

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

2021, Vol. 81 (3) 441–465

© The Author(s) 2020

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164420968630

journals.sagepub.com/home/epm



Kaiwen Man¹  and Jeffrey R. Harring²

Abstract

Many approaches have been proposed to jointly analyze item responses and response times to understand behavioral differences between normally and aberrantly behaved test-takers. Biometric information, such as data from eye trackers, can be used to better identify these deviant testing behaviors in addition to more conventional data types. Given this context, this study demonstrates the application of a new method for multiple-group analysis that concurrently models item responses, response times, and visual fixation counts collected from an eye-tracker. It is hypothesized that differences in behavioral patterns between normally behaved test-takers and those who have different levels of preknowledge about the test items will manifest in latent characteristics of the different data types. A Bayesian estimation scheme is used to fit the proposed model to experimental data and the results are discussed.

Keywords

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

¹University of Alabama, Tuscaloosa, AL, USA

²University of Maryland, College Park, MD, USA

Corresponding Author:

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

Email: kman@ua.edu

The rapid shift toward a digitally reliant society necessitates the continual updating of the environments in which students learn. To improve the effectiveness and efficiency of learning, technology-enhanced learning infrastructures and approaches, such as artificial intelligence-enhanced and virtual reality-based learning, are called with greater frequency to bridge this gap (see, e.g., Ercikan & Pellegrino, 2017; Hao et al., 2016; Jiao & Lissitz, 2018; Man & Harring, 2019; Mislevy, 2011). In the technology-enhanced learning system (TELS), students' learning status regarding biological and psychological reactions are continuously recorded in a formative manner via integrated detectors (e.g., eye-tracker, motion detectors, and virtual reality goggles) within the environment. This intensively collected, spontaneous biological and psychological information can be further used by practitioners to (1) better understand the learners, (2) improve their instruction and design, (3) monitor students' learning, and (4) secure online-delivered exams, which could in turn, promote increases in students' learning outcomes. And highly relevant to the recent global pandemic, the TELS can help transfer knowledge remotely like online-teaching, which makes the learning experience independent of space, pace, and time.

The tremendous amount of real-time data collected by the TELS are of different types and come in different forms, which can be modeled differently in terms of their functionalities. Outcome data (e.g., item responses and total scores) can be modeled to show students' responding accuracy either at the item- or test-level. Traditionally, item response theory (IRT) models were created to explain the association between a test-taker's observed binary correct/incorrect responses and their latent ability—the latter of which is considered either to have a unidimensional or multidimensional structure (see, e.g., Birnbaum, 1968; Lord, 1952; Rasch, 1960; Reckase, 1972).

Because of its availability with emergent technologies, process data (e.g., response times, keystrokes) can augment information about the test-taker above and beyond what item responses afford in isolation. As De Boeck and Jeon (2019) state, "abilities refer to levels of performance, whereas processes are the activities involved in reaching a performance outcome" (p. 1). Bergner and von Davier (2019) describe the *process* in process data and its relation to outcome measures in the following ways:

1. The process is irrelevant or at least ignorable given the outcome.
2. The process is auxiliary to the outcome.
3. The process is essential to understanding the outcome.
4. The outcome, and process scores are derived from an expert rubric, in the sense of a holistic rating.
5. The process is the outcome, and process scores are derived from a measurement model that accounts for dependencies in sequential data.

Among various types of process data, response times (RTs), the time a respondent takes to answer an individual item or task, have been frequently modeled to either show individuals' working efficiency or to account for the speed-accuracy trade-off by jointly modeling them with item responses (e.g., Bolsinova et al., 2017; Man

et al., 2019; Molenaar et al., 2015; Man et al., 2020; van der Linden, 2006). In addition, RTs can be used as a motivation indicator to identify aberrant item-responding behaviors such as carelessness and guessing (Guo et al., 2016; Wise & DeMars, 2006). By collecting, analyzing, and reporting process data (e.g., RTs) in large-scale assessment, process data, as a means to an end of understanding how a outcome had been reached as a result of sequences of actions, can be essential in evaluating and diagnosing task-performers' weaknesses and strengths in solving problems (Wang et al., 2018).

Biometric data (e.g., eye-tracking, heart rate recording, electroencephalography), a subcategory of process data, are only beginning to be used in educational assessments. Although its importance in understanding the complexities of the learning process, notwithstanding integration and modeling of biometric data in practice has been slow because these data must be captured concurrently in real-time with other more conventional data types (Man, 2020). One type of biometric data that is emerging is eye-tracking data (Bergner & von Davier, 2019). Eye-tracking data has been used in various disciplines for some time and different attributes of eye-movement such as visual fixation counts (VFCs) have proven to be conducive in understanding many cognitive processes (Poole et al., 2004). The collected eye-tracking data can be used to address questions related to cognition such as: Where does a test-taker or task performer gaze? When does blinking occur? How does the pupil react to different stimuli? What information does a task-taker ignore during the performance causing failure? The answers to these questions (among the many others that could be asked) can potentially provide finer-grained diagnostic information regarding how high-order cognitive constructs are used in performing a task—information that would be unattainable by depending solely on the analysis of item responses or/and RTs.

Many methods have been proposed to analyze the different types of process data independently (e.g., Fox & Marianti, 2016; Lu et al., 2020). However, few attempts have been made to evaluate, in a more panoramic manner, task-takers' abilities with process information, which analyzes outcome and process indicators in a single model. A two-factor hierarchical structure model proposed by van der Linden (2007) for jointly modeling item responses and RTs with random item and person parameters could be an ideal foundation for jointly modeling various process indicators. This method can provide interpretable parameter estimates, which can reveal not only the underlying behavioral patterns reflecting the trade-off between responding accuracy and working efficiency but item characteristics such as item difficulty. Moreover, this modeling framework has been used to timely calibrate online rendered items in the computer-based adaptive learning system with marginal maximum likelihood estimation making this type of joint-modeling method computationally feasible with large datasets (Kang et al., 2020). An interesting extension of this joint modeling is to a model with general linear factors that have a multilevel, multigroup structure. The multilevel–multigroup (ML-MG) model provides a general framework that considers more latent constructs than ability and work speed, and as we will demonstrate shortly, this ML-MG structure allows for the comparison of differences

in both item characteristics and behavioral patterns across groups such as cheaters and noncheaters.

In the field of test security, many methods (e.g., Lu et al., 2020; van der Linden, 2007) have been proposed to evaluate cheating behavior by modeling item response and RTs jointly or separately. Yet no study that we are aware of has proposed a modeling framework that also incorporates eye-tacking indicators (e.g., VFCs) to assess the behavioral pattern differences between the cheaters and noncheaters. Studies have shown that VFCs can be used to demonstrate cognitive information process efficiency and difficulty (Schaeffer et al., 2019). In addition, VFCs can show how familiar a person is with a visual target (Constantinides et al., 2019). Coupled with item responses and RTs, it is just this type of biometric data that we hope to demonstrate helps uncover aberrant test-taking behavior.

In this study, a ML-MG three-way factor model is proposed for jointly modeling item responses, RTs, and VFCs across three groups using experimental data from a study in which participants were randomly assigned to treatment conditions. The models allows for the investigation of the association among latent factors: ability, working speed, and test engagement, underlying item responses, RTs, and VFCs, respectively. The proposed ML-MG joint modeling approach is an extension of the Bayesian multilevel modeling framework proposed by van der Linden (2007). In this three-way ML-MG joint modeling approach, the Rasch model, an RT model, and a VFCs model are specified at the measurement level. The variance-covariance structures of the person-side and item-side parameters are specified at level two. Bayesian estimation is used to estimate the proposed three-way ML-MG joint model. An empirical example using data collected in an eye-tracking lab is provided. The findings from the real data analyses are discussed.

