Tsukuba Psychological Research
2023, **61**, 75-87

Bayesian estimation of test engagement behavior models with response times[1)]

Kazuhiro Yamaguchi[2)] (*Faculty of Human Sciences, University of Tsukuba, 305-8572, Japan*)
Kazuya Fujita[3)] (*Graduate School of Informatics, Nagoya University, 464-8601, Japan*)

Detecting non-engagement test answering behavior is a crucial task in situations where the tests are low-stakes for the individuals but the scores are employed in decision-making. Nagy and Ulitzsch (2022) proposed four test engagement models, but their estimations were conducted using maximum likelihood estimation. This study applied the Bayesian estimation method to the test engagement models. Bayesian formulation of the test engagement models was introduced and estimation was conducted using "just another Gibbs sampler" language. Real data analysis was conducted and problems were discussed regarding the future orientation of the Bayesian framework in test answering behavior modeling.

Key words: test engagement model, response time, Bayesian analysis, Markov chain Monte Carlo method

1. Introduction

Test scores may be considered as a compound of individual proficiencies, test items characteristics, and several uncontrivable error factors. Item response theory models (IRT; e.g., Embretson & Reise, 2000) are popular frameworks used to analyze academic tests. Several classical unidimensional IRT models only assume a continuous latent proficiency and item characteristic parameters such as difficulty, discrimination, and guessing parameters. Unidimensional IRT models have been employed in large scale assessments because of their simplicity.

However, such simple IRT models idealize individual item responses, and so, their model assumptions may not uphold in all testing situation. For example, test takers do not seriously take a test if the results do not affect their lives. In other words, test takers are not engaged in the test in low-stakes settings; for example, when a test is conducted to investigate current student learning states for deciding educational policies. In such cases, the test results are used by a government but do not directly affect individual test takers, and test takers have no incentive to seriously answer the items. Therefore, item responses are affected not only by academic proficiency, but also motivation for the test (e.g., Finn, 2015).

Without considering such disengaged responses and applying classical IRT models, the model parameters estimates may be biased because test taking motivation is a nuisance factor and it is neglected in the classical IRT models. If impactful policy decisions rely on such distorted results, it would be problematic because the results do not purely reflect actual individuals' proficiency. Therefore, considering test engagement behaviors is

1) Data analysis code is available in Open Science Framework page (https://osf.io/v4zk3/). We have no conflicts of interest to declare. This work was supported by JSPS KAKENHI 19H00616, 20H01720, 21H00936, and 22K13810.
2) Correspondence concerning this article should be addressed to Kazuhiro Yamaguchi, Faculty of Human Science, University of Tsukuba, Institutes of Human Sciences A314, 1-1-1 Tennnodai, Tsukuba-shi, Ibaraki-ken, 305-0006, Japan
Email: yamaguchi.kazuhir.ft@u.tsukuca.ac.jp
ORCID: https://orcid.org/0000-0001-8011-8575
3) ORCID: https://orcid.org/0000-0002-1062-3215

important to avoid misleading results.

One issue of classical IRT models is that the item responses are dichotomous and have limited information; therefore, external variables are required to model test engagement behaviors. One well-studied information regarding this is response time, and De Boeck and Jeon (2019) provided a review on the use of response times. Response times can be easily gathered in computer-based tests without huge data collecting cost (Ferrando & Lorenzo-Seva, 2007) and Programme for International Student Assessment (PISA) also recorded response time data (Organisation for Economic Co-operation and Development; OECD, n.d.). Further, response times can be used in adaptive testing in which test items are selected according to participant's responses. Considering the response time and information per time unit, measurement time can be minimized. However, this research does not study this method, but instead, focused on response times for modeling tests engagement behavior.

Response times and item responses were simultaneously modeled in the hierarchical Bayesian framework by van der Linden (2007). In addition, Ulitzsch et al. (2020) extended a joint modeling of item response and response times and modeled skip behavior. Pohl et al. (2019) employed a response time to model unreached responses. Ulitzsch et al. (2020) modeled missing responses due to low engagement with response times, and Ulitzsch, Penk, et al. (2021) provided a test-taking effort model using response times. Furthermore, Nagy and Ulitzsch (2022) formalized four types of test engagement models. While other variables, such as click streams (Ulitzsch, He, et al., 2021), may be used to provide another insight on test engagement behavior, we focused on response time in this study.

Nagy and Ulitzsch's (2022) models were based on previous theoretical assumption of test engagement behaviors, making every part of the model interpretable. The four models were classified according to dependency of engagement, latent variables, and assumption of latent continuous variables. Therefore, these models were called 1) dependent latent class IRT model with single-level relationships of response times (DLC-SL-IRT), 2) dependent latent class IRT model with two-level relationships of response times (DLC-TL-IRT), 3) independent latent class IRT model with a random effect of the latent class variable on response times (ILC-RE-IRT), and 4) independent latent class IRT model with a random intercept of response times (ILC-RI-IRT). They also provided estimation scripts with Mplus (Muthén & Muthén, 1998–2017). One common important feature of their models was the introduction of a latent engagement indicator variable; this allowed modeling of item responses and response times under engaged and disengaged cases. As a statistical model, their models were a mixture of latent variable model, and were flexible in understanding engaged behavior. However, these models required maximum likelihood estimation.

Bayesian estimation has several benefits for Nagy and Ulitzsch's (2022) models. First, disengagement may be rare and estimating parameters under disengagement with maximum likelihood could become difficult. Bayesian estimation could easily incorporate domain knowledge as prior distributions (e.g., Lee & Wagenmakers, 2013). In addition, test engagement behaviors were modeled as hierarchical models with several random effects; therefore, the maximum likelihood estimation was unstable because of a multiple integration problem. Bayesian estimation is feasible even when multiple latent factors are contained. The original parameter estimation with Mplus in Nagy and Ulitzsch's (2022) models required special data structure and did not directly represent data generating structure. Bayesian estimation, with Markov chain Monte Carlo (MCMC) which was implemented in this study, directly expressed data generating mechanisms and was easily programmed with "just another Gibbs sampler" (JAGS; Plummer, 2003) language. This also allowed model extensions to include variables that could be another information source, but it might be difficult under current maximum likelihood estimation. Finally, Bayesian formulation provided several model check methods, such as posterior predictive model check or widely applicable information criterion (WAIC; Watanabe, 2018) that have practically and theoretically sound.

This study extends test engagement models developed by Nagy and Ulitzsch (2022) to Bayesian formulation and implements Bayesian estimation with JAGS language. The next section provides

Bayesian formulation of the four models. The four types of Bayesian models were applied to real data that was analyzed by Nagy and Ulitzsch (2022) and the four models' parameter estimates were compared in the fourth section. The conclusion and further discussion are provided in the fifth section.

2. Model Formulation

2.1. Item Response Function

We borrowed basic notations from Nagy and Ulitzsch's (2022) models. The essential idea of the modeling framework of Nagy and Ulitzsch (2022) was to introduce a latent engagement indicator variable to represent the connection among the test engagement, response time, and item responses. Therefore, Nagy and Ulitzsch's (2022) models could be classified as mixture models.

