Article

Applied Psychological Measurement
2023, Vol. 47(1) 19–33

DOI: 10.1177/01466216221125177
journals.sagepub.com/home/apm
SAGE

Modeling Rapid Guessing Behaviors in Computer-Based Testlet Items

Kuan-Yu Jin[1], Chia-Ling Hsu[1], Ming Ming Chiu[2], and Po-Hsi Chen[3]

Abstract

In traditional test models, test items are independent, and test-takers slowly and thoughtfully respond to each test item. However, some test items have a common stimulus (dependent test items in a testlet), and sometimes test-takers lack motivation, knowledge, or time (speededness), so they perform rapid guessing (RG). Ignoring the dependence in responses to testlet items can negatively bias standard errors of measurement, and ignoring RG by fitting a simpler item response theory (IRT) model can bias the results. Because computer-based testing captures response times on testlet responses, we propose a mixture testlet IRT model with item responses and response time to model RG behaviors in computer-based testlet items. Two simulation studies with Markov chain Monte Carlo estimation using the JAGS program showed (a) good recovery of the item and person parameters in this new model and (b) the harmful consequences of ignoring RG (biased parameter estimates: overestimated item difficulties, underestimated time intensities, underestimated respondent latent speed parameters, and overestimated precision of respondent latent estimates). The application of IRT models with and without RG to data from a computer-based language test showed parameter differences resembling those in the simulations.

Keywords

multidimensional item response theory, mixture model, response time, rapid guessing, testlet

In the traditional view of testing, students slowly and thoughtfully try to answer test items (*solution behaviors*). However, sometimes test-takers lack motivation, knowledge, or time (*speededness*) and give quick, unreflective responses (*rapid guessing*, RG). Students' RG might occur on not only *independent* test items but also *dependent* test items associated with a common

[1]Assessment Technology and Research Division, Hong Kong Examinations and Assessment Authority, Wan Chai, Hong Kong
[2]The Education University of Hong Kong, Hong Kong
[3]National Taiwan Normal University, Taiwan

Corresponding Author:
Kuan-Yu Jin, Assessment Technology and Research Division, Hong Kong Examinations and Assessment Authority, 7/F, Dah Sing Financial Centre, 248 Queen's Road East, Wan Chai, Hong Kong.
Email: kyjin@hkeaa.edu.hk

stimulus (e.g., story for reading comprehension items, *testlets*; Wainer et al., 2007). Ignoring RG by fitting a simpler item response theory (IRT) model can bias the results, and ignoring the dependence in test-taker responses to testlet items can negatively bias standard errors of measurement. Among computer-based testing's (CBT) many benefits (textured computer test items via video or human–computer interaction, standardization of testing with fewer confounding human factors, greater flexibility in testing times and locations, quick reporting of test results, etc.), it notably records the amount of time taken by a test-taker before answering a test item (*response time*, RT), which enables the modeling of test-takers' RG on independent items and testlet items.

Because standard IRT models assume that test-takers exert maximal effort on each item, RG can bias the estimates of ability and item parameters, threaten test validity and reliability, and yield improper conclusions. Thus, many IRT models have been developed for both response accuracy and RG (Wang & Xu, 2015; Wang et al., 2018; Wise & DeMars, 2006). Properly modeling them improves the accuracy of test-taker ability estimates, enhances test validity and reliability, and informs the design of high (vs. low) time pressure tests.

Many tests have items with a common stimulus (item bundle or testlet), such as several multiple-choice (MC) items for a reading comprehension passage. Such test design usually violates the local independence assumption of a unidimensional IRT model. Specifically, responses to items within a testlet likely resemble one another more than the responses to items in different testlets. Failing to consider the local dependence among testlet items often overestimates reliability of the test (Sireci et al., 1991). Hence, testlet response models have been developed to fit assessment data with testlets (Wainer et al., 2007; Wang & Wilson, 2005).[1]

A CBT can display testlet items in two formats: one item per page (one/page) or all items on a page (all/page). Like an independent item, each one/page testlet item is sequentially shown, enabling the recording of its RT. By contrast, all/page testlet items are presented together; thus, only their collective RT is available. Hence, detecting RG under the two testlet formats might require different methods.

In this study, we review the literature of modern IRT models for RT, RG, and testlets before presenting a new class of IRT models that captures RG on tests with the two types of testlets. Then, we conduct two simulation studies to evaluate the parameter recovery for testlet data and the consequences of ignoring RG. Next, we apply our model to a computer-based reading test before discussing the results and concluding the study.

Literature Review

For any test item, a test-taker might exert effort to correctly answer it (*solution behavior*), each respectively resulting in progressively longer RT (Wise & Kong, 2005), or perform RG. Ideally, test-takers can carefully read the test item and spend time and effort to find the correct answer, resulting in the longest RT. However, some test-takers have little time, especially at the end of a test, so they might perform RG without fully reading the items, yielding extremely short RTs. Kong et al.'s (2007) study of four RT thresholds for RG yielded similar results; therefore, they suggested a 3-second threshold for all items. Studies show that RT has a bimodal distribution for most test-takers, differentiating RG behaviors from solution behaviors (e.g., Wise & Kong, 2005).

Standard IRT models for RT assume that test-takers exert maximal effort to find solutions for each test items; thus, their responses and RTs reflect their latent abilities and speed levels, respectively. However, because test-takers with shorter RTs than others might have superior ability, RT data can improve the estimation of the targeted ability, as shown in De Boeck and Jeon's (2019)

summary of joint IRT models of item responses and RT. Notably, in van der Linden's (2007) commonly used hierarchical model, a log transformation enables the statistical methods developed on normal models to be directly applied to RT. Similar in concept to a standard IRT model, the logarithmized RT for an item is regressed on two main effects at the first level of the hierarchical model: the speed parameter of a test-taker and the time intensity parameter of an item. At the second level, a bivariate normal distribution describes the relation between the latent ability and latent speed.

