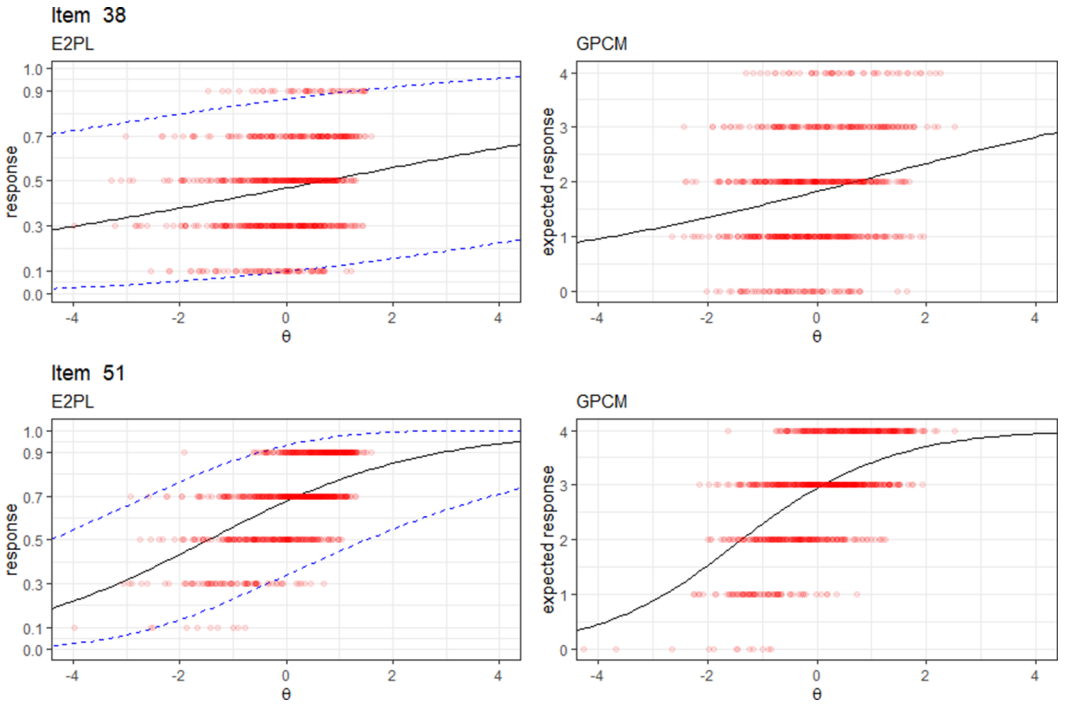


Figure 10. Visual comparison between the E2PL and GPCM for the selected items.
Note: Items 38 and 51 are selected for illustrative purposes.

6.2.2. *Quantitative comparison*

Next, the model-data fit of the E2PL is quantitatively compared to the GPCM using K-fold cross-validation. The unit of prediction for this comparison is the item response. For both models, raw scores ($x = 0, 1, \ldots, 4$) are predicted. The E2PL, which generates predictions on the unit interval, transforms these into the original scale ($x = 0, 1, \ldots, 4$), which is the inverse process of the one described in Figure 6. Given the size of the dataset, a 10-fold cross-validation approach is employed. On each iteration of the 10-fold cross-validation, RMSE is computed to assess the predictive performance of both models. RMSE is calculated on the raw score scale to avoid favoring either model.

RMSEs from the two models are listed in Table 4. RMSEs of the E2PL were consistently lower than those of the GPCM. Overall, the average RMSE of the E2PL was 0.92 and that of the GPCM was 1.00, indicating that the E2PL had a better predictive performance than the GPCM. Additionally, as a measure of the overall model-data fit, the pseudo R^2 (see Section 3.1.4) based on the E2PL is 0.393.

6.3. *Discussion*

It is notable from Figure 10 that the model-data fits of the two models appear visually comparable, making it difficult to judge the superiority of one model over the other based on visual inspection alone. This highlights the potential advantage of the E2PL in reducing the number of item parameters while maintaining an equivalent model-data fit, a conclusion supported by the 10-fold cross-validation results.