Multilevel–Multigroup Model Specification

Level-1: Measurement Models Across Different Groups

Item Response Model. A one-parameter logistic (1-PL MG; or Rasch MG) model, the multiple group version of the conventional 1-PL model (Lord, 1952), was selected to model the relation between latent ability reflecting the responding accuracy and item responses, and was fitted to each group. The model is specified as

$$P(u_{ijg} = 1 | \theta_{jg}; b_{ig}) = \frac{1}{1 + e^{-(\theta_{jg} - b_{ig})}}, \quad (1)$$

where $P(u_{ijg} = 1 | \theta_{jg}; b_{ig})$ is the probability of a correct response to item i ($i = 1, \dots, I$), by person j ($j = 1, \dots, J$), in group g ($g = 1, \dots, G$); b_{ig} is the difficulty parameter (location) for item i answered by persons in group g , and θ_{jg} is a general latent trait for person j in group g . The item slopes (discrimination parameters) are each fixed to unity.

On the person-side, using the I-PL MG model can manifest an individual's latent ability to solve the test items and reveal systematic differences between groups in

responding accuracy. On the item-side, item difficulties can be estimated for different groups of test-takers.

Response Time Model. In addition to the 1-PL MG model, a log-normal RT model (van der Linden, 2006) with MG structure is used to describe a test-taker’s working speed. Specification of the log-normal MG RT model extends the basic form outlined by van der Linden (2006) to G groups,

$$f(t_{ijg} | \tau_{jg}, \nu_{ig}, \beta_{ig}) = \frac{\nu_{ig}}{t_{ijg} \sqrt{2\pi}} \left(-\frac{1}{2} [\nu_{ig} \{\ln t_{ijg} - (\beta_{ig} - \tau_{jg})\}]^2 \right), \tag{2}$$

where t_{ijg} denotes the RT of test-taker j in group g on item i . The latent parameter, $\tau_{jg} \in \mathfrak{R}$, represents working speed for test-taker j in group g . The item parameter $\beta_{ig} \in \mathfrak{R}$ denotes time intensity, or simply, the average of $\ln(t_{ijg})$ when τ_{jg} is 0. Parameter $\nu_{ig} \in \mathfrak{R}$ is an item time discrimination parameter reflecting the dispersion of t_{ijg} for item i answered by persons in group g . The mean value of $\ln(t_{ijg})$ is parameterized as $\mu_{ijg} = \beta_{ig} - \tau_{jg}$.

Similar to the 1-PL MG model, the person-side parameters can be used to demonstrate how efficient a person was working on the test items. And, this parameter would be jointly modeled with latent ability and visual engagement at level-two, the structural model. Besides, the differences in overall working efficiencies between groups can be manifested as well. In terms of item parameters, time intensities could be used to show how much time effort was required for each item.

Visual Fixation Counts Model. VFCs are fitted using a negative binomial fixation (NBF) model proposed by Man and Harring (2019), which describes the relation between observed VFCs and latent test visual engagement.¹ The NBF model is specified as:

$$P(C = c_{ijg} | s_{ig}, m_{ig}, \omega_{jg}) = \frac{\Gamma(c_{ijg} + s_{ig})}{c_{ijg}! \Gamma(s_{ig})} \times \left(\frac{s_{ig}}{\exp(m_{ig} + \omega_{jg}) + s_{ig}} \right)^{s_{ig}} \times \left(\frac{\exp(m_{ig} + \omega_{jg})}{s_{ig} + \exp(m_{ig} + \omega_{jg})} \right)^{c_{ijg}}, \tag{3}$$

where parameter m_{ig} , an item-side parameter, denotes the visual intensity for item i answered by persons in group g . The presumption is that this parameter reflects the averaged amount of visual engagement for a group of test-takers to finish answering an item. A person-specific parameter, ω_{jg} , for each of the j ($j = 1, \dots, J$) test-takers in group g ($g = 1, \dots, G$), denotes the overall test engagement level and is assumed to be constant across all the items. Furthermore, a discrimination parameter, α_{ig} , for item i in group g is defined as $\alpha_{ig} = 1 / \sqrt{\mu_{.ig} + \mu_{.ig}^2 / s_{ig}}$, where $u_{.ig} = \sum_{i=1}^I u_{ijg} / I$, reflecting the dispersion of the fixation counts on item i .

Level 2: Multigroup Item Domain and Person Domain Models

The second-level models incorporate two variance–covariance structures, named as person-domain and item-domain structures separately, to account for the dependencies of both item and person parameters jointly. These are estimated from the Level-1 models for different groups.

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

$$\Theta_{pg} = (\theta_g, \tau_g, \omega_g)^T \sim MVN(\boldsymbol{\mu}_{pg}, \boldsymbol{\Sigma}_{pg}), \quad (4)$$

with mean vector, $\boldsymbol{\mu}_{pg} = (\mu_{\theta_g}, \mu_{\tau_g}, \mu_{\omega_g})^T$, and covariance matrix

$$\boldsymbol{\Sigma}_{pg} = \begin{pmatrix} \sigma_{\theta_g}^2 & & \\ \sigma_{\theta\tau_g} & \sigma_{\tau_g}^2 & \\ \sigma_{\theta\omega_g} & \sigma_{\tau\omega_g} & \sigma_{\omega_g}^2 \end{pmatrix}. \quad (5)$$

The parameters on the diagonal of the $\boldsymbol{\Sigma}_{pg}$ denote the variances of the latent constructs. The off-diagonal parameters represent the covariances between any pairs of latent constructs. For example, the parameter, $\sigma_{\theta\tau_g}$ is the covariance between latent ability and speediness of test-takers in group g .

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

$$\Xi_{I_g} = (b_g, \beta_g, m_g)^T \sim MVN(\boldsymbol{\mu}_{I_g}, \boldsymbol{\Sigma}_{I_g}). \quad (6)$$

The mean vector and symmetric covariance matrix, $\boldsymbol{\mu}_{I_g}$ and $\boldsymbol{\Sigma}_{I_g}$, are defined respectively as $\boldsymbol{\mu}_{I_g} = (\mu_{b_g}, \mu_{\beta_g}, \mu_{m_g})^T$ and

$$\boldsymbol{\Sigma}_{I_g} = \begin{pmatrix} \sigma_{b_g}^2 & & \\ \sigma_{b\beta_g} & \sigma_{\beta_g}^2 & \\ \sigma_{bm_g} & \sigma_{\beta m_g} & \sigma_{m_g}^2 \end{pmatrix}. \quad (7)$$

By estimating the two structural variance–covariance matrices, the associations among item parameters and the relationships among person parameters can be manifested across groups. Those structural nuances across groups represent distinct test-taking behavioral patterns among individuals who have preknowledge of test items.

The impact of having preknowledge can be evaluated by measuring the magnitude of drifts in the associated parameters, such as item difficulties and time intensities. The model constraints will be illustrated in the estimation section.

Testing Item Parameter Drift Across Groups

Item drift occurs when items function differently for various groups of test-takers. Usually, item drift was caused by the presence of new construct or irrelevant traits (e.g., test-takers have preknowledge on items) affecting the individuals' responding accuracy, which violates the unidimension assumption of testing (Hambleton et al., 1991; Smith & Prometric, 2004). Therefore, the distributions of additional traits/constructs differ across groups. One group of test-takers might have lower probabilities to answer the items correctly. To evaluate whether item parameter drift exists across experimental groups, pairwise differences among the same set of item parameters are defined as follows:

The drift in item difficulties: $D_{b_i(m,n)} = D_{b_i(m)} - D_{b_i(n)} \sim Normal(\mu_{D_{b_i(m,n)}}, \sigma_{D_{b_i(m,n)}}^2)$, $i = 1, \dots, I$, $m \neq n$, $m \in (1, \dots, G)$, $n \in (1, \dots, G)$, G means group. For instance, $D_{b_1(1,2)}$ describes the drift in difficulties for Item 1 between experimental conditions 1 and 2.