Solution and RG behaviors can be jointly taken into account in a measurement model. For example, Meyer (2010) integrated a two-class mixture Rasch model (Rost, 1990) with a lognormal mixture model of RT (Schnipke & Scrams, 1997) to classify a test-taker as either only RG or only showing solution behaviors, without allowing different behaviors by the same person. Wang and Xu's (2015) model has a latent indicator, allowing a test-taker to pursue either solution behavior or RG for each item. This indicator can depend on either each test-taker's RG propensity or an item-level feature (Wang et al., 2018).

The above methods for analyzing RT can model RG behaviors on independent items but not on a testlet's dependent items, whose responses and RTs might be correlated. For example, a test-taker might read the common stimulus of a testlet (e.g., a story on a reading comprehension test), have difficulty understanding it, and thereby perform RG on all the items in the testlet. Because the RT for the first item on a testlet includes reading the story, it will be long; RTs therefore might not detect RG on the first item in a testlet. (If a test-taker does not read the story and performs RG on the first testlet item, the RT will still be short). In contrast, a test-taker might only guess on the later items in a testlet, which are often more difficult than the first few items. These examples show why the existing models for RG might not apply to testlet response data. Hence, we propose a new model for RG on testlets.

New Models for RG in Testlets

A CBT can comprise independent items and testlet items and automatically record test-takers' RTs on each item. Hence, we propose a general framework for solution and RG behaviors on independent, one/page, and all/page items. We begin with the item response function for independent items

$$P_{ij} = \frac{\exp\left(\theta_i - b_j\right)}{1 + \exp\left(\theta_i - b_j\right)}, \tag{1}$$

where P_{ij} is the probability of test-taker i ($i = 1,\ldots, I$) correctly answering item j ($j = 1,\ldots, J$), θ_i is the latent ability of test-taker i, and b_j is the difficulty of item j. Equation (1) represents the one-parameter logistic model (Rasch, 1960); this model can be replaced by any other IRT model. According to past studies (Meng et al., 2015; Meyer, 2010; van der Linden, 2006; 2007), a lognormal function suitably describes the distribution of RTs in cognitive tests. Thus, we can isolate the two major facets that influence the observed response time (RT_{ij}): test-taker's speed and item's time intensity

$$\log\left(RT_{ij}\right) \sim N\left(-\tau_i + \beta_j, 1/\alpha_j^2\right), \tag{2}$$

where τ_i is the speed of test-taker i; β_j is the time intensity of item j; and α_j is the time discrimination parameter of item j, indicating the dispersion of the distribution of logarithmized RT. To model the testlet effects of the one/page items, Equations (1) and (2) are extended to the following

$$P_{ij} = \frac{\exp\left[\theta_i + \gamma_{id(j)} - b_j\right]}{1 + \exp\left[\theta_i + \gamma_{id(j)} - b_j\right]}, \tag{3}$$

$$\log(RT_{ij}) \sim N(-\tau_i - \lambda_{id(j)} + \beta_j, 1/\alpha_j^2), \tag{4}$$

where $\gamma_{id(j)}$ and $\lambda_{id(j)}$ are the interactions of test-taker i and item j within testlet d. Equation (3) presents the testlet IRT model in the Rasch framework (Wang & Wilson, 2005). Thus, $\gamma_{id(j)}$ and $\lambda_{id(j)}$ are the testlet effects on response accuracy and RT, respectively. Because the distinct RT for all/page items within a testlet is not available, quantifying the item-person interaction on the RT for all/page items is not feasible. To accommodate all/page items, equation (4) can be modified

$$\log(RT_{id}) \sim N(-\tau_i + \beta_d, 1/\alpha_d^2), \tag{5}$$

where β_d and α_d are the mean time intensity and time discrimination of all/page items within testlet d, respectively. Together, Equations (3), (4), (5) represent the testlet model for RT (TM-RT).

The TM-RT requires the estimation of several random variables: θ_i, τ_i, $\gamma_{id(j)}$, and $\lambda_{id(j)}$. As the number of testlets increases, the TM-RT becomes increasingly complex, as shown in Equations (3), (4), (5). According to the experiences of many past studies (e.g., Huang, 2020; Kim & Bolt, 2021; Man et al., 2019), using the following semi-informative priors enables efficient estimation of the TM-RT parameters via Bayesian methods with Markov chain Monte Carlo (MCMC):

$$\begin{aligned}
\begin{bmatrix} \theta_i \\ \tau_i \end{bmatrix} &\sim MVN[\mathbf{0}, \mathbf{\Sigma}], \\
\gamma_{id(j)} &\sim N\left(0, \sigma_{\gamma d}^2\right), \\
\lambda_{id(j)} &\sim N\left(0, \sigma_{\lambda d}^2\right), \\
b_j &\sim N(0,\ 0.1), \\
\beta_j &\sim N(0,\ 0.1), \\
\alpha_j &\sim N(0,\ 0.1).
\end{aligned}$$

Constraining the means of the four types of latent continuous variables (i.e., θ_i, τ_i, $\gamma_{id(j)}$, and $\lambda_{id(j)}$) to zero makes the TM-RT identifiable. The priors for the hyper-parameters are specified as follows

$$\begin{aligned}
\mathbf{\Sigma} &\sim \text{Inverse-Wishart}(\mathbf{I},\ 2), \\
\sigma_{\gamma d}^2 &\sim \text{Inverse-Gamma}(0.1,\ 0.1), \\
\sigma_{\lambda d}^2 &\sim \text{Inverse-Gamma}(0.1,\ 0.1),
\end{aligned}$$

where $\mathbf{I}$ is a 2×2 identity matrix.

