frontiers in Psychology

ORIGINAL RESEARCH
published: 29 January 2019
doi: 10.3389/fpsyg.2019.00051

Modeling Response Time and Responses in Multidimensional Health Measurement

Chun Wang[1]*, *David J. Weiss*[2] *and Shiyang Su*[3]

[1] College of Education, University of Washington, Seattle, WA, United States, [2] Department of Psychology, University of Minnesota, St. Paul, MN, United States, [3] Department of Psychology, University of Central Florida, Orlando, FL, United States

This study explored calibrating a large item bank for use in multidimensional health measurement with computerized adaptive testing, using both item responses and response time (RT) information. The Activity Measure for Post-Acute Care is a patient-reported outcomes measure comprised of three correlated scales (Applied Cognition, Daily Activities, and Mobility). All items from each scale are Likert type, so that a respondent chooses a response from an ordered set of four response options. The most appropriate item response theory model for analyzing and scoring these items is the multidimensional graded response model (MGRM). During the field testing of the items, an interviewer read each item to a patient and recorded, on a tablet computer, the patient's responses and the software recorded RTs. Due to the large item bank with over 300 items, data collection was conducted in four batches with a common set of anchor items to link the scale. van der Linden's (2007) hierarchical modeling framework was adopted. Several models, with or without interviewer as a covariate and with or without interaction between interviewer and items, were compared for each batch of data. It was found that the model with the interaction between interviewer and item, when the interaction effect was constrained to be proportional, fit the data best. Therefore, the final hierarchical model with a lognormal model for RT and the MGRM for response data was fitted to all batches of data via a concurrent calibration. Evaluation of parameter estimates revealed that (1) adding response time information did not affect the item parameter estimates and their standard errors significantly; (2) adding response time information helped reduce the standard error of patients' multidimensional latent trait estimates, but adding interviewer as a covariate did not result in further improvement. Implications of the findings for follow up adaptive test delivery design are discussed.

Keywords: response time, hierarchical model, health measurement, multidimensional graded response model, item response theory (IRT)

OPEN ACCESS

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Lihua Yao,
United States Department of Defense, United States
Fernando Marmolejo-Ramos,
University of Adelaide, Australia

***Correspondence:**
Chun Wang
wang4066@uw.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *22 August 2018*
Accepted: *09 January 2019*
Published: *29 January 2019*

Citation:
Wang C, Weiss DJ and Su S (2019) Modeling Response Time and Responses in Multidimensional Health Measurement. Front. Psychol. 10:51. doi: 10.3389/fpsyg.2019.00051

INTRODUCTION

When assessments are delivered via computer-based devices, collecting persons' response times (RTs) at the item level is straightforward. The analysis of item-level RTs on assessments has attracted substantial interest recently. For example, in personality assessments, RTs have been used to measure attitude strength (Bassili, 1996), to detect social desirability (Holden and Kroner, 1992), and to enhance criterion validity (Siem, 1996). In achievement testing, RTs have been used to

evaluate the speededness of the test (Van Der Linden et al., 1999), to detect aberrant behavior (e.g., Wang and Xu, 2015; Wang et al., 2018b,c), and to design a more efficient test (Bridgeman and Cline, 2004; Van der Linden and Guo, 2008; van der Linden, 2009; Fan et al., 2012). RTs have also been used to evaluate response data quality in web-based surveys (Galesic and Bosnjak, 2009).

In the health measurement domain, response time (sometimes called reaction time) is often used to measure cognitive functioning, particularly in research on aging (e.g., Pearson, 1924; Braver and Barch, 2002; Hultsch et al., 2002; Anstey et al., 2005; Osmon et al., 2018). Similar to the speed test in educational assessments, RTs are usually collected from timed, target stimuli tasks, in which respondents are instructed to respond as quickly as possible. In this case, only RTs, not response accuracy, is of interest. For example, in a study using the United Kingdom Health and Lifestyle Survey (Cox et al., 1987; Der and Deary, 2006), person-level reaction times were examined across different age and gender groups. Another example is using RTs from a stop-signal reaction time task to study response inhibition from patients with Parkinson's disease and other brain disorders (Gauggel et al., 2004; Verbruggen et al., 2013). Despite these widespread applications of RTs, little attention has been paid to the usefulness of item-level response times as collateral information for improving measurement precision. These previous studies have primarily used scale-level, aggregated RTs, such as its mean and standard deviation. However, item-level RTs, routinely collected during computer-based assessment delivery, provide richer information. Only a recent didactic review by Osmon et al. (2018) demonstrated the advantages of examining the entire RT distribution rather than only its mean and standard deviation to understand the efficacy of mental speed assessment in clinical neuropsychology. Therefore, it was of interest to apply advanced psychometric models for item-level RTs in the assessment of reported health behaviors and evaluate if RTs help better estimate the main constructs of interest.