The cross-validation results imply that the better predictive performance of the E2PL may be due to the number of item parameters required by each model. The GPCM uses five item parameters per item, whereas the E2PL requires only three. This reduction in the number of parameters could have contributed to the E2PL's superior performance, as the GPCM includes 188 additional parameters (2 parameters ×94 items) in this example.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

Table 4. RMSEs from the 10-fold cross-validation

Iteration	RMSE	
(*k*)	E2PL	GPCM
1	0.93	1.02
2	0.91	0.98
3	0.93	1.00
4	0.92	1.00
5	0.92	1.00
6	0.92	0.99
7	0.91	0.98
8	0.92	0.98
9	0.92	1.01
10	0.93	1.01
Overall	0.92	1.00

The results support the application of the E2PL to sparse polytomous item response data. Although the results of this section may not be generalizable, the results show that the E2PL can be an alternative to polytomous IRT models when a parsimonious model is preferred, especially for sparse data. The rating scale model (Andrich, 1982) can be another option for model parsimony. However, the rating scale model assumes that its threshold parameters are equally spaced on the θ scale. For some situations, including the current example, the scale conversion of this section can be more justifiable than the rating scale model's assumption.

7. Conclusion

This article proposed an IRT model that is suitable for dealing with continuous and sparse polytomous item response data. While existing models can provide useful psychometric insight, they often fall short in terms of model assumptions, interpretations, and parameter estimation to be used in areas, such as CAT and test linking/equating. The model presented in this study extends the widely-used 2PL model (Birnbaum, 1968) by adding an error term that follows a shifted beta distribution. The formulation of the model is well aligned with the generalized latent variable model framework (McCullagh & Nelder, 1989; Skrondal & Rabe-Hesketh, 2004), incorporating the same mean structure as the 2PL through the logit link function, along with an additional precision parameter to model the dispersion structure. Consequently, researchers and practitioners familiar with IRT models can easily apply and interpret this extended model. The model's structural resemblance to the 2PL also makes the model preferable when a mixed-format test is constructed with continuous response items.

The mean–precision parameterization of the beta distribution allows the separation of the item discrimination and difficulty parameters from the precision parameter ν. This separation enables the discrimination and difficulty parameters to control the expected response μ, while the precision parameter controls response variability conditional on μ. Interestingly, the statistical dependence between μ and ν allows the model to yield communality and unique dispersion indices that reflect item information, with these indices remaining invariant under scale transformation. This characteristic parallels the practical distinction between FA and IRT, where item thresholds (i.e., item difficulty) play a more central role in IRT than in FA. Additionally, the beta distribution can model asymmetric or zero-one-inflated item response distributions.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

Unless the precision parameter ν is close to or below 3, the information function of the model is a symmetric, bell-shaped curve peaking at $\theta = b$. This feature is especially useful in CAT for administering items that provide maximum information. Meanwhile, when the item discrimination and difficulty parameters are shared between the 2PL and the E2PL, the E2PL offers greater item information than the 2PL. This suggests that continuous item responses can yield more information than binary responses, as demonstrated in Section 5 with items providing substantial information over a wide θ range.

However, because observed item responses are likely to be discrete in the strict sense (e.g., from 0% to 100% with the increment of 1%), the continuity assumption on item responses may overestimate item information to some extent. Future studies can investigate the effect of this assumption violation and apply continuity correction methods to appropriately adjust the amount of information. Additionally, especially in survey data collected using slider items, the impact of response biases toward extreme values can be more detrimental than with polytomous items.

Model parameters can be estimated using the MML-EM procedure (Bock & Aitkin, 1981), and the simulation study demonstrated robust parameter recovery and stable estimation even with sample sizes as small as 250. All MML-EM procedures converged successfully, and the computational time was within acceptable limits for practical applications.

The model's application to sparse polytomous data was further tested via 10-fold cross-validation on empirical data, where it outperformed the GPCM (Muraki, 1992) in predictive accuracy. The improvement in performance can be attributed to the model's fewer item parameters compared to the GPCM, which is advantageous when sample sizes are limited. However, the results from the empirical example may not guarantee a parallel effect when the model is applied to a different dataset. In addition, the application of the model to polytomous data entails a scale conversion of item responses. While the conversion can be justified in many cases, including the example of this article, it may not be justified for other cases. Nonetheless, the model proposed in this article can be a useful alternative to polytomous IRT models when observed responses per score categories are insufficient.