Let an item response variable of an individual $i \in \{1,\cdots,I\}$ for an item $j \in \{1,\cdots,J\}$ be y_{ij}. The correct item response is represented as $y_{ij}=1$ and wrong answer is $y_{ij}=0$. In addition, latent dichotomous variable, C_{ij}, represents an engagement indicator; if the i^{th} individual was engaged for responding the j^{th} item, then $C_{ij}=1$, otherwise $C_{ij}=0$. Engaged correct item response function was modeled as a two-parameter logistic IRT model:

$$\mathrm{P}(y_{ij}=1 \mid \theta_i, a_j, b_j, C_{ij}=1) = \frac{\exp(a_j\ (\theta_i - b_j))}{1\ +\exp(a_j\ (\theta_i - b_j))}\ , \quad (1)$$

where θ_i is an individual's latent proficiency, a_j is an item discrimination parameter taking positive value, and b_j is a difficulty parameter. In this study, probability mass and density functions are represented by the same notation, $P(\cdot)$, and are distinguished based on their arguments. We also use $P(\cdot)$ to define a distribution of its argument. The item response function of Equation (1) is conditioned on latent engagement indicator. A disengaged response was modeled as random guessing:

$$P(y_{ij}=1 \mid g_j,\ C_{ij}=0) = g_j, \quad (2)$$

where $0 \le g_j \le 0.5$. This means that the item response is not dependent on latent proficiency at all in disengaged cases and the value of g_j should be sufficiently small. For simplicity, it is possible to think a common guessing: $g_j = g, \forall j$. Combining Equations (1) and (2), the item response function is defined as follows:

$$\mathrm{P}(y_{ij}=1 \mid \theta_i,\ a_j,\ b_j,\ g,\ C_{ij}) = \left\{ \frac{\exp\ (a_j(\theta_i - b_j))}{1\ +\exp(a_j(\theta_i - b_j))} \right\}^{C_{ij}} g^{1-C_{ij}}. \quad (3)$$

The complete data likelihood of an item response is expressed as follows:

$$\mathrm{P}(y_{ij} \mid \theta_i,\ a_j,\ b_j,\ g,\ C_{ij}) = \mathrm{P}(y_{ij}=1 \mid \theta_i, a_j, b_j, g, C_{ij})^{y_{ij}} \{1 - \mathrm{P}(y_{ij}=1 \mid \theta_i, a_j, b_j, g, C_{ij})\}^{1-y_{ij}}. \quad (4)$$

Assuming random sampling of individuals and conditional independence (also known as local independence), the joint likelihood of item response matrix Y whose i^{th} row and j^{th} column is y_{ij} is as follows:

$$\mathrm{P}(Y \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, g, \boldsymbol{C}) = \prod_i \prod_j \mathrm{P}(y_{ij} \mid \theta_i,\ a_j,\ b_j,\ g,\ C_{ij}), \quad (5)$$

where $\boldsymbol{\theta}$, $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{C}$ are sets of latent variables and item parameters: $\{\theta_1,\cdots,\theta_I\}$, $\{a_1,\cdots,a_J\}$, $\{b_1,\cdots,b_J\}$, and $\{C_{11},\cdots,C_{IJ}\}$. In addition to the likelihood function, prior distributions should be specified for Bayesian estimation. In this study, the prior of item difficulty parameter, $P(b_j)$, is a normal distribution with mean μ_{b_j} and variance $\sigma^2_{b_j}$, denoted as $N(\mu_{b_j},\ \sigma^2_{b_j})$. Here, subscripts on hyper-parameters represent corresponding model parameters. Similarly, the prior of item discrimination parameter, $P(a_j)$, is a truncated normal distribution whose mean and variance are μ_{a_j} and $\sigma^2_{a_j}$: $N(\mu_{a_j},\sigma^2_{a_j})I(a_{ij}>0)$, where $I(\cdot)$ is an indicator function for restricting the support of a_j as positive. The prior for guessing parameter is a truncated beta distribution: $P(g) = Beta(\alpha,\beta)I(g < 0.3)$, where $Beta(\alpha,\beta)$ is beta distribution with parameter α and β with the upper limit 0.3. This upper limit shows that the correct response under the disengaged condition should be small. Note that prior means and variances can set other values if there are sufficient empirical knowledge. The differences among the models are assumption of the distributions of laten proficiency $P(\theta_i)$, the latent engagement indicator $P(C_{ij})$, and modeling of the logarithm of item response time of an individual i for an item j, which is denoted by l_{ij}.

2.2. DLC-SL-IRT model

The DLC-SL-IRT model assumes that latent engagement is determined by response time. More formally, latent engagement probability is modeled as a logistic regression form:

$$\mathrm{P}(C_{ij}=1\,|l_{ij},\gamma,\tau_j)=\frac{\exp(\gamma(l_{ij}-\tau_j))}{1+\exp(\gamma(l_{ij}-\tau_j))}\,. \tag{6}$$

The parameter γ is a common slope parameter over items and similar to a discrimination parameter in IRT models; therefore, the γ indicates engagement sensitively associated with the log item response time. τ_j is an item-specific threshold parameter, which is analog to the difficulty parameter in IRT models, determining difficulty of engagement.

The latent proficiency parameter θ in the DLC-SL-IRT model only affects item response and does not relate to engagement. Therefore, prior latent proficiency in the DLC-SL-IRT model is the standard normal distribution for model identification: $P(\theta_i) = N(0,1)$. In addition, individual engagement tendency is not assumed in the DLC-SL-IRT model; therefore, all the information of an individual to determine engagement is contained in a single item response time. Similar to item response function, priors for γ and τ_j are assumed to be a truncated normal distribution and a normal distribution, respectively: $P(\gamma) = N(\mu_\gamma, \sigma^2_\gamma)\ I(\gamma>0)$ and $P(\tau_j) = \mathrm{N}(\mu_\tau, \sigma^2_\tau)$.

We can marginalize latent class indicator in the DLC-SL-IRT model because latent engagement indicator C_{ij} is defined for a single item response. This marginalization helps to improve the convergence of MCMC iterations. Then, the marginalized correct item response probability can be written as follows:

$$\begin{aligned}
&\mathrm{P}(y_{ij}=1\,|l_{ij},\ \theta_i,\ a_j,\ b_j,\ g,\ \gamma,\ \tau_j)\\
&=\sum_{c_{ij}=0}^{1}\mathrm{P}(y_{ij}=1\,|\theta_i,a_j,b_j,\mathrm{g},C_{ij}=c_{ij})\mathrm{P}(C_{ij}=c_{ij}|l_{ij},\gamma,\tau_j)\\
&=\mathrm{P}(y_{ij}=1\,|\theta_i,\ a_j,\ b_j,\ C_{ij}=1)\mathrm{P}(C_{ij}=1\,|l_{ij},\ \gamma,\ \tau_j)\\
&\quad+\mathrm{P}(y_{ij}=1\,|g,\ C_{ij}=0)\mathrm{P}(C_{ij}=0\,|l_{ij},\ \gamma,\ \tau_j),\\
&=\left\{\frac{\exp(a_j(\theta_i-b_j))}{1+\exp(a_j(\theta_i-b_j))}\right\}\left\{\frac{\exp(\gamma(l_{ij}-\tau_j))}{1+\exp(\gamma(l_{ij}-\tau_j))}\right\}\\
&\quad+g\left\{1-\frac{\exp(\gamma(l_{ij}-\tau_j))}{1+\exp(\gamma(l_{ij}-\tau_j))}\right\}.
\end{aligned} \tag{7}$$

In the MCMC procedure, latent engagement indicators can be gained as generated quantities if they are required.