MODELS

Multidimensional Graded Response Model

The most appropriate measurement model for ordered polytomous responses is the graded response model (GRM; Samejima, 1969). The item response function of the unidimensional GRM model is

$$P_{jk}(\theta) = P^{+}_{jk}(\theta) - P^{+}_{j,k+1}(\theta) = \frac{e^{[Da_j(\theta - b_{jk})]}}{1 + e^{[Da_j(\theta - b_{jk})]}} - \frac{e^{[Da_j(\theta - b_{j,k+1})]}}{1 + e^{[Da_j(\theta - b_{j,k+1})]}} \quad (1)$$

where $P_{jk}(\theta)$ is the probability of a randomly selected person with a latent trait θ selecting category k of item j (k−1 … K). $P^{+}_{jk}(\theta)$ is the boundary response function, interpreted as the probability of responding to category k and above for item j given θ. a_j is the item discrimination parameter for item j. b_{jk} is the boundary location parameter for item j in category k ($k = 0, \ldots, K$). D = 1.7 is the normalizing constant. Because by definition, $P^{+}_{j0}(\theta) \equiv 1$ and $P^{+}_{jK+1}(\theta) \equiv 0$, neither b_{j0} nor b_{jK+1} are estimable parameters. Therefore, for an item with four response categories, only three boundary parameters are estimated.

When the instruments include multiple scales measuring different constructs or different aspects of the same construct (e.g., Zickar and Robie, 1999; Fraley et al., 2000; Fletcher and Hattie, 2004; Zagorsek et al., 2006; Pilkonis et al., 2014), the multidimensional extension of the GRM, namely, the MGRM (Hsieh et al., 2010; Jiang et al., 2016), is appropriate. Let $\boldsymbol{\theta}$ be a vector of length H representing the latent traits of interest, and let $h = 1, 2, \ldots, H$. Similar to the unidimensional case, $P^{+}_{j0}(\boldsymbol{\theta}) \equiv 1$ and $P^{+}_{j(K+1)}(\boldsymbol{\theta}) \equiv 0$. When the test displays a simple structure, the boundary response function takes the form of

$$P^{+}_{jk}(\boldsymbol{\theta}) = \frac{1}{1 + \exp\left[-\mathrm{D}a_{jh}\left(\theta_h - b_{jk}\right)\right]} = \frac{1}{1 + \exp\left[-\mathrm{D}\left(a_{jh}\theta_h + c_{jk}\right)\right]}, \quad (2)$$

assuming item j measures dimension h only so that a_{jh} is the item discrimination parameter on the hth dimension of item j. In Equation 2, $c_{jk} = -a_{jh}b_{jk}$ and this a-c parameterization with $D = 1$ is consistent with *flex*MIRT's (Cai, 2013) default parameterization; the c parameter is interpreted as the "intercept." Equation 2 could also be modified to accommodate complex structure; for details, see Reckase (2009).

Bivariate Models of Responses and RTs

Given that RTs carry useful collateral information about both item and person characteristics, the bivariate model of responses and RTs (Molenaar et al., 2015) was considered. The measurement model for responses was as specified in Equation 2, and the measurement model for RTs takes the form

$$\ln t_{ij} = \lambda_j + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \omega_{ij} \quad (3)$$

Here, t_{ij} denotes the RT of patient i on item j, τ_i is the latent speed parameter of patient i, λ_j and φ_j are the time-intensity and time-discrimination parameters of item j, and ω_{ij} is the residual. If the residuals are assumed to be normally distributed, then Equation 3 suggests that the response time t_{ij} follows a log-normal distribution. Other more flexible types of residuals can also be assumed if the data warrants (e.g., Wang et al., 2013a,b).

The term, $\varphi_j\rho_d\theta_{id}$, is called a cross-relation function (Ranger, 2013; Molenaar et al., 2015), and it is assumed that item j measures the dth dimension. Different from van der Linden's (2007) hierarchical model in which a covariance structure is assumed on θ and τ at a second level, this cross-relation term directly models the relationship between the latent ability and observed log-transformed RTs (log-RTs). Certainly, the cross-relation term based on τ_i could alternatively enter into the measurement model of responses; for example, Molenaar et al. (2015) argued that incorporating the cross-relation term in the RT model had unique advantages. That is, when the purpose of including RT information is to improve the measurement

precision of θ, it is preferable to leave the measurement model for the responses unchanged while modeling the information about θ (if any) in the RTs. In this regard, θ accounts for the shared ability variance in the responses and RTs and τ accounts for the additional, unique variance in the RTs. This joint model is termed as Model 0 and its diagram is shown in **Figure 1**.

To ensure model identifiability, several constraints need to be in place. First, regarding the MGRM model, the mean and variance of $\boldsymbol{\theta}$s are restricted to be 0 and 1, respectively. Second, the mean and variance of τ is also constrained to be 0 and 1 such that the residual variance of ω_{ij} is freely estimated[1]. The three θ components are assumed to be correlated, and the correlation matrix is freely estimated. However, all three θs are assumed uncorrelated with τ due to the inclusion of the cross-relation term. The same set of constraints was assumed for all other models introduced hereafter.

Molenaar et al. (2015) suggested identifiability constraints that are similar to those listed, except that $\text{var}(\tau) = 1 - \rho^2$, instead of 1. Both constraints are sufficient, and their choice conveniently allows the interpretation of ρ as a correlation coefficient. Note that in van der Linden's (2007) model the variance of τ is estimable (Equation 22, p. 294). This is because the lognormal model for RT in van der Linden (2007) takes the form

$$f\left(t_{ij};\tau_i,\alpha_j,\beta_j\right) = \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_j\left(\ln t_{ij} - (\beta_j - \tau_i)\right)\right]\right\} \quad (4)$$

where α_j is interpreted as the *dispersion* parameter that quantifies the variance of the lognormal distribution, rather than the *discrimination* parameter as in Equation 3.