In terms of time intensities, the drift in time intensities are defined as: $D_{\beta_i(m,n)} = D_{\beta_i(m)} - D_{\beta_i(n)} \sim Normal(\mu_{D_{\beta_i(m,n)}}, \sigma_{D_{\beta_i(m,n)}}^2)$, $i = 1, \dots, I$, $m \neq n$, $m \in (1, \dots, G)$, $n \in (1, \dots, G)$. Similarly, the drifts in visual intensities are defined as:

$$D_{m_i(m,n)} = D_{m_i(m)} - D_{m_i(n)} \sim Normal(\mu_{D_{m_i(m,n)}}, \sigma_{D_{m_i(m,n)}}^2), i = 1, \dots, I, m \neq n, m \in (1, \dots, G), n \in (1, \dots, G).$$

To summarize the uncertainty of the posterior distribution of the different item drift differences, standardized Wald-statistics are defined for difficulties, time intensities, and visual intensities, separately:

$$\sqrt{W_{D_{b_i(m,n)}}} = \frac{\hat{\mu}_{D_{b_i(m,n)}}}{\hat{\sigma}_{D_{b_i(m,n)}}}, \quad \sqrt{W_{D_{\beta_i(m,n)}}} = \frac{\hat{\mu}_{D_{\beta_i(m,n)}}}{\hat{\sigma}_{D_{\beta_i(m,n)}}}, \quad \sqrt{W_{D_{m_i(m,n)}}} = \frac{\hat{\mu}_{D_{m_i(m,n)}}}{\hat{\sigma}_{D_{m_i(m,n)}}}.$$

The calculated W statistics will be compared with corresponding critical values: ± 1.96 based on that the probability of incorrectly rejecting the true hypothesis equals .05. If a calculated W statistic is less extreme than the critical values, it indicates that the posterior probability of existing item drift is less than .05. Otherwise, it indicates existence of item drift with high probability.

Figure 1 displays the graphical representation of the ML-MG joint model of item response, response time, and VFCs.

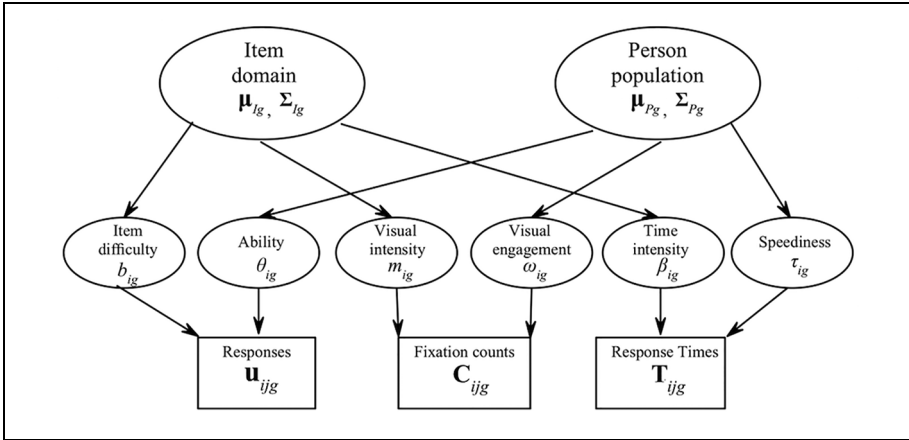


Figure 1. Multilevel-multigroup (ML-MG) three-way joint model approach of item response, response time, and visual fixation counts. μ_{ig} , mean vector of item parameters; Σ_{ig} , covariance of item parameters; μ_{pg} , mean vector of person parameters; Σ_{pg} , covariance of person parameters. g indicates different groups: $g = 1, \dots, G$.

Bayesian Estimation Using MCMC Sampling

Bayesian estimation was used for model parameter estimation in Just Another Gibbs Sampler (Plummer, 2015), which is housed in the R2jags package (Su & Yajima, 2015). Convergence is assessed via the coda package. Four chains using 60,000 total iterations with thinning of 4 to reduce autocorrelation among draws were executed. Model parameter estimates and standard deviations were summed up dependent on the posterior densities using the final 10,000 iterations after burning-in 50,000. The potential scale reduction factor was used for assessing convergence for all model parameters (Gelman et al., 2003). For the present study, a potential scale reduction factor value of 1.2 or less for each model parameter was used as the arbiter indicating convergence.

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 thresholds to zero) or the person-side (fixing the expectation of the latent ability parameter to zero). In this study, the model identification scales were fixed on the person-side by following the convention used for IRT model estimation (Volodin & Adams, 1995; Wu et al., 1998).

Within each group g , the population mean of the latent ability for the 1-PL model, θ , was set to 0 (Lord, 1952), and, the item discrimination parameter for each item was fixed as one. 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 population mean of the latent person-side visual engagement parameter Ω is fixed as 0 as well (Man & Harring, 2019).

$$\mu_{\omega_g} = \mu_{\theta_g} = \mu_{\tau_g} = 0 \quad \text{for } g = 1, \dots, G \tag{8}$$

Prior Distributions

Weak informative priors are preferentially used in this study to increase the generalizability of our code by imposing vague prior beliefs on estimating parameters. The setting of priors in this way was also implemented in Man et al. (2020) and Man and Harring (2019). $u_{ijg} \sim \text{Bernoulli}(P(u_{ijg} = 1))$, $\log(T_{ijg}) \sim N(\beta_{ig} - \tau_{jg}, \nu_{ig})$ $C_{ijg} \sim \text{NB}(\exp(m_{ig} + \omega_{jg}), s_{ig})$.

The prior distribution of item parameters, Ξ_{I_g} referring to Equation 6, for the proposed model is assumed to be trivariate normal. And, ν_{ig} , defined as the inverse of the variances of the log-times on different items, follows $IG(1, 1)$. In addition, the fixation dispersion parameter for each item [i.e., $s_{ig} \sim IG(1, 1)$, $i = 1, \dots, I$] is assumed to follow an inverse Gamma distribution as well. Hyperpriors are defined as

$$\mu_{d_g} \sim N(0, 0.5), \quad \mu_{\beta_g} \sim N(4.0, 0.5), \quad \mu_{m_g} \sim N(3.5, 1) \quad \Sigma_{I_g} \sim IW(\mathbf{I}_{I_g}, \nu),$$

where \mathbf{I}_{I_g} is an 3 by 3 identity matrix, and ν is the degree of freedom, which in this case is equal to 3.

Similarly, the prior specification for the person parameters, Θ_{p_g} referring to Equation 4, of the three-way joint model follows a trivariate normal distribution. Where the μ_{I_g} fixed as 0s. And,

$$\Sigma_{p_g} = \begin{pmatrix} \sigma_{\theta_g}^2 & & \\ \sigma_{\theta_g \tau_g} & \sigma_{\tau_g}^2 & \\ \sigma_{\theta_g \omega_g} & \sigma_{\tau_g \omega_g} & \sigma_{\omega_g}^2 \end{pmatrix} \sim IW(\mathbf{I}_{p_g}, \nu_g).$$

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

$$p(\Theta_{p_g}, \Xi_{I_g} | \mathbf{u}_g, \log(\mathbf{T}_g), \mathbf{c}_g) \propto \prod_{i=1}^I \prod_{j=1}^J p(\mathbf{u}_{ijg}, \log(\mathbf{T}_{ijg}), \mathbf{c}_{ijg} | \Theta_{jg}, \Xi_{ig}) p(\Theta_{jg} | \mu_{p_g}, \Sigma_{p_g})$$

$$p(\Xi_{ig} | \mu_{I_g}, \Sigma_{I_g}) p(\mu_{d_g}) p(\mu_{\beta_g}) p(\mu_{m_g}) p(\mathbf{S}_{I_g} | \nu_g) p(\mu_{p_g} | \mathbf{0}, \mathbf{S}_{p_g}) p(\mathbf{S}_{p_g} | \nu_g),$$

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

Posterior Predictive Model Checking

In this study, posterior predictive model checking (PPMC) was used for evaluating whether the proposed model adequately accounted for the variability existing in the

data. Specifically, PPMC was used to check our model-data fit (see, e.g., Gelman et al., 1996; Levy et al., 2009; Rubin, 1996; Sinharay et al., 2006).