Because the TM-RT does not account for test-takers' RG behaviors, we extend it to the mixture testlet model for RG behaviors (MTM-RG). For independent items, Equations (1), (2) are extended to the following:

$$P_{ij} = \Delta_{ij} \times P_{0j} + \left(1 - \Delta_{ij}\right) \times \frac{\exp\left(\theta_i - b_j\right)}{1 + \exp\left(\theta_i - b_j\right)}, \tag{6}$$

$$\log\left(RT_{ij}\right) \sim N\left(\Delta_{ij} \times \beta_0 + \left(1 - \Delta_{ij}\right) \times \left(-\tau_i + \beta_j\right), 1/\left(\Delta_{ij} \times \alpha_0^2 + \left(1 - \Delta_{ij}\right) \times \alpha_j^2\right)\right), \tag{7}$$

where P_{0j} is the uniform probability of getting the correct answer for test-takers who perform RG rather than use solution behaviors (i.e., $1/k_j$, where k_j is the number of options for MC item j); Δ_{ij} indexes the latent group membership of test-taker i on item j (i.e., 1 = RG behaviors and 0 = solution behavior); and β_0 and α_0 are the mean time intensity and time discrimination parameters for RG behaviors, respectively. Equation (7) also suggests that the observed RT for solution and RG behaviors forms a bimodal distribution. For one/page items, Equations (3) and (4) can be extended as follows

$$P_{ij} = \Delta_{ij} \times P_{0j} + \left(1 - \Delta_{ij}\right) \times \frac{\exp\left[\theta_i + \gamma_{id(j)} - b_j\right]}{1 + \exp\left[\theta_i + \gamma_{id(j)} - b_j\right]}, \tag{8}$$

$$\log\left(RT_{ij}\right) \sim N\left(\Delta_{ij} \times \beta_0 + \left(1 - \Delta_{ij}\right) \times \left(-\tau_i - \lambda_{id(j)} + \beta_j\right), 1\Big/\left(\Delta_{ij} \times \alpha_0^2 + \left(1 - \Delta_{ij}\right) \times \alpha_j^2\right)\right). \tag{9}$$

For individual and one/page items, Δ_{ij} is subject to the likelihood of item response function and observed RT. Because the total (or mean) RT for all/page items within a testlet does not help identify RG on a specific item, the item response function for all/page items in the MTM-RG is identical to Equation (8), and Equation (5) can be applied to deal with RT. In this way, only the likelihood of the item response function helps identify RG on all/page items.

The latent membership Δ_{ij} in the MTM-RG is a binary result of a test-taker's decision on a specific item and depicts the dependency between response accuracy and speed. It can be modeled as either $\Delta_{ij} \sim$ Bernoulli(π_i) or $\Delta_{ij} \sim$ Bernoulli(π_j). The person-specific guessing proportion parameter π_i represents the RG propensity of test-taker i (Wang & Xu, 2015). In contrast, the item-specific parameter π_j denotes the marginal probability of test-takers performing RG on item j (Jin et al., 2022; Wang et al., 2018), which helps detect test-takers' RG on items near the end of a test (*speededness*). We discuss these two methods further in the simulation studies.

The label switching problem in mixture modeling can be resolved by distinguishing between solution and RG behaviors in the MTM-RG (Equations (6), (7), (8), (9)). We add a conservative constraint of $\beta_j \geq \beta_0$ to ensure the identification of RG only when its RT is significantly lower than the expected RT of a solution behavior for each item. To realize this constraint, β_0 is restricted to follow a truncated normal distribution with an upper bound of the minimum of β_j during calibration.

Many factors may influence the performance of the MTM-RG. For example, the precisions of the γ- and λ-parameters are linked to the number of items within a testlet. Among testlets, those with more items often yield greater precision. In addition, a larger sample size can improve estimation of π_j for each item. Likewise, capturing test-takers' marginal probabilities of RG (i.e., π_i) requires more test items.

Item properties also influence the estimation of Δ_{ij}. To distinguish normal and RG test-takers, ideally, the probability of the correct answer for an independent or testlet item differs sharply for their corresponding responses, and their RTs show a bimodal distribution. If their item response functions are similar or their RTs are not clearly bimodal, these new models are unlikely to perform well.

Simulation I

Design. We evaluated the proposed model's parameter recovery and the consequences of ignoring RG and testlet effects via a simulation of 2000 test-takers responding to 30 dichotomously scored items; these include 10 independent items, 2 one/page testlets with 5 items each, and 2 all/page testlets with 5 items each. Because the relationship between the ability and speed latent traits (θ, τ) could vary in practice (Bolsinova et al., 2017), in the simulation, we assumed that the two random

variables were weakly and positively correlated. Thus, the person parameters were generated from the following distributions: $[\theta_i, \tau_i]' \sim MNV(\mathbf{0}, \mathbf{\Sigma})$. The variances of θ and τ were set at 1 and 1.44, respectively, and the correlation of the two variables was .2. The additional factors for the 2 one/page testlets were set at $\gamma_{i1(j)} \sim N(0, 0.8)$, $\gamma_{i2(j)} \sim N(0, 0.8)$, $\lambda_{i1(j)} \sim N(0, 0.8)$, and $\lambda_{i2(j)} \sim N(0, 0.8)$, and those for the 2 all/page testets were set at $\gamma_{i3(j)} \sim N(0, 0.8)$ and $\gamma_{i4(j)} \sim N(0, 0.8)$. The item parameters were randomly generated from the following distributions: $b_j \sim N(0, 1)$, $\beta_j \sim N(4, 0.25)$, and $\alpha_j \sim N(1, 0.0625)$. In the no RG condition, item responses were generated via the TM-RT. In the RG condition, two marginal probabilities of RG (π_j) were jointly manipulated: 10% for items 1–5, 11–15, and 21–25 and 20% for items 6–10, 16–20, and 26–30. For RG behaviors, $\beta_0 = 2$ and $\alpha_0 = 0.5$, following the parameter estimates in our empirical example. For each condition, 100 data sets were generated. The TM-RT and MTM-RG were fit to the simulated data. Two constrained models of the MTM-RG were also examined. The first constrained model without local dependence parameters (denoted as MM-RG) was fit to the data to determine whether ignoring the dependence among testlet items affects parameter estimation. The second constrained model, in which only item responses are analyzed (denoted as No-RT; equation (8) only), was fit to the data to determine whether the π-parameters can be precisely recovered when RT is not available.