Bivariate Model With Interviewer as a Covariate

Because more than one interviewer was used for data collection, three variations of the bivariate model with interviewer as a covariate were considered. The first model is

$$\ln t_{ij} = \lambda_j + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \sum_{p=1}^{P}\gamma_{jp}x_p + \omega_{ij}, \quad (5)$$

where x_p is a binary indicator variable indicating if interviewer p recorded the RTs for patient i, and P is the total number of interviewers in the data. P equaled 6 for batch 1 and 5 for batches 2–4. Because each patient interacted with only one interviewer, only one non-zero element in the summation $\sum_{p=1}^{P}\gamma_{jp}x_p$ enters into the regression equation for patient i. The model in Equation 5 (Bivariate Model 1) assumes that interviewer effects differed per item, i.e., there is an interaction between interviewer and items.

Model 2 is a slightly restricted version of Model 1, and the measurement model for RT becomes

$$\ln t_{ij} = \lambda_j + \varphi_j\tau_i - \varphi_j\rho_d\theta_{id} + \varphi_j\sum_{p=1}^{P}\gamma_p x_p + \omega_{ij}, \quad (6)$$

where all parameters have the same interpretations as in Equation 5 except τ_i, which can be interpreted as the individual "residual"

[1]By default, M*plus* sets the factor mean to be 0 for both $\boldsymbol{\theta}$ and τ.

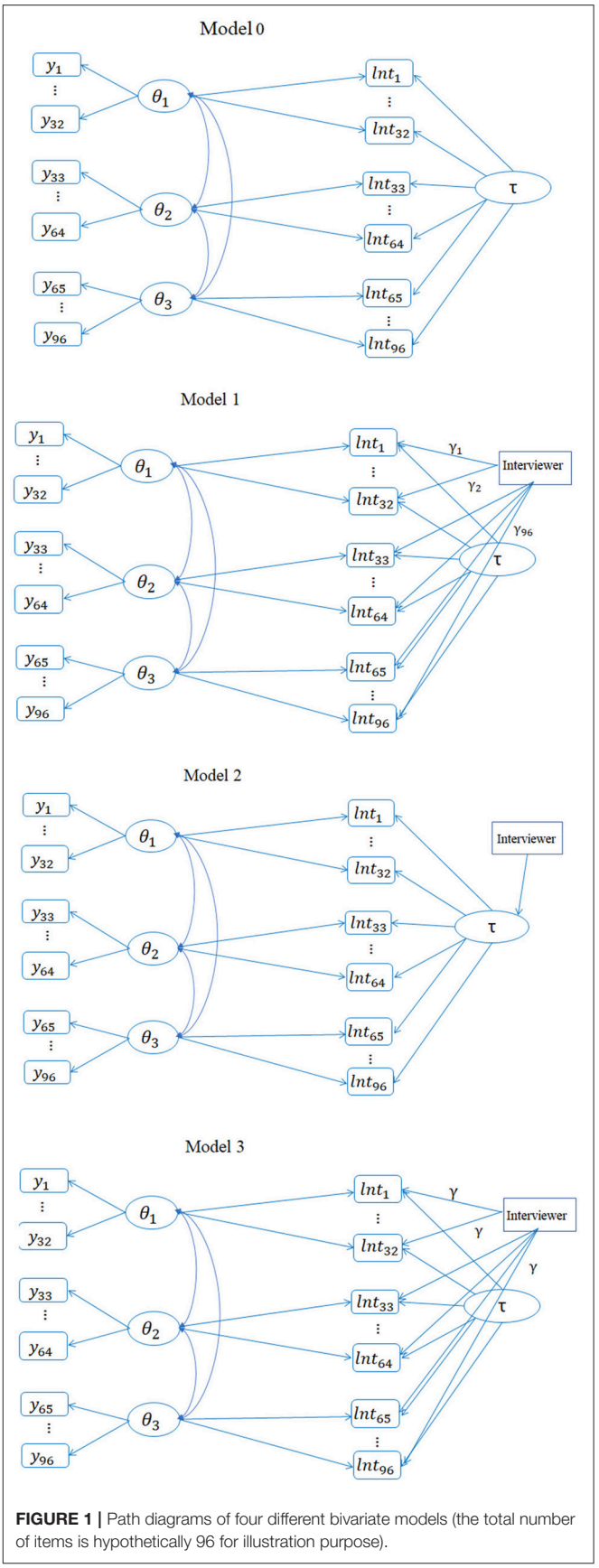

FIGURE 1 | Path diagrams of four different bivariate models (the total number of items is hypothetically 96 for illustration purpose).

speed after removing the interviewer effect. The MGRM model is still used for polytomous responses. In Equation 6, the interviewer effect differs across items but by the same amount, denoted as φ_j. This Model 2 can also be viewed as a hierarchical model in which the interviewer variable predicts the speed at the second level, as follows:

$$\ln t_{ij} = \lambda_j + \varphi_j \tau_i - \varphi_j \rho_d \theta_{id} + \omega_{ij}$$
$$\tau_i = \sum_{p=1}^{P} \gamma_p x_p + \varepsilon_i, \quad (7)$$

where ε_i is the individual residual speed. Compared to Model 1, Model 2 greatly reduces the number of parameters and hence is a more parsimonious model. When fitting the hierarchical model in M*plus* (Muthén and Muthén, 1998-2015), the variance of τ cannot be fixed directly but instead the variance of ε_i is fixed at 1.