Finally, the full conditional posterior is proportional to a product of the complete likelihood and priors:

$$\begin{aligned}
&P(\boldsymbol{\theta},\ \boldsymbol{a},\ \boldsymbol{b},\ g,\ \gamma,\ \boldsymbol{\tau}|Y,\ L)\propto\\
&\quad\mathrm{P}(Y]L,\boldsymbol{\theta},\boldsymbol{a},\boldsymbol{b},g,\gamma,\boldsymbol{\tau})P(\boldsymbol{\theta})P(\boldsymbol{a})P(\boldsymbol{b})P(g)P(\gamma)P(\boldsymbol{\tau})\\
&=\left\{\prod_i\prod_j\mathrm{P}(y_{ij}|l_{ij},\ \theta_i,\ a_j,\ b_j,\ g,\ \gamma,\ \tau_j)\right\}\left\{\prod_i P(\theta_i)\right\}\\
&\qquad\times\left\{\prod_j P(a_j)P(b_j)P(\tau_j)\right\}P(g)P(\gamma),
\end{aligned} \tag{8}$$

where $\boldsymbol{\tau}$ is a set of threshold parameter $\{\tau_1,\cdots,\tau_J\}$, and L is a log item response time matrix whose element of the i^{th} row and j^{th} column is l_{ij}.

2.3. DLC-TL-IRT model

To relax the DLC-SL-IRT model assumptions, the DLC-TL-IRT model introduces the individual engagement difficulty parameter ζ_i in Equation (6) and the engagement probability is defined as follows:

$$\mathrm{P}(C_{ij}=1\,|l_{ij},\gamma,\tau_j,\zeta_i)=\frac{\exp(\gamma[l_{ij}-(\tau_j+\zeta_i)])}{1+\exp(\gamma[l_{ij}-(\tau_j+\zeta_i)])}\,. \tag{9}$$

In this model, the individual engagement difficulty parameter ζ_i represents an engagement difference among individuals and is treated as a random effect. Small ζ_i indicates that the i^{th} individual tends to engage in test items easily. In addition, the ζ is connected to latent proficiency θ via a multivariate normal distribution:

$$\begin{aligned}
&\mathrm{P}\left(\theta_i,\ \zeta_i|\mu_{\theta\zeta}=\begin{bmatrix}\mu_\theta\\ \mu_\zeta\end{bmatrix},\ \Sigma_{\theta\zeta}=\begin{bmatrix}1 & \rho_{\theta\zeta}\sigma_\zeta\\ \rho_{\theta\zeta}\sigma_\zeta & \sigma^2_\zeta\end{bmatrix}\right)\\
&\quad=MVN(\mu_{\theta\zeta},\ \Sigma_{\theta\zeta}),
\end{aligned} \tag{10}$$

where mean vector of latent variable $\mu_{\theta\zeta}$ is set to the zero-vector for identifiability. Covariance matrix of person parameters is denoted as $\Sigma_{\theta\zeta}$. Elements of the covariance matrix are the variance parameter σ^2_ζ, representing deviation of engagement tendency, and the correlation parameter $\rho_{\theta\zeta}$ between θ and ζ which may be negative in this case because a high-proficiency person tends to easily engage in test items. In addition, the variance of θ needs to be fixed at one for the model identification.

In a traditional setting, an inverse Wishart

distribution is assumed for a prior covariance matrix, but the variance of θ needs to be fixed in this model. Therefore, the inverse Wishart distribution is not appropriate. Instead, the covariance matrix is decomposed in two lower-triangle matrices, called Cholesky decomposition (Zhan et al., 2019):

$$\Sigma_{\theta\zeta} = \Delta_{\theta\zeta}\, \Delta_{\theta\zeta}^{\top}, \tag{11}$$

$$\Delta_{\theta\zeta} = \begin{bmatrix} 1 & 0 \\ \phi & \psi \end{bmatrix}, \tag{12}$$

where, priors are set for ϕ and ψ, and so, $P(\phi) = N(0,1)I(\phi<0)$ and $P(\psi) = Gamma(1,1)$ are assumed priors, and where $Gamma(\alpha,\beta)$ is a gamma distribution with shape α and rate β.

The same marginalization in the DLC-SL-IRT model is applied for the item response probability in the DLC-TL-IRT model and we get a correct response probability without latent engagement indicator which is represented as follows:

$$\begin{aligned} &\mathrm{P}(y_{ij}=1\,|l_{ij},\,\theta_i,\,\zeta_i,\,a_j,\,b_j,\,g,\,\gamma,\,\tau_j) \\ &= \sum_{c_{ij}=0}^{1} \mathrm{P}(y_{ij}=1\,|\theta_i,a_j,b_j,g,C_{ij}=c_{ij})\mathrm{P}(C_{ij}=c_{ij}|l_{ij},\zeta_i,\gamma,\tau_j) \end{aligned} \tag{13}$$

Finally, the joint posterior probability of the parameters and latent variables is expressed as follows:

$$\begin{aligned} &P(\boldsymbol{\theta},\,\boldsymbol{\zeta},\,\boldsymbol{a},\,\boldsymbol{b},\,\boldsymbol{g},\,\gamma,\,\boldsymbol{\tau}|Y,\,L) \propto \mathrm{P}(Y|L,\,\boldsymbol{\theta},\,\boldsymbol{\zeta},\,\boldsymbol{a},\,\boldsymbol{b},\,\boldsymbol{g},\,\gamma,\,\boldsymbol{\tau}) \\ &\quad \times P(\boldsymbol{\theta},\,\boldsymbol{\zeta})P(\boldsymbol{a})P(\boldsymbol{b})P(g)P(\gamma)P(\boldsymbol{\tau}) \\ &= \left\{\prod_i\prod_j \mathrm{P}(y_{ij}|l_{ij},\,\theta_i,\,\zeta_i,\,a_j,\,b_j,\,\mathrm{g},\,\gamma,\,\tau_j)\right\}\left\{\prod_i P(\theta_i,\,\zeta_i)\right\} \\ &\quad \times \left\{\prod_j P(a_j)P(b_j)P(\tau_j)\right\}P(g)P(\gamma)P(\phi)\mathrm{P}(\psi), \end{aligned} \tag{14}$$

where ζ is a set of individual engagement difficulty parameters; $\{\zeta_1,\cdots,\zeta_I\}$.

2.4. ILC-RE-IRT model

The DLC-IRT models assume that latent engagement indicators are dependent on response times. In the ILC-IRT models, latent engagement indicators are not directly dependent on response times but connected through individual latent variables. One latent variable, η_i, that is represented as an individual engagement tendency and an item engagement difficulty parameter, κ_j, define latent engagement probability with one parameter logistic IRT model:

$$\mathrm{P}(C_{ij}=1\,|\eta_i,\kappa_j) = \frac{\exp(\eta_i-\kappa_j)}{1+\exp(\eta_i-\kappa_j)}. \tag{15}$$

Similar to IRT difficulty parameters, the prior for κ_j is a normal distribution denoted as $P(\kappa_j) = N(\mu_\kappa,\sigma_{\kappa_j}^2)$. Prior for η_i is defined for later.

The latent engagement status determines not only different item responses but also different response times. In this case, two different one-factor analysis models are assumed for the response times. In other words, we assume a mixture distribution of two normal distributions for response times:

$$\begin{cases} \mathrm{P}(l_{ij}|\tilde{\nu}_j,\,\tilde{\lambda}_j,\,\sigma_{\tilde{\epsilon}_j}^2,\,C_{ij}=0) = N(\tilde{\nu}_j+\tilde{\lambda}_j\xi_i,\,\sigma_{\tilde{\epsilon}_j}^2), \\ \mathrm{P}(l_{ij}|\nu_j,\,\lambda_j,\,\sigma_{\epsilon_j}^2,\,C_{ij}=1) = N(\nu_j+\lambda_j\xi_i,\,\sigma_{\epsilon_j}^2), \end{cases} \tag{16}$$

where ν_j and λ_j are an intercept and a factor loading parameter for engaged status, and $\tilde{\nu}_j$ and $\tilde{\lambda}_j$ are an intercept and a factor loading parameter for disengaged status. The unique factor variance parameters for engaged and disengaged statuses are $\sigma_{\epsilon_j}^2$ and $\sigma_{\tilde{\epsilon}_j}^2$. A factor score ξ_i, which can be thought of as a basic individual response speed, is common in both engaged and disengaged statuses.

