Psychometrika (2026), **91**, 312–334
doi:10.1017/psy.2025.10019

APPLICATION AND CASE STUDIES - ORIGINAL

Bayesian Joint Modeling of Response Times with Dynamic Latent Ability in Educational Testing

Xiaojing Wang[1], Abhisek Saha[2] and Dipak K. Dey[1]

[1]Department of Statistics, University of Connecticut, Storrs, United States; [2]Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, Massachusetts, United States

Corresponding author: Xiaojing Wang; Email: xiaojing.wang@uconn.edu

(Received 16 October 2024; revised 16 October 2024; accepted 9 January 2025)

Abstract

In educational testing, inferences of ability have been mainly based on item responses, while the time taken to complete an item is often ignored. To better infer the ability, a new class of state space models, which conjointly model response time with time series of dichotomous responses, is developed. Simulations for the proposed models demonstrate that the biases of ability estimation are reduced as well as the precisions of ability estimation are improved. An empirical study is conducted using EdSphere datasets, where the two competing relationships (i.e., monotone and inverted U-shape) for the distance between ability and difficulty are investigated in modeling response times. The results of model comparison support that the inverted U-shape relationship better captures the behaviors and psychology of examinees in exams for EdSphere datasets.

Keywords: computerized testing; dynamic item response models; local dependence; Markov chain Monte Carlo (MCMC); response times

1. Introduction

Item response theory (IRT) models, also known as latent trait (analysis) models, have been widely used in measurement testing for several decades. They originated from analyzing dichotomous items (Lord, 1953; Rasch, 1961), soon extended to modeling polychotomous items (Darrell Bock, 1972; Samejima, 1970). The applications of IRT models became diverse from education and psychology to political science, clinical and health studies, marketing, and so on. One of the most famous IRT models is the Rasch model (Rasch, 1961), belonging to one-parameter IRT models, which is typically specified as

$$\Pr(X_{i,l} = 1 \mid \theta_i, d_l) = \mathrm{F}(\theta_i - d_l), \tag{1.1}$$

where the subscript (i,l) is used to index i-th person and l-th item, $X_{i,l}$ then represents the correctness of the answer (1 if correct, 0 otherwise), θ_i is one's ability, d_l denotes the item difficulty, and $\mathrm{F}(x)$ is a cumulative distribution function (cdf) for continuous random variables and $\mathrm{F}^{-1}(\cdot)$ is called the link function. For the Rasch model, the link function is chosen to be logistic. Often, IRT models are based on the *local independence* assumption, which means, conditionally on θ_i, d_l, (as in (1.1)), the item response variables $X_{i,l}$'s are independent.

© 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 in any medium, provided the original work is properly cited.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

1.1. EdSphere testbed

Our study is motivated by the EdSphere dataset provided by High Road Learning Company. EdSphere is a personalized literacy learning platform that continuously collects data about student performance and strategic behaviors each time when he/she reads an article. The data were generated during each session when a student read an article selected from a large bank of articles. A session begins like this: a student selects from a generated list of articles having text complexities in a range targeted to his/her current ability estimate. The text complexity is measured in other platforms (cf. Swartz et al. (2016) and Stenner (2022) for more details). Once the article is chosen, the computer, following a prescribed protocol, randomly selects a sample of the eligible words to be "clozed," that is, to be removed and replaced by blanks, and presents the article to the student with these words clozed. When a blank is encountered while reading the article, the student clicks it and then the true removed word along with three incorrect options called foils are presented. As with the target word, the foils are selected randomly according to a prescribed protocol. The student has to select a word to fill in the blank from the four choices before he/she can move to the next question and an immediate feedback is provided in the form of the correct answer. The dichotomous items produced by this procedure are called "Auto-Generated-Cloze" items, which are *randomized items*. This type of item implies that even if two students select the same article to read, the sets of target words and foils will be different. As a consequence, it is not feasible to obtain data-based estimates of item parameters.

The EdSphere dataset consists of 16,949 students who registered over five years in the EdSphere learning platform at a school district in Mississippi. The students were in different grades and could enter and leave the program at different times between 2007 and 2011. They can take tests on different days and have different time lapses between tests, which indicates the responses observed are *longitudinal* at individually-varying and irregularly-spaced time points. Thus, a dynamic structure to modeling changes of latent traits is needed. In addition, as mentioned in Wang et al. (2013), in the environment of EdSphere, the factors, such as test random effects (e.g., an overall comprehension of the article) and daily random effects (such as the person's emotional status and other factors), might undermine the *local independence* assumption of classic IRT models.

To summarize, the EdSphere dataset has several distinctive features that often do not appear in a classic paper-and-pencil test, i.e., *randomized items*, *longitudinal observations*, and *local dependence*, making the use of classic IRT models face great challenges.

1.2. Recent developments of IRT models

To address these challenges in the advent of the modern computerized (adaptive) testing, there have been many developments of IRT models in the literature. For the generalization of IRT models to *longitudinal data*, some researchers used parametric function of time to model the changes of latent traits (Albers et al., 1989; Johnson & Raudenbush, 2006; Verhagen & Fox, 2013) and Wang & Nydick (2020) unified various longitudinal IRT models in terms of expressions using the function of time. Recently, Liu & Wang (2020) further developed a flexible nonparametric function of the time to model the latent trajectory. On the other hand, some researchers applied a Markov chain model to describe the time dependence of latent traits (Martin & Quinn, 2002; Park, 2011). To take account of *local dependence* issues, there have been parallel developments for the procedures of detecting it (Chen & Thissen, 1997; Christensen et al., 2017; Liu & Maydeu-Olivares, 2013; Yen, 1984) and ways of modeling it (Bradlow et al., 1999; Cai, 2010; Jannarone, 1986; Olsbjerg & Christensen, 2015). To deal with *randomized items*, often introduce random effects for item parameters (De Boeck, 2008; Sinharay et al., 2003).

However, the current literature that focuses on three challenges simultaneously is very limited. Wang et al. (2013) developed a new class of state space models, called dynamic item response (DIR) models, to address the three challenges within one unified framework. In this regard, their work is pioneering but they ignored the usage of the response time information, which is often easily obtained during computerized tests. When response accuracy and response time are both available in a test, Thissen (1983) argued that the separate analysis of them might yield misleading inferences. Their points are

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

further guarded by other researchers (Ferrando & Lorenzo-Seva, 2007; Ranger & Kuhn, 2012; Van der Linden et al., 2010; Wang et al., 2019), who showed that using response times as auxiliary information has improved the estimation of certain parameters in IRT models. In this article, we have considered the response times with the joint analysis of the response accuracy in a computerized (adaptive) testing. From our simulation and real data analysis, we also demonstrate that there are improvements in terms of accuracy and bias for the estimates of ability in the longitudinal study. This is a pretty important point because the algorithm for running a computerized (adaptive) testing is often focused on matching the difficulty of the test material with the current ability of an examinee. Hence, with more accurate estimates of one's ability, the practitioners and educators can obtain better designs in the computerized (adaptive) testing. In addition, teachers can better assign the learning materials for students to study according to their respective capacities.

1.3. Recent developments for modeling response times in educational testing

To model the response time of an item, one way is to treat it as a causal factor for the accuracy of that item (Roskam, 1997; Wang & Hanson, 2005). Another idea is to regard the response accuracy as a causal factor for the response time (Gaviria, 2005). However, both ideas have been criticized since the response time and accuracy of a test may not be directly related. Instead, the third way is to jointly model response times and item responses in a hierarchical fashion.

There are two popular classes of joint modeling. One popular choice is Thissen's (1983) model, i.e., taking the natural logarithm of response times and modeling as

$$\log R_{i,l} = \mu - \nu_i + \tau_l + \beta L(\theta_i - d_l) + \zeta_{i,l}, \tag{1.2}$$

where $R_{i,l}$ indicates the time used for answering the lth question by the ith person, ν_i is the speediness parameter, which takes account of the time that a person spends for infinitely easy set of problems, τ_l is the slowness intensity of a question, which dictates the time taken due to the nature of the problem, μ is the overall mean, $\zeta_{i,l}$ is the residual, β is a slope, and $L(x)$ denotes a linear function mapping how the distance of ability and item difficulty connects with response times.

For Equation (1.2), there are two popular choices for $L(x)$, one is a monotone mapping (Gaviria, 2005; Thissen, 1983), reflecting the idea that the larger the distance between one's ability and item difficulty is, the more time one spends to finish a question; the other one is an inverted U-shaped mapping, originating from the findings (Wang, 2006; Wang & Zhang, 2006) that examinees generally spend more time on items that match their ability levels, while spend less time on items either too easy or too hard. Besides, there are several papers using the inverted U-shape for regressing response times in the analysis of personality and psychology tests (Ferrando & Lorenzo-Seva, 2007; Molenaar et al., 2021; Ranger & Kuhn, 2012). Intuitively, the negative β in front of $L(x)$ for either monotone or inverted U-shaped mapping makes more sense in reality.

Another popular choice of joint models utilizes a hierarchical framework to conjointly model response times and accuracy but without specifying any explicit relationship between them (Klein Entink, 2009; Loeys et al., 2011; Van der Linden, 2007). Instead, they assigned joint multivariate normal priors to link different parameters in the joint models. Recently, this type of joint model has been extended to conjointly modeling with omission behavior (Ulitzsch et al., 2020) or paper-and-pencil tests (Liu et al., 2022); to consider varying speed across dimensions of ability in the model (Zhan et al., 2021); to incorporate a covariance structure to explain the local dependency between speed and accuracy (Meng et al., 2015); and to adopt a hidden Markov structure to separate between-subjects ability and speed variables from the within-subjects states variables (Molenaar et al., 2016).

However, all existing joint models are centered on a one-time exam for test takers without considering the time frequency feature of computerized testing. In this article, we aim to fill in this gap. Motivated by EdSphere datasets, we propose a joint model of response times with response accuracy for testing data collected at irregular and individual varying time points. Of course, our proposed joint model can

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

easily be modified for use in the analysis of longitudinal data with a simpler structure than we discuss here.

Furthermore, we have proposed two empirical methods to test which linkage function $L(x)$ in the joint models is more helpful in providing additional information to estimate ability. First, we propose Lindley's method (Lindley, 1965) to check whether the regression slope β in front of $L(x)$ is significant or not. Second, we develop a model selection method, called the partial DIC criterion, to compare different $L(x)$'s in the joint model for modeling the relationships between the response time and the ability–difficulty distance.

1.4. Preview

In Section 2, we put forward a new class of joint models for IRT models with response times, called DIR and response time models (DIR-RT models). Because of the complexity of the model considered, Bayesian inference and Markov chain Monte Carlo (MCMC) computational techniques will be presented in Section 3. In Section 4, we validate the proposed Bayesian inference procedure via some simulations and compare the performance of DIR-RT models with respect to DIR models. We illustrate the application of DIR-RT models to EdSphere testbed datasets in Section 5, where we further provide an empirical justification of goodness of fit among DIR-RT models with different choices of $L(x)$ in order to examine the better linkage to jointly model response times with IRT parameters. In Section 6, we point out some significant psychological implications from the analysis of the EdSphere dataset and discuss future research directions.

2. Joint models of dynamic item responses and response times

Clearly, modeling (1.1) and (1.2) jointly can maximize the information on inferring one's ability θ_i and the item difficulty d_l. Thus, we propose a *two-level* joint model. The first level has two sub-models to concurrently model the observations of response time and response accuracy with certain shared parameters, and the second level introduces a dynamic model to capture changes of latent traits over time. Although our investigation begins with an extension of the one-parameter IRT, it would be straightforward to generalize our proposed models to a two-parameter or three-parameter IRT model.

2.1. First level: The observation equations in DIR-RT models

Equations (1.1) and (1.2) are based on a one-time exam for each test taker. To consider a much complex data structure in the computerized test (as in the EdSphere testbed), we first expand the labels of notations. Let $X_{i,t,s,l}$ be the item response to indicate the correctness for the answer of the lth item in the sth test on the tth day given by the ith person, where $i = 1, \ldots, n$ (number of subjects); $t = 1, \ldots, T_i$ (number of test dates); $s = 1, \ldots, S_{i,t}$ (number of tests in a day); and $l = 1, \ldots, K_{i,t,s}$ (number of items in a test). Likewise, denote the difficulty of the lth item as $d_{i,t,s,l}$. It is ideal to record the time for each examinee spending on a single item, however, in practice, more often the time spent on the entire exam is merely stored for each individual (a case for reading comprehension tests in EdSphere datasets). Thus, in our proposed models, the response time is defined at the test level, i.e., $R_{i,t,s}$, implying the time spent on the sth test for the ith individual on the tth day; whereas, our models can be easily revised to cope with the response time stored for each item whenever such data are available.