Model 3 considers only the interviewer main effect and it assumes that the interviewer effect does not differ across items. Again the MGRM stays the same, and the model for RTs becomes

$$\ln t_{ij} = \lambda_i + \varphi_j \tau_i - \varphi_j \rho_d \theta_{id} + \sum_{p=1}^{P} \gamma_p x_{ip} + \omega_{ij}. \quad (8)$$

Although this Model 3 has essentially the same number of parameters as Model 2, it assumes no interactions between interviewers and items. The path diagrams for the four models are presented in **Figure 1**.

METHODS

Instrument and Subjects

Responses and RTs from the Activity Measure for Post-Acute Care (AM-PAC) were analyzed (Yost et al., 2018). The AM-PAC is the first multi-domain patient reported outcomes measure with the capability to direct care in a hospital rehabilitation environment. The scores from the AM-PAC are intended to be linked to the widely understood stages of the Functional Independence Measure (O'Dell et al., 1998; Huang et al., 2000) such that appropriate rehabilitative care plans can be immediately identified. It is anticipated that the AM-PAC will provide an inexpensive and accurate alternative to clinician assessments. The three domains covered in the AM-PAC include Applied Cognition, Daily Activity, and Mobility. A sample question from the Applied Cognition domain is: "*How much difficulty do you currently have reading a long book (over 100 pages) over a number of days*?", and the four response options are "Unable" (coded as 1), "A lot" (coded as 2), "A little" (coded as 3), and "None" (coded as 4). Items were administered to hospital inpatients via a computer-assisted personal interview using Qualtrics® web survey software. During the field testing of the items, an interviewer read each item to a patient and recorded, on a tablet computer, the patient's responses and the software recorded RTs. A total sample of 2,270 hospitalized patients were recruited to the study; their mean age was 65 years. Roughly 54% were male and 96% were non-Hispanic white, and 78% had two or more comorbidities (Yost et al., 2018).

Questions were grouped into blocks according to domain, and the order of item administration within a block was randomized. Given that there were 324 items in total in the bank, data collection proceeded in four batches to reduce patient burden. The first batch of 109 items was administered to patients, and 24 linking items were selected with eight items in each domain. The number of linking items was determined based on Kolen and Brennan (2004)'s recommendation that at least 20% of the items need to be shared between different test forms to have enough information to link the scale (Wang et al., 2016). These linking items in each domain were selected to produce a composite information function that was closest in shape to the domain information function. Linking items were assembled using the linear programming solver "lp_solve version 5.5" (Diao and van der Linden, 2011). Then, the set of linking items was carried forward in subsequent data collection batches. **Table 1** presents the number of items per domain for the four batches.

Preliminary Data Cleaning

Table 2 presents the summary descriptive statistics for the four batches of data. The cleaned Batch 1 dataset contained 563 respondents after deleting 67 (10.6%) respondents with at least 20 missing items. The cleaned Batch 2 dataset contained 490 respondents after deleting 52 (9.6%) respondents with more than 10 missing items. The cleaned Batch 3 dataset contained 500 respondents after deleting 55 (9.9%) respondents with more than 9 missing items. The cleaned Batch 4 dataset contained 507 respondents after deleting 36 (6.6%) respondents with more than nine missing items. Although each item contained four response categories, for some items, category 1 and/or category 2 received no responses or very few responses. These items were then recoded to ensure that the lowest response category for each item was always 1, but the highest response category could be 4 or less. As shown in **Table 2**, the response time distribution exhibited extreme skewness (ranging from 29.08 to 41.84), and therefore the distribution was truncated by removing the top 2.5% and removing the RTs smaller than 3 s, resulting in skewness from 1.48 to 1.66. The resulting data was entered into modeling analysis. Recent research by Marmolejo-Ramos et al. (2015a) suggested that Box-Cox transformation outperformed the elimination method in normalizing positively skewed data. However, the extremely long and short RTs were trimmed in these data because those RTs were considered as outliers. Extremely long RTs happened when the patient took a break such as "service came in to discuss plans" or "patient lunch came and wanted to stop." The row for the missing proportion of RTs in

TABLE 1 | Number of unique items per domain for the four batches.

Batch	Applied cognition	Daily activity	Mobility	Total	
				Unique	Linking
1	28	27	30	85	24
2	24	24	24	72	24
3	24	24	24	72	24
4	23	23	25	71	24
Linking	8	8	8	24	24
Total	107	106	111	324	—

TABLE 2 | Descriptive statistics of the observed data, by batch.

Variable	Batch 1	Batch 2	Batch 3	Batch 4
SAMPLE SIZE				
Before cleaning	630	542	555	543
After cleaning	563	490	500	507
Trimmed proportion of RTs	6.24%	3.21%	3.16%	4.59%
NUMBER OF ITEMS				
2 categories	1	0	0	0
3 categories	31	26	22	23
4 categories	77	70	94	72
RT BEFORE TRUNCATION				
Mean	9.27	9.79	9.82	8.39
SD	21.28	17.39	25.37	17.62
Skewness	41.84	32.40	35 .85	29.08
RT AFTER TRUNCATION				
Mean	8.21	8.44	8.06	7.18
SD	4.14	4.92	4.80	4.05
Skewness	1.48	1.66	1.65	1.53

Table 2 refers to the proportion of RTs at the person-by-item level, out of the cleaned sample size (e.g., 563 for batch 1), that was deleted either because they were extremely short (<3 s) or extremely long (upper 2.5%).

