

Bridging Models of Biometric and Psychometric Assessment: A Three-Way Joint Modeling Approach of Item Responses, Response Times, and Gaze Fixation Counts

Applied Psychological Measurement
2022, Vol. 46(5) 361–381
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01466216221089344
journals.sagepub.com/home/apm



Kaiwen Man¹ , Jeffrey R. Harring², and Peida Zhan³ 

Abstract

Recently, joint models of item response data and response times have been proposed to better assess and understand test takers' learning processes. This article demonstrates how biometric information such as gaze fixation counts obtained from an eye-tracking machine can be integrated into the measurement model. The proposed joint modeling framework accommodates the relations among a test taker's latent ability, working speed and test engagement level via a person-side variance-covariance structure, while simultaneously permitting the modeling of item difficulty, time-intensity, and the engagement intensity through an item-side variance-covariance structure. A Bayesian estimation scheme is used to fit the proposed model to data. Posterior predictive model checking based on three discrepancy measures corresponding to various model components are introduced to assess model-data fit. Findings from a Monte Carlo simulation and results from analyzing experimental data demonstrate the utility of the model.

Keywords

technology enhanced assessment, joint modeling, item response theory, response times, gaze-fixation counts, eye-tracking

Introduction

Recently, the innovative technology-enhanced learning system (ITELS) has drawn a great deal of attention by researchers in educational and psychological assessment (see, e.g., Ercikan &

¹University of Alabama, Tuscaloosa, AL, USA

²University of Maryland, College Park, MD, USA

³Zhejiang Normal University, Jinhua, China

Corresponding Authors:

Peida Zhan, Department of Psychology, Zhejiang Normal University, Jinhua, Zhejiang, China.

Email: pdzhan@gmail.com

Kaiwen Man, Educational Research Program, Educational Studies in Psychology, Research Methodology, and Counseling, 313 Carmichael Box 870231, University of Alabama, Tuscaloosa, AL 35487, USA.

Email: kman@ua.edu

Pellegrino, 2017; Hao et al., 2016; Jiao & Lissitz, 2018; Man & Harring, 2019). ITELs can provide detailed feedback about test-takers' learning processes in a formative manner through built-in sensors. These sensors (e.g., eye-trackers, motion detectors, heart rate monitors) can record learners' psychological and biological reactions in real-time as they perform tasks on computer-based assessments. The collected instantaneous psychological and biological data can be used by educators in tracking individual learning status and monitoring individual/group attainment differences (e.g., gender, social and economic status)—with the primary purpose of enhancing student learning by modifying their teaching activities and re-emphasizing particular content.

Many measurement models have been proposed for concurrent investigation of different types of product (i.e., item responses) and process data (i.e., response times) in order to understand an individual's latent cognitive performance and account for the corresponding quality of measurements. Modeling item responses is typically carried out using various item response theory (IRT) models (see, e.g., Birnbaum, 1968; Rasch, 1960; Samejima, 1996). To model RTs, various response time models (see, e.g., Fox & Mariani, 2016; Rijn & Ali, 2017; van der Linden, 2007) have been developed and utilized in practice.

Item responses and RTs can also be modeled jointly. For instance, van der Linden (2007) proposed a two-factor structure model for item responses and RTs with random item and person parameters. Later, Molenaar et al. (2015) extended van der Linden (2007)'s hierarchical model into a general linear factor hierarchical model by imposing a few constraints on item parameters. This modeling extension enabled applied researchers to estimate the model using existing mainstream software for latent variable modeling like Mplus (Muthén & Muthén, 1998–2017). Recently, Man et al. (2019) proposed a joint modeling approach for multidimensional item responses and RTs. Numerous other extensions have been proposed (see, e.g., Bolsinova et al., 2017; De Boeck et al., 2017; Fox & Mariani, 2016; Klein Entink et al., 2009; Zhan et al., 2017), which have discussed the added value of modeling product and process data simultaneously.

Among the statistical benefits, these joint-modeling methods facilitate the modeling of the associations between latent cognitive abilities and reaction speeds of learners and the dependencies among item parameters—providing insights regarding learning processes of test takers and measurement features that could not be accessed from modeling each type of data independent of one another. Conventionally, these joint models follow a multilevel modeling framework (van der Linden, 2007) in which measurement components are modeled at level one. The structural association of the person-side and the item-side parameters are jointly estimated at level two. Notably, the associations of responding accuracy and working speed can be operationalized by estimating the correlations of person-side parameters; while simultaneously estimating the correlations among item parameters. Besides the obvious benefit of connecting direct product data and process data through their covariances, van der Linden et al. (2010) showed that precision of both person-side and item-side parameters was enhanced through simultaneously modeling the distinct data types as well.

Recently, Man and Harring (2021) demonstrated how eye fixation counts could be jointly modeled with RTs and item responses in a multiple-group analysis to evaluate preknowledge cheating of test-takers. Their supposition was that gaze fixation counts could be used to measure visual attention of test-takers, and in conjunction with latent ability as measured by item responses and working speed as measured by RTs, could provide insights into aberrant test-taking behavior. Generally, fixations are defined as the moment of uptake of visual information when the eyes look to be fairly motionless within a small locale of a visual target (Goldberg & Wichansky, 2003; Rayner, 1998). Because of its connection to visual attention and cognitive processing, gaze fixation has been utilized across numerous disciplines. For example, in human-computer interaction research, increased gaze fixation counts on an intriguing visual

region indicated that it was more important, more perceptible to the subject than other visual zones (Poole et al., 2004; Shagass et al., 1976). In human development, fixations have been considered as a proxy for cognitive process load in the study of infants (e.g., Aslin, 2012). In the video game industry, fixations have been used to evaluate user interface design’s effectiveness based on where and how intensive the gazes are located on the screen (e.g., Corcoran et al., 2012). Finally, fixation counts have been utilized to show word awareness in reading performance studies (e.g., Justice & Lankford, 2002).

To jointly model visual fixations with RTs and item responses, the estimated model parameters represent essential relations allowing practitioners to more fully understand how learners decode tasks in a virtual based ITELS such as game-based testing, scenario-based testing, and virtual reality based remote learning. In this study, a three-way factor analysis model is proposed for jointly modeling item responses, RTs, and visual fixation counts, which depicts the relation between responding accuracy, working speed, and test engagement. The proposed modeling approach is an extension of the Bayesian multilevel modeling framework proposed by van der Linden (2007) in which a Rasch model, an RT model, and a visual fixation counts model are specified at the measurement level. The variance-covariance structures of person-side and item-side parameters are specified at level two. Markov chain Monte Carlo methods are used to facilitate Bayesian estimation of the model and posterior predictive model checking using three discrepancy measures corresponding to each of modeling components are presented to facilitate model-data fit. An empirical example using data collected in an eye-tracking lab is provided and the estimated parameters from the real data analyses are then used to inform data generation for a small-scale simulation study to examine parameter recovery.

Multilevel Model Specification

Measurement Models

One-parameter (1-PL) logistic model. The 1-PL, or Rasch model (Lord, 1952) describes the relation between item responses and ability. This is typically specified as

$$P(u_{ij} = 1 | \theta_j; b_i) = \frac{1}{1 + e^{-(\theta_j - b_i)}}, \tag{1}$$

where $P(u_{ij} = 1 | \theta_j; b_i)$ is the probability of a correct response to item i , $i = 1, \dots, I$, by person j , $j = 1, \dots, J$; b_i is the location (difficulty) parameter for item i , and θ_j is a general latent trait for person j . The item slopes (discrimination parameters) are each fixed to unity. The proposed 1-PL model may seem overly restrictive, but is implemented here given the modest sample size of the dataset used in the empirical example. Of course, other IRT models besides the Rasch model could be employed as the sample size and item characteristics warrant.

Log-Normal RT Model. In addition to the 1-PL model, the log-normal RT model is utilized to reflect a test-taker’s working speed (van der Linden, 2006). The log-normal RT model is specified as

$$f(t_{ij} | \tau_j, \nu_i, \beta_i) = \frac{\nu_i}{t_{ij} \sqrt{2\pi}} \left(-\frac{1}{2} [\nu_i \{ \ln t_{ij} - (\beta_i - \tau_j) \}]^2 \right), \tag{2}$$

where t_{ij} denotes the RT of test-taker j on item i . The latent parameter, $\tau_j \in \Re$, represents working speed for test-taker j . The item parameter $\beta_i \in \Re$ denotes time intensity, or simply, the mean

of $\ln(t_{ij})$ when τ_j is 0. the parameter $\nu_i \in \mathfrak{R}$ is an item time discrimination parameter reflecting the dispersion of t_{ij} for item i . The mean value of $\ln(t_{ij})$ is parameterized as $\mu_{ij} = \beta_i - \tau_j$.