For greater parsimony, the model could be simplified by fixing the item discrimination parameter a across all items, thereby aligning the expected response μ with the ICF of the Rasch model (Rasch, 1960). This simplification retains the core properties of the E2PL while reducing the number of item parameters. Additionally, future work could extend the E2PL to multidimensional applications by adopting a vectorized model equation.

The parameter estimation software developed for this article is publicly available in `R` via the `IRTest_Cont` function of the `IRTest` package (Li, 2025; R Core Team, 2024). Also, further application studies of the model may motivate item writers to develop and use more flexible item types, including continuous response items, when those items are expected to enhance the validity and reliability of assessments.

Acknowledgements. There are no conflicts of interest or financial interests, including funding from commercial entities, that could be perceived as impacting the integrity or objectivity of this research.

Competing interests. The authors declare none.

References

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105–113.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis*. (3rd ed.) John Wiley & Sons.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, D. R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.
Cai, L. (2012). Latent variable modeling. *Shanghai Archives of Psychiatry*, *24*(2), 118–120.
Casabianca, J. M., Donoghue, J. R., Shin, H. J., Chao, S.-F., & Choi, I. (2023). Using linkage sets to improve connectedness in rater response model estimation. *Journal of Educational Measurement*, *60*(3), 428–454. https://doi.org/10.1111/jedm.12360
Chen, Q., Zheng, H., Fan, H., & Mo, L. (2023). Construction of a reading literacy test item bank for fourth graders based on item response theory. *Frontiers in Psychology*, *14*, 1–8.
Chen, Y., Silva Filho, T., Prudencio, R. B., Diethe, T., & Flach, P. (2019). β^3-IRT: A new item response model and its applications. In *The 22nd international conference on artificial intelligence and statistics* (pp. 1013–1021).
Davey, T., & Pitoniak, M. J. (2011). Designing computerized adaptive tests. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (pp. 557–588). Routledge.
Espinheira, P. L., da Silva, L. C. M., Silva, A., de Oliveira Silva, A., & Ospina, R. (2019). Model selection criteria on beta regression for machine learning. *Machine Learning and Knowledge Extraction*, *1*(1), 427–449. https://doi.org/10.3390/make1010026
Espinheira, P. L., Ferrari, S. L., & Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, *35*(4), 407–419. https://doi.org/10.1080/02664760701834931
Ferrando, P. J. (2001). A nonlinear congeneric model for continuous item responses. *British Journal of Mathematical & Statistical Psychology*, *54*(2), 293–313. https://doi.org/10.1348/000711001159573
Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.
García-Pérez, M. A. (2024). Are the steps on likert scales equidistant? Responses on visual analog scales allow estimating their distances. *Educational and Psychological Measurement*, *84*(1), 91–122. https://doi.org/10.1177/00131644231164316
Gerdes, A., Jeuring, J. T., & Heeren, B. J. (2010). Using strategies for assessment of programming exercises. In *SIGCSE 10: Proceedings of the 41st ACM technical symposium on computer science education* (pp. 441–445).
Gu, P. Y. (2018). Validation of an online questionnaire of vocabulary learning strategies for ESL learners. *Studies in Second Language Learning and Teaching*, *8*(2), 325–350. https://doi.org/10.14746/ssllt.2018.8.2.7
Jones, P., Smith, R. W., & Talley, D. (2011). Developing test forms for small-scale achievement testing systems. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (pp. 487–525). Routledge.
Kallinger, S., Scharm, H., Boecker, M., Forkmann, T., & Baumeister, H. (2019). Calibration of an item bank in 474 orthopedic patients using Rasch analysis for computer-adaptive assessment of anxiety. *Clinical Rehabilitation*, *33*(9), 1468–1478.
Kloft, M., Hartmann, R., Voss, A., & Heck, D. W. W. (2023). The Dirichlet dual response model: An item response model for continuous bounded interval responses. *Psychometrika*, *88*(3), 888–916. https://doi.org/10.1007/s11336-023-09924-7
Kolen, M. (2004). *Test equating, scaling, and linking*. Springer.
Li, S. (2025). IRTest: An R package for item response theory with estimation of latent distribution. *The R Journal*, *16*(4), 23–41. https://doi.org/10.32614/RJ-2024-033
Macat International. (2024). Critical thinking assessment. Macat International Ltd. Retrieved 2024-07-12, from https://www.macat.com/critical-thinking-assessments
MacCallum, R. C. (2009). Factor analysis. In R. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 123–147). SAGE Publications Ltd. https://doi.org/10.4135/9780857020994
Maiorana, F., Giordano, D., & Morelli, R. (2015). Quizly: A live coding assessment platform for app inventor. In *2015 IEEE blocks and beyond workshop (blocks and beyond)* (pp. 25–30). https://doi.org/10.1109/BLOCKS.2015.7368995
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. (2nd ed.) Chapman and Hall.
Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223–236. https://doi.org/10.1207/s15327906mbr2903_2
Mitchell, S., Kallen, M. A., Troost, J. P., Bragg, A., Martin-Howard, J., Moldovan, I., & Carlozzi, N. E. (2023). Development and calibration data for the illness burden item bank: A new computer adaptive test for persons with type 2 diabetes mellitus. *Quality of Life Research*, *32*(3), 797–811.
Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*(2), 165–181. https://doi.org/10.1007/BF02294232
Molenaar, D., Cúri, M., & Bazán, J. L. (2022). Zero and one inflated item response theory models for bounded continuous data. *Journal of Educational and Behavioral Statistics*, *47*(6), 693–735. https://doi.org/10.3102/10769986221108455
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206
Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika*, *79*(4), 647–674.
Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*(1), 47–73. https://doi.org/10.1177/0146621605287691
O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of sample size on common item equating using the dichotomous Rasch model. *Applied Measurement in Education*, *33*(1), 10–23.
Open-Source Psychometrics Project. (2020). Data from the statistical "which character" personality quiz. Retrieved 2024-08-04, from https://openpsychometrics.org/tests/characters/data/