The label extension illustrates two major features of computerized (adaptive) testing: 1) there are few replications of items among different time, tests, and test takers and 2) the observations are recorded at individually-varying and irregularly-spaced time points. In our study, only $X_{i,t,s,l}$'s and $R_{i,t,s}$'s are observed. Usually, the response time is naturally bounded away from zero, and a logarithmic transformation of $R_{i,t,s}$ is taken to remove its skewness (Fox & Marianti, 2016; Liu et al., 2022; Van der Linden et al., 2010).

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

2.1.1. The observation equations of item responses

Often in a design of computerized tests, item difficulty, i.e., $d_{i,t,s,l}$, is a randomized parameter, assumed to be randomly selected from a bank of items with a certain ensemble mean. Thus, we model $d_{i,t,s,l} = a_{i,t,s} + \varepsilon_{i,t,s,l}$ as a measurement error model with $a_{i,t,s}$ being an ensemble mean difficulty of items in the sth test, and $\varepsilon_{i,t,s,l} \sim \mathcal{N}(0, \sigma^2)$ with σ^2 known according to the test design. Here, $\mathcal{N}(\cdot,\cdot)$ denotes a normal distribution. Similarly, as Wang et al. (2013) did, we extend classic one-parameter IRT models as

$$\Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, \varphi_{i,t}, \eta_{i,t,s}, a_{i,t,s}) = \mathrm{F}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \varepsilon_{i,t,s,l}), \tag{2.1}$$

where $\theta_{i,t}$ represents the ith person's ability on day t, with the assumption that one's ability is constant over a given day; $\varphi_{i,t}$ and $\eta_{i,t,s}$ take account of daily and test random effects, respectively, to explain the possible local dependence of item responses. Further, assume $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$ with its precision unknown and different for each person. To make $\eta_{i,t,s}$ well separate from $\varphi_{i,t}$ in inference, we denote $\eta_{i,t} = (\eta_{i,t,1}, \ldots, \eta_{i,t,S_{i,t}})'$ as the vector of test random effects on day t for the individual i and let $\mathbf{I}_{S_{i,t}}$ be an $S_{i,t} \times S_{i,t}$ identity matrix, then assume $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\mathbf{I}_{S_{i,t}} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$. The reason for letting $\eta_{i,t}$ be a singular multivariate normal (by setting the sum of test random effects to be zero on a day t) is to remove any possibility of unidentifiability issues between daily and test random effects. Notice that we use precision parameters instead of variance parameters for normal distributions as it is more convenient to draw precision parameters in the MCMC. Here, we choose $\mathrm{F}(x)$ to be a logistic link in the analysis as per the convention in the EdSphere platform.

2.1.2. The observation equations of response times

When working on a task, a subject has the choice to work faster or slower. However, when we work faster, we often incline to make more unnoticeable mistakes. On the other hand, when one's ability is not comparable to the test taken, it always takes a longer time for him/her to finish the test. Thus, we model the response time as below:

$$\log(R_{i,t,s}) = \mu_i - \nu_{i,t} + \beta L(\theta_{i,t} - a_{i,t,s}) + \zeta_{i,t,s}. \tag{2.2}$$

Equation (2.2) is an extension of Thissen's model (Thissen, 1983) and its variations (Ferrando & Lorenzo-Seva, 2007; Ranger & Kuhn, 2012). In Equation (2.2), μ_i reflects the average response time for the ith respondent in general. $\nu_{i,t}$ implies the variation for the speed of the respondent i on the tth day, with the negative sign indicating that the slower the speed is, the more time one needs to spend on the exam. We further assume that the speed for an examinee will not change much during one day, thus the index of the speed only varies according to individuals and days and $\nu_{i,t} \sim \mathcal{N}(0, \kappa_i^{-1})$, which has an individual-specific precision parameter κ_i and zero mean (to ensure identifiability in the presence of μ_i). In the third term, $L(\theta_{i,t} - a_{i,t,s})$ is mapping how the distance of $\theta_{i,t} - a_{i,t,s}$ in relationship to the response time; and β is a regression coefficient to adjust this relationship. Further, we assume the residual term $\zeta_{i,t,s}$, $\sim \mathcal{N}(0, \varrho^{-1})$ with ϱ as a common precision parameter to borrow the strength of the information across different tests and individuals. Although ϱ varying across individuals may be an alternative, such an assumption might cause unidentifiability issue with precision parameter κ_i when we encounter the situation that an examinee only takes one test per day.

For $L(\cdot)$, there are two popular choices in the literature. One is $L(\cdot) = \cdot$, the monotone relationship and the other is $L(\cdot) = |\cdot|$, the inverted U-shaped relationship. These two relationships represent two different psychological behaviors of test takers during an exam as discussed in Section 1.2. As far as we know, there is no formal statistical comparison of these two relationships in real examples of educational tests. We aim to fill in this gap by developing statistical technique for such comparison and applying to EdSphere datasets.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

2.2. Second level: System equations in the DIR-RT models

Following the idea of Wang et al. (2013), we combine both parametric growth models and Markov chain models to capture one's ability growth over time, that is,

$$\theta_{i,t} = \theta_{i,t-1} + c_i(1-\rho\theta_{i,t-1})\Delta^+_{i,t} + w_{i,t}. \tag{2.3}$$

The first term in Equation (2.3) denotes the ability at the previous time point, $\theta_{i,t-1}$. The second term represents a parametric growth model with c_i as the average growth rate of the ith person's ability over time; $\Delta^+_{i,t} = \min\{\Delta_{i,t}, T_{\max}\}$ is the time lapse between two test dates for the ith individual (i.e., $\Delta_{i,t}$) but truncated by a pre-specified maximum time interval $T_{\max}$ ($T_{\max} = 14$ is used in the application, reflecting typical holiday time for students in school); ρ is the parameter to control the rate of one's growth, which reduces the growth rate when the ability becomes mature. Note that ρ is known from empirical experiments in EdSphere datasets (Swartz et al., 2016; Wang et al., 2013). In principle, ρ should be individual-specific, but it is distinguishable from c_i only when one's ability level is reaching maturation; our investigation of ability growth in the EdSphere data focuses on early age students, so only the c_i are made individual-specific. Lastly, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$ represents the uncertainty that cannot be explained by the first two terms in Equation (2.3), where ϕ is a common unknown parameter to help borrow information and avoid substantial risks of confounding in the likelihood between δ_i's and $\phi^{-1}\Delta_{i,t}$ when the time lapses between tests are equally spaced for a student. The variance of $w_{i,t}$ presumes that one's ability changes become more uncertain, if he/she is absent for a long period. Moreover, we can write Equation (2.3) as a first-order Markov process (see *Step 2* of Appendix A.1), which is beneficial for conducting MCMC later.

2.3. A summary of DIR-RT models

To summarize, the proposed one-parameter DIR-RT model has two levels:

$$\begin{aligned}
\textit{2nd level:}\quad & \theta_{i,t} = \theta_{i,t-1} + c_i(1-\rho\theta_{i,t-1})\Delta^+_{i,t} + w_{i,t},\\
\textit{1st level:}\quad & \Pr(X_{i,t,s,l}=1 \mid \theta_{i,t}, \varphi_{i,t}, \eta_{i,t,s}, a_{i,t,s}, \varepsilon_{i,t,s,l})\\
& \qquad = \frac{\exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \varepsilon_{i,t,s,l})}{1+\exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \varepsilon_{i,t,s,l})},\\
& \log(R_{i,t,s}) = \mu_i - \nu_{i,t} + \beta L(\theta_{i,t} - a_{i,t,s}) + \zeta_{i,t,s},
\end{aligned}$$

where $R_{i,t,s}$ and $X_{i,t,s,l}$ are observed; $a_{i,t,s}$'s, ρ, $\Delta^+_{i,t}$'s, and $\Delta_{i,t}$'s are known and $\varepsilon_{i,t,s,l} \sim \mathcal{N}(0,\sigma^2)$ with known σ^2. Also, we have the assumptions $\varphi_{i,t} \sim \mathcal{N}(0,\delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\mathbf{I}_{S_{i,t}} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\zeta_{i,t,s} \sim \mathcal{N}(0, \varrho^{-1})$, and $\nu_{i,t} \sim \mathcal{N}(0, \kappa_i^{-1})$. Here, $L(\theta_{i,t} - a_{i,t,s})$ can either monotone relationship $(\theta_{i,t} - a_{i,t,s})$ or inverted U-shaped relationship $|\theta_{i,t} - a_{i,t,s}|$.

3. Statistical inference

In order to estimate the uncertainties of unknowns in such a complex model structure, we employ Bayesian inference and MCMC computation in this section.

3.1. Prior distributions for the unknown parameters

Prior choice is crucial in any Bayesian analysis. In the absence of expert's knowledge or historical information, objective priors are used for the unknown parameters to avoid the large impacts of priors on the inference but to have some good frequentist properties (Berger, 2006). Whenever scientific knowledge is available, we instead want to incorporate the information into the prior specifications.

Following these rules, a natural choice for the prior of one's initial latent ability is $\theta_{i,0} \sim \mathcal{N}(\mu_{G_{j_i}}, V_{G_{j_i}})$, where $\mu_{G_{j_i}}$ and $V_{G_{j_i}}$ are the mean and the variance of the subpopulation (j) to which an individual i

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

belongs. Since c_i's in Equation (2.3) are the average growth rates (or the average learning rate in an educational context), it is often assumed to be positive, thus we specify c_i as $\pi(c_i) \propto \mathcal{I}(c_i \geq 0)$, for all i, where $\mathcal{I}(\cdot)$ is an indicator function. The speed κ_i's and the random error ϱ are the precision parameters in the log-normal model for response times, we follow the suggestion of Sun et al. (2001), using $\pi(\varrho) \propto 1/\varrho^{3/2}$, for ϱ and for all i, $\pi(\kappa_i) \propto 1/\kappa_i^{3/2}$ for κ_i. In addition, for the prior choice of scale parameters δ_i's, τ_i's, and ϕ, we follow the discussion of Wang et al. (2013) and use $\pi(\phi) \propto 1/\phi^{3/2}$, $\pi(\delta_i) \propto 1/\delta_i^{3/2}$, and $\pi(\tau_i) \propto 1/\tau_i^{3/2}$ for all i. A natural objective prior for μ_i, the average response time of each individual, is a constant prior, $\pi(\mu_i) \propto 1$, for all i. Similarly, assume the prior of β is $\pi(\beta) \propto 1$. Intuitively, the regression coefficient β in Equation (2.2) with a negative sign makes more sense though, we prefer to let the data determine the value and the sign of β.

3.2. Posterior distribution and data augmentation scheme

Equation (2.1) introduces a logit link for modeling the dichotomous item responses $X_{i,t,s,l}$, and according to Andrews & Mallows (1974), the density of a standard logistic distribution can be represented as a scale mixture of normals (Andrews & Mallows, 1974; Wang et al., 2013). Then, applying the data augmentation idea, we can introduce a latent variable $Y_{i,t,s,l}$ for each response variable $X_{i,t,s,l}$. Notice that in the data augmentation, $Y_{i,t,s,l}$ follows a normal distribution with mean $\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \varepsilon_{i,t,s,l}$ and variance $4\gamma_{i,t,s,l}^2$, where $\gamma_{i,t,s,l}$ is a scale parameter assumed to have a Kolmogorov–Smirnov (K-S) distribution (i.e., $\pi(\gamma) = 8\sum_{\alpha=1}^{\infty}(-1)^{(\alpha+1)}\alpha^2\gamma\exp(-2\alpha^2\gamma^2)$, $\gamma \geq 0$. Here, we use the fact that the K-S mixture of normals will be the logistic distribution and we can verify that $\Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \varepsilon_{i,t,s,l}) = \Pr(Y_{i,t,s,l} > 0 \mid \theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \varepsilon_{i,t,s,l})$. Though introducing more unknowns $Y_{i,t,s,l}$'s, it facilitates the MCMC computation. Then, we can rewrite the one-parameter DIR-RT models (2.1)–(2.3) as

$$\begin{aligned}
\theta_{i,t} &= \theta_{i,t-1} + c_i(1 - \rho\theta_{i,t-1})\Delta_{i,t}^{+} + w_{i,t},\\
\log(R_{i,t,s}) &= \mu_i - \nu_{i,t} + \beta L(\theta_{i,t} - a_{i,t,s}) + \zeta_{i,t,s},\\
Y_{i,t,s,l} &= \theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \xi_{i,t,s,l},
\end{aligned}$$

where $\xi_{i,t,s,l} \sim \mathcal{N}(0, \psi_{i,t,s,l}^{-1})$ with $\psi_{i,t,s,l}^{-1} = 4\gamma_{i,t,s,l}^2 + \sigma^2$ and $\gamma_{i,t,s,l} \sim$ K-S distribution, $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\mathbf{I}_{S_{i,t}} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, $\nu_{i,t} \sim \mathcal{N}(0, \kappa_i^{-1})$, and $\zeta_{i,t,s} \sim \mathcal{N}(0, \varrho^{-1})$.