Analysis

We estimated the parameters of the four IRT models, TM-RT, MTM-RG, MM-RG and No-RT, using JAGS (Plummer, 2017) with the R2jags package (Su & Yajima, 2015) in R using three chains with 20,000 iterations. The first 10,000 burn-in iterations were dropped, and the subsequent 10,000 iterations were retained to form the posterior distributions of the estimated parameters. The expected posteriori value (i.e., the mean) of each parameter served as its point estimate. The potential scale reduction factor (PSRF) (Gelman & Rubin, 1992) helped monitor the convergence of the posterior distributions. The JAGS code to run the MTM-RG in R is publicly accessible on the Open Science Framework (https://osf.io/fkcva/).

To evaluate parameter recovery, the bias and root mean square error (RMSE) of item estimates were computed across 100 replications

$$Bias(\omega) = \sum_{r=1}^{100} \left(\widehat{\omega}_r - \omega\right)/100, \tag{10}$$

$$RMSE(\omega) = \sqrt{\sum_{r=1}^{100} \left(\widehat{\omega}_r - \omega\right)^2/100}, \tag{11}$$

where ω and $\widehat{\omega}_r$ are the true and estimated values in the r-th replication, respectively.

We expect the following results: When RG behaviors do not occur (i.e., data generated via the TM-RT), the MTM-RG and TM-RT show similar parameter recovery. When RG behaviors occur (i.e., data generated via the MTM-RG), the MTM-RG recovers the parameters well, whereas the TM-RT yields biased parameter estimates.

Results

Throughout the 100 replications for estimated parameters under the TM-RT or MTM-RG across the RG and no RG conditions, the PSRF values were between 1.00 and 1.05, suggesting satisfactory convergence of the two models. Under the no RG condition, both the TM-RT (data

generating model) and the MTM-RG yielded similar and accurate parameter estimates (see Appendices 1-4). Hence, fitting the MTM-RG to data generated from the TM-RT yielded good parameter recovery.

Appendices 5-8 list the parameter recovery for the four models under the RG condition. Fitting the TM-RT yielded biased estimates for item difficulty ($M = 0.229$; range: −0.206 to 0.844), time intensity (−0.206; −0.591 to −0.009), and time discrimination (−0.200; −0.486 to −0.024). Likewise, compared to the MTM-RG, the TM-RT showed larger RMSE values for item difficulty (0.271; 0.079–0.847), time intensity (0.223; 0.048–0.595), and time discrimination (0.204; 0.037–0.486). These results show that when a measurement model ignores substantial RG behaviors, test items can be mistakenly evaluated as more difficult, less time-consuming, or less sensitive to differences in respondents' speeds. Moreover, the TM-RT uniformly underestimated the variances of random variables, which tends to lead to an overestimation of testlet effects.

By contrast, the MTM-RG, compared to the TM-RT, yielded accurate parameter estimates with smaller biases (for item difficulty [−0.004; −0.135 to 0.044], time intensity [−0.010; −0.021 to 0.002], and time discrimination [0.003; −0.006 to 0.013]) and RMSE values (for item difficulty [0.122; 0.080–0.252], time intensity [0.055; 0.043–0.071], and time discrimination [0.033; 0.025–0.050]). Notably, the bias and RMSE of π_j for the independent and one/page items in the MTM-RG neared zero. Without RT information to identify RG responses to the all/page items (21–30), their parameter recovery was slightly poorer.

Fitting the MM-RG underestimated the time discrimination parameters for one/page test items under both conditions. On the contrary, fitting the No-RT based on limited information from response patterns yielded poor estimation for π-parameters and other item parameters. In brief, measurement models should jointly consider the occurrence of RG based on RT (if available) and dependence among items within a testlet.

Simulation 2

Design. Simulation 2 investigated the impacts of RG on person estimates on the same test as that in Simulation 1. Instead of specifying the proportion of RG for each item, the items were classified into five groups (500 respondents per group), with each group having a different level of RG (π_i): 0%, 20%, 40%, 60%, and 80%. The 0% RG (no RG) group served as the baseline for comparison. Respondents in other RG groups rapidly guessed on their specified proportion (e.g., 20%) of both independent and testlet items.

Analysis

To evaluate parameter recovery, we computed the person estimates of θ and τ, the correlations of their estimates and their true values, and their standard errors. For both θ and τ, we expected higher correlations among their estimates and their true values (a) for lower RG levels and (b) in the MTM-RG than in the TM-RT. For both θ and τ, we also expected larger standard errors for larger proportions of RG in the MTM-RG only, not the TM-RT.

Results

The results largely confirmed our expectations. Figures 1a and 1b show the separation of correlation between their estimates and true values for the five RG groups in the MTM-RG. Achieving an unbiased θ estimate became more difficult as RG increased. For both θ and τ, the correlations were generally lower in the TM-RT than in the MTM-RG across different RG levels. Also, when the proportion of RG was higher, the precisions of the θ and τ estimates were