Negative Binomial Model. Visual fixation counts are summarized using a negative binomial fixation (NBF) model (Man & Harring, 2019; Wang, 2010), which captures the association between observed visual fixation counts and latent test visual engagement. The NBF model is expressed as

$$P(C = c_{ij}|s_i, m_i, \omega_j) = \frac{\Gamma(c_{ij} + s_i)}{c_{ij}!\Gamma(s_i)} \left(\frac{s_i}{\exp(m_i + \omega_j) + s_i} \right)^{s_i} \left(\frac{\exp(m_i + \omega_j)}{s_i + \exp(m_i + \omega_j)} \right)^{c_{ij}}, \tag{3}$$

where parameter m_i is associated with the test and can be interpreted as the visual intensity for item i . The assumption is that this parameter represents the averaged amount of cognitive engagement for test takers to finish answering an item. Person-specific parameter, ω_j for each of the J test takers ($j = 1, \dots, J$), denotes the overall test engagement level for test taker j , and is assumed, at least initially, to be constant across all the items. Furthermore, a discrimination parameter, α_i , for item i is defined as $\alpha_i = 1/\sqrt{\mu_{.i} + \mu_{.i}^2/s_i}$, where $\mu_{.i} = \sum_{j=1}^J \mu_{ij}/J$, reflecting dispersion of the fixation counts on item i .

Item Domain and Person Domain Models

The second-level models incorporate two correlational structures to account for the dependencies of both item and person parameters jointly.

Modeling Person Domain Parameters. In this joint modeling approach, the person domain covers three latent person-side variables: (1) latent ability θ , (2) working speed τ , and (3) visual engagement ω . These three latent variables for the population of test takers is posited to follow a multivariate normal distribution such that

$$\Theta_p = (\theta, \tau, \omega)^T \sim MVN(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \tag{4}$$

where superscript T is the transpose operator. The mean vector is $\boldsymbol{\mu}_p = (\mu_\theta, \mu_\tau, \mu_\omega)^T$ and the covariance matrix is formulated as

$$\boldsymbol{\Sigma}_p = \begin{pmatrix} \sigma_\theta^2 & & \\ \sigma_{\theta\tau} & \sigma_\tau^2 & \\ \sigma_{\theta\omega} & \sigma_{\tau\omega} & \sigma_\omega^2 \end{pmatrix}. \tag{5}$$

The parameters on the diagonal of the $\boldsymbol{\Sigma}_p$ show the variances of the latent constructs. The off-diagonal parameters indicate the covariances between any pairs of latent constructs. For example, the parameter, $\sigma_{\theta\tau}$ presents the covariance between latent ability and speediness of test-takers.

Modeling Item Domain Parameters. A multivariate normal distribution is also assumed for the item parameters such that

$$\Xi_I = (b, \beta, m)^T \sim MVN(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I). \tag{6}$$

The mean vector and symmetric covariance matrix, $\boldsymbol{\mu}_I$ and $\boldsymbol{\Sigma}_I$, are defined, respectively, as $\boldsymbol{\mu}_I = (\mu_b, \mu_\beta, \mu_m)^T$ and

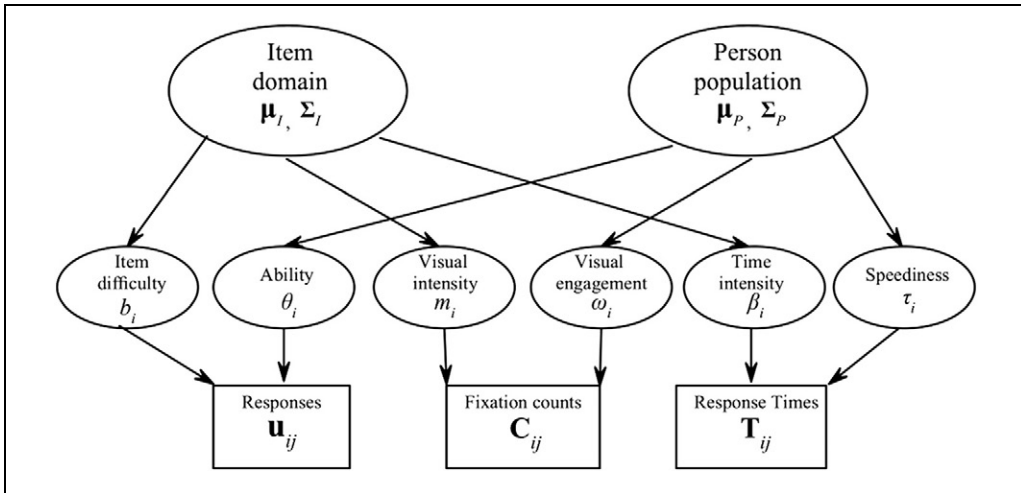


Figure 1. Three-way joint model approach of item response, response time, and visual fixation counts. μ_I , mean vector of item parameters; Σ_I , covariance of item parameters; μ_P , mean vector of person parameters; Σ_P , covariance of person parameters.

$$\Sigma_I = \begin{pmatrix} \sigma_b^2 & & & & & \\ \sigma_{b\beta} & \sigma_\beta^2 & & & & \\ \sigma_{b\omega} & \sigma_{\beta m} & \sigma_m^2 & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix}, \tag{7}$$

Figure 1 displays the graphical representation of the three-way joint modeling of item responses, response times, and visual fixation counts where μ_I denotes the mean vector of item parameters, Σ_I denotes the variances of and covariances between item parameters, μ_P is the mean vector of person parameters, and Σ_P denotes the matrix of variances of and covariances between person parameters.

Bayesian Estimation Using MCMC Sampling

A Bayesian approach was employed to facilitate model parameter estimation using *Just Another Gibbs Sampler* (JAGS; Plummer, 2015), which is housed in the *R2jags* package (Su & Yajima, 2015). The *coda* package was utilized to evaluate convergence. Two chains using 96,000 total iterations with thinning of 2 to alleviate auto-correlation among draws, were executed. Model parameter estimates and standard deviations were summarized based on the posterior densities using the final 4000 iterations after burning-in 92,000. The potential scale reduction factor (PSRF) was used for assessing convergence for all model parameters (Gelman et al., 2003). Convergence was declared when a PSRF value of at most 1.1 was reached for each model parameter.

Constraints for Model Identification

To properly identify the scales of the latent variables, model constraints are needed either on the item side (fixing the summation of item difficulties to zero) or the person side (fixing the

expectation of the latent ability parameter to zero). For model identification in this study, the scales were fixed on the person side by following the convention used for IRT model estimation (C. Wang, Fan, Chang, & Douglas, 2013; Wu et al., 1998).

For the 1-PL model, the population mean of the latent ability, θ , was set to 0 (Lord, 1952), and, the item discrimination parameter for each item was fixed to unity. For the log-normal RT model, the population mean of latent speediness, τ , was constrained to 0 as well (van der Linden, 2006). For the NBF model, the mean of the latent person side visual engagement parameter ω was also set to zero (Man & Harring, 2019).

$$\mu_\omega = \mu_\theta = \mu_\tau = 0. \tag{8}$$

Prior Distributions

In reference to equation (6), the prior distribution for item parameters, Ξ_I in the proposed model is assumed to be trivariate normal. A Gamma distribution is assumed for the time discrimination parameter [i.e., $\nu_i \sim \Gamma(1, 1)$]. This is the inverse of the variances of the log-times on different items (σ_e^2) based on the RT model: $\log(T_{ij}) \sim N(\beta_i - \tau_j, \sigma_{e_i}^2)$. In addition, the fixation dispersion parameter for each item [i.e., $s_i \sim \text{IG}(1, 1), i = 1, \dots, I$] is assumed to follow an inverse Gamma distribution as well (see, e.g., Luo & Jiao, 2018; Zhan et al., 2018). Hyperpriors are defined as

$$\mu_d \sim N(0, 0.5), \mu_\beta \sim N(4.0, 0.5), \mu_m \sim N(3.5, 1) \Sigma_I \sim IW(\mathbf{I}_I, \nu),$$

where \mathbf{I}_I is a 3×3 identity matrix, and ν is the degree of freedom, which in this case is equal to 3.