In a regression formulation, Equation (16) is rewritten as follows:

$$l_{ij} = C_{ij}\;(\nu_j+\lambda_j\,\xi_i+\epsilon_{ij}) + (1-C_{ij})\,(\tilde{\nu}_j+\tilde{\lambda}_j\,\xi_i+\tilde{\epsilon}_{ij}), \tag{17}$$

where two sets of residual terms $\{\epsilon_{1j},\cdots,\epsilon_{Ij}\}$ and $\{\tilde{\epsilon}_{1j},\cdots,\tilde{\epsilon}_{Ij}\}$ are independently and identically distributed random variables followed different normal distributions: $N(0,\sigma_{\epsilon_j}^2)$ and $N(0,\sigma_{\tilde{\epsilon}_j}^2)$. Here, residual variances can be different among items. Priors are $P(\nu_j) = N(\mu_{\nu_j},\,\sigma_{\nu_j}^2)$, $P(\tilde{\nu}_j) = N(\mu_{\tilde{\nu}_j},\,\sigma_{\tilde{\nu}_j}^2)$, $P(\lambda_j) = N(\mu_{\lambda_j},\,\sigma_{\lambda_j}^2)$ $I(\lambda_j>0)$, $P(\tilde{\lambda}_j) = N(\mu_{\tilde{\lambda}_j},\,\sigma_{\tilde{\lambda}_j}^2)I(\tilde{\lambda}_j>0)$, $P(\sigma_{\epsilon_j}^2) = Gamma(\alpha_j,\,\beta_j)$, and $P(\sigma_{\tilde{\epsilon}_j}^2) = Gamma(\tilde{\alpha}_j,\tilde{\beta}_j)$. The gamma distributions here can be replaced by an inverse gamma distribution with shape α and rate β, which may be another standard choice. In this study, the hyper parameters of the gamma distributions are set as the same value, $\alpha_j=\beta_j=\tilde{\alpha}_j=\tilde{\beta}_j=1/2$. This gamma distribution is equivariant to χ^2 distribution with one degree of freedom.

An important point of the ILC-IRT models is that three types of individual parameters, θ_i, η_i, and ξ_i, are related each other. The major distribution that represents a connection among three continuous random variables is a multivariate normal distribution is as follows:

$$\mathrm{P}\left(\theta_i,\ \eta_i,\ \xi_i|\mu_{\theta\eta\xi}=\begin{bmatrix}\mu_\theta\\ \mu_\eta\\ \mu_\xi\end{bmatrix},\ \Sigma_{\theta\eta\xi}=\begin{bmatrix}1 & \rho_{\theta\eta} & \rho_{\theta\xi}\\ \rho_{\theta\eta} & 1 & \rho_{\eta\xi}\\ \rho_{\theta\xi} & \rho_{\eta\xi} & 1\end{bmatrix}\right)$$
$$=MVN(\mu_{\theta\eta\xi},\ \Sigma_{\theta\eta\xi}), \tag{18}$$

where $\mu_{\theta\eta\xi}$ is a mean vector and $\Sigma_{\theta\eta\xi}$ is a 3×3 positive definite covariance matrix. Again, μ_θ, μ_η, and μ_ξ are zero and diagonal elements of $\Sigma_{\theta\eta\xi}$ that are variances of three latent factors are set to one to identify the model. Therefore, $\Sigma_{\theta\eta\xi}$ is a correlation matrix rather than a covariance matrix here. Off-diagonal elements of $\Sigma_{\theta\eta\xi}$ are correlation parameters among three latent variables that are denoted as $\rho_{\theta\eta}$, $\rho_{\theta\xi}$, and $\rho_{\eta\xi}$, whose subscripts represent a combination of variables to be considered. Priors of correlation parameters are directly specified and are uniform distributions: $P(\rho_{\theta\eta})=Uniform(0,1)$, $P(\rho_{\theta\xi})=Uniform(-1,1)$, and $P(\rho_{\eta\xi})=Uniform(0,1)$. We assumed positive correlations between θ and η, and η and ξ; however, no-strong assumption was assumed between θ and ξ. Note that label switching problem need to be prevented for the general ILC-IRT models to put ordered constraints on intercepts or factor loadings.

Under conditional independence assumptions, the conditional distribution of model parameters of the general ILC-IRT models is represented as assembling the likelihood functions, individual parameters, and priors as follows:

$$\begin{aligned}
&P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\xi},\boldsymbol{C},\boldsymbol{a},\boldsymbol{b},g,\boldsymbol{\kappa},\boldsymbol{\nu},\tilde{\boldsymbol{\nu}},\boldsymbol{\lambda},\tilde{\boldsymbol{\lambda}},\boldsymbol{\sigma}_\epsilon^2,\boldsymbol{\sigma}_{\tilde\epsilon}^2,\rho_{\theta\eta},\rho_{\theta\xi},\rho_{\eta\xi}|Y,L)\\
&\propto \mathrm{P}(Y|\boldsymbol{\theta},\boldsymbol{a},\boldsymbol{b},g,\boldsymbol{C})\mathrm{P}(L|\boldsymbol{\xi},\boldsymbol{\nu},\tilde{\boldsymbol{\nu}},\boldsymbol{\lambda},\tilde{\boldsymbol{\lambda}},\boldsymbol{\sigma}_\epsilon^2,\boldsymbol{\sigma}_{\tilde\epsilon}^2,C)\\
&\quad\times P(\boldsymbol{C}|\boldsymbol{\eta},\boldsymbol{\kappa})P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\xi})\\
&\times P(\boldsymbol{a})P(\boldsymbol{b})P(\boldsymbol{\kappa})P(\boldsymbol{\nu})P(\tilde{\boldsymbol{\nu}})P(\boldsymbol{\lambda})P(\tilde{\boldsymbol{\lambda}})P(\boldsymbol{\sigma}_\epsilon^2)P(\boldsymbol{\sigma}_{\tilde\epsilon}^2)\\
&\quad\times P(g)P(\rho_{\theta\eta})P(\rho_{\theta\xi})P(\rho_{\eta\xi}),\\
&=\left\{\prod_i\prod_j \mathrm{P}(y_{ij}|\theta_i,a_j,b_j,g,C_{ij})\mathrm{P}(l_{ij}|\xi_i,\nu_j,\tilde\nu_j,\lambda_j,\tilde\lambda_j\sigma_{\epsilon_j}^2,\sigma_{\tilde\epsilon_j}^2,C_{ij})\right.\\
&\qquad\left.P(C_{ij}|\eta_i,\kappa_j)\right\}\\
&\times\left\{\prod_i P(\theta_i,\eta_i,\zeta_i)\right\}\left\{\prod_j P(a_j)P(b_j)P(\tau_j)P(\kappa_j)P(\nu_j)P(\tilde\nu_j)\right.\\
&\qquad\left.P(\lambda_j)P(\tilde\lambda_j)P(\sigma_{\epsilon_j}^2)P(\sigma_{\tilde\epsilon_j}^2)\right\}\\
&\times P(g)P(\rho_{\theta\eta})P(\rho_{\theta\xi})P(\rho_{\eta\xi}),
\end{aligned}\tag{19}$$

where $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\nu}$, $\tilde{\boldsymbol{\nu}}$, $\boldsymbol{\lambda}$, $\tilde{\boldsymbol{\lambda}}$, $\boldsymbol{\sigma}_\epsilon^2$, and $\boldsymbol{\sigma}_{\tilde\epsilon}^2$ are parameter sets corresponding to individuals' and item parameters. C_{ij} is a conditional variable on both y_{ij} and l_{ij}, and so marginalization of engagement indicator will generate dependency between y_{ij} and l_{ij}. Therefore, C_{ij} remains in the likelihood functions of the general ILC-IRT model. The general ILC-IRT model has two likelihood functions: one from item responses and the other from response times, which is different from the DLC-IRT models.