Further, we define $\theta = (\theta_1, \ldots, \theta_n)'$ with $\theta_i = (\theta_{i,0}, \theta_{i,1}, \ldots, \theta_{i,T_i})'$, $c = (c_1, \ldots, c_n)'$, $\tau = (\tau_1, \ldots, \tau_n)'$, $\delta = (\delta_1, \ldots, \delta_n)'$, $\mu = (\mu_1, \ldots, \mu_n)'$, and $\kappa = (\kappa_1, \ldots, \kappa_n)'$. For $l = 1, \ldots, K_{i,t,s}$, $s = 1, \ldots, S_{i,t}$, $t = 1, \ldots, T_i$, and $i = 1, \ldots, n$, we have $Y = \{Y_{i,t,s,l}\}$, $\gamma = \{\gamma_{i,t,s,l}\}$, and $X = \{X_{i,t,s,l}\}$; $\varphi = \{\varphi_{i,t}\}$ and $\nu = \{\nu_{i,t}\}$; $\log R = \{\log R_{i,t,s}\}, \eta = \{\eta_{i,t,s}\}$ and particularly, we use $\eta_{i,t}^{*} = (\eta_{i,t,1}, \ldots, \eta_{i,t,S_{i,t}-1})'$. Then, given the data $(X, \log R)$, the joint posterior of $(\theta, Y, c, \tau, \eta, \varphi, \delta, \phi, \beta, \nu, \kappa, \mu, \varrho, \gamma)$ for the proposed DIR-RT models is

$$\begin{aligned}
&\pi(\theta, Y, c, \tau, \eta, \varphi, \delta, \phi, \beta, \nu, \kappa, \mu, \varrho, \gamma \mid X, \log R)\\
&\propto \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}\prod_{l=1}^{K_{i,t,s}} \pi(\gamma_{i,t,s,l})\right\}\left\{\prod_{i=1}^{n} \pi(\theta_{i,0})\pi(c_i)\pi(\delta_i)\pi(\tau_i)\pi(\kappa_i)\pi(\mu_i)\right\}\pi(\beta)\pi(\phi)\pi(\varrho)\\
&\times\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}\prod_{l=1}^{K_{i,t,s}} [\mathcal{I}(Y_{i,t,s,l} > 0)\mathcal{I}(X_{i,t,s,l} = 1)\mathcal{I}(Y_{i,t,s,l} \leq 0)\mathcal{I}(X_{i,t,s,l} = 0)]\right.\\
&\left.\times\sqrt{\frac{\psi_{i,t,s,l}}{2\pi}}\exp\left(-\frac{\psi_{i,t,s,l}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2}\right)\mathcal{I}\left(\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1}\eta_{i,t,s}\right)\right\}\\
&\times\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\delta_i}{2\pi}}\exp\left(-\frac{\delta_i\varphi_{i,t}^2}{2}\right)\right\}\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\left(\frac{\tau_i}{2\pi}\right)^{(S_{i,t}-1)/2}\exp\left(-\frac{\tau_i\eta_{i,t}^{*'}\Sigma_{i,t}^{-1}\eta_{i,t}^{*}}{2}\right)\right\}
\end{aligned}$$

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

$$\times\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\phi}{2\pi\Delta_{i,t}}}\exp\left(-\frac{\phi[\theta_{i,t}-\theta_{i,t-1}-c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}^{+}]^2}{2\Delta_{i,t}}\right)\right\}$$
$$\times\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}\sqrt{\frac{\varrho}{2\pi}}\exp\left(-\frac{\varrho(\log(R_{i,t,s})-\mu_i+\nu_{i,t}-\beta L(\theta_{i,t}-a_{i,t,s}))^2}{2}\right)$$
$$\times\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\kappa_i}{2\pi}}\exp\left(-\frac{\kappa_i\nu_{i,t}^2}{2}\right), \tag{3.1}$$

where $\Sigma_{i,t}^{-1}=\mathbf{1}_{S_{i,t}-1}\mathbf{1}_{S_{i,t}-1}'+\mathbf{I}_{S_{i,t}-1}$, with $\mathbf{1}_{S_{i,t}-1}$ being a $(S_{i,t}-1)$-dimensional unit vector. Note that $\pi(\theta_{i,0})$, $\pi(c_i)$, $\pi(\delta_i)$, $\pi(\tau_i)$, $\pi(\kappa_i)$, $\pi(\mu_i)$, $\pi(\beta)$, $\pi(\phi)$, and $\pi(\varrho)$ are the priors specified in Section 3.1 and $\pi(\gamma_{i,t,s,l})$ is the K-S density. To verify the posterior propriety of DIR-RT models under the objective priors, we can follow the similar steps as Wang et al. (2013) did in the DIR models. The major difference in the proof is that DIR-RT models contain an additional part, which is to conjointly model with response times. Since the logarithm of response times is modeled as a normal linear mixed regression model, then the well-established facts for the posterior propriety of normal linear mixed models in Bayesian literature (cf. Sun et al., 2001) can be coupled with results from Appendix C in Wang et al. (2013) to show the posterior propriety of DIR-RT models.

3.3. MCMC computation of DIR-RT models

The computation is carried out by the MCMC scheme, where we sample the posterior (3.1) via block Gibbs sampling schemes. The difficulty of the sampling scheme arises in sampling the posterior distribution of latent ability $\theta_i=(\theta_{i,0},\ldots,\theta_{i,T_i})'$ for each individual i, for $i=1,\ldots,n$, where coordinates of θ_i are typically high dimensional and strongly correlated. When $L(x)=x$, using the data augmentation idea, the proposed model is transformed so that θ_i (the vector) could be block sampled (see Step 2 of Appendix A.1)—within a single Gibbs sampling step conditional on the other parameters (excluding $\theta_{i,t}$'s)—by the highly efficient forward filtering and backward sampling algorithm (West & Harrison, 1997). However, if $L(x)=|x|$, the computation becomes more challenging as θ_i cannot be drawn as a block. Instead, we utilize the fact that the full conditional distribution of each component of θ_i, i.e., $\theta_{i,t}$, follows a mixture of truncated Gaussians, then $\theta_{i,t}$'s are drawn one at a time (cf. Step 2 of Appendix A.1).

The details of MCMC steps are given in Appendix A.1. The Gibbs sampling starts at *Step 1* in Appendix A.1, with initial values $\theta^{(0)}=\vec{0}$, $c^{(0)}=\vec{0}$, $\phi^{(0)}=1$, $\varphi^{(0)}=\vec{0}$, $\eta^{(0)}=\vec{0}$, $\delta^{(0)}=\vec{1}$, $\tau^{(0)}=\vec{1}$, $\gamma^{(0)}=\vec{1}$, $\mu^{(0)}=\vec{1}$, $\nu^{(0)}=\vec{0}$, and $\beta^{(0)}=0$ in the applications and simulations, then loops through *Step 15* in Appendix A.1, until the MCMC has converged. The convergence is evaluated informally by looking at trace plots.

The statistical inferences are made straightforward from the MCMC samples. For example, an estimate and 95% credible interval (CI) for the latent trajectory of one's ability $\theta_{i,t}$ can be plotted from the median, 2.5%, and 97.5% empirical quantiles of the corresponding MCMC realizations. In examples, ability will be graphed as a function of t, so that the dynamic changes of an examinee are apparent.

4. Simulation study

To validate the inference procedure and compare the benefits of jointly modeling response times with item responses, we conduct some simulations. Due to space limitations and motivated by the analysis of empirical data in Section 5, we only illustrate the results when $L(x)$ follows an inverted U-shape linkage in the joint model. Similar conclusions are yielded when we proceed with monotone linkage, thus we omit the details here.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

Table 1. Values of unknowns used in the simulation

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$	$i=10$
c_i	0.0055	0.0065	0.0026	0.0037	0.0061	0.0047	0.0035	0.0043	0.0039	0.0015
δ_i	2.0408	1.3333	1.8182	1.2346	1.5873	1	2.2222	1.0526	1.1494	2
τ_i	4	3.1250	4.3478	2.7027	3.7037	2.8571	4	2.2222	9.0909	4.5455
κ_i	2.3256	1.5873	1.6949	0.5495	1.2658	0.9346	1.3889	1.8182	2.7027	1.2195
μ_i	1.6	1.47	1	1.92	1.45	1.73	1.5	1.35	0.81	1.23

4.1. DIR-RT models simulation

Following the simulation study of DIR models in Wang et al. (2013), we consider multiple individuals taking a series of tests scheduled at individually-varying and irregularly-spaced time points. Assume there are 10 individuals, each of them taking four tests on 50 different test dates, where each test contains 10 items. This specification means $K_{i,t,s}=10$, for $s=1,\ldots,S_{i,t}$, $t=1,\ldots,T_i$, $i=1,\ldots,n$ with $S_{i,t}=4$, $T_i=50$, and $n=10$. The time lapse between two consecutive test dates $\Delta_{i,t}=t+10$ if $t\le T_i/2$ and $\Delta_{i,t}=t-10$ otherwise, creating an irregularly spaced gap between two test dates.

To compare the results of DIR-RT models with DIR models, we assign same values of the parameters, ϕ, δ_i, τ_i, c_i as used in Wang et al. (2013), where $\phi=1/0.0218^2$, leading to standard deviation of $w_{i,t}$ in Equation (2.3) being $0.0218\sqrt{\Delta_{i,t}}$ and the values of δ_i, τ_i, and c_i are specified in Table 1. For the modeling part of response times, the parameter values of κ_i and μ_i are listed in Table 1, $\beta=-0.17$ and $\varrho=1.25$. The parameters of DIR-RT models are chosen in such a way that they mimic the magnitude of EdSphere datasets.

We consider the inverted U-shape linkage, i.e., $L(x)=|x|$. Then, we simulate random effects and latent variables in DIR-RT models using the assigned parameter values above. Once we generated the values of $\theta_{i,t}$ using Equation (2.3), we set the test difficulties, $a_{i,t,s}$'s in Equation (2.1) to be $\theta_{i,t}+\zeta^{\star}$, where $\zeta^{\star}$ is a random variable with uniform distribution on $(-0.1,0.1)$. Next, the values of $\varepsilon_{i,t,s,l}$ are drawn from $\mathcal{N}(0,\sigma^2)$ with $\sigma=0.7333$ and we choose $\rho=0.1180$. Notice that the values of σ and ρ are the same values as used in the EdSphere platform. Finally, the dichotomous data of item responses and continuous data of response times generated from the joint model are our observations, and we use the Bayesian machinery from Section 3 to estimate the model parameters of DIR-RT models. To analyze the simulation data, we utilize the priors for unknowns as specified in Section 3.1 except for β, where we use $\pi(\beta)\propto\mathbf{1}(\beta<0)$ to make the MCMC computation converge faster.