In reference to equation (4), the prior distribution for the person parameters, Θ_p follows a tri-variate normal distribution, where the μ_p are fixed to zero. And,

$$\Sigma_p = \begin{pmatrix} \sigma_\theta^2 & & \\ \sigma_{\theta\tau} & \sigma_\tau^2 & \\ \sigma_{\theta\omega} & \sigma_{\tau\omega} & \sigma_\omega^2 \end{pmatrix} \sim IW(\mathbf{I}_P, \nu).$$

The joint posterior probability for the proposed model can be represented as

$$p(\Theta_p, \Xi_I | \mathbf{u}, \log(\mathbf{T}), \mathbf{c}) \propto \prod_{i=1}^I \prod_{j=1}^J p(\mathbf{u}_{ij}, \log(\mathbf{T}_{ij}), \mathbf{c}_{ij} | \Theta_j, \Xi_i) p(\Theta_j | \mu_p, \Sigma_p) p(\Xi_i | \mu_I, \Sigma_I) p(\mu_d) p(\mu_\beta) p(\mu_m) p(\Sigma_I | \nu) p(\mu_p | 0, \Sigma_p) p(\Sigma_p | \nu)$$

where $p(\cdot | \cdot)$ indicates the conditional density function.

Posterior Predictive Model Checking based on Discrepancy Measures

Posterior predictive model-checking (PPMC; see Gelman et al., 1996; Levy, 2009), a popular Bayesian model-checking tool, will be used to evaluate whether the proposed model adequately accounts for variability (uncertainty) that exists in the data. Of its many advantages, PPMC has a strong theoretical basis and has an intuitive appeal and can be applied in a straightforward manner (Sinharay et al., 2006). To quantify the differences, a discrepancy measure, $T(\cdot)$, a function of data and model parameters, is usually computed, which summarizes the data and the corresponding model parameters (Gelman et al., 1996). A small difference is indicative of satisfactory data-model fit.

Discrepancy Measures for the Proposed Models

Three statistics are now introduced in this section. Each of the three statistics will be used as a different discrepancy measure, $T(\cdot)$, to evaluate item by person-level data-model fit for item responses, response times, and visual fixation counts, separately. Specifically, the values of $T(y^{pred}, \psi)$ and $T(y, \psi)$ will be calculated based on the predicted dataset and observed dataset based on three statistics. Then, PPP-values will be calculated based on the discrepancies between $T(y^{pred}, \psi)$ and $T(y, \psi)$. The three item-fit statistics are: (1) the W index (Wright & Stone, 1979); (2) the L index (Marianti et al., 2014); and, newly proposed (3) M index, which will be discussed in detail subsequently.

Item Response Based W Statistic. The W index is computed from performing a residual analysis from applying the Rasch model (Rasch, 1960) to a set of examinees' item responses (Wright & Stone, 1979). As a consequence of this parsimoniously parameterized model, analyses require relatively small sample sizes (i.e., the number of examinees) to produce reasonable data-model fit (Linacre & Wright, 1994). The computation of the W index follows

$$W_{ij} = \frac{[u_{ij} - P_i(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \tag{9}$$

where $P_i(\theta)$ is the probability of correctly answering item i given the ability estimate, θ , u_{ij} is the dichotomous response (0, 1) of item i for a specific person j .

RT-Based L Statistic. Marianti et al. (2014) suggested an RT-based item-fit statistic, named the L statistic. Parameters used for calculating the L statistic are estimated based on the RT model proposed by van der Linden (2006). The L statistic is formulated as

$$L_{ij} = \frac{[\ln(t_{ij}) - \beta_i + \tau_j]^2}{\sigma_{e_i}^2}, \tag{10}$$

where t_{ij} is the response time for test taker j on an item i , β_i is the time-intensity parameter that is the averaged population time required for answering that item, τ_j is the speediness parameter for each test taker, and σ_{e_i} is defined as $1/\nu_i$.

Visual Fixations Based M Statistic. To evaluate the data model fit based on the visual fixation counts, a visual fixation counts based item-fit M statistic is proposed. The M statistic is a residual-based model-fit measure, which is constructed from a summation of the variance weighted squared residuals defined as the differences between the observed outcome, c_{ij} , and predicted value, $E(c_{ij})$. (Cochran, 1952; Fox & Marianti, 2017; van der Linden & Hambleton, 1997). The M statistic is formulated as

$$M_{ij} = \frac{[c_{ij} - \exp(m_i + \omega_j)]^2}{\sigma_{ij}^2}, \tag{11}$$

where c_{ij} is the visual fixation counts for test taker j on an item i , m_i is the visual-intensity parameter that is the averaged population visual efforts required for answering that item, ω_i is the individualized visual engagement parameter, and σ_{ij}^2 is the variance of the visual fixation counts, which is defined as $\sigma_{ij}^2 = \exp(m_i + \omega_j) + (\exp(m_i + \omega_j))^2/s_i$.

Having a PPP-value close to 0 based on a discrepancy measure would indicate problematic data-model fit implying that the proposed model fails to sufficiently regenerate the data (Sinharay et al., 2006).

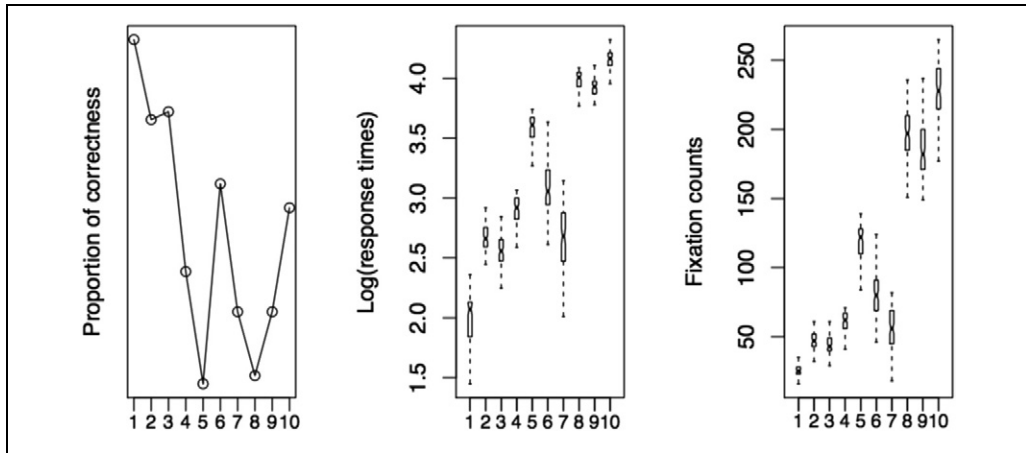


Figure 2. Visualization of item responses, response times, and visual fixation counts of 10 items.

Real Data Analysis

The proposed three-way joint model of item responses, response times and visual fixation counts were fitted to the data. Parameter estimates of the level-1 measurement models were reported. In addition, the associations of the person-side and item-side parameters at the level-2 were discussed by summarizing the corresponding variance-covariance estimates.

Data Description

Data were collected in an eye-tracking lab setting at a large university with IRB approval. A total of $n = 93$ university students who had normal or corrected vision were recruited. Students were invited to a room and seated approximately 80 cm away from a 17" monitor with an eye-tracking device, Gazepoint, placed under the screen. Gazepoint is an accessible and reliable experimental eye-tracker with 60 Hz sampling rate and 0.5–1 degree of visual angle accuracy, which is commonly used for conducting eye-tracking research. Students were asked to take a test consisting of $I = 10$ questions related to verbal reasoning. The test structure followed the structure of one section of a high-stakes credentialing exam. Item responses, response times and gaze fixation counts of the area of interest were recorded simultaneously as the participants answered the assessment questions. The position-variance method (Jacob & Karn, 2003) was the default algorithm for processing the fixation counts¹. Figure 2 displays the collected item response (transferred into the proportions of correctness), item response times and visual fixation counts side-by-side for the 10 items.