The general ILC-IRT model is over-parameterized and loses meaning of the parameters. One simple constraint is to add a disengagement situation in which set factor loadings are 0 and residual variance are the same across items:

$$\begin{aligned}\tilde\lambda_j&=0,\forall j,\\ \sigma_{\tilde\epsilon_j}^2&=\sigma_{\tilde\epsilon}^2,\forall j.\end{aligned}\tag{20}$$

This constraint means that the disengaged response times are not affected by individual response speed because a disengage response is a quick response and does not different across items. This constraint prove response times equation as follows:

$$l_{ij}=\tilde\nu_j+C_{ij}\ (\delta_j+\lambda_j\ \xi_i)+C_{ij}\ \epsilon_{ij}+(1-C_{ij})\tilde\epsilon_{ij},\tag{21}$$

where $\delta_j=\nu_j-\tilde\nu_j$. The second term in Equation (21) represents the effect of engagement on a log response time and contains the random effect ξ. The third and fourth terms are residual corresponding engagement and disengagement situations. In other representation, conditional distributions of log response time given the engagement status are as follows:

$$\begin{cases}P(l_{ij}|C_{ij}=0,\ \tilde\nu_j,\ \sigma_{\tilde\epsilon}^2)=N(\tilde\nu_j,\ \sigma_{\tilde\epsilon}^2),\\ P(l_{ij}|C_{ij}=1,\tilde\nu_j,\delta_j,\lambda_j,\xi_j\sigma_{\epsilon_j}^2)=N(\tilde\nu_j+\delta_j+\lambda_j\xi_i,\sigma_{\epsilon_j}^2).\end{cases}\tag{22}$$

The ILC-RE-IRT model assumes that an engaged response time takes longer than a disengaged response.

Finally, posterior distribution of the ILC-RE-IRT model is slightly simplified version of Equation (19):

$$\begin{aligned}
&P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\xi},\mathbf{C},\boldsymbol{a},\boldsymbol{b},g,\boldsymbol{\kappa},\boldsymbol{\delta},\tilde{\boldsymbol{\nu}},\boldsymbol{\lambda},\boldsymbol{\sigma}_\epsilon{}^2,\sigma_{\tilde\epsilon}^2,\rho_{\theta\eta},\rho_{\theta\xi},\rho_{\eta\xi}|Y,L)\\
&\propto \mathrm{P}(Y|\boldsymbol{\theta},\boldsymbol{a},\boldsymbol{b},g,\boldsymbol{C})\mathrm{P}(L|\boldsymbol{\xi},\boldsymbol{\delta},\tilde{\boldsymbol{\nu}},\boldsymbol{\lambda},\boldsymbol{\sigma}_\epsilon{}^2,\sigma_{\tilde\epsilon}^2,\boldsymbol{C})\\
&\quad\times P(\boldsymbol{C}|\boldsymbol{\eta},\boldsymbol{\kappa})P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\xi})\\
&\times P(\boldsymbol{a})P(\boldsymbol{b})P(\boldsymbol{\kappa})P(\boldsymbol{\delta})P(\tilde{\boldsymbol{\nu}})P(\boldsymbol{\lambda})P(\boldsymbol{\sigma}_\epsilon^2)P(g)P(\sigma_{\tilde\epsilon}^2)\\
&\quad\times P(\rho_{\theta\eta})P(\rho_{\theta\xi})P(\rho_{\eta\xi})
\end{aligned}$$

$$= \left\{ \prod_i \prod_j \mathrm{P}(y_{ij}|\theta_i, a_j, b_j, g, C_{ij}) \mathrm{P}(l_{ij}|\delta_j, \tilde{\nu}_j, \lambda_j \sigma^2_{\epsilon_j}, \sigma^2_{\tilde{\epsilon}}, C_{ij}) P(C_{ij}|\eta_i, \kappa_j) \right\} \times \left\{ \prod_i P(\theta_i, \eta_i, \zeta_i) \right\} \left\{ \prod_j P(a_j) P(b_j) P(\tau_j) P(\kappa_j) P(\delta_j) P(\tilde{\nu}_j) P(\lambda_j) P(\sigma^2_{\epsilon_j}) \right\} \times P(g) P(\sigma^2_{\tilde{\epsilon}}) P(\rho_{\theta\eta}) P(\rho_{\theta\xi}) P(\rho_{\eta\xi}), \quad (23)$$

where $\boldsymbol{\delta} = \{\delta_1, \cdots, \delta_J\}$.

2.5. ILC-RI-IRT model

Different constraints posed on the general ILC-IRT model provide a different ICL-IRM model. For example, equality constraints on factor loadings between engagement and disengagement conditions (i.e., $\lambda_j = \tilde{\lambda}_j, \forall j$) are possible. This provides following regression formulation of a log response time:

$$l_{ij} = \tilde{\nu}_j + \lambda_j \xi_i + C_{ij} \delta_j + C_{ij} \epsilon_{ij} + (1 - C_{ij}) \tilde{\epsilon}_{ij}, \quad (24)$$

In this formula, the random effect ξ is outside of the regression coefficient and thought of as a random intercept. In the ILC-RI-IRT model, the individual response speed has an effect even in the disengagement situation. The effect of engagement δ_j is a fixed effect and does not vary among individuals. Conditional distributions given the latent engagement indicator are normal distributions whose means and variances are different:

$$\begin{cases} P(l_{ij}|C_{ij} = 0, \tilde{\nu}_j, \lambda_j, \xi_j, \sigma^2_{\tilde{\epsilon}}) = N(\tilde{\nu}_j + \lambda_j \xi_i, \sigma^2_{\tilde{\epsilon}}), \\ P(l_{ij}|C_{ij} = 1, \tilde{\nu}_j, \delta_j, \lambda_j, \xi_j, \sigma^2_{\epsilon_j}) = N(\tilde{\nu}_j + \delta_j + \lambda_j \xi_i, \sigma^2_{\epsilon_j}). \end{cases} \quad (25)$$

Priors are the same as in ILC-RE-IRT model. Conditional posterior distribution is a representation of Equation (23) but with Equation (25) for corresponding terms.

3. Application to Real Data

3.1. Data Analysis Setup

Example data analyzed in Nagy and Ulitzsch (2022) were gained from the Programme for the International Assessment of Adult Competencies (PIAAC), which is an international large-scale assessment for adults. More detailed explanations were shown in Nagy and Ulitzsch (2022). The sample size was 637, and 20 item responses and log-response times were included the data set. The items used open response format and the correct item response probability in disengagement was expected to be close to zero. Log response times were standardized in this study.