Introduction of the Method

Let $\psi = (\Theta_p^T, \Xi_r^T)^T$ be the vector of unique parameters we are interested in estimating, and let \mathbf{y} be the set of observed data (e.g., item responses, RTs, and VFCs). Thus, the likelihood based on the conditional distribution of the data given model parameters could be expressed as $p(\mathbf{y}|\psi)$, and the prior distributions of all the model parameters could be denoted as $p(\psi)$. By applying Bayes' rule, the posterior distribution for a given set of parameters could be expressed as

$$p(\psi|\mathbf{y}) \equiv \frac{p(\mathbf{y}|\psi)p(\psi)}{\int_{\psi} p(\mathbf{y}|\psi)p(\psi)d\psi}. \quad (9)$$

To check the model-data fit by PPMC, predicted data are generated from the joint posterior distribution. The generated replicated dataset is denoted as \mathbf{y}_r^{pred} for $r = 1, 2, \dots, R$; where R indicates the number of draws from the joint posterior distribution. The distribution of predicted data, named as the posterior predictive distribution of predicted data (see, Equation 9), could be used for checking the data model fit.

$$p(\mathbf{y}^{pred}|\mathbf{y}) = \int p(\mathbf{y}^{pred}|\psi)p(\psi|\mathbf{y})d\psi. \quad (10)$$

Model fit is evaluated by comparing the differences between the predicted data \mathbf{y}_r^{pred} for $r = 1, 2, \dots, R$, and the observed data, \mathbf{y} . A small difference would be indicative of satisfactory data-model fit. Instead of directly comparing the predicted data and the observed data, 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).

The model-data fit can be evaluated by comparing the difference between the $T(\mathbf{y}^{pred}, \psi)$ and $T(\mathbf{y}, \psi)$, which are calculated based on predicted and realized data, respectively. In practice, a posterior predictive p value (PPP-value) is defined as the probability of obtaining the predicted data that is more extreme than the observed data. The estimated PPP-value is the proportion of $T(\mathbf{y}^{pred}, \psi)$ equal to or larger than $T(\mathbf{y}, \psi)$ over the R draws. A PPP-value close to 0 or 1 is indicative of poor model-data fit since the predicted data \mathbf{y}_r^{pred} is more extreme than the observed data, \mathbf{y} . The PPP-value is defined as

$$p = P(T(\mathbf{y}^{pred}, \psi) \geq T(\mathbf{y}, \psi)) = I_{T(\mathbf{y}^{pred}, \psi) \geq T(\mathbf{y}, \psi)} p(\mathbf{y}^{pred}|\psi)p(\psi|\mathbf{y})d\mathbf{y}^{pred}d\psi, \quad (11)$$

where I is the indicator function. To compute the data-model fit for the proposed model by applying the PPMC method, Sinharay et al. (2006) suggested the following three-step procedure outlined in Patz and Junker (1999):

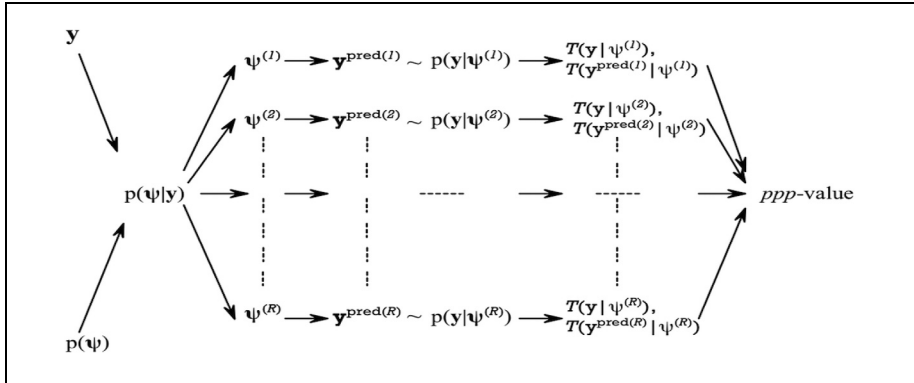


Figure 2. Graphical demonstration of posterior predictive model checking (PPMC) method. y , observed data; y^{pred} , predicted data; ψ , model parameters; $p(\psi)$, prior distributions of model parameters; $p(\psi|y)$, posterior distributions of model parameters; $T(\cdot)$ discrepancy measures.

1. Draw the item parameter and person parameter estimates for the proposed model from the posterior distribution (see, Equation 9).
2. Draw y^{pred} from the proposed model given by Equation 10 based on the drawn item parameter and person parameter estimates in Step 1.
3. Compute the values of observed and predictive discrepancy measures (e.g., item-fit statistics or descriptive statistics only based on data) from the above draws of parameters and data set.

The data–model fit can be evaluated based on the computed PPP-values, which are given by the Equation 11. Figure 2, a modification of a schematic presented by Sinharay et al. (2006), graphically demonstrates the detailed procedure of using the PPMC method to evaluate the data-model fit.

A posterior predictive probability (PPP) value near .5 indicates that there are no systematic differences between the realized and predictive values, and thus an adequate fit of the model (Sinharay et al., 2006). In the results section, the item-wise data–model fit for the item responses, RTs, and VFCs will be calculated by averaging over all the persons’ PPP-values for each item, and the results will be reported in a table later.

Real Data Analysis

The proposed ML-MG three-way joint model of item responses, RTs, and VFCs were fitted to the data. Parameter estimates of the Level-1 measurement models were reported. In addition, the trade-offs of the person-side and item-side parameters at

Table 1. Number of Subjects in Each Condition.

	Condition 1	Condition 2	Condition 3
Number of subjects	93	98	107

Note. Condition 1: Participants in the control condition who did not receive any test preparation materials. Condition 2: Participants received items that were similar to their exam. Condition 3: Participants in the third condition would receive similar exam questions and the answer key.

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 $N = 335$ university students who had normal or corrected vision were recruited for the study. Students were asked to take a test consisting of $I = 10$ questions related to verbal reasoning. The test material used for the current study followed the structure of a high-stakes credentialing exam. Data from subjects who did not complete the designed tasks were excluded from the following analysis, leaving $N = 298$ participants in the study. Table 1 lists the numbers of subjects in each condition.

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 were recorded, and RTs and gaze fixation counts of the area of interest, were measured simultaneously as the participants answered the assessment questions. The position–variance method (Jacob & Karn, 2003) was the default algorithm for processing fixation counts. The interested reader can visit the Gazepoint 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. Figure 3 displays the collected item response (transferred into the proportions of correctness), item response times, and VFCs side-by-side for the ten items.

Data Visualization Across Different Experimental Conditions

Based on the descriptive statistics of the collected data, it is not hard to gain insights about the group differences by comparing the means for each variable across three

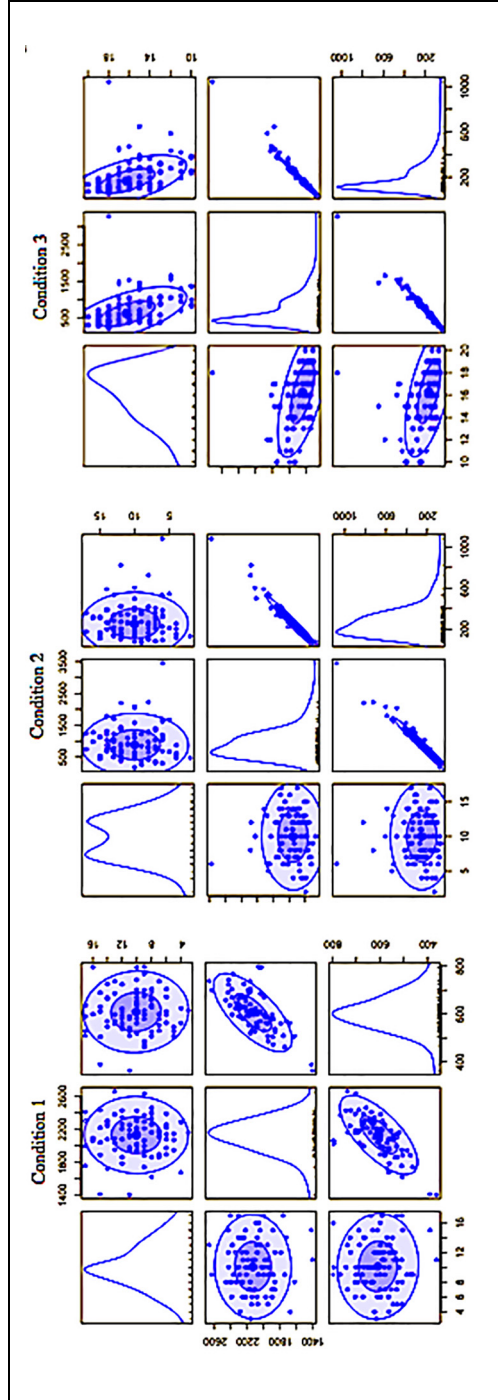


Figure 3. Scatterplots of essential variables across three conditions. The variable names showing in the matrix from the top left to the bottom right are: total.score, total.gaze, total.time. The distribution of each variable is listed on the diagonal of the plot matrix. The bivariate scatterplots are listed on the off-diagonal.

experimental conditions. To have better understanding about the data and to properly model it for accurate inferences, the collected data was explored by showing the bivariate scatterplots of the major variables, which are quite useful and straightforward for interpreting trends and the associations among the key variables. All scatterplots were created based on the total scores for each individual, see Figure 3. For instance, on the top left of Figure 3, the total scores were calculated by summing up the 10 item scores. Visualizing the key variables is helpful to understand the most appropriate means for answering our research questions.