Accessing Data Model Fit Based on PPMC Method

The PPP-values were calculated based on 2000 iterations after dropping burn-in iterations with thinning of 2. The PPP-values for the three model components across the 10 items were all larger than 0.05 level indicating satisfactory data model fit. Notably, for the IRT model, the PPP-values over the 10 items were systematically lower than the ones calculated based on the RT model and NBF model although all values met the 0.05 threshold. Figure 3 summarizes the PPP-values computed over the 2000 iterations across the 10 items. The three dashed

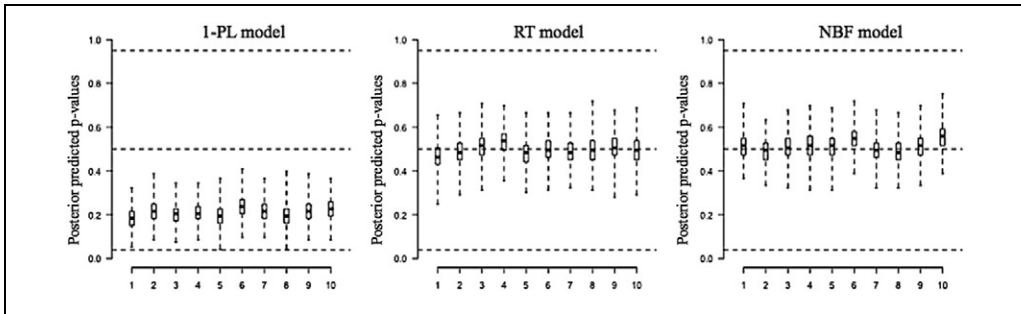


Figure 3. Posterior predictive p -values for 1-PL IRT model, log-normal response time model, and negative binomial visual fixation counts model over 10 items. The three dash horizontal lines denote 0.05, 0.5, and 0.95, respectively. The box-plots represent the item by person-level PPP values.

horizontal lines in each plot denote 0.05, 0.5, and 0.95, respectively. PPP-values lying below the dashed line at the 0.05 level would indicate non-satisfactory data-model fit. All the PPP-values in Figure 3 were above 0.05, which suggests that the proposed joint model fits the data satisfactorily even at a modest sample size.

Model Parameter Estimates

Table 1 shows the 10 item parameter estimates across the three proposed models. For the 1-PL IRT model, the item difficulty parameter estimates, \hat{b}_i , ranged from -1.02 to 0.99 . Item 8 was the most challenging item. In contrast, Item 1 was the least difficult item on the test. For the RT model, the time intensity parameter estimates, $\hat{\beta}_i$, varied from 1.92 to 4.14 . Item 10 was the most time-consuming item for test takers. In contrast, Item 1 required the least amount of time on average for test takers to answer. For the NBF model, the most visually engaging item was item 10, while item 1 required the least visual effort. Notably, the last 3 items required more time and visual effort in solving them, which met our expectations since the last three items were reading comprehension questions.

Table 2 displays the results of the estimation of the variance-covariance of item domain and population domain at a higher level (see Figure 1). This is of interest because the higher-level item domain covariance components between different item parameters show the dependencies between responses, RT, and visual fixation counts. The estimated covariance between item difficulty and item visual intensity was 0.410 (Cor. = 0.574 with a 95% credible interval of 0.013 to 0.528) suggesting that the item difficulty is positively associated with item visual intensity for the given test (see Figure 4). The estimated covariance between item difficulty and item time intensity was 0.417 (Cor. = 0.572 with a 95% credible interval of 0.292 to 0.875), which is also significant indicating that the item difficulty was positively related with item time intensity (see Figure 5). Moreover, the estimated covariance between item visual intensity and item time intensity was 0.595 (Cor. = 0.813 with a 95% credible interval of 0.190 to 0.723), which is also significant showing that the item visual intensity was positively related with item time intensity (see Figure 5).

The person-level covariances, $\sigma_{\theta,\omega}$, $\sigma_{\theta,\tau}$, and $\sigma_{\omega,\tau}$ shown in Table 2, were estimated to be -0.001 (Cor. = 0.008 , 95% credible interval: -0.023 to 0.020); 0 (Cor. = 0.003 ; 95% credible interval: -0.024 to 0.027); and -0.003 (Cor. = 0.151 ; 95% credible interval: -0.007 to 0.001), respectively (see Figure 4). Notably, all the covariance estimates were not statistically

Table 1. Item Parameter Estimates.

Model	I-PL	RT		NBFM	
Item	<i>b</i>	β	ν	<i>m</i>	α
1	-1.02 (.250)	1.92 (.046)	0.42 (.031)	3.20 (.025)	0.19 (.012)
2	-0.51 (.227)	2.64 (.026)	0.21 (.017)	3.86 (.020)	0.14 (.005)
3	-0.54 (.226)	2.59 (.029)	0.24 (.016)	3.81 (.023)	0.12 (.011)
4	0.33 (.212)	3.02 (.042)	0.36 (.021)	4.28 (.040)	0.04 (.003)
5	0.99 (.217)	3.57 (.027)	0.21 (.014)	4.77 (.018)	0.07 (.007)
6	-0.10 (.215)	3.09 (.036)	0.32 (.024)	4.42 (.028)	0.05 (.004)
7	0.51 (.233)	2.67 (.040)	0.34 (.023)	4.02 (.032)	0.07 (.005)
8	0.99 (.242)	3.98 (.025)	0.17 (.017)	5.27 (.016)	0.06 (.007)
9	0.62 (.225)	3.96 (.026)	0.22 (.013)	5.26 (.021)	0.03 (.003)
10	0.09 (.206)	4.14 (.024)	0.19 (.015)	5.42 (.016)	0.05 (.005)

Standard error (standard deviation of the posterior distribution) is in parenthesis; *b*, item difficulty; β , item time intensity; ν , item time discrimination; *m*, item engagement intensity; α , item engagement discrimination; I-PL, One-parameter logistic model; RT, log-normal Response Time model; NBFM, Negative binomial visual fixation counts model.

Table 2. Variance-covariance Estimates.

Item parameters			Person parameters		
Variance-covariance parameters			Variance-covariance parameters		
	Mean	CI		Mean	CI
σ_b^2	0.687	(0.249,0.829)	σ_θ^2	0.488	(0.246,0.825)
σ_m^2	0.728	(0.285,0.866)	σ_ω^2	0.017	(0.013,0.023)
σ_β^2	0.717	(0.282,0.861)	σ_τ^2	0.021	(0.015,0.028)
$\sigma_{b, \beta}$	0.417	(0.292,0.875)	$\sigma_{\theta\omega}$	-0.001	(-0.023,0.020)
$\sigma_{b, m}$	0.410	(0.013,0.528)	$\sigma_{\theta\tau}$	0.000	(-0.024,0.027)
$\sigma_{\beta, m}$	0.595	(0.190,0.723)	$\sigma_{\omega\tau}$	-0.003	(-0.007,0.001)

Mean, mean value of the posterior distribution; CI, credible interval; *b*, item difficulty; β , item time intensity; *m*, item engagement intensity; θ , ability; ω , visual engagement; τ , speediness.

significant, indicating that the three latent dimensions were not statistically related to each other. The non-significant results could be a proxy for a lack of motivation for students required to take this low-stakes assessment (Wise & Kong, 2005).

Overall, the results showed a number of interesting findings. First, given the current data set, the fitted three-way model shows that the three latent dimensions were not statistically significantly related to each other, which demonstrates weak associations among accuracy, working speed, and visual engagement of test-takers when their eyes were being tracked. Similarly, Lee et al. (2019) found a weak correlation between speediness and abilities when test-takers solved a set of complex simulation-based visual tasks. The weak correlation among these latent constructs may be due to factors such as content display, item types, and testing conditions. Second, the proposed model presents the special structure of the measurement features. The results show that item difficulty, time intensity, and visual intensity are positively related to each other, which indicates that more difficult items require more time and visual effort by the examinee.

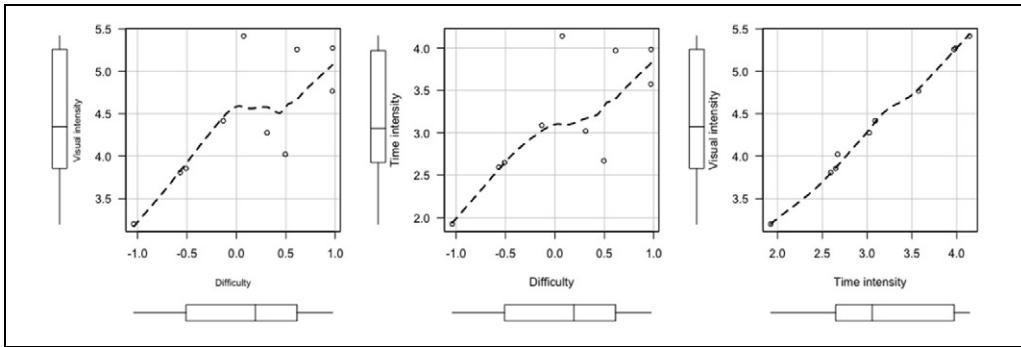


Figure 4. Scatter-plots for person parameter estimates. A loess non-parametric smoothed curve is plotted for each scatter-plot.