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Retrieved from https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97.

Samejima, F. (1973). Homogeneous case of continuous response model. *Psychometrika*, *38*(2), 203–219. https://doi.org/10.1007/BF02291114

Seo, B., & Cho, S. H. (2018). Design and implementation of students' coding assessment system for a coding puzzle game. *Journal of Korea Game Society*, *18*(1), 7–18.

Shin, H. J., Li, S., Ryoo, J. H., & von Davier, A. (2024). Harnessing artificial intelligence for generating items in critical thinking tests. In *Annual meeting of the national council on measurement in education (NCME)*. (Paper presented at the conference. This reference reflects the final order of authorship which is different from the program)

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71.

Toepoel, V., & Funke, F. (2018). Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies*, *25*(2), 112–122. https://doi.org/10.1080/08898480.2018.1439245

Vall-Llosera, L., Linares-Mustarós, S., Bikfalvi, A., & Coenders, G. (2020). A comparative assessment of graphic and 0–10 rating scales used to measure entrepreneurial competences. *Axioms*, *9*(1), 21.

van der Linden, W. J. (2016). *Handbook of item response theory*. CRC Press.

van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing* (Vol. 10). Springer.

Verhelst, N. D. (2019). Exponential family models for continuous responses. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 135–160). Springer Nature. https://doi.org/10.1007/978-3-030-18480-3_7

von Davier, A. A., Runge, A., Park, Y., Attali, Y., Church, J., & LaFlair, G. (2024). The item factory: Intelligent automation in support of test development at scale. In H. Jiao, & R. W. Lissitz (Eds.), *Machine learning, natural language processing, and psychometrics* (pp. 1–25). Information Age Publishing.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

Wang, T., & Zeng, L. (1998). Item parameter estimation for a continuous response model using an EM algorithm. *Applied Psychological Measurement*, *22*(4), 333–344.

Yan, D., von Davier, A. A., & Lewis, C. (2016). *Computerized multistage testing: Theory and applications*. CRC Press.

A. Appendix

A.1. Item parameter estimation

In the MML-EM procedures, the iteration of the expectation step (E-step) and the maximization step (M-step) can be terminated when changes in parameter estimates drop below a predefined threshold.