The MCMC estimation code was written in JAGS language. Normal priors replaced the standard normal distribution. Correct response probability in disengagement prior parameters was $\alpha = 1$ and $\beta = 4$ and the upper limit was set to 0.3 to represent low correct response probability. Additional constraints to the $\tilde{\nu}$ parameters were negative and the δ parameters were constrained as positive. The $\tilde{\nu}$ parameters were average log response time in the disengagement situation and the log response times were standardized in this study so we assumed the responses were faster than general average that was zero. This assumption provided previous negative constraints on the $\tilde{\nu}$ parameters. Similarly, the δ parameters were the effects of engagement on the log response times and the engagement ought to take several times. This consideration generated that he δ parameters were positive. The number of chains were three, total MCMC iterations were 40,000, burn-in period was the first 10,000 samples, and thinning number was five. Convergence criterion was Gelman-Rubin index ($\hat{R}$; Gelman & Rubin, 1992) lower than 1.10. WAIC and posterior predictive p-value (PPP) were employed for model comparisons. Employed data, JAGS model, and estimation codes are available from Open Science Framework page: https://osf.io/v4zk3/.

3.2. Results

The $\hat{R}$s of the model parameters were less than 1.10, and so MCMC iterations were judged to be converged. The DLC-TL-IRT model (WAIC = 11430.940, *SE* = 118.911, PPP = .652) indicated a lower WAIC value than the DLC-SL-IRT model (WAIC = 11504.543, *SE* = 117.988, PPP = .697). In addition, the PPP of the DLC-TL-IRT model was closer to. 5 than in the DLC-SL-IRT model. These results suggest that the DLC-TL-IRT model was better than the DLC-SL-IRT model. WAIC of the ILC-RI-IRT (WAIC = 36037.783, *SE* = 246.879, PPP for item response = .579, PPP for log response time = .524) was smaller than that of the ILC-RE-IRT model (WAIC = 36375.162, *SE* = 241.876, PPP for

item response = .586, PPP for log response time = .519), and the PPP values of the ILC-RI-IRT and ILC-RE-IRT models were almost the same. These results were consistent with Nagy and Ulitzsch (2022)'s findings.

Table 1 shows the posterior means of response times thresholds τ in the DLC-IRT models, engagement difficulties κ in the ILC-IRT models, and averages of the posterior means of engagement probability over individuals. The absolute values of response time thresholds τ estimates between Table 1 in this study and Table 4 in Nagy and Ulitzsch (2022) were similar. However, Nagy and Ulitzsch (2022, Table 4) reported several extremes in κ estimates (e.g., absolute values greater than 5) but the results shown in Table 1 are quite moderate because priors prevented extreme estimates. Engaged response results were also similar between the current and Nagy and Ulitzsch (2022)'s study. However, the values of current estimates were smaller. This means our Bayesian estimation showed that the individuals were less engaged than the maximum likelihood estimates in Nagy and Ulitzsch (2022).

The correct response probability for disengaged status in the DLC-SL-IRT and DLC-TL-IRT models was the same [$g = .002$ ($SD = .002$)] but that of the ILC-RE-IRT and ILC-RI-IRT models were $g = .025$ ($SD = .008$) and $g = .010$ ($SD = .005$), respectively, making the values larger than the DLC-IRT models. The latent class discrimination parameters γ of the DLC-SL-IRT and DLC-TL-IRT models were $\gamma = 5.376$ ($SD = 0.432$) and $\gamma = 4.663$ ($SD = 0.430$), respectively, providing congruent results. The correlation

Table 1
Posterior means of engagement-related parameters and engagement probabilities in the four engagement models

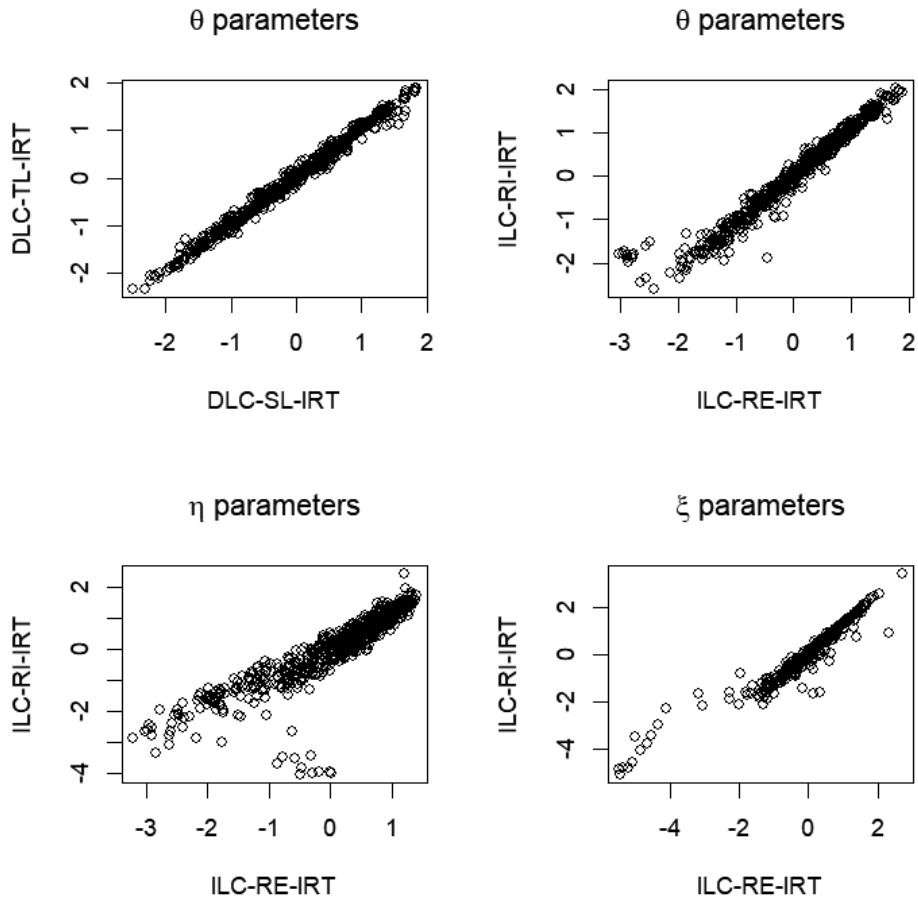
Item	Response time thresholds (τ parameters)		Engagement difficulties (κ parameters)		Average of posterior means of engagement probability			
	DLC-SL-IRT	DLC-TL-IRT	ILC-RE-IRT	ILC-RI-IRT	DLC-SL-IRT	DLC-TL-IRT	ILC-RE-IRT	ILC-RI-IRT
1	−2.429	−2.673	−3.283	−3.517	.986	.985	.938	.944
2	−2.551	−2.807	−3.329	−3.268	.989	.988	.939	.932
3	−1.583	−1.875	−3.232	−3.561	.961	.954	.935	.946
4	−2.186	−2.233	−2.676	−2.482	.978	.971	.900	.882
5	−1.331	−1.150	−1.869	−1.347	.904	.840	.825	.757
6	−1.581	−1.554	−3.010	−2.540	.943	.929	.923	.886
7	−0.572	−0.448	−1.989	−1.488	.811	.753	.839	.777
8	−1.814	−1.667	−3.032	−2.314	.956	.930	.924	.867
9	−1.452	−1.265	−2.635	−1.807	.930	.886	.897	.815
10	−1.885	−1.812	−2.705	−1.887	.957	.939	.902	.824
11	−1.902	−1.927	−2.370	−1.662	.956	.948	.876	.799
12	−1.872	−1.652	−2.819	−2.610	.946	.914	.911	.892
13	−1.637	−1.606	−2.952	−2.202	.941	.924	.919	.857
14	−1.269	−1.175	−2.582	−2.064	.905	.872	.894	.844
15	−1.421	−1.486	−2.603	−2.073	.909	.886	.895	.845
16	−0.761	−0.655	−1.588	−1.173	.809	.757	.791	.732
17	−1.034	−0.748	−1.561	−1.142	.828	.752	.787	.728
18	−0.688	−0.572	−1.543	−1.082	.812	.741	.785	.718
19	−1.469	−1.264	−2.272	−1.719	.918	.867	.867	.805
20	−1.606	−1.369	−2.612	−2.184	.929	.879	.896	.855