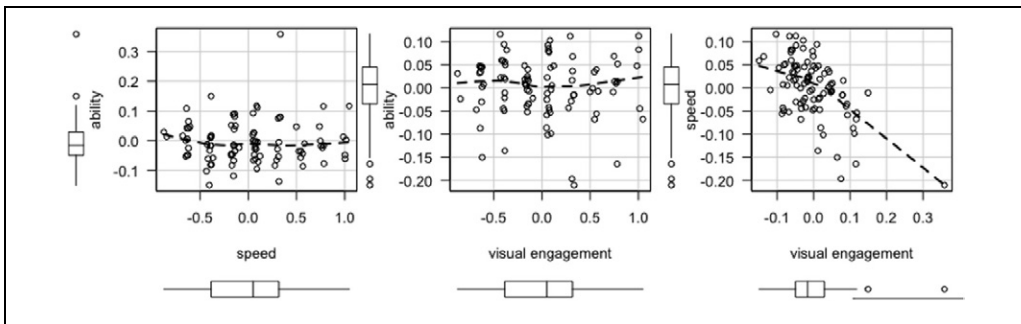


Figure 5. Scatter-plots for item parameter estimates. A loess non-parametric smoothed curve is plotted for each scatter-plot.

These results are based on a small sample size of $n = 93$ individuals, which may bring into question the statistical conclusion validity using parameters estimates and corresponding standard deviations of such a complex model. A small-scale Monte Carlo simulation study to investigate parameter recovery was executed to address this concern.

Simulation Study: Parameter Recovery

Three simulation factors were manipulated to assess parameter recovery in the three-way joint model: (1) test length, (2) sample size, and (3) correlations among person-side latent variables. Two levels of test length were considered, $I = 10$ and $I = 25$, matching previous simulation and experimental investigations (e.g., Luo, 2021). Sample sizes were set at $N = 100, 500,$ and 1000 . Correlations (ρ) among person-side latent variables were fixed as 0.3, 0.5, and 0.8, separately. These values were selected to follow past methodological investigations while simultaneously considering the possible scale of adoption of eye-tracking technology in practice. Three main item parameters were generated from a three-dimensional normal distribution based on the estimates obtained in the real data example,

$$\begin{pmatrix} b_i \\ \beta_i \\ m_i \end{pmatrix} \sim N_3 \left[\begin{pmatrix} 0 \\ 4.5 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & & \\ 0.25 & 0.25 & \\ 0.25 & 0.125 & 0.25 \end{pmatrix} \right]. \tag{12}$$

Following Man and Harring (2019), item time-discrimination parameter ν , was generated from a uniform distribution, $U(0.5, 0.8)$ and item visual dispersion parameters, s_i , were generated from an inverse-Gamma distribution, $IG(2, 6)$.

On the person-side, individual random effects were generated from the following the three-dimensional normal distribution

$$\begin{pmatrix} \theta_j \\ \tau_j \\ \omega_j \end{pmatrix} \sim N_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & & \\ 0.5 * \rho & 0.25 & \\ -0.5 * \rho & -0.25 * \rho & 0.25 \end{pmatrix} \right], \quad (13)$$

where ρ is the correlation between latent variables. Scalars (e.g., 0.25, 0.5) are chosen to assure that the variances of person-side variables are fixed as 1, 0.25, and 0.25, which mimics the empirical analysis results.

According to equation (13), the correlation between a person's ability and working speed is positive— $\rho_{\theta,\tau} > 0$ reflecting that capable test-takers were more likely to answer items correctly and faster than less able examinees. A positive correlation between a person's working speed and visual engagement level (i.e., $\rho_{\tau,\omega} > 0$) would indicate test takers who answer items for rapidly tend to be more visually engaged. Similarly, a positive correlation $\rho_{\theta,\omega} > 0$ indicates that capable test-takers who are working efficiently are more likely to gaze at items with any frequency. For instance, test-takers who are less able are more likely to reexamine items to gain insights into how to solve them. Thus, three factors were considered to evaluate parameter recovery in the proposed model: 1) sample sizes, which were set at $n = 100, 500$, and 1000 (e.g., Fox et al., 2014; Man & Harring, 2021; 2) test lengths, which were set at $I = 10$ and 25; and 3) magnitudes of the pairwise associations among person-side latent parameters, which were set at $\rho = 0.3, 0.5$, and 0.8. These levels were chosen to mimic the results from the real data analysis described previously as well as past studies (Fox & Marianti, 2016; Man & Harring, 2019)

Outcome Measures

Root mean square error (RMSE) and bias were adopted to assess model parameter recovery. There are two separate types of RMSE and bias values to be calculated: (1) one for item-side parameters $\hat{\varphi}$ and (2) one for person-side parameters $\hat{\xi}$. The RMSEs are defined as $RMSE(\hat{\xi}) = \frac{1}{R} \sum_{r=1}^R \sqrt{I^{-1} \sum_{i=1}^I (\hat{\xi}_{ir} - \xi_{ir})^2}$, and $RMSE(\hat{\varphi}) = \frac{1}{R} \sum_{r=1}^R \sqrt{J^{-1} \sum_{j=1}^J (\hat{\varphi}_{jr} - \varphi_{jr})^2}$, respectively. Bias values are denoted as $bias(\hat{\xi}) = \frac{1}{R} \sum_{r=1}^R \left\{ I^{-1} \sum_{i=1}^I (\hat{\xi}_{ir} - \xi_{ir}) \right\}$, and $bias(\hat{\varphi}) = \frac{1}{R} \sum_{r=1}^R \left\{ J^{-1} \sum_{j=1}^J (\hat{\varphi}_{jr} - \varphi_{jr}) \right\}$, where ξ is the true values generated for item i , $\hat{\xi}_{ir}$ is the parameter estimates for item i ($i = 1, \dots, I$) in replicate r , ($r = 1, \dots, R$). φ and $\hat{\varphi}_r$ are the true value of the person-side parameter and estimate from the r th replication for a pool of J test takers, respectively. In addition, distinct PPMC values were calculated to evaluate model-data fit for the 1-PL logistic model, the log-normal RT model, and the NBF model.

Simulation Results

To better understand which factors impacted model parameter recovery, a factorial three-way (sample size \times test length \times correlations among latent traits) analyses of variance (ANOVAs) were

Table 3. Factorial Three-way ANOVA with RMSE as the Outcome Variable.

Par	Number of items	Sample size	Cor
Item parameters			
<i>b</i>	0.005(0.138)	0.967(0.000)*	0.001(0.807)
β	0.004(0.305)	0.954(0.000)*	0.002(0.769)
ν	0.056(0.359)	0.104(0.450)*	0.050(0.702)
<i>m</i>	0.000(0.895)	0.970(0.000)*	0.001(0.848)
α	0.004(0.021)	0.990(0.000)*	0.001(0.825)
μ_b	0.054(0.016)*	0.831(0.000)*	0.034(0.124)
μ_β	0.317(0.000)*	0.562(0.000)*	0.002(0.888)
μ_m	0.378(0.000)*	0.394(0.001)*	0.037(0.350)
σ_b^2	0.732(0.000)*	0.029(0.047)*	0.195(0.000)*
$\sigma_{b,\beta}$	0.595(0.000)*	0.179(0.004)*	0.111(0.070)*
σ_β^2	0.916(0.000)*	0.000(0.079)	0.031(0.377)
$\sigma_{b,m}$	0.493(0.000)*	0.313(0.000)*	0.122(0.003)*
$\sigma_{\beta,m}$	0.727(0.000)*	0.181(0.001)*	0.450(0.074)*
σ_m^2	0.992(0.000)*	0.000(0.789)	0.001(0.471)
Person parameters			
θ	0.980(0.000)*	0.000(0.784)	0.014(0.001)
τ	0.512(0.002)*	0.064(0.416)*	0.018(0.772)
ω	0.828(0.000)*	0.008(0.645)	0.053(0.095)
σ_θ^2	0.001(0.711)	0.008(0.402)	0.941(0.000)*
$\sigma_{\theta\tau}$	0.042(0.107)	0.061(0.155)*	0.731(0.000)*
$\sigma_{\theta\omega}$	0.000(0.976)	0.045(0.350)	0.722(0.000)*
σ_τ^2	0.040(0.175)	0.057(0.267)*	0.671(0.000)*
$\sigma_{\omega\tau}$	0.006(0.397)	0.021(0.298)	0.880(0.000)*
σ_ω^2	0.025(0.004)	0.006(0.253)	0.946(0.000)*