The unknown parameters are estimated through the posterior median calculated from their corresponding MCMC samples. Each MCMC was run for 50,000 iterations with a 25,000 burn-in period. In nested or hierarchical models involving many precision parameters (δ_i's, τ_i's, $\psi_{i,t,s,l}$'s, ϕ, and others), a large number of burn-in iterations is often required to stabilize MCMC mixing. To run the MCMC with 50,000 iterations, it takes about 15 hours and 22 minutes by using a 3.40-GHz processor with 16-GB RAM. The MCMC convergence is assessed through a graphical examination of trace plots and through Geweke's Diagnostics test statistics (Geweke, 1991). The absolute values of all parameters' Geweke's test statistics (in both simulation and application) are below 2, suggesting our MCMC chain has been converged. Please refer to Sections 1and 3 of the Supplementary Material for more details and additional information of auto-correlation of MCMC samples and the effective samples size are also provided there. Figures 1a–d give posterior median estimates (red squares) along with 95% CIs (red bars) of c_i's, $\tau_i^{-1/2}$'s, $\delta_i^{-1/2}$'s, $\kappa_i^{-1/2}$'s, and μ_i's, respectively, and illustrate their true values (black dot). Clearly seen from Figure 1, the true values of parameters are contained within their corresponding 95% CIs. For the posterior median estimates of parameters $\phi^{-1/2}$, $\varrho^{-1/2}$, and β are 0.0190, 0.9075, and -0.1815, respectively, with their corresponding 95% CIs being $[0.0159,0.0229]$, $[0.8753,0.9427]$, and $[-0.7028,-0.0071]$, all of which contain their truth.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

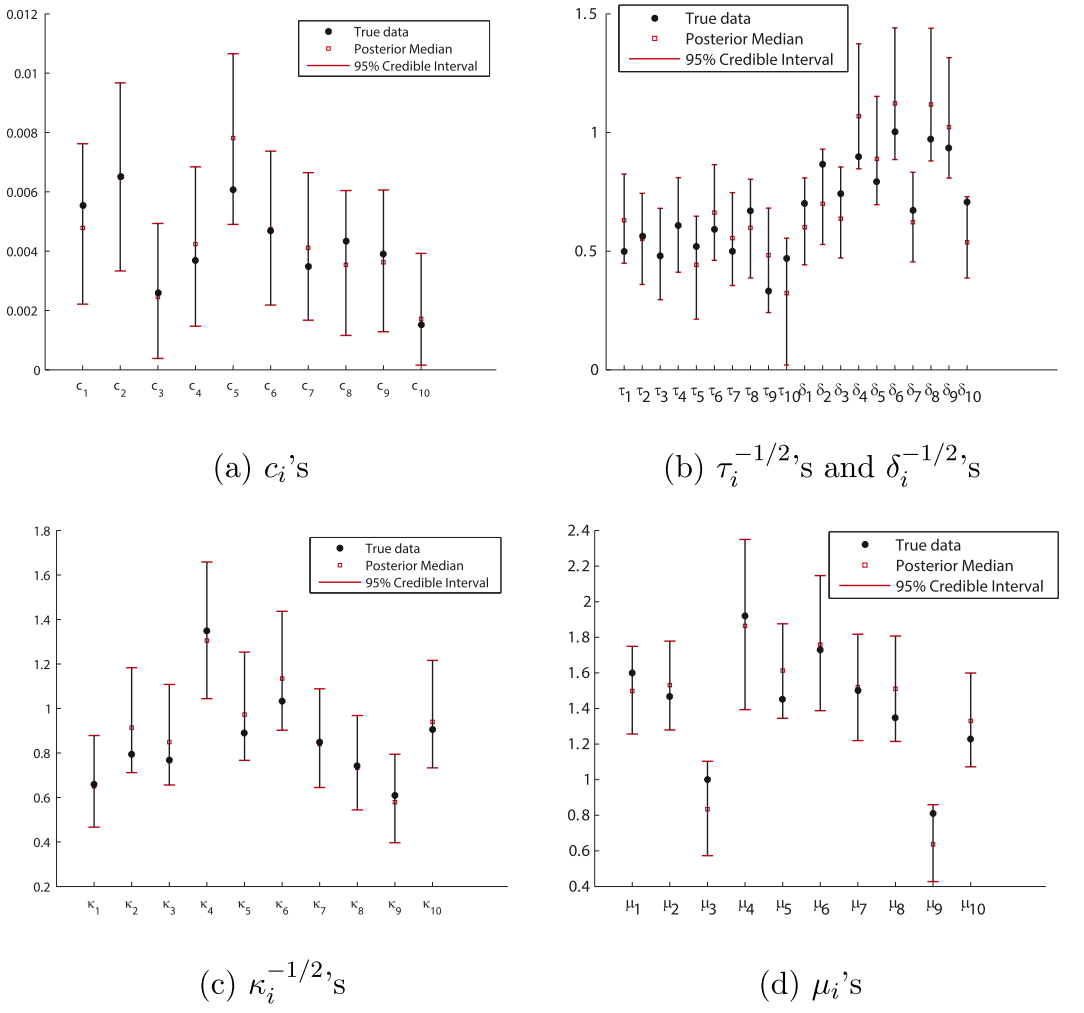


Figure 1. Posterior summary of c_i's, $\tau_i^{-1/2}$, $\delta_i^{-1/2}$'s, $\kappa_i^{-1/2}$'s, and μ_i's.
Note: The black dots represent truth, red squares are posterior median estimates, and red bars indicate 95% CIs.

Next, we discuss our primary interest of estimating latent ability trajectories. Figures 2a–d illustrate four types of growth curves in our simulation, where (a) represents an individual with steady growth; (b) indicates an individual with increasing growth but nearly flat region at the end; (c) shows an individual with interrupted growth (with true ability drops in certain period); and (d) displays monotonic growth with decreasing growth rate in the middle. In Figure 2, the true ability curves (black dots) have been plotted along with our posterior median estimates of ability (blue circles) and their corresponding 95% credible band (CB) (starred lines). Notice that in each subfigure, a very small proportion of true values (less than 5%) are outside of 95% CBs.

To better assess how well the Bayesian model actually captures the truth, we can calculate the frequency of true values falling in the corresponding CIs of MCMC runs over different random trials, i.e., the frequentist coverage probability (CP) (Wang et al., 2013). Thus, to evaluate CPs, we conduct the simulation with the same set-up values of parameters specified earlier but generate datasets with 100 different random seeds. The CPs for ϱ, β, and ϕ are 96%, 97%, and 87%, respectively, and the average CPs over all individuals for κ_i's, c_i's, τ_i's, δ_i's, and μ_i's are 95.4%, 92.1%, 95.4%, 93.8%, and 95%, respectively (refer to the CPs of these individual parameters in additional Table S.6 in the Supplementary Material).

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

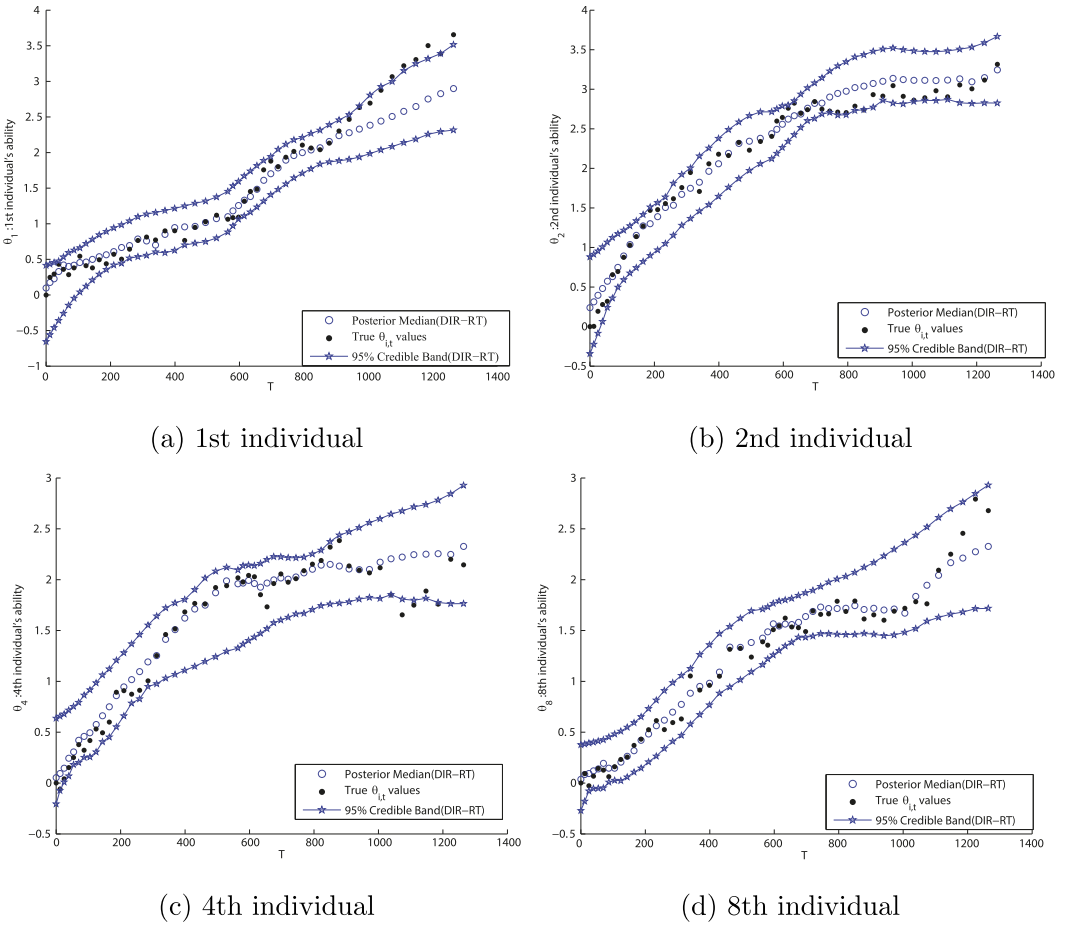


Figure 2. The latent trajectory of one's ability growth, where black dots, blue circles, and starred lines represent true ability, the posterior median estimates, and the 95% CBs, respectively.

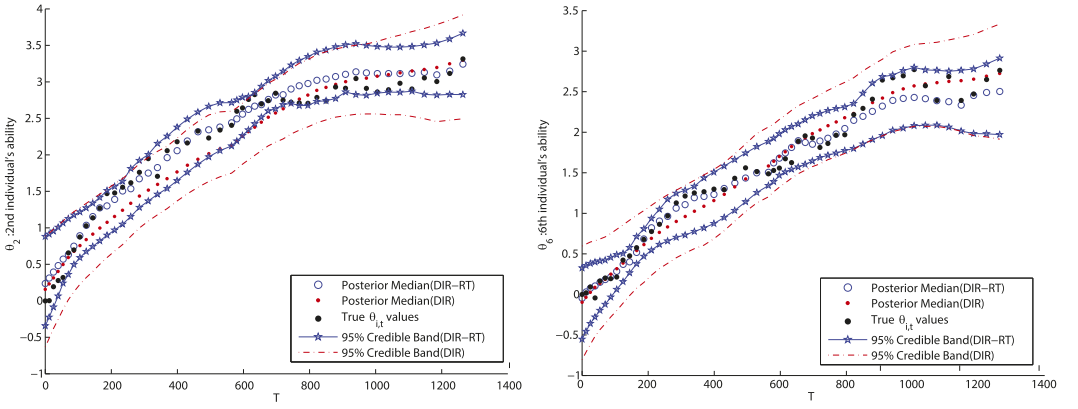
In addition, the average CPs for one's ability over the study period, i.e., for $\theta_1, \ldots, \theta_{10}$, are 94.88%, 94.90%, 95.08%, 95.62%, 95.40%, 95.54%, 95.50%, 96.22%, 95.78%, and 95.22%, respectively. Thus, while the inferential method is Bayesian, it seems to yield sets that have good frequentist coverage.

4.2. Benefits of joint modeling response times with item responses

To the best of our knowledge, the two-level DIR-RT model is the first attempt to jointly model response times and accuracy in the analysis of longitudinal data for latent traits. We are, thus, interested in knowing the benefits of introducing response times as an extra source of information. An interesting investigation is to compare the estimates of one's ability trajectory relative to the truth with and without using response times.

Figure 3 displays the growth curve of two selected individuals, where the statistical inference is based on the simulated example in Section 4.1. For other individuals, results are similar and thus, their plots are omitted due to space limitations. In Figures 3a,b, 95% CIs of DIR models (dashed red lines) encompass 95% CIs of DIR-RT models (starred blue lines), both 95% CIs contain the true values (black dots). The average length of the 95% CB of ability estimates for DIR-RT models is much shorter than that of DIR models (0.6454 versus 1.0370 for the 2nd individual; and 0.6772 versus 1.1401 for the 6th individual, respectively). In addition, in Figure 3, both graphs of DIR-RT for ability estimates (blue circles) adhere

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press



(a) 2nd individual (b) 6th individual

Figure 3. The comparison of ability estimates between DIR-RT and DIR models, where black dots, blue circles, and red dots represent true ability, DIR-RT ability estimates, and DIR ability estimates, respectively; starred lines (blue) and dashed lines (red) represent 95% CBs for DIR-RT and DIR models, respectively.

Table 2. Characteristics of the first three students randomly sampled from the EdSphere data

	Total tests	Days	Max. tests/days	Range of items/test	Max. gap	Initial grade
No.1	150	74	9	4–22	79	4
No.2	203	128	15	6–24	107	2
No.3	211	107	9	5–24	79	3

more closely to true ability (black dots) in relative to DIR ability estimates (red dots). The average mean-squared distance between the truth and the posterior median ability estimates over time for DIR-RT models are 0.0240 for θ_2 and 0.0187 for θ_6, in comparison to that of DIR models are 0.0711 for θ_2, and 0.0653 for θ_6, both of which are at least three times larger than DIR-RT models. To conclude, the results of DIR-RT models illustrate that we can largely improve the precision and remarkably reduce the bias of the estimates of one's ability by incorporating response times.