To further test the normality of item-level RT distributions, the Kolmogorov-Smirnov (K-S) test (Smirnov, 1948) was conducted for all item-level log-RTs. The K-S statistic quantifies the distance between the empirical distribution function of a sample and the cumulative distribution function of a reference distribution, and it is a non-parametric test of the equality of two distributions. For the present purpose, the K-S test was done with response times that were at least 3 s and were below the 97.5% percentile. This item-level K-S test compared the log-RTs of that item to the theoretical normal distribution with the mean and variance computed for the item. The null hypothesis is that the log-RTs follow a normal distribution. Hence, a significant p-value (i.e., $p < 0.05$) indicated that the log-RTs distribution was significantly different from normal. Results showed that in Batch 1, 54 out of 110 items exhibited statistically significant p-values, but the K-S statistics for those items were very small (ranged from 0.05 to 0.1). In Batches 2, 3, and 4, 30 out of 96, 16 out of 96, 11 out of 95 items, respectively, had significant p-values, but again, the K-S statistics were small.

The K-S test was chosen because of its wide popularity. For instance, it was used to evaluate the item RT distributions from computer-based licensure examinations (Qian et al., 2016). However, other tests, such as the Shapiro-Wilk (S-W) test (Royston, 1982) has been found to be more powerful than the K-S test to detect departure from normality (Marmolejo-Ramos and González-Burgos, 2013). Unsurprisingly, using the S-W tests on the same data set showed that 99.1% of Batch 1 items, 90.6% of Batch 2 items, 95.8% of Batch 3 items and 92.6% of Batch 4 items had significant p-values. However, the lognormal model was still used as the parametric model for RTs in the following analysis because the skewness (shown in **Table 2**) after truncation was not high, and the lognormal distribution was a convenient choice that most software packages can handle.

Collapsing Response Categories

In the data analysis, response categories for some items were collapsed due to lack of observations in those categories. Specifically, for a given item, if a category received no response or only one response, the response of this option, if any, was combined into the responses of the next higher category. Therefore, as shown in **Table 2**, some items had fewer than four response categories. The treatment of collapsing response categories is legitimate for the graded response model because it does not substantially change the item parameter estimates[2]. For instance, a 4-category GRM item ($k = 1, 2, 3, 4$) item will have four parameters, i.e., a_j, b_{j1}, b_{j2}, b_{j3}. When collapsing the lowest two response categories, the parameters of the same item become $a_j^* \approx a_j$, $b_{j2}^* \approx b_{j2}$, $b_{j3}^* \approx b_{j3}$. This is because the GRM is essentially a difference model (see Equation 1), and the same discrimination parameter is assumed across all boundary response functions [i.e., $P_{jk}^{+}(\theta)$].

Model Fitting and Item Calibration

Bivariate Model Fitting

All four models in **Figure 1** were fit with marginal maximum likelihood estimation (MML) using the Expectation-Maximization (EM) algorithm in M*plus*[3] These models were fitted to each batch of data separately to evaluate global model fit via AIC, BIC and−2Log-likelihood. The M*plus* source code of Batch 4 is provided in the **Appendix**. The same source code was used for other batches, as well. As shown in **Table 3**, Model 2 was the best-fitting model across all four batches of data based on BIC, but Model 1 was preferred based on AIC. In addition, Model 2 and Model 3 are respectively nested within Model 1. The deviance test (i.e., likelihood ratio test) revealed that there was a significant difference between Model 1 and Model 2, Model 1 and Model 3, implying that Model 1 should be preferred. However, Model 2 was used in the following analysis for two reasons: (1) Model 2 is a much more parsimonious model than Model 1 and it is conceptually more reasonable because the interviewer effect should not interact with items, i.e., the interviewer's speed should be relatively static across items; (2) when fitting Model 1 in the concurrent calibration described below, it failed to converge due to complexity and data sparsity.

Concurrent Calibration

When data are collected in different batches, linking items are used to place the items from the different batches onto a common scale. Concurrent calibration has been demonstrated to be more effective than separate calibration plus *post-hoc* linking (Kolen and Brennan, 2004) because the latter approach suffers from linking error. Three models were compared in the concurrent calibration: the MGRM model for responses only, Model 0, and

[2]A separate study (Jiang and Wang, 2019) was conducted that provided analytic and simulation evidence for this claim.

[3]M*plus* was chosen because it is widely used in social science research. M*plus* plotting using R is available via the "rhdf5" package. For details, refer to http://www.statmodel.com/mplus-R/.

Model 2. Models 1, and 3 were not considered because of their poorer fit compared to Model 2. Both the item and person parameters and their standard errors were compared across the three models. The main research question was whether including RTs and interviewer information helped improve the estimation accuracy of both item and person parameters.

When pooling data from the four batches together, the concurrent calibration of Model 0 and Model 2 failed to converge due to the sparsity of data and model complexity. Therefore, a two-stage approach was implemented. In the first stage, data from Batches 2–4 were pooled and a concurrent calibration was conducted on the pooled data. Data from Batch 1 was left out because this batch had the largest number of items flagged under the K-S test. By shrinking the sample size, all models successfully converged. Then in the second stage, Batch 1 data were calibrated using the fixed parameter calibration approach (Kim, 2006). That is, the linking item parameters (i.e., a, b, λ, and φ) were fixed at their estimated values obtained from Stage 1 for each of the three models such that the remaining items were estimated on the same scale as the linking items. Hence, no further linking procedure was needed.