Note. The four models are the dependent latent class IRT model with single-level relationships of response times (DLC-SL-IRT), dependent latent class IRT model with two-level relationships of response times (DLC-TL-IRT), independent latent class IRT model with a random effect of the latent class variable on response times model (ILC-RE-IRT), and independent latent class IRT model with a random intercept of response times (ILC-RI-IRT) model.

between θ and ζ of the DLC-TL-IRT model was $\rho_{\theta\zeta} = -.971$ ($SD = .036$), implying that the high proficiency individuals could easily engage in the test items. The variance of individual engagement difficulty was $\sigma^2_\zeta = 0.234$ ($SD = 0.073$); there were individual differences of engagement difficulty. Correlation among three latent variables in the ILC-RE-IRT model were $\rho_{\theta\eta} = .483$ ($SD = .050$), $\rho_{\theta\xi} = .341$ ($SD = .044$), and $\rho_{\eta\xi} = .131$ ($SD = .045$), showing that the latent proficiency and engagement tendency, and the engagement tendency and response speed were correlated. The correlations in the ILC-RI-IRT model were $\rho_{\theta\eta} = .676$ ($SD = .050$), $\rho_{\theta\xi} = .008$ ($SD = .058$), and $\rho_{\eta\xi} = .414$ ($SD = .040$). Both ILC-RE-IRT and ILC-RI-IRT models showed correlations that were greater than .45 between the latent proficiency and engagement tendency. However, the other two correlation results were not consistent with each other. The correlation between latent proficiency and response speed $\rho_{\theta\xi}$ in the ILC-RI-IRT model was approximately zero, but it was positive for one in the ILC-RE-IRT model. The test engagement tendency and response speed $\rho_{\eta\xi}$ in the ILC-RI-IRT model was larger than in the ILC-RI-IRT model.

Figure 1 indicates scatter plots for latent proficiency θ between the DLC-SL-IRT and DLC-TL-

Figure 1. Scatter plots of latent individual parameters (θ,η, and ξ) in the four engagement models

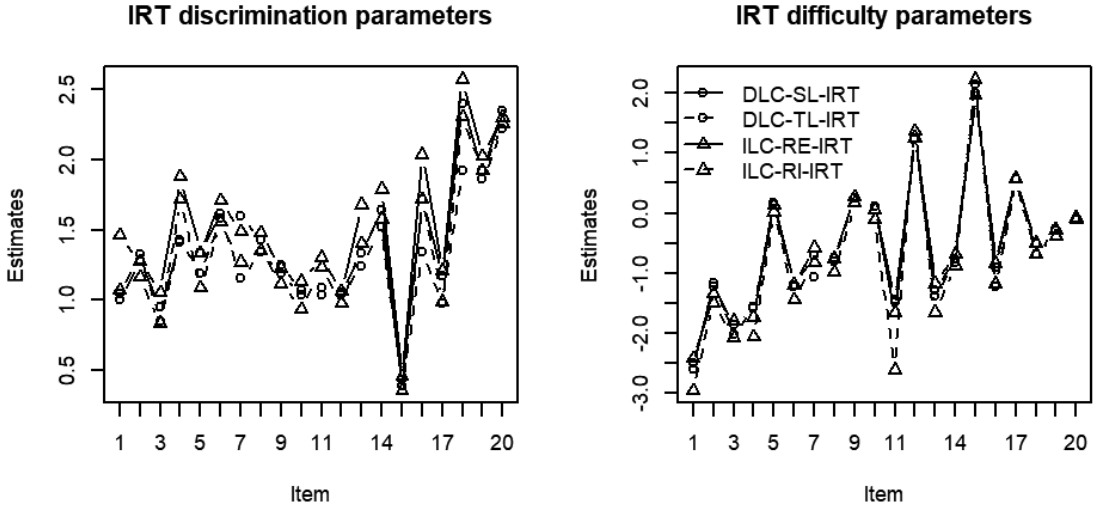


Note. The four models are the dependent latent class IRT model with single-level relationships of response times (DLC-SL-IRT), dependent latent class IRT model with two-level relationships of response times (DLC-TL-IRT), independent latent class IRT model with a random effect of the latent class variable on response times model (ILC-RE-IRT), and independent latent class IRT model with a random intercept of response times (ILC-RI-IRT) model.

IRT models (upper left panel), and three proficiencies (θ,η, and ξ) between the ILC-RE-IRT and ILC-RI-IRT models (upper right, lower left, and lower right panels, respectively). The latent proficiency between the DLC-SL-IRT and DLC-TL-IRT models was consistent and provided similar results. Comparison of the θ between the ILC-RE-IRT and ILC-RI-IRT models indicated that lower proficiency (less than −2) in the ILC-RE-IRT model was slightly highly estimated in the ILC-RI-IRT model. The range of latent engagement tendency η in the ILC-RI-IRT model was wider than that in the ILC-RE-IRT model. This implies that the ILC-RI-IRT model could capture detailed individual engagement tendency. Some individuals who took −1 to 0 values in the ILC-RE-IRT model were much smaller values (less than −3) in the ILC-RI-IRT model. The lower values (less than −2) in the response speed factor ξ in the ILC-RE-IRT model took higher values in the ILC-RI-IRT model.