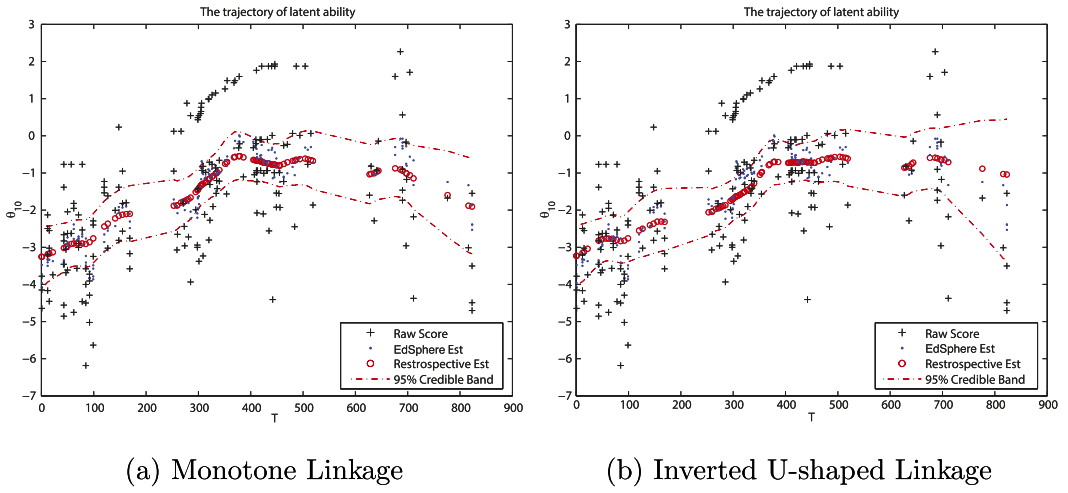
5. EdSphere testbed application

For illustration purposes, we randomly select a sample of 25 individuals from the EdSphere testbed, where different characteristics for each student are shown in Table 2. Due to space limitations, we only show the details of the first three selected individuals. The primary focus of this application is to study the following goals: 1) to assess the appropriateness of the local independence assumption for this type of data; 2) to understand the growth in ability of students by retrospectively producing the estimated growth trajectory for each student; and 3) to investigate which linkage between response times and the distance of ability–difficulty is more suitable to model the behaviors and psychology for students in the exam via empirical justifications.

5.1. DIR-RT models with inverted U-shape vs. monotone linkage

To the best of our knowledge, there has not been any empirical work to check the goodness of fit of the two linkages (i.e., monotone or inverted U-shape) for which one indeed fits the data better in conjointly modeling response times and item responses, especially when the testing data are collected at irregular

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press



(a) Monotone Linkage (b) Inverted U-shaped Linkage

Figure 4. The posterior summary of ability growth for the 10th individual in two linkages, where red circles, black plus, and blue dots are posterior median estimates of the ability, raw score, and EdSphere estimates, respectively, and red dashed lines are 95% CBs of our estimates.

and individual varying time points for a series of computerized (adaptive) testing. The lack of this research motivates us to propose some statistical measures that can be used to conduct an empirical study of EdSphere datasets to identify a better linkage and in this study, we still use the prior assigned for unknowns in Section 3.1.

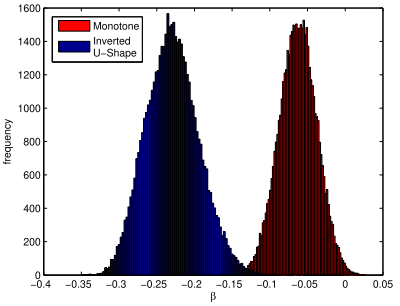
Figure 4 illustrates the ability trajectories using monotone (left) and inverted U-shape (right) linkages side by side, where red dots present the posterior median trajectory of ability estimation, red dashed lines correspond to their CBs and black plus points are the "raw score," which is a rough estimate of one's ability obtained by solving the equation that the expectation of expected score for a person's ability is equivalent to the observed score (Stenner, 2022; Swartz et al., 2016). As a side for reference, we also include the EdSphere estimates (blue dots), the estimates employed in the design of the EdSphere learning platform, in Figure 4, which can be viewed as a preliminary version of DIR models without considering local dependence. However, EdSphere estimates use the data only available to that date. Thus, EdSphere estimates (blue dots) are much varying than ours in Figure 4.

In comparison of (a) and (b) in Figure 4, the estimates of ability growth are comparatively more robust (in an underlying increasing trend) for the inverted U-shape than that for a monotone linkage. This phenomenon is shown, for example, in the estimation of ability trajectory in Figure 4, during the period of 350–500 days and 700–800 days, where the ability trajectory estimated by an inverted U-shape linkage in the joint model is more reluctant to change its increasing trend unless there is strong information from the data (noting raw scores (black plus) in Figure 4 can be regarded as the raw data since they have the same scale as $\theta_{i,t}$). In the following, we have proposed two statistical methods to rigorously compare the two linkages in fitting the joint model.

5.1.1. Comparison of two linkages using Lindley's method

The regression slope β plays a key role in controlling the influence of the ability–difficulty distance function $L(\cdot)$ to the response time. When $\beta = 0$, it implies the distance between ability and difficulty does not affect the time an individual spends on a test and the corresponding linkage function $L(\cdot)$ can be ignored. Thus, we are interested in testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. Lindley's method (see Lindley 1965, Section 5.6), advocated by authors including Zellner (1971), is an ad hoc way to test this hypothesis. According to Lindley's method, we reject the hypothesis of $\beta = 0$ at the α level of significance if the $100(1-\alpha)\%$ highest posterior density interval does not include 0. The posterior density of β is

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press



Model(β)	Inverted U-shape	Monotone
PM	-0.2305	-0.0627
95% CI	$(-0.2940, -0.1571)$	$(-0.1125, -0.0137)$
99% CI	$(-0.3105, -0.1345)$	$(-0.1317, 0.0003)$

Figure 5. Posterior histogram (left) and posterior summary (right) for β under monotone linkage and inverted U-shape, where "PM" in the table is short for "posterior median."

in bell shapes (clearly seen from histograms in Figure 5). Thus, $100(1-\alpha)\%$ highest posterior density interval of β is the same as its $100(1-\alpha)\%$ CI, but the latter one is much easier to obtain from the MCMC samples. From the table in Figure 5, it suggests $\beta = 0$ cannot be rejected at $\alpha = 1\%$ for monotone linkage since 99% CI of β under monotone linkage includes 0, while $\beta = 0$ is rejected at both $\alpha = 1\%$ and $\alpha = 5\%$ for inverted U-shape. This indicates that the monotone linkage has a weaker correlation with response times than the inverted U-shape linkage for EdSphere datasets.

5.1.2. Comparison of two linkages using deviance information criterion (DIC)

The hypothesis of $\beta = 0$ can also be viewed as a Bayesian model selection question and then, we can employ DIC, one of the most popular model selection criteria in the Bayesian literature, for the comparison. Let $\boldsymbol{y}$ denote the data and $\boldsymbol{\Theta}$ indicate parameters in the model. Define $\pi(\boldsymbol{\Theta} \mid \boldsymbol{y})$ as the posterior distribution of $\boldsymbol{\Theta}$ and $\overline{\boldsymbol{\Theta}} = \mathrm{E}_{\pi(\boldsymbol{\Theta}|\boldsymbol{y})}[\boldsymbol{\Theta} \mid \boldsymbol{y}]$ represents the posterior mean of $\boldsymbol{\Theta}$. Then, $\mathrm{Dev}(\Theta) = -2\log f(\boldsymbol{y} \mid \boldsymbol{\Theta})$ is the deviance function for the likelihood $f(\boldsymbol{y} \mid \boldsymbol{\Theta})$. Following Spiegelhalter et al. (2002), we have

$$\mathrm{DIC} = \mathrm{Dev}(\overline{\boldsymbol{\Theta}}) + 2p_D, \tag{5.1}$$

where $p_D = \mathrm{E}_{\pi(\boldsymbol{\Theta}|\boldsymbol{y})}[\mathrm{Dev}(\boldsymbol{\Theta}) \mid \boldsymbol{y}] - \mathrm{Dev}[\bar{\boldsymbol{\Theta}}]$ represents the effective number of model parameters. DIC is a measure to trade off between model adequacy ($\mathrm{Dev}(\bar{\boldsymbol{\Theta}})$ part) and model complexity (p_D part), and thus the DIC can be used to compare the linkage choices in DIR-RT models. The smaller the DIC value is, the better the model does fit the data.

There are many variants of DIC, such as conditional DIC and integrated DIC. Nevertheless, recent studies have cautioned against the use of the conditional DIC, whose computation is based on the likelihood conditioning on the latent variables and is sensitive to transformations of latent variables and distributions (Millar, 2009). Zhang et al. (2019) also found that the conditional DIC often selects a model that is more complex than the true model. Instead, the integrated DIC by marginalizing out latent variables in the likelihood often performs better. Without doubt, the more the latent variables can be integrated out, the more efficiency of DIC can be achieved. Several papers showed the integrated DIC often had much smaller Monte Carlo errors compared to the conditional DICs (Celeux et al., 2006; Chan & Grant, 2016; Merkle et al., 2019). But usually, the analytical expressions for the integrated likelihood are hard to obtain and rather often, one resorts back to numerical integration, which is typically time-consuming.

In the proposed DIR-RT models, when $L(\cdot)$ is a monotone linkage, DIR-RT models can be written as a framework of dynamic linear models (see *Step 2* of Appendix A.2) by introducing K-S random variables ($\gamma_{i,t,s,l}$'s) in the data augmentation step. Then, following an analogy of Chan & Grant (2016) for derivations of dynamic linear models, we can show that the integrated likelihood of DIR-RT models under monotone linkage can be approximated by taking one-dimensional Monte Carlo

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

integration over $\gamma_{i,t,s,l}$'s drawn from the K-S prior distribution. However, when $L(\cdot)$ is an inverted U-shape linkage, the simplification of the integrated likelihood of DIR-RT models becomes much harder since given $\gamma_{i,t,s,l}$'s, the DIR-RT models cannot be written as dynamic linear models. Thus, to evaluate the integrated likelihood under an inverted U-shape becomes much more computationally intensive.

However, conditioning on ability $\theta_{i,t}$'s, the modeling of response times part and item response part in DIR-RT models are independent. Thus, no matter what the linkage is chosen, Equations (2.1) and (2.3) of the DIR-RT models are the same under different linkages. Then, motivated by the partial DIC idea of Yao et al. (2015) in meta-analysis, an alternative to integrated DIC for our joint DIR-RT models is to consider the integrated DIC based on the part of response times only. Based on this partial DIC derived (see details in Appendix B), the inverted U-shape linkage turns out to fit the EdSphere data better, where the partial DIC for the inverted U-shape is 5661.4 in comparison to that of the monotone linkage, which is 5775.3.

5.2. Retrospective estimation of ability growth under inverted U-shaped linkage

Both Lindley's method and partial DIC criterion support the choice of an inverted U-shape linkage for the analysis of EdSphere data using DIR-RT models. Therefore, to investigate the first two goals mentioned in the beginning of Section 5, we are going to use an inverted U-shape throughout the rest of the article. After we run MCMC for the EdSphere datasets using the inverted U-shape, we have conducted a posterior predictive check (Gelman et al., 1996) using a mean test statistic for the response time and its p-value is 0.4978, which shows the inverted U-shape indeed fit the response time well.

Figure 6 presents a retrospective analysis of the reading ability for 3rd, 12th, 18th, and 23rd individuals (using all data recorded for each individual during the study period). In Figure 6, the red circles are the posterior median estimates of one's ability, the red dashed lines correspond to the 2.5% and 97.5% quantiles of the posterior distributions of the abilities, and the black plus points are raw scores. Similar to Wang et al. (2013), we find all these growth trajectories have an overall increasing trend but such kind of growth can be interrupted. In particular, when there is a large time gap between subsequent tests, the ability appears to drop for some individuals, which is clearly seen from Figure 6. A natural explanation might be that during vacations, students do not read and could actually lose ability or they become less familiar with computerized tests after a long break.