Due to the collapsing of response categories, a side note for the two-stage approach is worth mentioning. Specifically, for the linking items, the threshold parameters of Batch 1 did not always match those in Batches 2–4. For example, an item had four categories (three threshold parameters) in Batches 2–4, but only three categories (two threshold parameters) in Batch 1. The linking items always had the same or fewer categories in Batch 1 as compared to the combined data due to the smaller sample size of Batch 1. In this case, only the corresponding threshold parameters and discrimination parameter for an item were input into the fixed calibration. The rationale is the same as before—collapsing response categories does not substantially change the item parameters.

TABLE 3 | Global fit results (AIC, BIC,−2Log-likelihood) for the four bivariate models, by batch.

Batch and model	Number of free parameters	AIC	BIC	-2Log-likelihood
BATCH 1				
Model 0	736	133566	136755	132094
Model 1	1281	133174	138725	130612
Model 2	**741**	**133316**	**136527**	131834
Model 3	741	133409	136620	131926
BATCH 2				
Model 0	652	102468	105202	101164
Model 1	940	102049	105992	100170
Model 2	**655**	**102235**	**104982**	100924
Model 3	655	102339	105086	101030
BATCH 3				
Model 0	656	111384	114149	110072
Model 1	1040	110613	114996	108532
Model 2	**660**	**111001**	**113783**	109682
Model 3	660	111323	114105	110004
BATCH 4				
Model 0	648	108550	111290	107254
Model 1	1028	107733	112080	105676
Model 2	**652**	**108174**	**110931**	106870
Model 3	652	108364	111121	107060

Bold values highlighted the best-fitting model based on the information criteria.

RESULTS

Global Model Fit

Table 4 presents the global model fit statistics for the three models in both stages. Note that the AIC and BIC from the MGRM are smaller because they are on a different scale compared to Model 0 and Model 2 due to its exclusion of RT information. Consistent with the separate calibration results, Model 2 fit the data better than Model 0, reflected by smaller AIC and BIC values.

Item Parameter Estimates

Figure 2 presents the scatterplots of item discrimination parameters (a_j) across the three models; all points fall along the 45° line, implying a close alignment of item parameter estimates from the three models. This is unsurprising because the variance of θ was fixed at 1 across all models, which fixed the scale of a_j. Means of SEs of estimates of a_j were 0.188 in the MGRM, 0.190 in Model 0, and 0.190 in Model 2. A simple t-test showed no significant differences of SEs between the different models.

The correlations of boundary parameters $\hat{b}_{jk}$ between different models were all 1, and therefore the scatterplots (**Figure 3**) show that the estimates of $\hat{b}_{jk}$ from the different models fall tightly on the 45° line. Moreover, t-tests showed no significant differences of mean SEs of $\hat{b}_{jk}$ between different models. Thus, the results suggest that, in these data, estimation of MGRM item parameters a_j and b_j, and their SEs were not affected by the addition of RT information.

With respect to the item time discrimination parameter, φ_j, the correlation between their estimates from Model 0 and Model 2 was 0.99. The scatterplot (**Figure 4**) shows that these estimates of $\hat{\varphi}_j$ from the two models fell on a line that was not 45°, indicating that there was a linear relationship between $\hat{\varphi}_j$s from the two models. The explanation is as follows: Focusing on the two terms in Model 0 and Model 2, respectively, $\varphi_j(\tau_i - \rho_d\theta_{id})$

TABLE 4 | Global model fit results.

Stage and Model	AIC	BIC
STAGE 1: CONCURRENT CALIBRATION (BATCHES 2-4)		
MGRM	185303.933	190099.963
Model 0	327447.703	336067.811
Model 2	**326589.532**	**335230.884**
STAGE 2: FIXED PARAMETER CALIBRATION		
MGRM	74013.471	75391.453
Model 0	134224.604	136824.572
Model 2	**133924.461**	**136546.095**

Bold values highlighted the best-fitting model based on the information criteria.