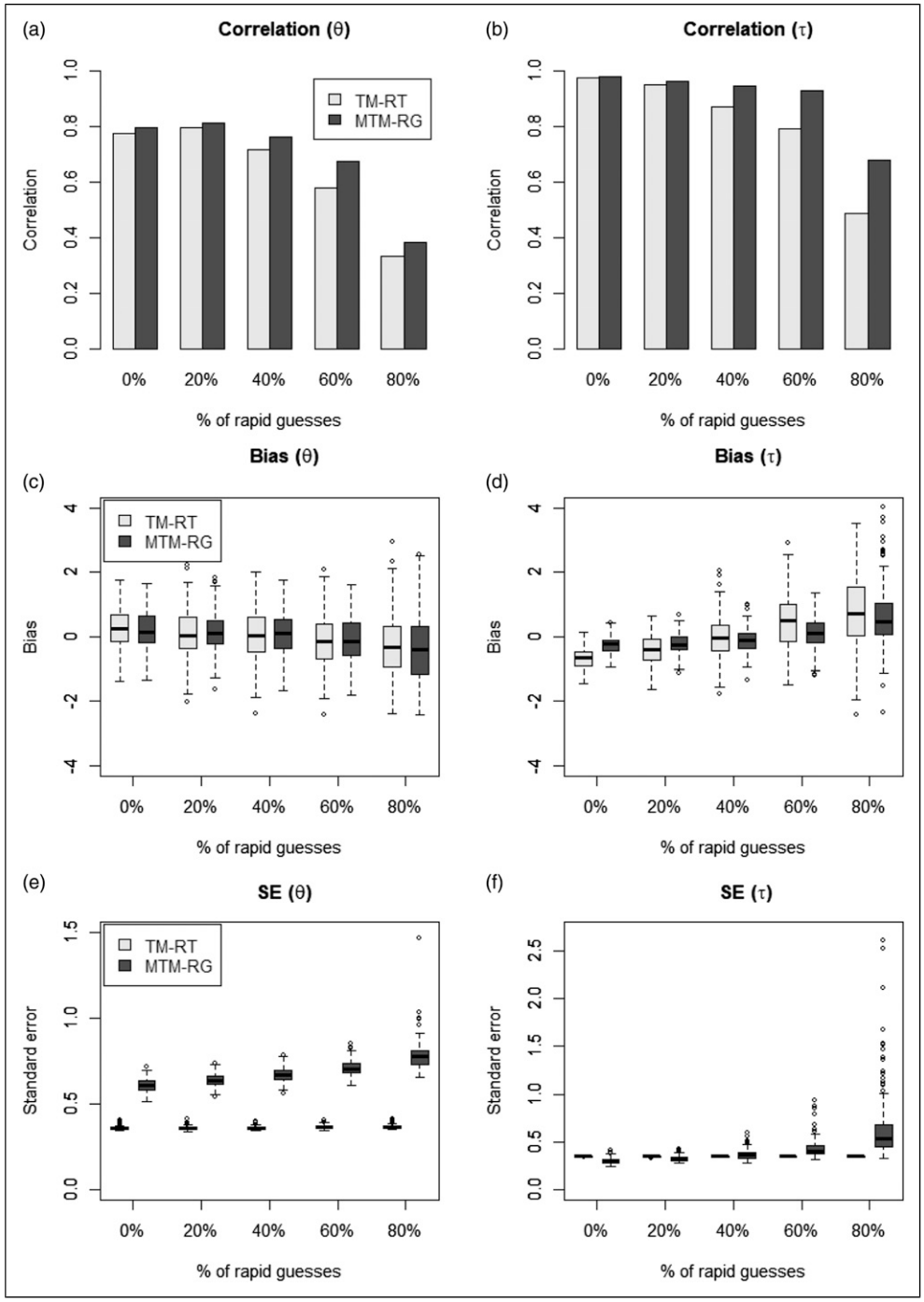

Figure 1. Correlation, bias, and standard error of person estimates for the TM-RT and MTM-RG in Simulation 2.

either lower or not substantially different. Notably, adding test-takers with higher RG decreased the correlations between these estimates and their true values. For θ, when 0% and 20% RG groups were jointly considered, the correlation under the MTM-RG was .86. When other RG groups were sequentially included from low to high, the correlations were .84, .79, and .74, respectively. The correlations for τ showed a similar pattern.

Figures 1c and 1d show the separation of bias for the different RG groups in the MTM-RG. Achieving unbiased θ and τ estimates became more difficult as RG increased. For both θ and τ, higher proportions of RG yielded larger standard errors for the MTM-RG but similar standard errors for the TM-RT (see Figures 1e and 1f). Because the standard errors of the TM-RT do not account for RG, fitting the TM-RT to data with RG can overestimate test reliability.

An Empirical Example

We fit the TM-RT and MTM-RG to the test of Chinese as a foreign language (TOCFL) measuring non-Chinese speakers' reading proficiency in Taiwan. A computer-based system randomly administers reading items and testlets within 50 min and records test-takers' item responses and RT. Test-takers can return to previous items (or testlets) and modify their answer, in which case the RT of the last attempt is recorded and analyzed.

We analyzed 907 test-takers' responses to one booklet of the reading test with 50 items. The booklet comprised 3 independent items, 11 one/page testlets, and 3 all/page testlets, with 2–6 items per testlet. The data matrix was complete (i.e., no missing responses). Figure 2 shows the histograms of RT for four selected items. Three of them show a bimodal distribution, but item 17 does not. The RT on item 17 (and the subsequent items of the one/page testlets) resembles an exponential distribution, indicating that the separation in RT is not salient for solution and RG behaviors. These results are prima facie evidence of test-takers' RG behaviors.

Analysis

Using the same priors as in Simulation 1, we identified the superior model with the lower deviance information criterion (DIC) (Spiegelhalter et al., 2002). We compared the item difficulties, time intensities, and time discriminations of the TM-RT with those of the MTM-RG. We also computed the sizes of the testlet effects on response accuracy and RT, $\sigma_\gamma^2/\sigma_\theta^2$ and $\sigma_\lambda^2/\sigma_\tau^2$, respectively.

Because the test-takers likely performed RG, we expected the MTM-RG to outperform the TM-RT. The extent to which the 50 reading items triggered RG was of interest; therefore, the item-specific guessing proportion parameter π_j was specified in the MTM-RG. Based on Simulation 1, a measurement model that ignores substantial RG behaviors, namely, the TM-RT (rather than the MTM-RG), tends to assess items to (a) be more difficult, (b) be less time intensive for individual items and one/page testlets, and (c) have less time discrimination for individual items and one/page testlets. Compared with the MTM-RG, the TM-RT can also (d) underestimate the variances of random variables, thereby overestimating testlet effects. We also expect (e) greater response accuracy to accompany shorter RT, (f) substantial and varied RG, and (g) varied answering strategies.