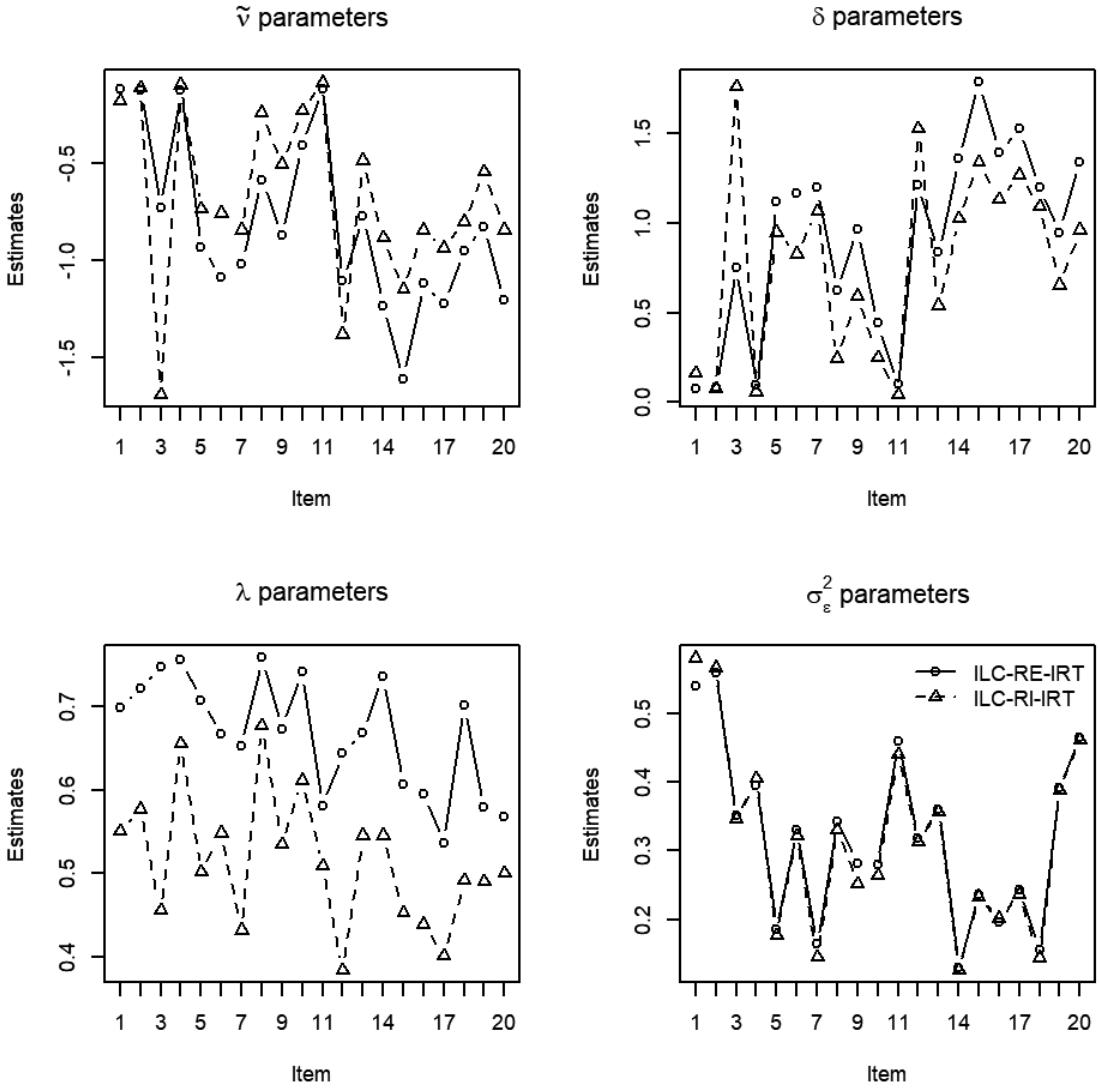
Figure 2 presents the IRT discrimination and difficulty parameters (left and right panels, respectively) estimates with four engagement models. Four models did not show significant differences in difficulty parameters. Discrimination parameters indicated slightly different estimates across several items (e.g., item 1, 4, 7, 13, 16, and 18) but systematic tendency was not shown. Figure 3 depicts response time model parameters ($\tilde{\nu}$ parameter: upper left; δ parameter: upper right; λ parameter: lower left; σ_ϵ^2 parameter: lower right) estimates with two engagement models. The $\tilde{\nu}$ parameters showed a different tendency in the ILC-RI-IRT model, which showed larger values than the ILC-RE-IRT model except for items 3 and 12. The δ parameters showed the opposite tendency: the ILC-RI-IRT model showed lower estimates than the ILC-RE-IRT model except for items 3 and 12. The λ parameters of the ILC-RE-IRT model were greater than the ILC-RI-IRT model's. The σ_ϵ^2 parameter estimates were almost the same between the two ILC-IRT models. Finally, the residual variance in the discontinuous parameters of the ILC-RE-IRT model was $\sigma_\epsilon^2 = 54.663$ ($SD = 0.430$), which was much larger than that of the ILC-RI-IRT model [$\sigma_\epsilon^2 = 1.319$ ($SD = 0.052$)].

Figure 2. IRT discrimination parameters (left panel) and difficulty parameters (right panel) estimates with the four engagement models



Note. The four models are the dependent latent class IRT model with single-level relationships of response times (DLC-SL-IRT), dependent latent class IRT model with two-level relationships of response times (DLC-TL-IRT), independent latent class IRT model with a random effect of the latent class variable on response times model (ILC-RE-IRT), and independent latent class IRT model with a random intercept of response times (ILC-RI-IRT) model.

Figure 3. Response time model parameters ($\tilde{\nu}$ parameter: upper left, δ parameter: upper right, λ parameter: lower left, σ_ε^2 parameter: lower right) estimates with the independent latent class IRT model with a random effect of the latent class variable on response times (ILC-RE-IRT), and independent latent class IRT model with a random intercept of response times (ILC-RI-IRT) models



4. Conclusion and Discussion

This study provided Bayesian formulation of four test engagement models and their likelihood functions with explicitly described priors. Furthermore, Bayesians estimation method with JAGS language was applied to PIAAC data. The real data example showed that the parameter estimates did not have extreme values and showed stable estimates. Parameter estimates similarity and differences among the models were shown.

Maximum likelihood estimation is difficult if the parameters are close to the boundaries. In such cases, maximum likelihood estimation procedure may provide unreasonable solutions. In the context of test engagement behavior, correct response probabilities in dis-engagement situation and engagement probabilism can be close to zero or

one, and it is possible for the maximum likelihood estimation to not work well. In addition, test engagement models considered in this study combined multilevel and mixture models, which are known as difficult models to estimate. Sample size of lower-stakes tests may not be large, which would make parameter estimation harder.

The Bayesian MCMC method can handle these problems. Even if the parameters are close to their boundaries, MCMC provides appropriate approximated posteriors. In addition, prior distributions work as regularization terms and prevent irregular solution. These benefits are especially important in cases with small sample sizes. Additionally, JAGS codes for estimation are simple and naturally represent data generating functions, making model extension and parameter restriction easy. For adaptive testing, posterior distribution rather than point estimates was proposed (Chang & Ying, 1996). When researchers can specify engagement or disengagement, they can consider the stopping rule (when to stop measurement) more precisely because observations with disengagement have little information regarding proficiency.

One disadvantage of the MCMC procedure is that it takes a longer time for parameter estimation and requires powerful computers that was pointed out by Nagy and Ulitzsch (2022). The estimation times of the real data example were several hours in the authors' computational environment. If the number of individuals and test items increased, Bayesian MCMC procedure will not be a good choice. Another approximation technique, such as variational Bayesian inference (Nakajima et al., 2019), will be required for larger datasets.

References

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213–229. https://doi.org/10.1177/014662169602000303

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*(FEB). https://doi.org/10.3389/fpsyg.2019.00102

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologist. Erlbaum.

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*, 525–543. https://doi.org/10.1177/0146621606295197

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series, 2015*(2), 1–17. https://doi.org/10.1002/ets2.12067

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472. https://doi.org/10.1214/ss/1177011136

Lee, M. D., & Wagenmakers, E.-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Muthén, L. K., & Muthén, B. O. (1998–2017). Mplus: Statistical analysis with latent variables: User's guide (version 8). Muthén & Muthén.

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement, 82*(5), 845–879. https://doi.org/10.1177/00131644211045351

Nakajima, S., Watanabe, K., & Sugiyama, M. (2019). *Variational Bayesian learning theory*. Cambridge University Press. https://doi.org/10.1017/9781139879354

OECD (n.d.). PISA 2018 Technical Report. Retrieved September 16, 2022, from https://www.oecd.org/pisa/data/pisa2018technicalreport/

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *The 3rd International Workshop on Distributed Statistical Computing, 124*, 1–8. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika, 84*(3), 892–920. https://doi.org/10.1007/s11336-019-09669-2

R Core Team (2022). R: A language and environment for statistical computing. R

Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika, 86*(1), 190-214. https://doi.org/10.1007/s11336-020-09743-0

Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (2021). Model meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment, 26*(2), 104-124. https://doi.org/10.1080/10627197.2020.1858786

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*(S1), 83-112. https://doi.org/10.1111/bmsp.12188

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287-308. https://doi.org/10.1007/s11336-006-1478-z

Watanabe, S. (2018). Mathematical theory of Bayesian statistics. Chapman and Hall/CRC.

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

(Received Sep. 26 : Accepted Oct. 25)