In Figure 7, we summarize the posterior median (red square) and 95% CI (red bars at two ends) of the average growth rates c_i's, the standard deviations of test random effects $\tau_i^{-1/2}$'s, the standard deviations of the daily random effects $\delta_i^{-1/2}$'s, the standard deviations of speediness, $\kappa_i^{-1/2}$'s, and the average response time for each individual, μ_i, for $i = 1, \ldots, 25$. Moreover, the estimated posterior median of $\phi^{-1/2}$ is 0.0708 and its 95% CI is $[0.0608, 0.0831]$ and the result of β is summarized in the table of Figure 5.

Figures 7b,c show that the standard deviations of test and daily random effects (i.e., $\eta_{i,t,s}$'s and $\varphi_{i,t}$'s) are almost all quite large with 95% CIs that are well separated from zero except 22nd and 23rd individuals. Recall that these random effects were included in the model to account for a possible lack of the local independence; the evidence is thus strong that the local independence is, indeed, not tenable for this data and that both types of random effects are presented for most individuals. Similarly, Figure 7d illustrates that speediness of individuals is different on the daily basis except that of the 22nd individual (whose speed is almost steady during the studying period). Moreover, there are clearly some patterns in the variation of speediness across individuals; as for some of them, the difference of their speediness on a daily basis is more crucial than that of the others. The differentiation of the average response time in Figure 7e suggests that some individuals incline to take a longer time to finish a test than others no matter what the difficulty level of the test is. As well, it is not surprising the average growth rates are quite different among individuals, as shown in Figure 7a.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

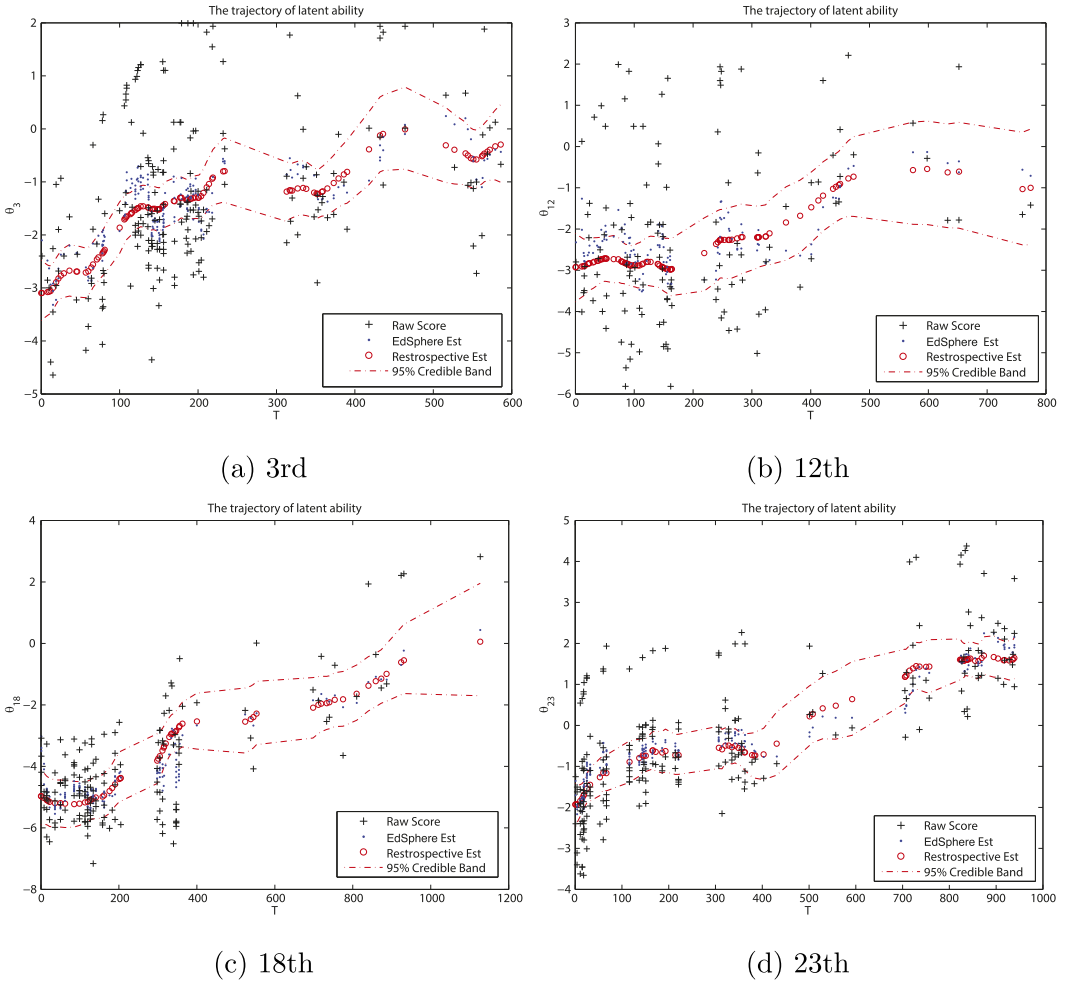


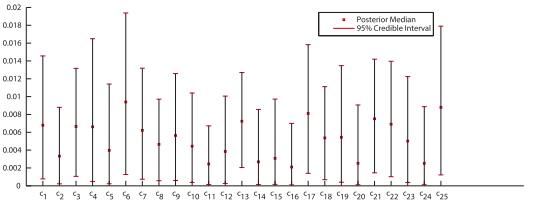
Figure 6. The posterior summary of the ability growth for θ_3, θ_{12}, θ_{18}, and θ_{23}, where red circles, black plus, and blue dots are posterior median estimates of the ability, raw score, and EdSphere estimates, respectively, and red dashed lines represent 95% CBs of our estimates.

6. Conclusions and discussion of future work

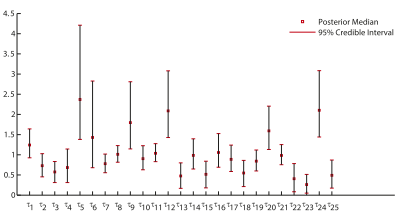
Our proposed DIR-RT models can jointly model the observations of response times and item responses through sharing ability parameters and can accommodate the complex longitudinal data observed at individually-varying and irregularly-spaced time points. Of course, by simplifying the indices of DIR-RT models, they can also be applied to longitudinal data with a simpler structure, such as when the observations of each participant are equally spaced and examinees are given the same repeated tests over the study period. Thus, according to the structure of the longitudinal data, the practitioners can flexibly employ DIR-RT models at their discretion.

From our simulation study, we have noticed that the incorporation of response time into the item response model in the analysis of longitudinal data has both significantly improved the precision and reduced the bias for the ability estimation. As is known, the enhancement of ability accuracy is vital in the design of computerized (adaptive) testing. For example, the tests provided in the EdSphere learning platform are tailored to the current ability estimation of examinees. With more accurate estimates of ability (in the sense of less bias and higher precision), the assigned tests in the EdSphere platform will

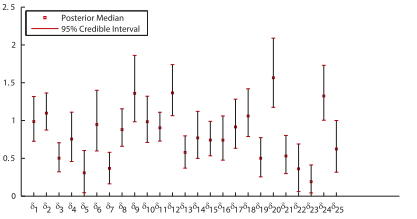
https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press



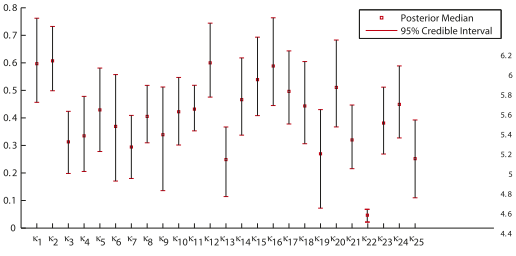
(a) The posterior median and 95% CI of c_i's



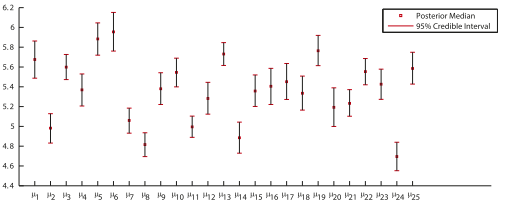
(b) The posterior median and 95% CI of $\delta_i^{-1/2}$'s



(c) The posterior median and 95% CI of $\tau_i^{-1/2}$'s



(d) The posterior median and 95% CI of $\kappa_i^{-1/2}$'s



(e) The posterior median and 95% CI of μ_i's

Figure 7. The posterior summary of c, $\tau_i^{-1/2}$'s, $\delta_i^{-1/2}$'s, $\kappa_i^{-1/2}$'s, and μ_i's.

better match the students' ability and further, it enables teachers to better assist students based on their respective capacities.

Using the proposed DIR-RT models to analyze EdSphere datasets, it further supports the findings in Wang et al. (2013). For example, the evidence of violation of the local independence assumption is generally strong in DIR-RT models, and the use of test and daily random effects to model the local dependence seems to be necessary and successful; and the retrospective analysis of ability estimation is useful in understanding population behavior, such as the frequently observed drops in ability after a long vocation in testing.

More importantly, our analysis is the first empirical study in testing to evaluate the choice of linkage function to describe the relationship between the ability–difficulty and response time in a joint model for longitudinal data. The empirical result favors the inverted U-shape linkage, which is quite significant and meaningful, since it supports that in a series of computerized (adaptive) tests, students intend to spend more time on tests that match their ability levels and to spend less time on those either too easy or too hard. Our discovery using the EdSphere dataset is consistent with the theoretical findings of Wang (2006). Further, the partial DIC criterion is a new model assessment method, which makes the comparison feasible to different linkage functions in the analysis of complex longitudinal data for the proposed DIR-RT model.

Many extensions of current DIR-RT models are possible, such as aforementioned extensions to two-parameter and three-parameter DIR-RT models or including a dynamic structure on the speediness

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

parameter which can conjointly model with one's latent ability. Additionally, investigating any undesired or unplanned dependencies in the response time data—beyond the substantively meaningful latent factors, especially those with dynamic structures—represents another interesting research direction with significant practical implications. Moreover, Figure 7 clearly illustrates some patterns among individuals for the average growth rate c_i's, the variation of speediness κ_i's, and the average response time μ_i's. In the next step, we can use either model-based or distance-based clustering methods to analyze the psychological behaviors of students reflected in the patterns shown in Figure 7. Since the MCMC computation of DIR-RT models is time-consuming, we limit our application to a small sample of EdSphere datasets. We plan to develop parallel computing schemes to improve the computation efficiency and then, we can conveniently apply our approach to the entire dataset.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2025.10019.

Funding statement. Dr. Wang's research was fully supported by the National Science Foundation CAREER Award No. 1848451.

Competing interests. The authors declare no competing interests exist.

References

Albers, W., Does, R., Imbos, T., & Janssen, M. (1989). A stochastic growth model applied to repeated tests of academic knowledge. *Psychometrika*, *54*(3), 451–466.

Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(1), 99–102.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402.

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*(4), 651–673.

Chan, J. C. C., & Grant, A. L. (2016). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics & Data Analysis*, *100*, 847–859.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194.

Darrell Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525–543.

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.

Gaviria, J.-l. (2005). Increase in precision when estimating parameters in computer assisted testing using response time. *Quality and Quantity*, *39*(1), 45–69.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. (Tech. Rep.). Federal Reserve Bank of Minneapolis.

Jannarone, R. (1986). Conjunctive item response theory kernels. *Psychometrika*, *51*(3), 357–373.

Johnson, C., & Raudenbush, S. W. (2006). A repeated measures, multilevel Rasch model with application to self-reported criminal behavior. In C. S. Bergeman and S. M. Boker (Eds.), *Methodological issues in aging research* (pp. 131–164). Psychology Press.

Klein Entink, R. H. (2009). *Statistical models for responses and response times* [Unpublished doctoral dissertation]. University of Twente.

Lindley, D. V. (1965). *Introduction to probability and statistics from a bayesian viewpoint, part 2, inference*. (Vol. *2*). Cambridge University Press.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

Liu, F., Wang, X., Hancock, R., & Chen, M.-H. (2022). Bayesian model assessment for jointly modeling multidimensional response data with application to computerized testing. *Psychometrika*, *87*(4), 1290–1317.

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254–274.

Liu, Y., & Wang, X. (2020). Bayesian nonparametric monotone regression of dynamic latent traits in item response theory models. *Journal of Educational and Behavioral Statistics*, *45*(3), 274–296.

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, *76*(3), 487–503.

Lord, F. (1953). The relation of test score to the trait underlying the test. *Educational Psychology Measurement*, *13*(4), 517–549.

Martin, A. D., & Quinn, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis*, *10*(2), 134–153.

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.

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*(3), 802–829.