Results

The simulation results generally supported our expectations. Because the DIC was lower for the MTM-RG than for the TM-RT (376,770 < 425,635), the MTM-RG fit the data better, suggesting substantial RG and testlet effects. The estimated item difficulties were generally higher in the TM-RT than in the MTM-RG ($b_{\text{TM-RT}} - b_{\text{MTM-RG}}$: $M = 0.27$, range: −0.18–1.72; see Figure 3a).

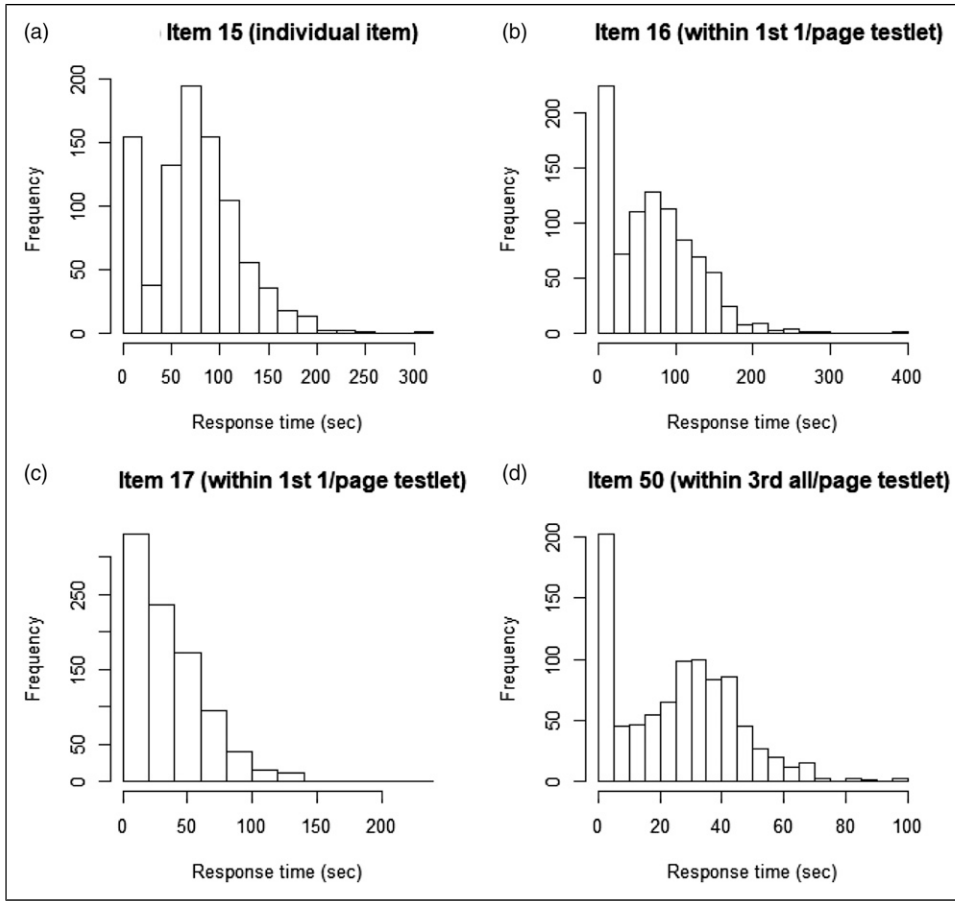

Figure 2. Histogram of RTs in the reading test. *Note.* The mean RT on the 3rd all/page testlet was used to draw Figure 2d.

Moreover, the time intensity estimates were generally lower in the TM-RT than in the MTM-RG (see Figure 3b). Because the item intensity of the first one/page item within a testlet was uniformly the highest, test-takers likely needed more time to understand the contents of the testlets before responding to the first testlet item. The estimated time discriminations were generally lower in the TM-RT than in the MTM-RG (see Figure 3c). In brief, these reading test results confirm our expectations (a)–(c), namely, a measurement model that ignores substantial RG behaviors (TM-RT) tends to assess items to (a) be more difficult, (b) be less time intensive for individual items and one/page testlets, and (c) have less time discrimination for individual items and one/page testlets.

The TM-RT and MTM-RG showed testlet effects on response accuracy and RT (see Figure 3d). Both the TM-RT and MTM-RG showed that the 4th, 12th, and 13th testlets had larger effects than the other testlets on response accuracy. The 11 one/page testlets showed moderately larger testlet effects on RT in the TM-RT than in the MTM-RG. By contrast, the three all/page testlets showed slightly larger testlet effects on RT in the MTM-RG than in the TM-RT.[2] Thus, these results also confirm expectation (d), that TM-RT underestimates the variances of random variables, which causes overestimates of testlet effects.