b, item difficulty; β , item time intensity; ν , item time discrimination; *m*, item engagement intensity; α , item engagement discrimination; θ , ability; ω , visual engagement; τ , speediness; Cor., Correlations among person-side latent variables. Numbers listed are the main effects across all parameters and corresponding *p*-values in parentheses. Asterisk notes the ones identified to be statistically significant (*p*-value < .05)

conducted where the RMSEs were treated as the outcome variables regressed on the manipulated factors and their interactions. To examine the effects of manipulated factors, η^2 was calculated as the effect size measure. In addition to reporting the *p*-values for each effect, an effect size was interpreted to be nontrivial when $\eta^2 \geq 0.06$ (see, e.g., Bakeman, 2005; Cohen, 1988).

Table 3 summarized the effects of the manipulated factors on the item and person-side parameters. None of the interaction effects were practically important, thus, only the main effects were reported. In terms of the item parameters, the ANOVA results showed that all manipulated factors (number of items, sample size, and correlations among item parameters) were highly influential on the recovery of item means and variances. Moreover, different sample sizes showed meaningful effects on item parameter estimates. The correlations among item parameters had large effects on the recovery of variance-covariances of item parameters. All nontrivial effects ($\eta^2 \geq 0.06$) for the item parameter are plotted in Figure 6.

Figure 6 reveals that the recovery of item parameters improved as the sample size increases. For instance, the average RMSE of *b* parameters decreased from 0.239 to 0.081 as the sample size increased from 100 to 1000. Also, the recovery of time-related boundaries (i.e., β and ν) and that of visual-related parameters (i.e., *m* and α) had smaller RMSE values across conditions than did item difficulty (*b*). This result is not too surprising and enlarging the sample size would clearly yield item parameters with less bias and variability (see, e.g., Man et al., 2019; Molenaar et al.,

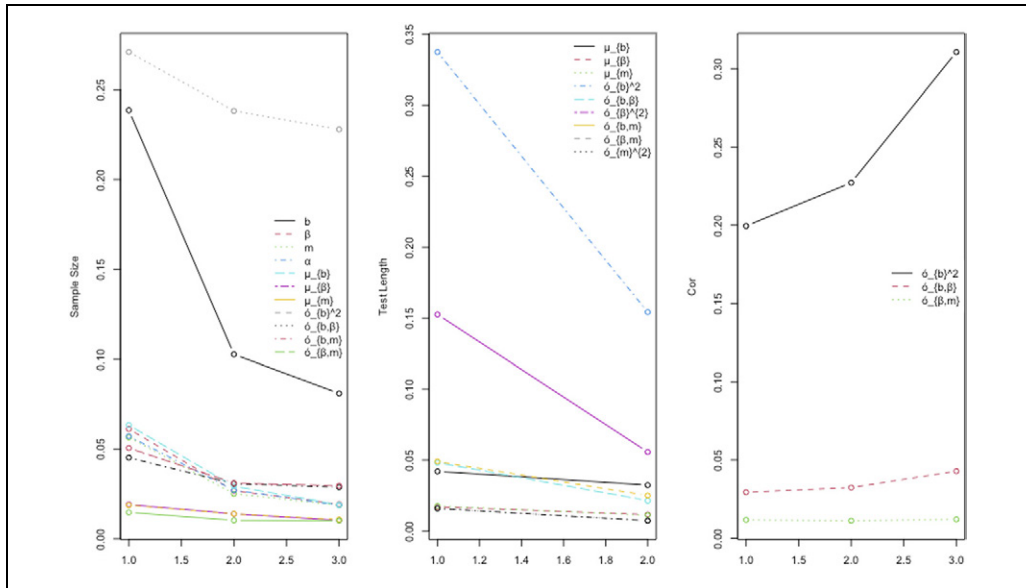


Figure 6 . RMSE values of item-side parameters across different levels of manipulated factors. b , item difficulty; β , item time intensity; ν , item time discrimination; m , item engagement intensity; α , item engagement discrimination; μ ., item mean; σ ., item variance; Cor., correlations among person-side latent variables.

2015). In addition, having a longer test improves item parameter recovery, especially regarding the item variances. And, the recovery of time- and visual engagement-related parameters (e.g., $\mu\beta$, μ and $\sigma_{\zeta,m}$) was better than that of item difficulty (e.g., σ_b^2). In addition, it appears that correlations among item parameters had limited impact on recovery of the item-side covariances.

Regarding the recovery of latent ability, working speed, and visual attention parameters, the ANOVA results indicated that test length had a larger impact than did sample size and correlations among latent traits on recovery of person parameters. Outcome measures of person-side parameters improved as test length increased (e.g., the averaged RMSE of $\theta = 0.568$ when test length = 10, and the averaged RMSE of $\theta = 416$ when test length = 25). Additionally, latent working speed and visual engagement were recovered better than general ability from the IRT model. Moreover, correlations among person-side latent traits impacted the recovery of variance/covariances of the person-side parameters. Relatively higher RMSEs for the latent ability related parameters were observed in the conditions where latent traits were highly correlated (cor. = 0.8). However, for the conditions with a longer test length ($I = 25$), parameters were recovered better than those with shorter test length ($I = 10$). One reason for having better parameter recovery with longer tests is that more information was provided from the data to estimate the nuances among test-takers regarding their responding accuracy, working-efficiency, and visual attentiveness. General trends of RMSE according to the sample size and correlations among latent traits are shown in Figure 7.

To assess data model fit, test-level PPP-values were calculated and averaged over all simulated datasets for each condition. The PPP-values were summarized based on 2000 iterations after dropping burn-in iterations with thinning of 2. A PPMC value within the range of 0.05 and 0.95 indicates satisfactory data-model fit. In general, all models showed adequate fit with the simulated datasets. Compared to the RT and VFC models, PPMC values for the 1-PL IRT model

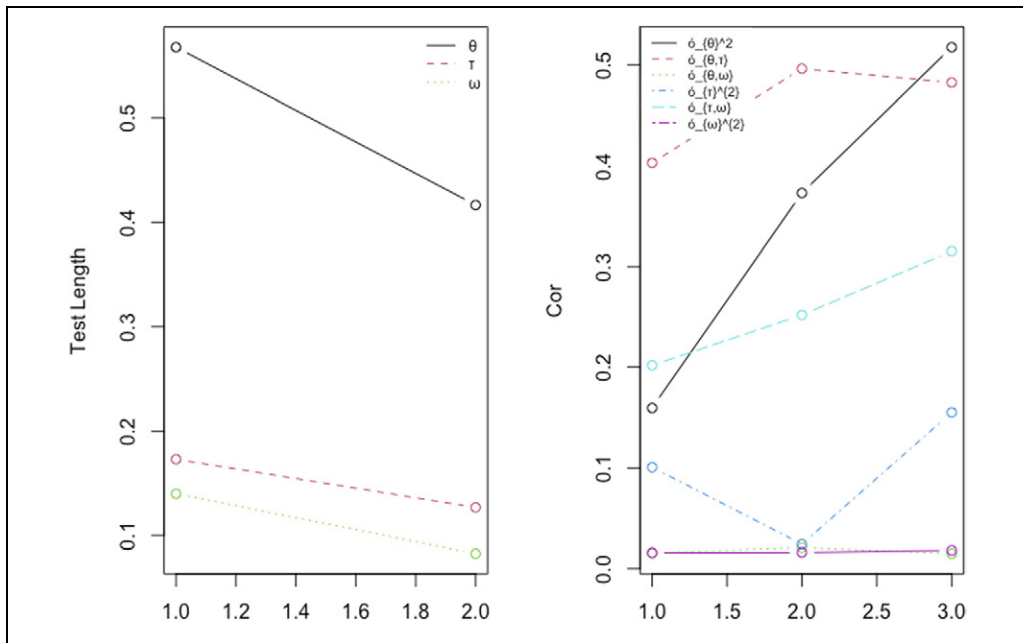


Figure 7. RMSE values of person-side parameters across different levels of manipulated factors. θ , ability; ω , visual engagement; τ , speediness; Cor., Correlations among person-side latent variables.

were relatively smaller, especially for the conditions with shorter test-length and smaller sample size. However, all the PPP-values were above 0.05 indicating acceptable fit. Detailed results can be found in Table 8 in the Supplemental file.