A.1.1. Marginal log-likelihood

Let $\log \mathcal{L}$ denote the marginal log-likelihood, $g(\theta)$ be the latent distribution, and θ_q^* be the qth quadrature point within the quadrature scheme of the MML-EM procedure ($q = 1, 2, \ldots, Q$). Then, the marginal log-likelihood can be expressed as follows:

$$\begin{aligned}\log \mathcal{L} &= \log\left(\prod_{j=1}^{N}\int_{\theta} g(\theta)\prod_{i_j\in I_j} f\left(x_{ji_j}|\theta, a_{i_j}, b_{i_j}, v_{i_j}\right)d\theta\right)\\ &= \sum_{j=1}^{N}\log\left(\int_{\theta} g(\theta)\prod_{i_j\in I_j} f\left(x_{ji_j}|\theta, a_{i_j}, b_{i_j}, v_{i_j}\right)d\theta\right)\\ &\approx \sum_{j=1}^{N}\log\left(\sum_{q=1}^{Q} g\left(\theta_q^*\right)\prod_{i_j\in I_j} f\left(x_{ji_j}|\theta_q^*, a_{i_j}, b_{i_j}, v_{i_j}\right)\right),\end{aligned} \tag{A.1}$$

where the integration is approximated by the summation and $g\left(\theta_q^*\right)$ is a normalized density after discretization.

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

A.1.2. E-step of the MML-EM procedure

In the E-step of the MML-EM procedure, expected values are calculated. The quantity γ_{jq} implies the expected probability of the jth examinee's ability parameter belonging to the qth grid in the latent space.

$$\gamma_{jq} = \frac{g\left(\theta_q^*\right)\prod_{i_j \in I_j} f\left(x_{ji_j}|\theta_q^*, a_{i_j}, b_{i_j}, v_{i_j}\right)}{\sum_{q=1}^{Q} g\left(\theta_q^*\right)\prod_{i_j \in I_j} f\left(x_{ji_j}|\theta_q^*, a_{i_j}, b_{i_j}, v_{i_j}\right)}. \tag{A.2}$$

In addition, $f_q = \sum_{j=1}^{N} \gamma_{jq}$ indicates the expected population frequency at the qth grid.

A.1.3. M-step of the MML-EM procedure

In the M-step, the marginal log-likelihood becomes more tractable using γ_{jq} and Jensen's inequality:

$$\begin{aligned}
\log \mathcal{L} &\approx \sum_{j=1}^{N} \log\left(\sum_{q=1}^{Q} g\left(\theta_q^*\right)\prod_{i_j \in I_j} f\left(x_{ji_j}|\theta_q^*, a_{i_j}, b_{i_j}, v_{i_j}\right)\right) \\
&= \sum_{j=1}^{N}\sum_{q=1}^{Q} \gamma_{jq} \log\left(g\left(\theta_q^*\right)\prod_{i_j \in I_j} f\left(x_{ji_j}|\theta_q^*, a_{i_j}, b_{i_j}, v_{i_j}\right)\right) - \sum_{j=1}^{N}\sum_{q=1}^{Q} \gamma_{jq}\log\gamma_{jq} \\
&= \sum_{j=1}^{N}\sum_{q=1}^{Q} \gamma_{jq} \sum_{i_j \in I_j} \log f\left(x_{ji_j}|\theta_q^*, a_{i_j}, b_{i_j}, v_{i_j}\right) + \sum_{j=1}^{N}\sum_{q=1}^{Q} \gamma_{jq}\log g\left(\theta_q^*\right) - \sum_{j=1}^{N}\sum_{q=1}^{Q} \gamma_{jq}\log\gamma_{jq}.
\end{aligned} \tag{A.3}$$

Since only the first term of the equation's last line is dependent on the item parameters, the item parameters can be estimated by maximizing it. Moreover, by the local independence assumption, the item parameters can be estimated separately for each item.