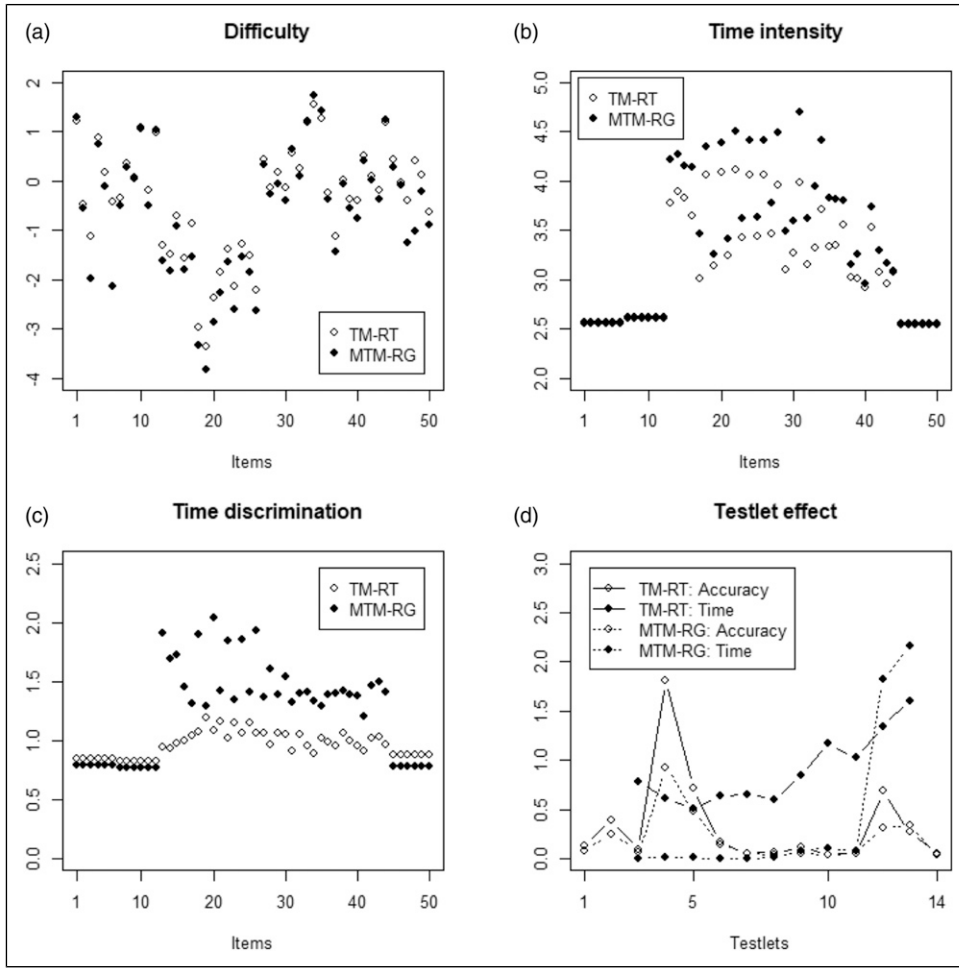

Figure 3. Item parameter estimates for the TM-RT and MTM-RG in the reading test. *Note*. The 1st, 2nd, and 14th testlets are presented in the one/page design. They are items 1–6, 7–12, and 45–50, respectively.

Appendix 9 displays the relationship between the parameter estimates of the 50 items for the MTM-RG. Because the RT is not informative for all/page items, they are not discussed here. Item difficulty and item intensity did not have a strong linear relationship. More difficult items often had more homogeneous time discrimination. Similarly, items with lower time intensity had more homogeneous time discriminations.

The MTM-RG showed a link between response accuracy and RT, substantial and varied RG, and varied answering strategies. Response accuracy (θ) and speed (τ) were weakly correlated (r = .24), indicating that test-takers with better Chinese reading abilities than others more quickly answered the questions. Furthermore, the estimated RG mean time intensity (β_0) was 2.05 (SE = 0.04) and time discrimination α_0 was 0.48 (0.01). The marginal probability of RG was substantial (M = 0.196) and varied widely across items (0.006–0.616, see Figure 4a). For one/page and individual items, the correlation between π and b estimates was .73, indicating that test-takers were more likely to randomly guess on items that were more difficult. In addition, as the RT information to identify RG responses was minimal, all/

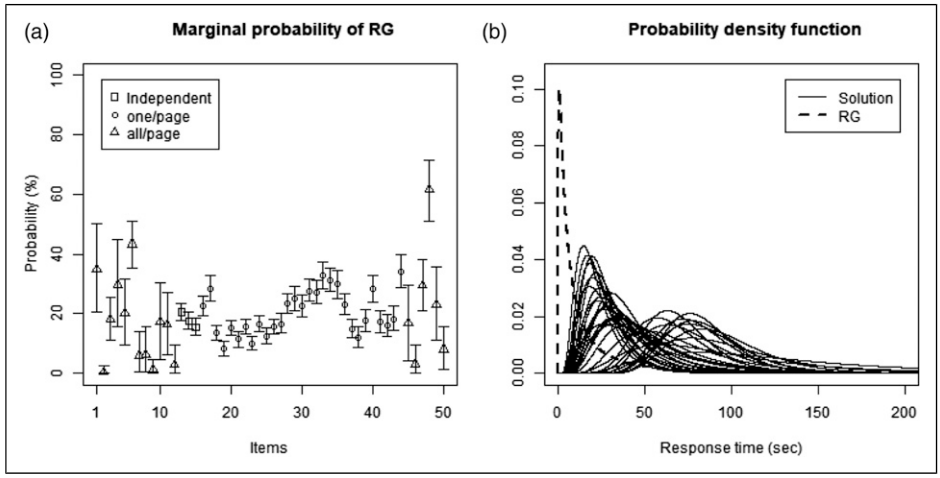

Figure 4. Marginal probability of RG and the probability density functions of RT under the MTM-RG in the reading test. *Note*. The three all/page testlets are items 1–6, 7–12, and 45–50. The upper and lower bounds define the 95% probability interval. The all/page testlets are excluded in Figure 4b.

page testlet items generally had a larger standard error for π estimates, echoing the finding in Simulation 1. Figure 4b shows the probability density functions of RT for solution (solid curves, one for each item) and RG (dashed curve) in the MTM-RG, which suggests that these test-takers used varied answering strategies to spend more time on some items and less time on others (including RG) to achieve a higher overall test score. Because different test-takers faced different item orders, we cannot test whether RG occurred more often on items near the end of the test. The above results were consistent with our expectations (e)–(g) (greater response accuracy to accompany shorter RT, substantial and varied RG, and varied answering strategies).

Because the MTM-RG estimates the time needed for effortful test-takers to complete each item, we can estimate the time that each test-taker needs to complete the entire test. According to the MTM-RG, only 69.8% of the test-takers can apply their full effort to all test items within the allotted 50 min of test time; thus, 30.2% of the test-takers face time pressures and likely guess on some items. Extending the test time to 60 min reduces the time pressure on only 23.5% of the test-takers. To sharply reduce the time pressure for almost all of these test-takers (say 99% of them), the MTM-RG indicates a time limit of 252 min (over 4 hours). Hence, the MTM-RG can specify how to change a high time-pressure exam into a low time-pressure one (and vice versa) for most test-takers.