Millar, R. B. (2009). Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics*, *65*(3), 962–969.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626.

Molenaar, D., Rózsa, S., & Kō, N. (2021). Modeling asymmetry in the time–distance relation of ordinal personality items. *Applied Psychological Measurement*, *45*(3), 178–194.

Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods*, *47*(4), 1413–1424.

Park, J. H. (2011). Modeling preference changes via a hidden Markov item response theory model. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 479–491). CRC Press.

Ranger, J., & Kuhn, J.-T. (2012). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*, *36*(3), 214–231.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In E. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 4: Contributions to biology and problems of medicine* (pp. 321–333). University of California Press.

Roskam, E. E. (1997). Models for speed and time-limit tests. In Wim J. Linden and Ronald K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). Springer.

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *35*(1), 139–139.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics*, *28*(4), 295–313.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stenner, A. J. (2022). Measuring reading comprehension with the lexile framework. In William P. Fisher Jr. and Paula J. Massengill (Eds.), *Explanatory models, unit standards, and personalized learning in educational measurement: Selected papers by a. Jackson stenner* (pp. 63–88). Springer.

Sun, D., Tsutakawa, R. K., & He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica*, *11*(1), 77–95.

Swartz, C. W., Hanlon, S. T., Childress, E. L., & Stenner, A. J. (2016). An approach to design-based implementation research to inform development of edsphere*: A brief history about the evolution of one personalized learning platform. In Y. Rosen, S. Ferrara and M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 284–318). IGI Global.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). Academic Press.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, *55*(3), 425–453.

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347.

Verhagen, J., & Fox, J.-P. (2013). Longitudinal measurement in health-related surveys. A bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine*, *32*(17), 2988–3005.

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

Wang, C., Weiss, D. J., & Su, S. (2019). Modeling response time and responses in multidimensional health measurement. *Frontiers in Psychology*, *10*, 51.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

Wang, T. (2006). *A model for the joint distribution of item response and response time using a one-parameter weibull distribution*. (Tech. Rep.). CASMA Research Report.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323–339.

Wang, T., & Zhang, J. (2006). Optimal partitioning of testing time: Theoretical properties and practical implications. *Psychometrika*, *71*(1), 105–120.

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, *7*(1), 126–153.

West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models*. Springer.

Yao, H., Kim, S., Chen, M.-H., Ibrahim, J. G., Shah, A. K., & Lin, J. (2015). Bayesian inference for multivariate meta-regression with a partially observed within-study sample covariance matrix. *Journal of the American Statistical Association*, *110*(510), 528–544.

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

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. Wiley.

Zhan, P., Jiao, H., Man, K., Wang, W.-C., & He, K. (2021). Variable speed across dimensions of ability in the joint model for responses and response times. *Frontiers in Psychology*, *12*, 909.

Zhang, X., Tao, J., Wang, C., & Shi, N.-Z. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-based indices. *Journal of Educational Measurement*, *56*(1), 3–27.

A. Appendix: Posterior sampling schemes

A.1.

We proceed with the MCMC steps by the Gibbs sampler, where we first derive the full conditional posteriors of each unknown parameter.

Step 1: **Sampling Y: Truncated normal distribution sampling**

Given θ, φ, η, γ, and X, the latent variables $\{Y_{i,t,s,l}\}$ are sampled from

$$Y_{i,t,s,l} \sim \mathcal{N}_+(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 1,$$
$$Y_{i,t,s,l} \sim \mathcal{N}_-(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 0,$$

where $\mathcal{N}_+(\cdot,\cdot)$ means the normal distribution truncated at the left by zero, while $\mathcal{N}_-(\cdot,\cdot)$ is the normal distribution truncated at the right by zero.

Step 2: **Sampling θ: Depending on the Choice of $L(\cdot)$**

Step 2.1: $L(\cdot) = \cdot$, **Forward Filtering and Backward Sampling (FFBS).**

Define $\lambda_{i,t} = \theta_{i,t} - \rho^{-1}$, $g_{i,t} = 1 - c_i\rho\Delta^+_{i,t}$, $Z_{i,t,s,l} = Y_{i,t,s,l} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s} - \rho^{-1}$, and $H_{i,t,s} = \log(R_{i,t,s}) - \mu_i + \nu_{i,t} + \beta(a_{i,t,s} - \rho^{-1})$, then the (conditional) one-parameter DIR-RT model will fit the framework of dynamic linear model (West & Harrison, 1997), i.e.,

$$\text{System Equation: } \lambda_{i,t} = g_{i,t}\lambda_{i,t-1} + w_{i,t}, \tag{A.1}$$
$$\text{Observation Equation: } Z_{i,t,s,l} = \lambda_{i,t} + \xi_{i,t,s,l}, \tag{A.2}$$
$$H_{i,t,s} = \beta\lambda_{i,t} + \zeta_{i,t,s}, \tag{A.3}$$

where $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\xi_{i,t,s,l} \sim \mathcal{N}(0, 4\gamma^2_{i,t,s,l} + \sigma^2)$, and $\zeta_{i,t,s} \sim \mathcal{N}(0, \varrho^{-1})$. Denote information available on the tth day as

$$\mathcal{F}_{i,t} = \left\{g_{i,e}, \phi, \beta, \psi_{i,e,1,1}, \ldots, \psi_{i,e,S_{i,e},K_{i,e,S_{i,e}}}, \varrho, H_{i,e,1}, \ldots, H_{i,e,S_{i,e}}, Z_{i,e,1,1}, \ldots, Z_{i,e,S_{i,e},K_{i,e,S_{i,e}}}\right\}_{e=1}^{t}.$$

The FFBS algorithm can be implemented to block update each $\lambda_i = (\lambda_{i,0}, \ldots, \lambda_{i,T_i})'$:

1. (Forward Filtering) For $t \geq 1$, it is not hard to show that $[\lambda_{i,t} \mid \mathcal{F}_{i,t}] \sim \mathcal{N}(\mu_{i,t}, V_{i,t})$, with $\mu_{i,t} = V_{i,t}(R^{-1}_{i,t}d_{i,t} + \sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}} Z_{i,t,s,l}\psi_{i,t,s,l} + \varrho\beta\sum_{s=1}^{S_{i,t}} H_{i,t,s})$ and $V_{i,t} = \left(R^{-1}_{i,t} + \sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}}\psi_{i,t,s,l} + \varrho\beta^2 S_{i,t}\right)^{-1}$. Notice that when $t = 0$, $\lambda_{i,0}$ follows $\mathcal{N}(\mu_{i,0}, V_{i,0})$ with $\mu_{i,0} = \mu_{G_j} - \rho^{-1}$ and $V_{i,0} = V_{G_j}$.
2. (Backward Sampling) Save all quantities of $\mu_{i,t}$ and $V_{i,t}$. Then, draw λ_{i,T_i} from $\mathcal{N}(\mu_{i,T_i}, V_{i,T_i})$. When $t = (T_i - 1)$ to 0, with some algebra, we can see $\lambda_{i,t}$ will be drawn from $[\lambda_{i,t} \mid \lambda_{i,t+1}, \mathcal{F}_{i,t}] \sim \mathcal{N}(h_{i,t}, m_{i,t})$, where $h_{i,t} = m_{i,t}(V^{-1}_{it}\mu_{i,t} + \phi g_{i,t+1}\Delta^{-1}_{i,t+1}\lambda_{i,t+1})$ and $m_{i,t} = (\phi g^2_{i,t+1}\Delta^{-1}_{i,t+1} + V^{-1}_{i,t})^{-1}$.

Thus, for $t = 0, \ldots, T_i$, set $\theta_{i,t} = \lambda_{i,t} + \rho^{-1}$ and θ_i is sampled as a whole by noticing the identity $\Pr(\theta_i \mid \mathcal{F}_{i,T_i}) = \Pr(\theta_{i,T_i} \mid \mathcal{F}_{i,T_i})\Pr(\theta_{i,T_i-1} \mid \theta_{i,T_i}, \mathcal{F}_{i,T-1}) \ldots \Pr(\theta_{i,0} \mid \theta_{i,1}, \mathcal{F}_{i,0})$.

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

Step 2.2: $L(\cdot) = |\cdot|$, **Conditional mixture of truncated normal distribution**

Consider ϕ, c, Y, φ, η, γ, μ, ν, ϱ, and β are given. Similarly, the (conditional) one-parameter DIR-RT model fits the framework of state-space models as shown in Equation (A.1), Equation (A.2) and with Equation (A.3) becomes

$$H^*_{i,t,s} = \beta|\lambda_{i,t} - q_{i,t,s}| + \zeta_{i,t,s},$$

where $q_{i,t,s} = a_{i,t,s} - \rho^{-1}$ and $H^*_{i,t,s} = \log(R_{i,t,s}) - \mu_i + \nu_{i,t}$. Then, a Gibbs algorithm to sample $\lambda_{i,0}, \ldots, \lambda_{i,T_i}$ is designed below.

Let M denote the number of MCMC iterations. After some mathematical derivations, for $t = 1, \ldots, T_i - 1$, $\lambda^{(M)}_{i,t}$ is drawn from a mixture of truncated normal distribution, i.e.,

$$\lambda^{(M)}_{i,t} \sim \Pr(\lambda_{i,t} \mid \lambda^{(M)}_{i,t-1}, \lambda^{(M-1)}_{i,t+1}, \mathcal{F}_{i,T_i}),$$

where $\Pr(\lambda_{i,t} \mid \lambda^{(M)}_{i,t-1}, \lambda^{(M-1)}_{i,t}, \mathcal{F}_{i,T_i}) = \sum_{s=0}^{S_{i,t}} p_{i,t,s} \mathcal{N}_{(q_{i,t,s}, q_{i,t,s+1}]}(d_{i,t,s}, R_{i,t})$ with $q_{i,t,0} = -\infty$, $q_{i,t,S_{i,t}+1} = \infty$, $p_{i,t,s}$ defined as $p_{i,t,s} = \frac{\Phi\left(\frac{q_{i,t,s}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right) - \Phi\left(\frac{q_{i,t,s+1}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right)}{\sum_{s=0}^{S_{i,t}} \left(\Phi\left(\frac{q_{i,t,s}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right) - \Phi\left(\frac{q_{i,t,s+1}-d_{i,t,s}}{\sqrt{R_{i,t}}}\right)\right)}$, and $m_{i,t} = \phi\Delta^{-1}_{i,t} g_{i,t} \lambda^{(M)}_{i,t-1} + \phi\Delta^{-1}_{i,t+1} g_{i,t+1} \lambda^{(M-1)}_{i,t+1} + \sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l} Z_{i,t,s,l} + \varrho \sum_{s=1}^{S_{i,t}} \beta^2 q_{i,t,s}$, $d_{i,t,s} = R_{i,t}\left(m_{i,t} + \varrho\beta\left(\sum_{j=0}^{s} H^*_{i,t,j} - \sum_{j=s}^{S_{i,t}} H^*_{i,t,j}\right)\right)$, $s = 0, \ldots, S_{i,t}$, $H_{i,t,0} = 0$, and $R_{i,t} = \left(\phi\Delta^{-1}_{i,t} + \phi\Delta^{-1}_{i,t+1} g^2_{i,t+1} + \sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l} + \varrho\beta^2 S_{i,t}\right)^{-1}$. The formula is almost the same for sampling λ_{i,T_i} with only deleting the terms involving the index of $t+1$ in $m_{i,t}$ and $R_{i,t}$ and similarly, we delete the terms involving the index of $t-1$ in $m_{i,t}$ and $R_{i,t}$ for $\lambda_{i,0}$. At the end, set $\theta_{i,t} = \lambda_{i,t} + \rho^{-1}$, for $t = 1, \ldots, T_i$.