In Condition 1, distribution of each variable (listed on the diagonal of the plot matrix) was relatively normal. In addition, by looking at the bivariate normal density contours (listed on the off diagonal), the correlation between total score and total time, and the one between total score and total gaze are expected to be relatively weak due to its round contours. In contrast, the correlation between total gaze and total time is expected to be positive due to its up-tilted elliptical contour. In terms of Condition 2, their panel plots show bimodal and skewed distributions, which are different from those shown in Condition 1. The bimodal distribution may indicate a mix of two groups of test-takers with different test-taking strategies, responding to the items in different ways. In addition, the total gaze and total RTs are skewed to the right, which means, on average, test-takers tend to spend a shorter time finishing the items on their tests. Regarding Condition 3, generally speaking, all the distributions listed on the diagonal are relatively more skewed with less variability. The distributions are very skewed with high peaks, which indicate the responding behavioral patterns of test takers under Condition 3 are dramatically different from test-takers in the other conditions. The results show that test-takers in this group correctly answered the items more rapidly with less visual attention. Also, all the test-takers in Condition 3 behaved more alike.

Accessing Data Model fit based on PPMC Method

Table 2 shows the item-wise PPP-values for assessing data model fits across conditions. The PPP-values were summarized based on 10,000 iterations after dropping burn-in iterations with thinning of 4. On comparison of the PPP-values for the three models across 10 items, in general, most of the PPP-values were close to .5 for I-PL, RT, and NBFM model, indicating satisfactory fit of all three measurement models. One thing to note, the PPP-values for Condition 3 were more extreme than the ones in Conditions 1 and 2. Yet all of them are within the range between 0.05 and 0.95.

To understand and evaluate the pattern differences in test-taking behaviors across distinct experimental conditions, a multiple-group joint three-way factor model of item responses, RTs, and VFCs were fitted separately to the data in different conditions. Parameter estimates of the Level-1 measurement models across the three conditions were reported. Moreover, the distinctions of the associations of the person-side and the item-side parameters were reported by showing the corresponding covariance estimates across the contrasting experimental conditions.

Table 2. Item-Wise PPP Values for Assessing Data-Model Fits Across Conditions.

Condition	PPP-VAL	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
C1	I-PL	0.464	0.468	0.461	0.592	0.685	0.485	0.671	0.611	0.521	0.397
	RT	0.540	0.364	0.554	0.551	0.559	0.469	0.696	0.517	0.567	0.419
	NBFM	0.225	0.393	0.459	0.419	0.368	0.455	0.416	0.352	0.479	0.461
C2	I-PL	0.507	0.505	0.487	0.508	0.495	0.500	0.500	0.506	0.489	0.490
	RT	0.547	0.531	0.549	0.532	0.524	0.533	0.540	0.511	0.518	0.501
	NBFM	0.748	0.751	0.680	0.626	0.714	0.713	0.728	0.677	0.692	0.674
C3	I-PL	0.492	0.510	0.503	0.481	0.521	0.488	0.529	0.521	0.511	0.506
	RT	0.851	0.773	0.602	0.186	0.454	0.449	0.571	0.396	0.324	0.234
	NBFM	0.939	0.950	0.577	0.472	0.453	0.733	0.822	0.293	0.531	0.452

Note. PPP-VAL = posterior predictive p value; I-PL = one-parameter logistic model; RT model = log-normal response time model; NBF model = negative binomial visual fixation counts model.

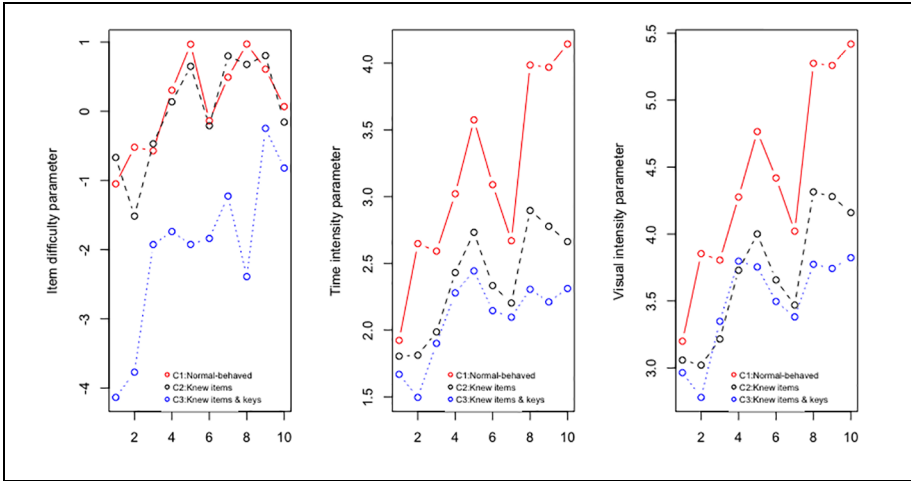


Figure 4. Scatterplots for item parameter estimates. A loess nonparametric smoothed curve is plotted for each scatterplot.

Impact of Having Preknowledge of Test Items on Item Characteristics

To evaluate the impact of having preknowledge of test questions on the properties of test items (see Figure 4), Table 3 displays a comparison of item parameter estimates of the proposed model with regard to the three experimental conditions. In general, item difficulties (\hat{b}), time intensities ($\hat{\beta}$), and visual intensities (\hat{m}), on average, tend to show lower values in the Condition 3 than the other two conditions (see Table 3). This is potentially attributable to the fact that test-takers tend to spend less time and less visual effort on a test with which they were more familiar by practicing the similar items in advance.

Item Difficulty Estimates Across Conditions. In general, items, on average, appeared to be much easier in Condition 3 than the other two conditions. Across item difficulties, \hat{b} ranged from -1.06 to 0.99 in the Condition 1, varied from -1.55 to 0.79 in the Condition 2, and fluctuated from -4.09 to -0.24 , which is demonstrated in Table 3. To test the significance of item drifts in difficulties due to the preknowledge effect across items, Table 4 presents the differences in difficulties $D_{b(i)}$, standard deviations $SD_{D_{b(i)}}$, and, Wald statistics $W_{D_{b(i)}}$ across items. Intriguingly, the difference ($D_{b(1,2)}$) in item difficulties between the Condition 1 and Condition 2 is not as large as the difference ($D_{b(1,3)}$) between Conditions 1 and 3 (see Figure 4). This was also supported by Wald statistics, which showed insignificant drifts in difficulties between Conditions 1 and 2, except the Item 2 ($W_{D_{b(1,2)}} = 2.904, Z_{critic} = 1.96$). This might indicate that practicing items beforehand without knowing the answer keys has limited impact on item difficulties. In contrast, the item difficulties ($D_{b(2,3)}$) would

Table 3. Item Parameter Estimates Across Different Experimental Conditions.