In summary, the simulation study showed that model parameters of the proposed model are well-recovered across all conditions with MCMC estimation. Particularly, for the conditions with small sample size ($n = 100$) and short test length ($I = 10$), model parameters were adequately recovered based on reasonably small RMSE values. RMSE was clearly smaller under larger sample size and longer test length conditions. The results from the empirical study are consistent with the simulation study, which indicate that the manifested patterns regarding test-takers’ behaviors and item characteristics in the empirical example appear to be valid.

Table 4 showed that all item parameters were well-recovered across with low biases. The recovery of time-related boundaries (i.e., $RMSE(\beta) = 0.065$ and $RMSE(\nu) = 0.141$) and that of visual-related parameters (i.e., $RMSE(m) = 0.054$ and $\alpha = 0.054$), have smaller RMSE values than does item difficulty ($RMSE(b) = 0.234$). And, the mean biases were lower than 0.028. Regarding the item mean vector and variance-covariance matrices. In general, the recovery of all model parameters was satisfactory. The RMSE values of item means are ranged from 0.023 to 0.072, and the RMSE values of the variance-covariance components are ranged from 0.021 to 0.302. And, the bias values are ranged from 0.012 to 0.225. In general, the recovery of time- and visual engagement-related parameters (e.g., μ_β , μ and $\sigma_{\zeta, m}$) was better than that of item difficulty (e.g., σ_b^2) with lower values in both RMSE and bias.

As also shown in Table 4, person-side parameters, latent working speed and visual engagement (i.e., RMSE values for $(\tau) = 0.176$ and $\omega = 0.138$) were recovered better than ability ($RMSE(\theta) = 0.586$) from the IRT model. All bias values were lower than 0.004 in absolute value. Moreover, the recovery of variance-covariance components related to working-speed and

Table 4. RMSE and Bias for Simulated Data with a Small Sample Size ($N = 100$) and a Short Test Length ($I = 10$).

Par	Bias	RMSE
Item parameters		
b	0.028	0.234
β	0.001	0.065
ν	0.001	0.141
m	-0.002	0.054
α	-0.002	0.054
μ_b	0.025	0.072
μ_β	0.010	0.018
μ_m	-0.012	0.023
σ_b^2	0.225	0.302
$\sigma_{b,\beta}$	0.020	0.056
σ_β^2	0.155	0.157
$\sigma_{b,m}$	0.017	0.057
$\sigma_{\beta,m}$	0.012	0.021
σ_m^2	0.165	0.167
Person parameters		
θ	0.001	0.586
τ	-0.001	0.176
ω	-0.004	0.138
σ_θ^2	0.188	0.205
$\sigma_{\theta,\tau}$	-0.403	0.406
$\sigma_{\theta,\omega}$	0.030	0.033
σ_τ^2	-0.094	0.105
$\sigma_{\tau,\omega}$	-0.188	0.199
σ_ω^2	0.031	0.037

visual engagement (e.g., σ_τ^2 and σ_ω^2) was better than that of latent ability related parameters (e.g., σ_θ^2 , and $\sigma_{\theta,\tau}$), in terms of both RMSE and bias. Overall, recovery of the person-side variance and covariances (Σ_p) was satisfactory in which all RMSE values were less than 0.406.

In addition, ANOVA results demonstrated nontrivial effects from sample size and test length. We expected the crossed condition with small sample size ($n = 100$) and short test length ($I = 10$) to be potentially challenging to recover model parameters. The RMSE and bias values of each parameter in that condition were summarized in Table 4. The detailed results of other conditions can be located in the Supplemental file.

Discussion

With increasing frequency, researchers are using multimodal data in order to better understand the interconnections of the myriad of complex behaviors and cognitive processes of test-takers. To this end, innovative assessments environments, like ITELS, make the concurrent collection of various types of data, possible. Gathering and assembling different data (i.e., item responses, response times, and gaze fixation counts) from these settings notwithstanding, the overall success of linking these data to underlying human biological and cognitive processes also relies on innovative psychometric models like the hierarchical factor model presented here. This work builds on models for simultaneously analyzing item responses and response times (van der Linden, Klein Entink, & Fox, 2010). These joint models have proven effective in providing valuable

information about item and person characteristics above and beyond what could be ascertained from analyzing item responses alone (Man & Haring, 2019). Two important caveats are worth noting. First, is that these disparate data types must be collected simultaneously if they are to be modeled jointly. That is, if relations and associations among model parameters reflecting item and person characteristics are of interest, then the environment in which data collection occurs, must be sufficiently equipped. Second, software must be available for merging and cleaning these data collected from different sources (i.e., log files, eye-tracking machine) so that modeling can even occur. These tasks are not inconsequential and require planning for successful execution.

A three-way joint model that integrates visual fixation counts into a traditional psychometric modeling framework was proposed and its utility was shown through an analysis of data gathered on a sample of $n = 93$ college students and via a small-scale Monte Carlo simulation. The joint modeling framework permits estimation of item and person parameters from a 1-PL IRT model, an RT model and a negative binomial model, simultaneously. Elements of item-side and person-side variance-covariance matrices represent the *glue* that connect the distinct model parameters together. An MCMC algorithm was used to facilitate parameter estimation. Results from the real data example showed that the proposed model captured the underlying patterns embedded in the data and showed satisfactory data-model fit—even though the sample size was quite modest.

We suggested interpretations of estimated item parameters (i.e., difficulty, time and visual intensity) and person parameters (i.e., latent ability, working speed, and visual engagement), particularly elements of the covariance matrices where associations between item characteristics and between person characteristics could be examined separately—each providing insights into the interconnected patterns of behaviors and attributes of the test-takers and characteristics of the task items. Our interpretations were admittedly superficial in nature as they were informed by past methodological studies² (see, e.g., Fox & Mariani, 2016; Man & Haring, 2019; van der Linden, 2006). We acknowledge that they also lacked depth, nuance, and understanding that a learning scientist or cognitive psychologist, who had an intimate knowledge of the underlying biological processes, could provide. Future projects would certainly benefit from collaborating within a multidisciplinary team comprised of methodologists and substantive researchers, each bringing content knowledge, complementary skills and expertise to the research enterprise—especially in the early stages where discussions among team members could inform data collection and research design.

In a virtual-based learning system, this joint modeling framework can help evaluate relations among responding accuracy, task decoding speed, and visual engagement providing educators a deeper understanding of test takers' cognitive process in reaching their final answers. This information, if presented in a digestible manner, could be subsequently used in a formative way—providing feedback to inform course design modifications or suggesting different teaching strategies aimed at enhancing student outcomes (Jiao & Lissitz, 2018). Besides those variables gathered as part of the current study, the ITELS environment facilitates the collection of a number of other eye-tracking related biometric information variables (e.g., blinking rates, pupil diameters) that could be jointly modeled to reflect other characteristics of test-taking behaviors. For example, pupil diameter has been reported to be negatively correlated with levels of fatigue (e.g., Morad et al., 2000; Yoss et al., 1970). Other biometric data (e.g., blood-oxygen-level dependent signal, Electroencephalography (EEG), or heart rate) could be integrated into the current modeling framework for systematically assessing the learners' learning progressions in ITELS. Other individual attributes and background variables, like gender, could also be added as covariates to show differences in model parameters between groups. As a consequence, practitioners could better understand the nuances in performance across groups or identify aberrant/gifted learning groups.

As was just argued, an ITELS clearly presents opportunities to collect a multitude of distinct data types. Whether this leads to useful information that can be tied directly to underlying theories of behavior, development, learning, and cognition has yet to be established and fully exploited. As one reviewer pointed out, test-takers might be less than enthusiastic about ever-increasing monitoring, and instead find it quite intrusive. Imagine being subjected to a polygraph while sitting for a licensing exam. While this may seem a little far-fetched, the slope may quickly become very slippery without standards guiding the ethical use of these assessment environments in practice. This said, we remain cautiously optimistic about the potential insights that can be garnered about underlying biological and cognitive processes by further connecting advanced modeling techniques, like the joint model of multimodal data presented here, with theory.


Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is funded by an Institute of Educational Sciences (IES) grant R305A210428.

ORCID iDs

Jeffrey R. Harring  <https://orcid.org/0000-0002-7102-0303>

Kaiwen Man  <https://orcid.org/0000-0002-9696-9726>

Peida Zhan  <https://orcid.org/0000-0002-6890-7691>

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The interested reader can visit the Gazept website for tutorials (<https://www.gazept.com/tutorials/>) about usage and setup as well as a listing of peer-reviewed publications (<https://www.gazept.com/about-us-page-2/publications/>) that used the eye-tracking hardware to gather data for the research projects. The test was delivered with a time limit. For the verbal reasoning section, students were asked to finish 10 questions within 20 minutes, where the last three questions required reading short passages.
2. Visual attention in cognitive psychology is a broader concept, often defined as a process that directs a tiny fraction of the information arriving at the visual cortex involved in visual working memory, pattern recognition, and motivation (Anderson et al., 2005). The “visual engagement” term was suggested to be used to only reflect the “visual cognitive process loading amount” endorsed by gaze counts rather than the entire visual attention process. In other words, fixation counts could be used to reflect the visual attention load, and the interpretation of visual engagement depends on the context.

References

- Anderson, C. H., Van Essen, D. C., & Olshausen, B. A. (2005). Directed visual attention and the dynamic control of information flow. In *Neurobiology of attention* (pp. 11–17). Elsevier. <https://doi.org/10.1016/b978-012375731-9/50007-0>

- Aslin, R. N. (2012). Infant eyes: A window on cognitive development. *Infancy*, 17(1), 126–140. <https://doi.org/10.1111/j.1532-7078.2011.00097.x>
- Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37(3), 379–384. <https://doi.org/10.3758/bf03192707>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring any examinees ability. In F. Lord, & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley
- Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, 82(4), 1126–1148. <https://doi.org/10.1007/s11336-016-9537-6>
- Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, 23(3), 315–345. <https://doi.org/10.1214/aoms/1177729380>
- Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement*, 12(4), 425–434. <https://doi.org/10.1177/014662168801200410>
- Corcoran, P. M., Nanu, F., Petrescu, S., & Bigioi, P. (2012). Real-time eye gaze tracking for gaming design and consumer electronics systems. *IEEE Transactions on Consumer Electronics*, 58(2), 347–355. <https://doi.org/10.1109/tce.2012.6227433>
- De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, 70(2), 225–237. <https://doi.org/10.1111/bmsp.12094>
- Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Taylor Francis
- Fox, J. P., Entink, R. K., & Avetisyan, M. (2014). Compensatory and non-compensatory multidimensional randomized item response models. *British Journal of Mathematical and Statistical Psychology*, 67(1), 133–152.
- Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, 51(4), 1–4. <https://doi.org/10.1080/00273171.2016.1171128>
- Fox, J. P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, 54(2), 243–262. <https://doi.org/10.1111/jedm.12143>
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. Chapman & Hall
- Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Journal of Educational and Behavioral Statistics*, 6(4), 733–760.
- Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In *The mind's eye* (pp. 493–516). Elsevier. <https://doi.org/10.1016/b978-044451020-4/50027-x>
- Hao, J., Smith, L., Mislevy, R., von Davier, A., & Bauer, M. (2016). Taming log files from game/simulation-based assessments: Data models and data analysis tools. *ETS Research Report Series*, 1(1), 1–17. <https://doi.org/10.1002/ets2.12096>
- Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *The Mind's Eye*, 2(3), 573–605. <https://doi.org/10.1016/b978-044451020-4/50031-1>
- Jiao, H., & Lissitz, R. (2018). *Technology enhanced innovative assessment development, modeling, and scoring from an interdisciplinary perspective*. Information Age Publishing
- Justice, M., & Lankford, C. (2002). Pilot findings. *Communication Disorders Quarterly*, 24(1), 11–21. <https://doi.org/10.1177/152574010202400103>
- Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14(1), 54–75. <https://doi.org/10.1037/a0014877>
- Lee, Y.-H., Hao, J. G., Man, K., & Ou, L. (2019). How do test takers interact with simulation-based tasks? A response-time perspective. *Frontiers in Psychology*, 10, 906. <https://doi.org/10.3389/fpsyg.2019.00906>
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33(7), 519–537. <https://doi.org/10.1177/0146621608329504>
- Linacre, J., & Wright, B. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, 8(2), 350.

- Lord, F. M. (1952). A theory of test scores (psychometric monograph no. 7). In *Annual meeting of the psychometric society*. Iowa City
- Luo, Y., & Jiao, H. (2018). Using the stan program for bayesian item response theory. *Educational and Psychological Measurement, 78*(3), 384–408. <https://doi.org/10.1177/0013164417693666>
- Luo, Y. (2021). A comparison of common IRT model-selection methods with mixed-format tests. *Measurement: Interdisciplinary Research and Perspectives, 19*(4), 199–212.
- Man, K., & Haring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement, 79*(4), 617–635. <https://doi.org/10.1177/0013164418824148>
- Man, K., & Haring, J. R. (2021). Assessing preknowledge cheating via innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement, 81*(3), 441–465. <https://doi.org/10.1177/0013164420968630>
- Man, K., Haring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement, 43*(8), 639–654. <https://doi.org/10.1177/0146621618824853>
- Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451. <https://doi.org/10.3102/1076998614559412>
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology, 68*(2), 197–219. <https://doi.org/10.1111/bmsp.12042>
- Morad, Y., Lemberg, H., & Dagan, Y. (2000). Pupillography as an objective indicator of fatigue. *Current Eye Research, 21*(1), 535–542. [https://doi.org/10.1076/0271-3683\(200007\)2111-zft535](https://doi.org/10.1076/0271-3683(200007)2111-zft535)
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.
- Plummer, M. (2015). *JAGS: Version 4.0. User Manual*. Retrieved from <http://www.uvm.edu/~bbeckage/Teaching/DataAnalysis/Manuals/manual.jags.pdf>
- Poole, A., Ball, L. J., & Phillips, P. (2004). In search of salience: A response-time and eye-movement analysis of bookmark recognition. In S. Fincher, P. Markopoulos, D. Moore, & R. Ruddle (Eds.), *People and computers xviii—design for life*. Springer
- Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. In *Danish institute for educational research* (Expanded edition). University of Chicago Press
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372–422. <https://doi.org/10.1037/0033-2909.124.3.372>
- Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology, 70*(2), 317–345. <https://doi.org/10.1111/bmsp.12101>
- Samejima, F. (1996). The graded response model. In W. van der Uden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer
- Shagass, C., Roemer, R. A., & Amadeo, M. (1976). Eye-tracking performance and engagement of attention. *Archives of General Psychiatry, 33*(1), 121–125. <https://doi.org/10.1001/archpsyc.1976.01770010077015>
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298–321. <https://doi.org/10.1177/0146621605285517>
- Su, Y. S., Yajima, M., Su, M. Y. S., & System Requirements, J. A. G. S. (2015). Package ‘R2jags’. *R package version 0.03–08*, URL <http://CRAN.R-project.org/package=R2jags>
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204. <https://doi.org/10.3102/10769986031002181>

- 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>
- van der Linden, W. J., Entink, R. H. K., & Fox, J. P. (2010). Irt parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327–347. <https://doi.org/10.1177/0146621609349800>
- Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer Science & Business Media
- Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417. <https://doi.org/10.3102/1076998612461831>
- Wang, L. (2010). Irtzip modeling for multivariate zero-inflated count data. *Journal of Educational and Behavioral Statistics*, *35*(6), 671–692. <https://doi.org/10.3102/1076998610375838>
- Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Psychological Measurement*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Rasch Measurement.
- Wu, M., Adams, R., Wilson, M., & Haldane, S. (1998). *Conquest: Generalized item response modeling software [computer software and manual]*. Australian Council for Educational Research
- Yoss, R. E., Moyer, N. J., & Hollenhorst, R. W. (1970). Pupil size and spontaneous pupillary waves associated with alertness, drowsiness, and sleep. *Neurology*, *20*(6), 545. <https://doi.org/10.1212/wnl.20.6.545>
- Zhan, P., Jiao, H., & Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 262–286. <https://doi.org/10.1111/bmsp.12114>