Conclusion and Discussion

To model RG in computer-based testlets with one/page or all/page testlet items, we proposed and tested two testlet IRT models with item responses and RT: TM-RT and MTM-RG. The TM-RT quantifies (a) the latent traits of response accuracy and speed and (b) the testlet effects on the probability of item correctness and the corresponding RT. Extending the TM-RT, the MTM-RG also accounts for RG behaviors.

The simulation results showed that the TM-RT and MTM-RG both fit data with no RG well but that the MTM-RG outperforms the TM-RT on data with RG. For data with RG, the TM-RT yields biased parameter estimates, assessing items to be excessively difficult, be insufficiently time intensive, and have insufficient time discrimination, along with insufficient variances of random

variables. With proportionately more RG, the measurement precision of the assessment was lower. For testlet items, the all/page design lacks RT for each item; therefore, the MTM-RG is less sensitive to RG responses for all/page items.

We also applied both the TM-RT and MTM-RG to 907 test-takers' responses to the reading test. The MTM-RG fit the data better than the TM-RT, indicating substantial RG and varied answering strategies across items (including selective RG). Similar to the findings of the simulations, the TM-RT generally yielded higher item difficulties, lower time intensities, and lower time discriminations than the MTM-RG. Moreover, the MTM-RG showed that the magnitudes of the testlet effects varied, highlighting the need to attend to local dependence among items within testlets.

For any new analysis, we recommend fitting different models to the data and comparing their item/person parameters (rather than only using MTM-RG). If the differences between models are small (i.e., tiny RG effects or testlet effects), reporting the results of the simpler model can aid audience understanding and hence be more practical.

Although this study shows the feasibility of the MTM-RG, future studies can generalize it along several axes. First, to help estimate latent membership, Δ_{ij} can be regressed on item and/or person covariates (if any) in a logistic regression function. Second, because some CBTs allows test-takers to change their answers on previous items (Jeon et al., 2017), future studies can model such behaviors. Third, the MTM-RG can be generalized to accommodate other aberrant response behaviors, such as preknowledge (Wang et al., 2017) or nonresponses (Ulitzsch et al., 2020).

Many factors (e.g., item administration, test-taker strategies) might affect accurate RT measurement for individual items within a testlet. For example, the administration of items on a paper-and-pencil test versus a computerized adaptive test might affect how test-takers perceive the items (e.g., degree of difficulty) and their attention to them. These perceptions in turn might affect test-taker strategies on how to understand and respond to each item (e.g., in which order do they answer the items on a testlet). Hence, future studies should test the robustness of their models against empirical data (preferably in diverse contexts).

In this study, we used JAGS to estimate the parameters, which required longer computation time. For example, in Simulation 1, each MTM-RG and TM-RT analyses took an average of 50 and 200 minutes to complete with a 3.6-GHz Intel Core i7 processor. Other software can also implement the MTM-RG (e.g., OpenBugs [Spiegelhalter et al., 2010] and Mplus [Muthén & Muthén, 1998–2017]). Future studies can compare the efficiencies of these software in fitting the MTM-RG across data sets.

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 was financially supported by the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

ORCID iDs

Kuan-Yu Jin https://orcid.org/0000-0002-0327-7529
Chia-Ling Hsu https://orcid.org/0000-0002-4267-0980
Ming Ming Chiu https://orcid.org/0000-0002-5721-1971
Po-Hsi Chen https://orcid.org/0000-0002-5377-2077

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Testlets can be also modeled with polytomous-item and item-bundle approaches. For the pros and cons of different approaches, see Wang and Jin (2016).
2. Because only some testlets showed substantial testlet effects, it is not surprising that the MM-RG fit the empirical data better than the MTM-RG did ($DIC_{MM\text{-}RG} < DIC_{MTM\text{-}RG}$: 353,908 < 376,770).

References

Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional dependence between response time and accuracy: An overview of its possible sources and directions for distinguishing between them. *Frontiers in Psychology*, *8*, 202. https://doi.org/10.3389/fpsyg.2017.00202

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

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Huang, H.-Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement*, *80*(6), 1168–1195. https://doi.org/10.1177/0013164420914711

Jeon, M., De Boeck, P., & van der Linden, W. J. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, *42*(4), 467–490. https://doi.org/10.3102/1076998616688015

Jin, K.-Y., Siu, W.-L., & Huang, X. (2022). Exploring the impact of random guessing in distractor analysis. *Journal of Educational Measurement*, *59*(1), 43–61. https://doi.org/10.1111/jedm.12310

Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*(1), 131–154. https://doi.org/10.1177/0013164420913915

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

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

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*(1), 1–27. https://doi.org/10.1111/jedm.12060

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538. https://doi.org/10.1177/0146621609355451

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus Version 8 User's Guide*. Muthén & Muthén

Plummer, M. (2017). *JAGS version 4.3 user manual*. Sourceforge. https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/jags_user_manual.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Institute of Education Research

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. https://doi.org/10.1177/014662169001400305

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*(4), 583–616. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2010). *OpenBUGS Version 3.1.1 user manual*. OpenBUGS. http://www.openbugs.net/Manuals/Manual.html

Su, Y. S., & Yajima, M. (2015). *R2jags: Using R to run 'JAGS'*. Cran. https://cran.r-project.org/web/packages/R2jags/index.html

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

van der Linden, W. J. (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

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501. https://doi.org/10.3102/1076998618767123

Wang, W.-C., & Jin, K.-Y. (2016). Analyses of testlet data. In Q. Zhang (Ed), *Pacific rim objective measurement symposium (PROMS) 2015 conference proceedings*. Springer. https://doi.org/10.1007/978-981-10-1687-5_13

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149. https://doi.org/10.1177/0146621604271053

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement*, *41*(4), 243–263. https://doi.org/10.1177/0146621616687285

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

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2