Step 3: **Sampling** c: **Truncated normal distribution sampling**

When θ and ϕ are given, the full conditional distribution of c_i is a truncated normal distribution

$$c_i \sim \mathcal{N}_+\left(\frac{\sum_{t=1}^{T_i}(1-\rho\theta_{i,t-1})(\theta_{i,t}-\theta_{i,t-1})\Delta^+_{i,t}\Delta^{-1}_{i,t}}{\sum_{t=1}^{T_i}(\Delta^+_{i,t}(1-\rho\theta_{i,t-1}))^2\Delta^{-1}_{i,t}}, \frac{1}{\phi\sum_{t=1}^{T_i}(\Delta^+_{i,t}(1-\rho\theta_{i,t-1}))^2\Delta^{-1}_{i,t}}\right).$$

Step 4: **Sampling** η: **Multivariate normal distribution sampling**

Provided θ, φ, τ, Y, and γ are given, if $S_{i,t} = 1$, $\eta_{i,t,S_{i,t}} = 0$, while if $S_{it} > 1$, then

$$\eta^*_{i,t} \sim \mathcal{N}_{S_{i,t}-1}\left((A^T_{i,t}\Sigma^{-1}_{\psi_{i,t}}A_{i,t} + \tau_i\Sigma^{-1}_{i,t})^{-1}A^T_{i,t}\Sigma^{-1}_{\psi_{i,t}}Y^*_{i,t}, (A^T_{i,t}\Sigma^{-1}_{\psi_{i,t}}A_{i,t} + \tau_i\Sigma^{-1}_{i,t})^{-1}\right),$$

is the multivariate normal distribution, where $Y^*_{i,t} = (Y^*_{i,t,1}, \ldots, Y^*_{i,t,S_{i,t}})'$ with $Y^*_{i,t,s} = (Y_{i,t,s,1} - h_{i,t,s}, \ldots, Y_{i,t,s,K_{i,t,s}} - h_{i,t,s})$ and $h_{i,t,s} = \theta_{i,t} - a_{i,t,s} + \varphi_{i,t}$, $\Sigma^{-1}_{\psi_{i,t}} = \mathrm{diag}((\psi_{i,t,1,1}, \ldots, \psi_{i,t,S_{i,t},K_{i,t,S_{i,t}}})')$, $A_{i,t} = \left(\bigoplus_{s=1}^{S_{i,t}-1} \mathbf{1}'_{K_{i,t,s}}, -\mathbf{J}_{K_{i,t,S_{i,t}} \times (S_{i,t}-1)}\right)'$, $\oplus$ is a direct sum and diag implies a diagonal matrix. Set $\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s}$.

Step 5: **Sampling** τ: **Gamma distribution sampling**

Let $\mathcal{G}a(a,b)$ denote the gamma distribution with shape parameter a and rate parameter b. Given η, the full conditional distribution of τ_i is the gamma distribution

$$\tau_i \sim \mathcal{G}a\left(\frac{\sum_{t=1}^{T_i} S_{i,t} - (T_i+1)}{2}, \frac{\sum_{t=1}^{T_i} \eta^{*\prime}_{i,t}\Sigma^{-1}_{i,t}\eta^*_{i,t}}{2}\right).$$

Step 6: **Sampling** φ: **Normal distribution sampling**

Given θ, η, Y, and γ, the full conditional distribution of $\varphi_{i,t}$ is the normal distribution

$$\varphi_{i,t} \sim \mathcal{N}\left(\frac{\sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \eta_{i,t,s})}{\sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l} + \delta_i}, \frac{1}{\sum_{s=1}^{S_{i,t}}\sum_{l=1}^{K_{its}} \psi_{i,t,s,l} + \delta_i}\right).$$

Step 7: **Sampling** δ: **Gamma distribution sampling**

When φ is given, the full conditional distribution of δ_i is the gamma distribution

$$\delta_i \sim \mathcal{G}a\left(\frac{T_i - 1}{2}, \frac{\sum_{t=1}^{T_i}\varphi^2_{i,t}}{2}\right).$$

Step 8: **Sampling** ϕ: **Gamma distribution sampling**

When θ, c are given, the full conditional distribution of ϕ is the gamma distribution

$$\phi \sim \mathcal{G}a\left(\frac{\sum_{i=1}^{n} T_i - 1}{2}, \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} \Delta^{-1}_{i,t}(\theta_{i,t} - \theta_{i,t-1} - c_i(1-\rho\theta_{i,t-1})\Delta^+_{i,t})^2}{2}\right).$$

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

Step 9: **Sampling γ: Metropolis–Hastings (MH) sampling**

Given Y, θ, φ, and η, the full conditional distribution of $\gamma_{i,t,s,l}$ is not in a closed form. Thus, we resort to an MH scheme to sample this distribution. A suitable proposal for sample γ is K-S distribution itself. We first sample γ from the K-S distribution and then let

$$\gamma_{i,t,s,l}^{(M)} = \begin{cases} \gamma^{*}, & \text{with probability} \quad \min(1, LR) \\ \gamma_{i,t,s,l}^{(M-1)}, & \text{otherwise} \end{cases}$$

where, given Y, θ, φ, and η,

$$LR = \sqrt{\frac{\sigma^2 + 4(\gamma_{i,t,s,l}^{(M-1)})^2}{\sigma^2 + 4(\gamma^{*})^2}} \exp\left\{ -\frac{(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2} \times \left(\frac{1}{\sigma^2 + 4(\gamma^{*})^2} - \frac{1}{\sigma^2 + 4(\gamma_{i,t,s,l}^{(M-1)})^2} \right) \right\}.$$

Step 10: **Sampling μ: Normal distribution**

Given ϱ, ν, θ, β, the full conditional distribution of μ_i is the normal distribution:

$$\mu_i \sim \mathcal{N}\left(\frac{\sum_{t=1}^{T_i} \sum_{s=1}^{S_{i,t}} [\log(R_{i,t,s}) + \nu_{i,t} - \beta L(\theta_{i,t} - a_{i,t,s})]}{\sum_{t=1}^{T_i} S_{i,t}}, \frac{1}{\varrho \sum_{t=1}^{T_i} S_{i,t}} \right).$$

Step 11: **Sampling ν: Normal distribution**

Given ϱ, μ, θ, β, the full conditional distribution of $\nu_{i,t}$ is the normal distribution:

$$\nu_{i,t} \sim \mathcal{N}\left(\frac{\sum_{s=1}^{S_{i,t}} - \varrho [\log(R_{i,t,s}) - \mu_i - \beta L(\theta_{i,t} - a_{i,t,s})]}{\varrho S_{i,t} + \kappa_i}, \frac{1}{\varrho S_{i,t} + \kappa_i} \right).$$

Step 12: **Sampling ϱ: Gamma distribution**

Given μ, ν, θ, β, the full conditional distribution of ϱ is the gamma distribution:

$$\varrho \sim \mathcal{G}a\left(\frac{\sum_{i=1}^{n} \sum_{t=1}^{T_i} S_{i,t} - 1}{2}, \frac{\sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{s=1}^{S_{i,t}} [\log(R_{i,t,s}) - \mu_i + \nu_{i,t} - \beta L(\theta_{i,t} - a_{i,t,s})]^2}{2} \right).$$

Step 13: **Sampling κ: Gamma distribution**

Given ν, the full conditional distribution of κ_i is the gamma distribution:

$$\kappa_i \sim \mathcal{G}a\left(\frac{T_i - 1}{2}, \frac{\sum_{t=1}^{T_i} \nu_{i,t}^2}{2} \right).$$

Step 15: **Sampling β: Normal distribution**

Given ϱ, ν, μ, and θ the full conditional distribution of β is

$$\beta \sim \mathcal{N}\left(\frac{\sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{s=1}^{S_{i,t}} L(\theta_{i,t} - a_{i,t,s}) [\log(R_{i,t,s}) - \mu_i + \nu_{i,t}]}{\sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{s=1}^{S_{i,t}} (\theta_{i,t} - a_{i,t,s})^2}, \frac{1}{\varrho \sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{s=1}^{S_{i,t}} (\theta_{i,t} - a_{i,t,s})^2} \right).$$

If we have the prior for β being positive, then β is drawn from the same normal distribution above but just truncated by zero on the left.

B. Appendix: Formula to compute partial DIC in DIR-RT models

In this section, we derive the formula of partial DIC discussed in Section 5.1. Let $\mathbf{\Theta} = vec(\varrho, \kappa, \beta, \mu)$ define the vectorized parameters and $R_{i,t}^{\star} = (\log R_{i,t,1} \ldots, \log R_{i,t,S_{i,t}})'$ and $\mu_{i,t} = (\mu_i + \beta L(\theta_{i,t} - a_{i,t,1}), \ldots, \mu_i + \beta L(\theta_{i,t} - a_{i,t,S_{i,t}}))'$. Then, for any $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$, we have $R_{i,t}^{\star} \overset{ind}{\sim} \mathcal{N}_{S_{i,t}}(\mu_{i,t}, \Omega_{i,t})$, where *ind* indicates independent and $\Omega_{i,t} = \kappa_i^{-1} \mathbf{1}_{S_{i,t}} \mathbf{1}_{S_{i,t}}' + \varrho^{-1} \mathbf{I}_{S_{i,t}}$. The partial likelihood of DIR-RT models based only on response times is $\mathcal{L}(\mathbf{\Theta} \mid R^{\star}, \theta, L(\cdot)) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} f(R_{i,t}^{\star} \mid \mathbf{\Theta}, \theta, L(\cdot))$, where $f(R_{i,t}^{\star} \mid \mathbf{\Theta}, \theta, L(\cdot))$ follows a multivariate normal probability density and $R^{*} = \{R_{i,t}^{*}\}$. According to (5.1), the partial DIC of the response times part in DIR-RT models is

$$\text{DIC}_{\text{P}} = 2\text{E}_{\pi(\mathbf{\Theta} \mid R^{\star}, \theta)} Q(\mathbf{\Theta}, R^{\star}, L(\cdot)) - Q(\bar{\mathbf{\Theta}}, R^{\star}, L(\cdot)), \tag{B.1}$$

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press

with $Q(\boldsymbol{\Theta}, R^{\star}, L(\cdot)) = -2\log \mathcal{L}(\boldsymbol{\Theta} \mid R^{\star}, \theta, L(\cdot))$, $\pi(\boldsymbol{\Theta} \mid R^{\star}, \theta)$ as the posterior distribution of Equation (2.2) given θ is known and $\overline{\boldsymbol{\Theta}}$ is the posterior median estimates of $\pi(\boldsymbol{\Theta} \mid R^{\star}, \theta)$. Using standard results on linear algebra, we have $|\Omega_{i,t}| = \varrho^{-S_{i,t}}(1 + \varrho\frac{S_{i,t}}{k_i})$, $\Omega_{i,t}^{-1} = \varrho \mathbf{I}_{S_{i,t}} - \frac{\varrho^2}{k_i + \varrho S_{i,t}} \mathbf{1}_{S_{i,t}} \mathbf{1}'_{S_{i,t}}$, and hence,

$$\begin{aligned} -2\log \mathcal{L}(\boldsymbol{\Theta} \mid R^{\star}, \theta, L(\cdot)) &= \sum_i \sum_t [(R^{\star}_{i,t} - \mu_{i,t})' \Omega^{-1}_{i,t} (R^{\star}_{i,t} - \mu_{i,t}) + log|\Omega_{i,t}| + (S_{i,t} log 2\pi)] \\ &= \varrho \sum_{i,t} (R^{\star}_{i,t} - \mu_{i,t})' (R^{\star}_{i,t} - \mu_{i,t}) - \varrho^2 \sum_{i,t} \frac{[(R^{\star}_{i,t} - \mu_{i,t})' \mathbf{1}_{S_{i,t}}]^2}{k_i + \varrho S_{i,t}} \\ &\quad + \log(\frac{2\pi}{\varrho}) \sum_{i,t} S_{i,t} + \sum_{i,t} \log(\kappa_i + \varrho S_{i,t}) - \sum_i T_i \log \kappa_i. \end{aligned} \tag{B.2}$$

Given θ as the posterior median estimates from DIR-RT models, then, employing the expression (B.2) and the MCMC samples drawn from $\pi(\boldsymbol{\Theta} \mid R^{\star}, \theta)$, we can easily calculate DIC_{P} in (B.1).

To further study on the sensitivity of the estimation of one's ability θ in the determination of the linkage, we generate DIR-RT datasets by using the same set-up as shown in Table 1 with $\varrho = 1.25$ and $\phi = 1/0.0218^2$ but with different linkage functions. Also, we vary random seeds to generate 10 different datasets under the same settings for two different linkage functions, respectively. We compare the values of θ_i equal to 1) the posterior median estimates from DIR-RT models; 2) the truth; and 3) zero. We found that the misclassification rates for the DIC_P over 1) and 2) are the same for 10 random trials (< 10%), but have much higher misclassification rates when we pick up an extreme wrong values, i.e., zero vector for θ_i. Thus, in general, with well-estimated θ_i, the proposed DIC_P should work well if the response time is assumed to be independent of the item response when given one's ability.

Cite this article: Wang, X., Saha, A. and Dey, D. K. (2026). Bayesian Joint Modeling of Response Times with Dynamic Latent Ability in Educational Testing. *Psychometrika*, 312–334. https://doi.org/10.1017/psy.2025.10019

https://doi.org/10.1017/psy.2025.10019 Published online by Cambridge University Press