Condition	Item	Model									
		I-PL		RT				NBFM			
		<i>b</i>	<i>SD</i>	β	<i>SD</i>	ν	<i>SD</i>	<i>m</i>	<i>SD</i>	α	<i>SD</i>
C1	1	-1.05	0.24	1.92	0.05	0.43	0.03	3.20	0.03	0.19	0.012
	2	-0.52	0.24	2.65	0.03	0.21	0.02	3.85	0.02	0.14	0.006
	3	-0.6	0.23	2.59	0.03	0.24	0.02	3.80	0.02	0.13	0.011
	4	0.31	0.22	3.02	0.04	0.36	0.03	4.27	0.04	0.04	0.003
	5	0.97	0.24	3.57	0.03	0.21	0.02	4.76	0.02	0.07	0.007
	6	-0.14	0.22	3.08	0.04	0.33	0.02	4.42	0.03	0.05	0.004
	7	0.50	0.22	2.67	0.04	0.36	0.03	4.02	0.03	0.07	0.005
	8	0.99	0.24	3.98	0.03	0.18	0.01	5.27	0.01	0.06	0.007
	9	0.6	0.22	3.97	0.02	0.21	0.02	5.26	0.02	0.03	0.003
	10	0.09	0.23	4.14	0.02	0.18	0.01	5.42	0.01	0.05	0.005
C2	1	-0.67	0.24	1.81	0.06	0.44	0.03	3.07	0.06	0.12	0.011
	2	-1.55	0.29	1.81	0.06	0.39	0.03	3.03	0.05	0.13	0.012
	3	-0.48	0.22	1.99	0.05	0.34	0.03	3.22	0.05	0.12	0.010
	4	0.13	0.23	2.43	0.06	0.45	0.03	3.73	0.06	0.05	0.005
	5	0.64	0.24	2.74	0.06	0.42	0.03	4.01	0.05	0.05	0.004
	6	-0.21	0.23	2.34	0.06	0.42	0.03	3.66	0.05	0.06	0.005
	7	0.78	0.24	2.21	0.06	0.42	0.03	3.48	0.06	0.07	0.006
	8	0.66	0.23	2.90	0.07	0.56	0.04	4.32	0.06	0.03	0.002
	9	0.79	0.24	2.78	0.08	0.63	0.05	4.38	0.07	0.02	0.002
	10	-0.17	0.23	2.66	0.08	0.62	0.05	4.17	0.07	0.02	0.002
C3	1	-4.09	0.55	1.69	0.06	0.4	0.03	2.98	0.06	0.13	0.012
	2	-3.76	0.51	1.51	0.06	0.41	0.03	2.80	0.06	0.15	0.014
	3	-1.91	0.31	1.92	0.08	0.59	0.04	3.36	0.07	0.05	0.005
	4	-1.72	0.28	2.29	0.08	0.73	0.05	3.82	0.08	0.03	0.003
	5	-1.92	0.29	2.46	0.06	0.41	0.03	3.77	0.06	0.05	0.004
	6	-1.81	0.30	2.16	0.06	0.43	0.03	3.52	0.06	0.07	0.006
	7	-1.23	0.28	2.11	0.06	0.41	0.03	3.40	0.06	0.08	0.007
	8	-2.39	0.33	2.32	0.07	0.52	0.04	3.80	0.06	0.04	0.004
	9	-0.24	0.25	2.23	0.07	0.56	0.04	3.76	0.07	0.04	0.004
	10	-0.81	0.28	2.33	0.07	0.55	0.04	3.84	0.07	0.04	0.004

Note. I-PL = one-parameter logistic model; RT = log-normal response time model; NBF = negative binomial visual fixation counts model.

decrease greatly if the test-takers practice the equivalent items with keys. All item drifts were significant between Conditions 1 and 3, and between Conditions 2 and 3 by comparing with the cutoff values: ± 1.96 .

Time Intensity Estimates Across Conditions. Similarly, test-takers who practiced the items or knew the answer keys beforehand tended to take less time to finish their tests. By averaging the time intensities across the 10 items, β (the averaged time intensity) is 3.21 in Condition 1; 2.367 in Condition 2, and 2.102 in Condition 3 (see

Table 4. Impact of Having Preknowledge of Test Items on Item Drifts.

Par.	Drift	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
b	$D_b(1,2)$	-0.381	0.996	-0.101	0.168	0.319	0.071	-0.308	0.294	-0.196	0.225
	$SD_{D_b}(1,2)$	0.337	0.343	0.319	0.325	0.339	0.317	0.331	0.330	0.324	0.314
	$W_{D_b}(1,2)$	-1.131	2.903	-0.317	0.518	0.941	0.225	-0.931	0.890	-0.606	0.718
	$D_b(1,3)$	3.088	3.253	1.356	2.042	2.892	1.700	1.717	3.362	0.856	0.888
	$SD_{D_b}(1,3)$	0.640	0.553	0.377	0.370	0.396	0.370	0.349	0.416	0.348	0.345
	$W_{D_b}(1,3)$	4.824	5.882	3.597	5.519	7.304	4.596	4.919	8.082	2.459	2.574
	$D_b(2,3)$	3.469	2.257	1.457	1.874	2.573	1.629	2.025	3.068	1.052	0.662
	$SD_{D_b}(2,3)$	0.630	0.569	0.381	0.378	0.384	0.374	0.368	0.410	0.346	0.347
	$W_{D_b}(2,3)$	5.506	3.967	3.825	4.956	6.702	4.355	5.503	7.483	3.041	0.191
	$D_m(1,2)$	0.141	0.833	0.590	0.547	0.764	0.762	0.553	0.960	0.979	1.258
	$SD_{D_m}(1,2)$	0.065	0.060	0.058	0.073	0.061	0.067	0.068	0.066	0.076	0.074
	$W_{D_m}(1,2)$	2.167	13.888	10.175	7.497	12.522	11.374	8.129	14.538	12.879	17.006
m	$D_m(1,3)$	0.235	1.074	0.458	0.479	1.011	0.923	0.641	1.501	1.516	1.595
	$SD_{D_m}(1,3)$	0.065	0.064	0.076	0.090	0.061	0.067	0.067	0.069	0.072	0.067
	$W_{D_m}(1,3)$	3.609	16.779	6.026	5.321	16.577	13.772	9.566	21.748	21.060	23.805
	$D_m(2,3)$	0.094	0.241	-0.132	-0.068	0.247	0.161	0.088	0.541	0.538	0.337
	$SD_{D_m}(2,3)$	0.087	0.086	0.093	0.104	0.085	0.087	0.087	0.096	0.103	0.099
	$W_{D_m}(2,3)$	1.078	2.797	-1.422	-0.658	2.910	1.847	1.014	5.637	5.219	3.399
	$D_\beta(1,2)$	0.118	0.837	0.606	0.591	0.842	0.756	0.467	1.089	1.191	1.479
	$SD_{D_\beta}(1,2)$	0.079	0.066	0.064	0.077	0.067	0.072	0.076	0.077	0.084	0.079
	$W_{D_\beta}(1,2)$	1.499	12.675	9.470	7.674	12.564	10.503	6.147	14.148	14.175	18.725
	$D_\beta(1,3)$	0.255	1.152	0.692	0.743	1.132	0.944	0.574	1.680	1.758	1.832
	$SD_{D_\beta}(1,3)$	0.077	0.067	0.081	0.094	0.067	0.074	0.075	0.073	0.076	0.075
	$W_{D_\beta}(1,3)$	3.308	17.195	8.539	7.903	16.894	12.757	7.651	23.015	23.135	24.421
β	$D_\beta(2,3)$	0.136	0.315	0.086	0.152	0.290	0.188	0.107	0.591	0.568	0.352
	$SD_{D_\beta}(2,3)$	0.092	0.088	0.098	0.110	0.092	0.092	0.093	0.104	0.110	0.106
	$W_{D_\beta}(2,3)$	1.481	3.585	0.873	1.382	3.154	2.041	1.147	5.680	5.160	3.324

Note. Par. = Parameter; b = difficulty; m = visual intensity; β = time intensity; $D_{b(\cdot)}$ = difference in item difficulty between two conditions; $D_{m(\cdot)}$ = difference in item visual intensity between two conditions; $D_{\beta(\cdot)}$ = difference in item time intensity between two conditions; $SD_{D_{b(\cdot)}}$ = standard deviation of posterior distribution of difference in item difficulty; $SD_{D_{m(\cdot)}}$ = standard deviation of posterior distribution of difference in item visual intensity; $SD_{D_{\beta(\cdot)}}$ = standard deviation of posterior distribution of difference in item time intensity; $W_{D_{b(\cdot)}}$ = Wald test statistic of difference in item difficulty; $W_{D_{m(\cdot)}}$ = Wald test statistic of difference in item visual intensity; $W_{D_{\beta(\cdot)}}$ = Wald test statistic of difference in item time intensity.