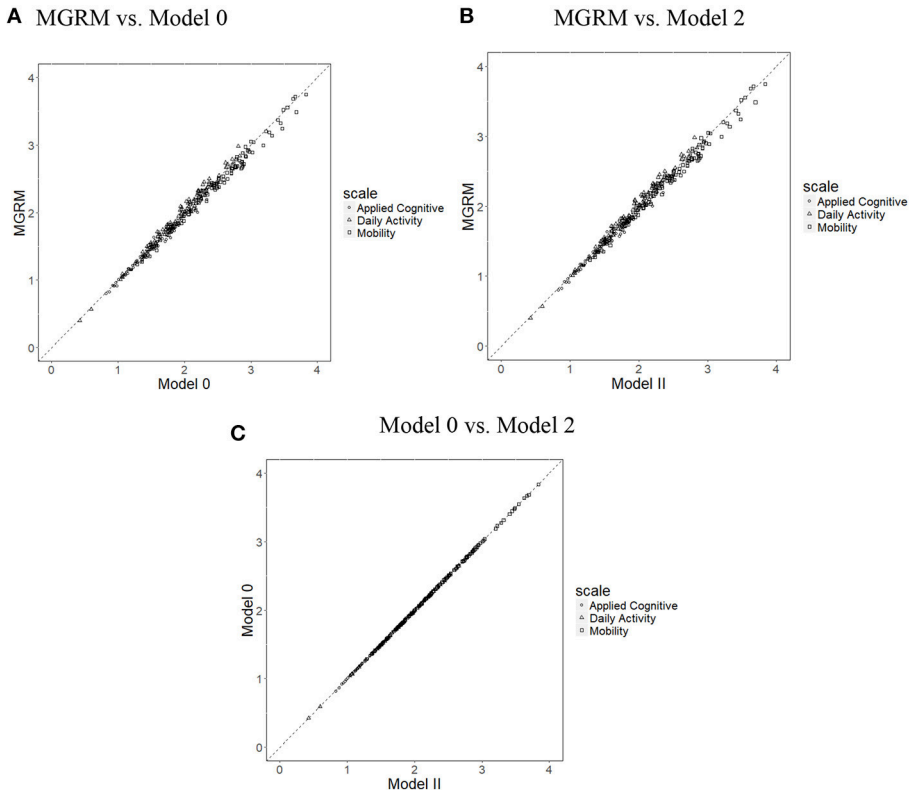

FIGURE 2 | Scatterplots of item discrimination parameters (a_j) across three models. **(A)** MGRM vs. Model 0, **(B)** MGRM vs. Model 2, and **(C)** Model 0 vs. Model 2.

(Equation 3) and $\varphi_j(\sum_{p=1}^{P}\gamma_p x_{ip} + \varepsilon_i - \rho_d\theta_{id})$ (Equation 7) the $(\tau_i - \rho_d\theta_{id})$ and $(\varepsilon_i - \rho_d\theta_{id})$ are the same across the two equations because both τ_i and ε_i are on the 0–1 scale. Due to the data collection design, the same interviewer went through all items in the batch each time, and each interviewer interviewed a portion of the sample. For instance, suppose the sample size is N, and there are n_1, $n_2 - n_1$, $n_3 - n_2$, $n_4 - n_3$, and $N - n_4$ patients interviewed by each interviewer, as shown in **Table 5**. Then, for item j, those patients assigned to Interviewer 1 all carry the same interviewer effect of γ_1, and similarly for the three other groups. Hence, the second and third columns are the same for every item, and also because the mean and variance of these two columns are different, there is a unique linear relationship between $\hat{\varphi}_j$s from the two models. On the other hand, SEs of $\hat{\varphi}_j$ of Model 2 are significantly lower ($p < 0.001$) than those of Model 0: Mean SE was 0.018 in Model 0 and 0.013 in Model 2.

Regarding the time intensity parameter, λ_j, again the correlation between their estimates from Model 0 and Model 2 was larger than 0.99. SEs of $\hat{\lambda}_j$ from Model 2 were significantly higher ($p < 0.001$) than those of Model 0: Mean SE was 0.020 in Model 0 and 0.023 in Model 2. The results suggest that the SE of item response time parameters $\hat{\varphi}_j$ and $\hat{\lambda}_j$ is affected in different directions by the addition of interviewers as covariates. However, the absolute difference in SEs was not too large to be concerning because the difference appeared mostly in the third decimal place.

Person Parameter Estimates

In terms of θ estimation, the $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ from all three models correlated as high as 0.99 to 1. Mean SEs from Models 0 and 2 (**Table 6**) were significantly lower than those from the MGRM ($p < 0.001$), and there was no significant difference of SE between Model 0 and Model 2. This result implies that adding response time decreased the SE of $\hat{\theta}$, which is consistent with prior findings (e.g., van der Linden et al., 2010; Wang et al., 2013a,b), but adding the interviewer variable did not further decrease the SEs.

Table 7 presents the estimated correlation parameters. Consistent with previous findings (e.g., Wang et al., 2018a), there were moderate to high correlations among the three latent traits. Moreover, the speed factor also played a modest role as reflected by the moderate size of ρ_1 to ρ_3. These correlations were higher in Model 2 than in Model 0, which is unsurprising because after removing the interviewer effects on RTs, the individual speed factor should correlate higher with individual latent traits.

The last column in **Table 7** refers to the fixed effects of interviewers. During Batches 2–4 data collection, the same five interviewers were recruited and one of them was randomly selected as the reference for dummy coding. It appears from the estimated $\hat{\gamma}_p$ that interviewers differed substantially and that is why including the interviewer variable in the model helped improve model data fit. For Batch 1 data collection, a different set of six interviewers was recruited; among them, three overlapped with the other set of five. However, because a different reference