Newton–Raphson Method The Newton–Raphson method can be used to find item parameter estimates that maximize the marginal log-likelihood of the M-step. Instead of v, $\xi = \log v$ is estimated for a more numerically feasible estimation, as v can take only positive values. For notational brevity, item indices are omitted and additional quantities are introduced:

$$\begin{aligned}
\omega_{1q} &= \sum_{j=1}^{N} \gamma_{jq} \log x_j, \\
\omega_{2q} &= \sum_{j=1}^{N} \gamma_{jq} \log\left(1 - x_j\right),
\end{aligned} \tag{A.4}$$

and

$$\mu_q = \frac{e^{a\left(\theta_q^* - b\right)}}{1 + e^{a\left(\theta_q^* - b\right)}}. \tag{A.5}$$

In addition, as omitted responses are not counted by the index j, the total number of test takers may vary across items.

Then, the first derivatives of the item parameters are as follows:

$$\frac{\partial \log \mathcal{L}}{\partial a} = \sum_{q=1}^{Q} (\theta_q^* - b) v\mu_q(1-\mu_q)\left[\omega_{1q} - \omega_{2q} - f_q\left(\psi(v\mu_q) - \psi(v(1-\mu_q))\right)\right], \tag{A.6}$$

$$\frac{\partial \log \mathcal{L}}{\partial b} = -a\sum_{q=1}^{Q} v\mu_q(1-\mu_q)\left[\omega_{1q} - \omega_{2q} - f_q\left(\psi(v\mu_q) - \psi(v(1-\mu_q))\right)\right], \tag{A.7}$$

and

$$\frac{\partial \log \mathcal{L}}{\partial \xi} = vN\psi(v)\sum_{q=1}^{Q}\left[\mu_q\left(\omega_{1q} - f_q\psi(v\mu_q)\right) + (1-\mu_q)\left(\omega_{2q} - f_q\psi(v(1-\mu_q))\right)\right], \tag{A.8}$$

where $\psi(\cdot)$ is a digamma function.

The expectations of the second derivatives are as follows:

$$E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a^2}\right] = -\sum_{q=1}^{Q}\left((\theta_q^* - b)v\mu_q(1-\mu_q)\right)^2 f_q\left(\psi_1(v\mu_q) + \psi_1(v(1-\mu_q))\right), \tag{A.9}$$

$$E\left[\frac{\partial^2 \log \mathcal{L}}{\partial b^2}\right] = -a^2\sum_{q=1}^{Q}\left(v\mu_q(1-\mu_q)\right)^2 f_q\left(\psi_1(v\mu_q) + \psi_1(v(1-\mu_q))\right), \tag{A.10}$$

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

$$E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \xi^2}\right] = \nu^2 N \psi_1(\nu) - \nu^2 \sum_{q=1}^{Q} f_q \left[\mu_q^2 \psi_1(\nu\mu_q) + (1-\mu_q)^2 \psi_1(\nu(1-\mu_q))\right], \tag{A.11}$$

$$E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a \partial b}\right] = a \sum_{q=1}^{Q} (\theta_q^* - b)\left(\nu\mu_q(1-\mu_q)\right)^2 f_q \left(\psi_1(\nu\mu_q) + \psi_1(\nu(1-\mu_q))\right), \tag{A.12}$$

$$E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a \partial \xi}\right] = -\nu^2 \sum_{q=1}^{Q} (\theta_q^* - b)\mu_q(1-\mu_q) f_q \left[\mu_q \psi_1(\nu\mu_q) - (1-\mu_q)\psi_1(\nu(1-\mu_q))\right], \tag{A.13}$$

$$E\left[\frac{\partial^2 \log \mathcal{L}}{\partial b \partial \xi}\right] = a\nu^2 \sum_{q=1}^{Q} \mu_q(1-\mu_q) f_q \left[\mu_q \psi_1(\nu\mu_q) - (1-\mu_q)\psi_1(\nu(1-\mu_q))\right], \tag{A.14}$$

where $\psi_1(\cdot)$ is a trigamma function.