Table 3). By taking the exponential of each averaged time intensity estimate, the unit of $\hat{\beta}$ were converted into seconds. On average, test-takers in Condition 1 took about 25 seconds to finish an item, those in Condition 2 used about 11 seconds, and those in Condition 3 took about 8 seconds. The results show that, on average, test-takers in Condition 3 who were practicing items beforehand with answer keys worked three times faster than those in Condition 1 who did not receive any test preparation materials on answering an item. In addition, the Wald statistics demonstrated in Table 4 of testing item drifts in time intensities $W_{D_{\beta(\cdot)}}$ confirm that the mean differences $D_{\beta(\cdot)}$

Table 5. Person-Side Correlation Matrix Estimates.

Parameter	Conditions					
	C1		C2		C3	
	Mean	CI	Mean	CI	Mean	CI
$Cor_{\theta, \omega}$	-0.011	(-0.244, 0.227)	-0.193	(-0.437, -0.108)	-0.678	(-0.812, -0.505)
$Cor_{\theta, \tau}$	0.005	(-0.239, 0.251)	0.24	(-0.020, 0.327)	0.672	(0.496, 0.810)
$Cor_{\omega, \tau}$	-0.152	(-0.359, -0.080)	-0.899	(-0.935, -0.886)	-0.91	(-0.942, -0.867)

Note. C1 = condition 1; C2 = condition 2; C3 = condition 3; CI = credible interval; Cor. = correlation.

are significant in average time used by test-takers on answering items between Conditions 1 and 2, and Conditions 1 and 3.

Visual Intensity Estimates Across Conditions. A trend of visual intensities similar to the summarized response patterns in the previous session was observed, which indicates test-takers familiar with the items tend to put less visual effort on searching for information to answer the questions (see Figure 4). By averaging the visual intensities across the 10 items, \bar{m} (the averaged visual intensity) is 4.427 in Condition 1; 3.707 in Condition 2, and 3.505 in Condition 3 (see Table 3). By taking the exponential of each averaged visual intensity estimate, the unit of \bar{m} were converted into counts. In general, test-takers in Condition 1 generated about 84 fixation counts to finish an item, those in Condition 2 produced about 40 fixation counts, and those in Condition 3 created about 33 fixations. The results show that, on average, that test-takers in Condition 3 put much less visual effort than participants in the other two conditions on solving questions. Moreover, the Wald statistics (see, Table 4) demonstrated the similar pattern in testing drifts in visual intensities as in time intensities $W_{D_{m(\cdot)}}$ support that the mean differences $D_{m(\cdot)}$ are significant in average visual effort put by test-takers on answering items between Conditions 1 and 2, and Conditions 1 and 3.

Impact of Having Preknowledge of Test Items on Test-Takers' Behavior

Table 5 shows the impact of having preknowledge of test items on the test-takers' behaviors. The behavioral pattern differences were demonstrated via comparison of the three person-side covariances, indicating association among the interested latent constructs (ability, working speed, and visual engagement) across the three experimental conditions. As a trend, as students gain more preknowledge of the test items the correlation between latent ability and working speed increased from 0.005 in Condition 1 (95% credible interval: -0.023 to 0.020) to 0.672 (95% credible interval: 0.496 to 0.621) in Condition 3. The increased correlation between latent ability and working speed might be caused by test-takers in Condition 3 receiving practice

items with answer keys. Therefore, they answered more items correctly than the ones who did not receive any test preparation materials.

In terms of changes in the trade-offs between the latent ability and visual engagement across conditions, Figure 5 shows that test-takers who were familiar with the test items tended to put less visual efforts on answering items. The correlation between those two latent constructs dropped from -0.011 in Condition 1 (95% credible interval: -0.244 to 0.227) to -0.678 in the Condition 3 (95% credible interval: -0.812 to -0.505). Similarly, negative trade-offs between the working speed and visual engagement were observed. The correlation ($r_{\theta, \omega}$) decreased from -0.152 in the Condition 1 (95% credible interval: -0.359 to 0.062) to -0.910 in the Condition 3 (95% credible interval: -0.942 to -0.867). This result infers that as test-takers knew the answer keys of practice items, they favored quickly answering the questions without elaborately paying attention to the content (see Figure 4.11).

Discussion

As is becoming increasingly evident, gaining a more comprehensive understanding of complex test-taking behaviors necessarily requires collecting and modeling supplementary information beyond conventional item responses. To this end, technology-enhanced assessments allow the collecting of response process data, such as RTs and gaze fixation counts, that can be used to systematically reflect the characteristics regarding item parameters and spatial patterns of test-takers' cognitive capacities (Fox & Marianti, 2016). Incorporating process data information has been demonstrated to facilitate estimation of person and item parameters in IRT (Man & Haring, 2019; van der Linden et al., 2010) while providing insights of test-takers' behaviors that is hard to be identified from item responses only.

The proposed ML-MG three-way joint model can help (1) integrate visual fixation—an eye-tracking indicator—into a traditional psychometric modeling framework and (2) investigate pattern differences in the trade-offs of visual attention, working speediness and accuracy across groups. With this modeling framework, some important test takers' cognitive processes can be evaluated in a virtual-based learning system by estimating the relations among the responding accuracy, task decoding speed, and visual engagement. Those manifested relations could facilitate practitioners to better understand and classify different types of responding behaviors. Especially now, due to the outbreak of pandemic, it is essential to have tools to differentiate cheaters from normally behaved test-takers, which can keep our online delivered tests as secure as possible (Jiao & Lissitz, 2018).

Results from the real data example show that the proposed model captures the underlying patterns of data set showing a satisfactory data model fit. In addition, the proposed ML-MG three-way joint model demonstrates additional benefits. Both the associations among item parameters and trade-offs among person parameters can be assessed across groups. This may help practitioners and substantive researchers to better understand behavior nuances and cognitive processes in test-takers'

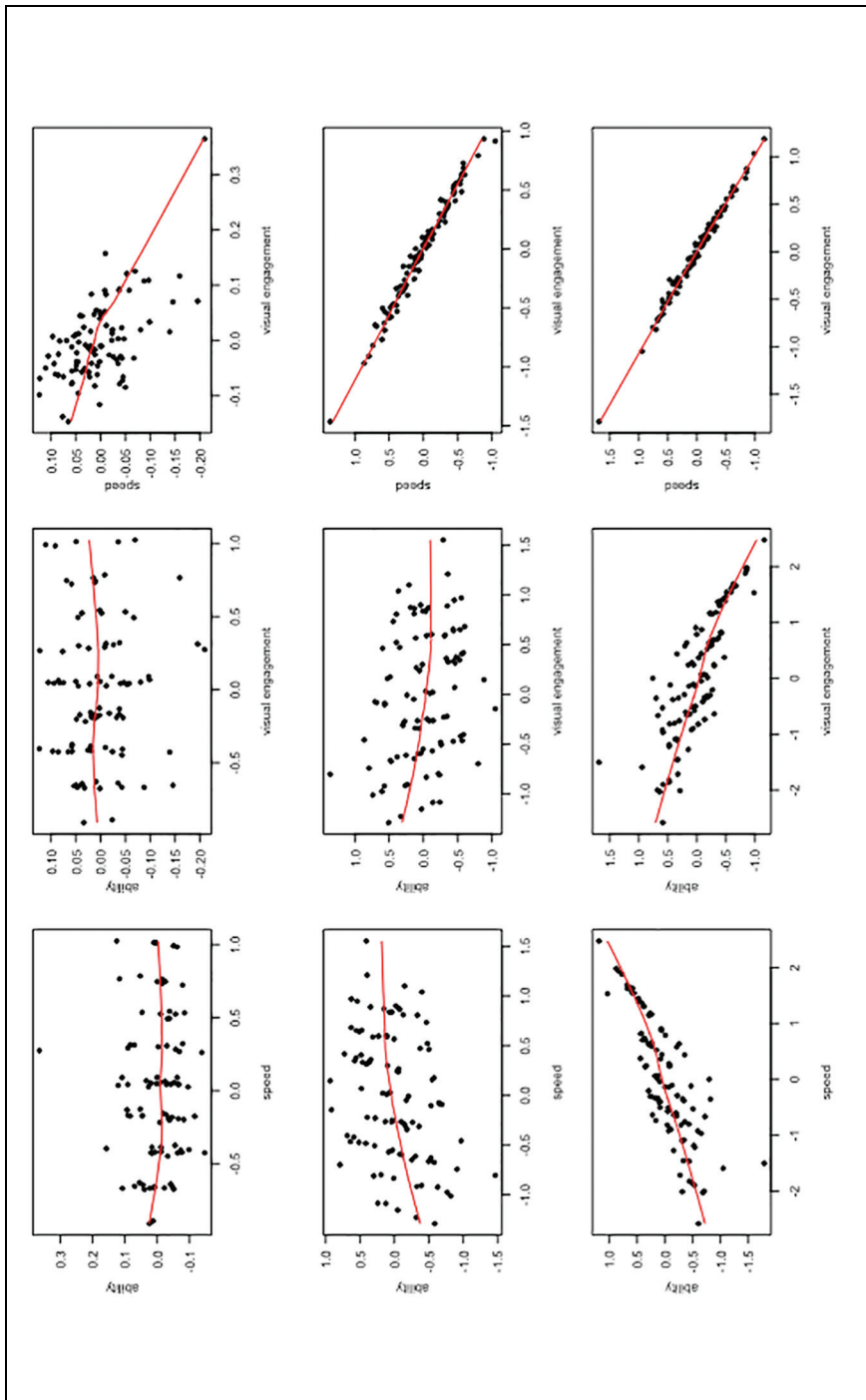


Figure 5. Scatterplots for person-side parameter estimates. A loess nonparametric smoothed curve is plotted for each scatterplot.

performance belonging to different groups in the technology-enhanced environment. For instance, with the proposed model, the impact of having preknowledge on items could be evaluated by quantifying the differences in working efficiency, visual engagement, and responding accuracy across groups.

Moreover, other eye-tracking related biometric information variables (e.g., blinking rates, pupil diameters) could be added as auxiliary information to reflect other characteristics of test-taking behaviors. For example, the diameter of the pupil has been reported to be negatively correlated with levels of fatigue (e.g., Morad et al., 2000; Yoss et al., 1970). Also, many other types of biometric information (e.g., blood oxygen level-dependent signal, electroencephalography, or heart rate) could be integrated into the current modeling framework to assess whether these involuntary bodily processes could provided any new, systematic insights into the learners' learning progressions in the ITELs. For instance, heart rate could be used to track test-takers' anxiety levels in ITELs (e.g., Friedman & Thayer, 1998). Furthermore, other background variables like gender could be added as covariates to show the difference between groups.

Lastly, the proposed model could be further expanded. An interesting next elaboration might be to model multidimensional compensatory responses (Molenaar et al., 2015) and its functional relation to RTs and VFCs rather than modeling unidimensional item responses. Of course, the measurement model can also be extended to two-parameter logistic IRT model (2PL), three-parameter logistic IRT model (3PL), or polytomous item responses in a straightforward manner as long as the sample size is sufficiently large to estimate item parameters with satisfactory precision. However, due to the budget and time constrain we had for conducting this study, 1-PL model was used to fit item responses. This is an essential elaboration as many educational and psychological tests and assessments include items that are Likert-type scaled. Finally, the current assumptions that visual engagement and working speed are constant over the entire test, and this assumption could be relaxed in a future study. This may provide *individualized* items-specific information regarding changes of behavioral patterns of test-takers over items.


Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

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

Note

1. For interpretation purposes, we here introduce a positive sign to the latent test engagement parameter which was used in the original NBF model (Man & Haring, 2019).

References

- Bergner, Y., & von Davier, A. A. (2019). Process data in naep: Past, present, and future. *Journal of Educational and Behavioral Statistics, 44*(6), 706-732. <https://doi.org/10.3102/1076998618784700>
- Birnbaum, Z. W. (1968). *On the importance of different components in a multicomponent system* (Tech. Rep.). Washington University Seattle Lab of Statistical Research.
- Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika, 82*, 1126-1148. <https://doi.org/10.1007/s11336-016-9537-6>
- Constantinides, A., Belk, M., Fidas, C., & Pitsillides, A. (2019, June). On the accuracy of eye gaze-driven classifiers for predicting image content familiarity in graphical passwords. *Proceedings of the 27th ACM conference on user modeling, adaptation and personalization* (pp. 201-205). ACM. <https://doi.org/10.1145/3320435.3320474>
- De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102. <https://doi.org/10.3389/fpsyg.2019.00102>
- 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., & Mariani, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4), 540-553. <https://doi.org/10.1080/00273171.2016.1171128>
- Friedman, B. H., & Thayer, J. F. (1998). Autonomic balance revisited: Panic anxiety and heart rate variability. *Journal of Psychosomatic Research, 44*(1), 133-151. [https://doi.org/10.1016/S0022-3999\(97\)00202-X](https://doi.org/10.1016/S0022-3999(97)00202-X)
- 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*, 733-807.
- Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173-183. <https://doi.org/10.1080/08957347.2016.1171766>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hao, J., Smith, L., Mislavy, 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, 2016*(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. In R. Radach, J. Hyona & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573-605). Elsevier.
- Jiao, H., & Lissitz, R. (2018). *Technology enhanced innovative assessment development, modeling, and scoring from an interdisciplinary perspective*. Information Age.

- Kang, H.-A., Zheng, Y., & Chang, H.-H. (2020). Online calibration of a joint model of item responses and response times in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 45*(2), 175-208. <https://doi.org/10.3102/1076998619879040>
- Levy, M., 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>
- Lord, F. M. (1952). *A theory of test scores*. Psychometric Corporation.
- Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology, 73*(2), 261-288. <https://doi.org/10.1111/bmsp.12175>
- Man, K. (2020). *Methods of integrating multi-modal data for detecting aberrant testing behaviors in large-scale assessments* [Unpublished doctoral dissertation]. University of Maryland, College Park.
- Man, K., Harring, J. R., & Liu, Y. (2020). Methods of Integrating Multi-Modal Data for Assessing Aberrant Test-Taking Behaviors. *Multivariate Behavioral Research, 55*(1), 155-156.
- Man, K., & Harring, 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., Harring, 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>
- Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment* (CRESST Report 800). National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- 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)
- Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342-366. <https://doi.org/10.3102/10769986024004342>
- Plummer, M. (2015). *JAGS: Just another Gibbs sampler* (V. 4.0.0). <http://mcmc-jags.sourceforge.net>
- 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* (pp. 363-378). Springer.
- Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Nielsen Lydiche.
- Reckase, M. D. (1972). *Development and application of a multivariate logistic latent trait model* [Unpublished doctoral dissertation]. Syracuse University.
- Rubin, D. B. (1996). Comment: On posterior predictive p-values. *Statistica Sinica, 6*, 787-792.

- Schaeffer, M., Tardel, A., Hofmann, S., & Hansen-Schirra, S. (2019). Cognitive effort and efficiency in translation revision. In E. Huertas-Barros, S. Vandepitte & E. Iglesias-Fernández (Eds.) *Quality assurance and assessment practices in translation and interpreting* (pp. 226-243). IGI Global.
- 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>
- Smith, R. W., & Prometric, T. (2004, April 2-6). *The impact of braindump sites on item exposure and item parameter drift* [Paper presentation]. Annual meeting of the American Education Research Association, San Diego, CA, United Staets.
- Su, Y. S., & Yajima, M. (2015). *R2jags: Using R to run JAGS*. (V. 0.5).
- 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., Klein Entink, R. H., & 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>
- Volodin, N., & Adams, R. (1995). *Identifying and estimating a D-dimensional item response model* [Paper presentation]. International Objective Measurement Workshop, University of California, Berkeley, CA, United States.
- Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 45-58. <https://doi.org/10.1080/15366367.2018.1435105>
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated irt model. *Journal of Educational Measurement, 43*(1), 19-38. <https://doi.org/10.1111/j.1745-3984.2006.00002.x>
- 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), 545-545. <https://doi.org/10.1212/WNL.20.6.545>