On the kth iteration, the parameters are updated using the following equation:

$$\begin{bmatrix} a_{(k+1)} \\ b_{(k+1)} \\ \xi_{(k+1)} \end{bmatrix} = \begin{bmatrix} a_k \\ b_k \\ \xi_k \end{bmatrix} - \begin{bmatrix} E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a_k^2}\right] & E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a_k \partial b_k}\right] & E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a_k \partial xi_k}\right] \\ E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a_k \partial b_k}\right] & E\left[\frac{\partial^2 \log \mathcal{L}}{\partial b_k^2}\right] & E\left[\frac{\partial^2 \log \mathcal{L}}{\partial b_k \partial \xi_k}\right] \\ E\left[\frac{\partial^2 \log \mathcal{L}}{\partial a_k \partial \xi_k}\right] & E\left[\frac{\partial^2 \log \mathcal{L}}{\partial b_k \partial \xi_k}\right] & E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \xi_k^2}\right] \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \log \mathcal{L}}{\partial a_k} \\ \frac{\partial \log \mathcal{L}}{\partial b_k} \\ \frac{\partial \log \mathcal{L}}{\partial \xi_k} \end{bmatrix}. \tag{A.15}$$

The Newton–Raphson iteration is terminated when changes in parameter estimates drop below a predefined threshold.

A.2. Ability parameter estimation

Assuming that test takers are independent, their ability parameter estimates can be estimated individually. Thus, the index j for the test takers is dropped for notational brevity and $i = 1, 2, \ldots, I$ indicates items. For the MLE, the likelihood to be maximized is simply the product of individual response probabilities:

$$\log \mathcal{L} = \sum_{i=1}^{I} \log f(x_i | \theta, a_i, b_i, \nu_i). \tag{A.16}$$

For the WLE, the likelihood is the sum of $\log \mathcal{L}$ and a specific function to obtain unbiased estimate. The first derivative of the WLE likelihood will be directly provided.

A.2.1. EAP

The γ_q can be regarded as a posterior distribution of the ability parameter. Therefore, EAP scores can be calculated as the posterior mean of the latent distribution.

$$EAP = \sum_{q=1}^{Q} \theta_q^* \gamma_q. \tag{A.17}$$

A.2.2. MLE & WLE

Again, the Newton–Raphson method is introduced to calculate MLE and WLE. The Newton–Raphson procedure is terminated when changes in parameter estimates drop below a predefined threshold.

MLE The first and second derivatives are as follows:

$$\left.\frac{\partial \log \mathcal{L}}{\partial \theta}\right|_{MLE} = \sum_{i=1}^{I} a_i \nu_i \mu_i (1-\mu_i) \left[\log \frac{x_i}{1-x_i} - \psi(\nu_i \mu_i) + \psi(\nu_i(1-\mu_i))\right] \tag{A.18}$$

and

$$I = -E\left[\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2}\right] = \sum_{i=1}^{I} \left(a_i \nu_i \mu_i (1-\mu_i)\right)^2 \left(\psi_1(\nu_i \mu_i) + \psi_1(\nu_i(1-\mu_i))\right). \tag{A.19}$$

On the kth iteration, MLE is updated using the following equation:

$$MLE_{(k+1)} = MLE_k + \frac{\left.\frac{\partial \log \mathcal{L}}{\partial \theta_k}\right|_{MLE}}{I}. \tag{A.20}$$

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press

WLE While the second derivative is the same as in MLE, a weight is added to the first derivative of MLE:

$$\left.\frac{\partial \log \mathcal{L}}{\partial \theta}\right|_{WLE} = \left.\frac{\partial \log \mathcal{L}}{\partial \theta}\right|_{MLE} + \frac{J}{2I}, \tag{A.21}$$

and

$$J = -\sum_{i=1}^{I} a_i^3 v_i^2 \mu_i^2 (1-\mu_i)^2 (1-2\mu_i) \left[\psi_1(\mu_i v_i) + \psi_1(v_i(1-\mu_i))\right]. \tag{A.22}$$

Likewise, on the kth iteration, WLE is updated using the following equation:

$$WLE_{(k+1)} = WLE_k + \frac{\left.\frac{\partial \log \mathcal{L}}{\partial \theta_k}\right|_{WLE}}{I}. \tag{A.23}$$

Cite this article: Li, S. and Shin, H. Jeong (2025). An Extended Two-Parameter Logistic Item Response Model to Handle Continuous Responses and Sparse Polytomous Responses. *Psychometrika*, 1705–1731. https://doi.org/10.1017/psy.2025.10044

https://doi.org/10.1017/psy.2025.10044 Published online by Cambridge University Press