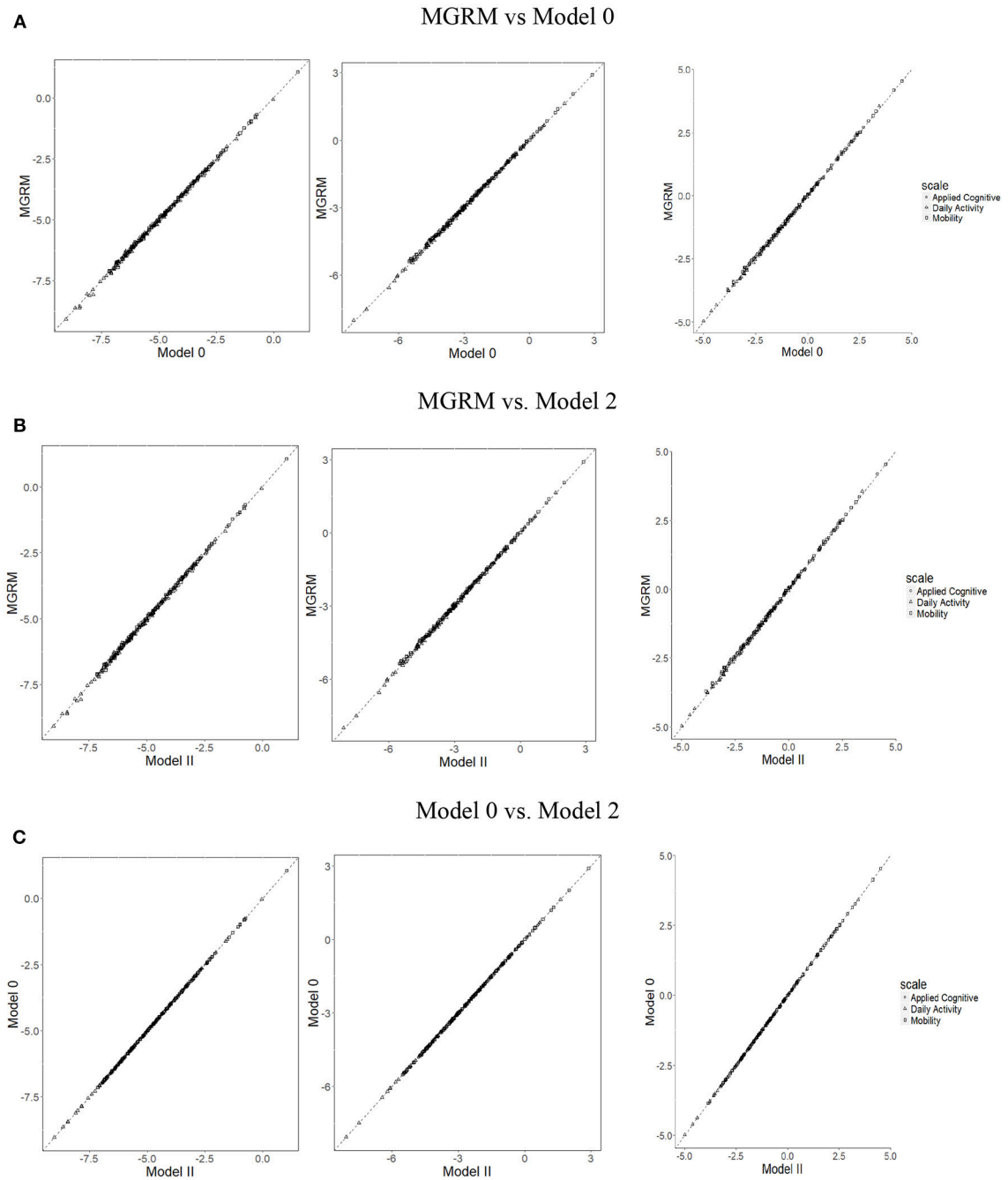

FIGURE 3 | Scatterplots of item boundary parameters (from left to right: $\hat{b}_{j1}$, $\hat{b}_{j2}$, $\hat{b}_{j3}$) across three models. **(A)** MGRM vs. Model 0, **(B)** MGRM vs. Model 2, and **(C)** Model 0 vs. Model 2.

interviewer was selected, the estimated $\hat{\gamma}_p$ from stage 1 and 2 model fitting were not directly comparable. Still, the results show that interviewers operated at different speeds and they contributed to the observed RT variabilities.

DISCUSSION AND CONCLUSIONS

Response time as part of the assessment process data has gained great popularity in recent decades in educational and psychological measurement. This is because collecting RTs has become easy, due to computer-based assessment delivery, and RTs provide an additional source of information for researchers to understand an individual's behavior as well as the characteristics of the items. More than a dozen IRT models have been proposed in the psychometrics literature, with an early focus on modeling the different shapes of RT distributions (e.g., Rouder et al., 2003; van der Linden, 2007; Loeys et al., 2011; Wang et al., 2013a,b) and a later focus on modeling within-subject

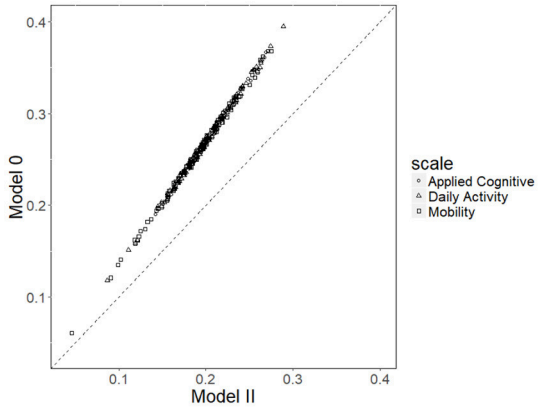

FIGURE 4 | Scatterplot of estimates of $\hat{\varphi}_j$ between Model 0 and Model 2.

TABLE 5 | An illustration of the linear transformation relationship of $\hat{\varphi}_j$ from Model 0 and Model 2.

Patients	Model 0 φ_j^0	Model 2 φ_j^2	Interviewer
1 to n_1	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_1$	1
n_1+1 to n_2	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_2$	2
n_2+1 to n_3	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_3$	3
n_3+1 to n_4	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id}) + \gamma_4$	4
n_4+1 to N	$(\tau_i - \rho_d\theta_{id})$	$(\tau_i - \rho_d\theta_{id})$	5 (reference)

TABLE 6 | Mean and SD of SE of $\hat{\theta}$ from three models.

θ	MGRM	Model 0	Model 2
SE θ_1			
Mean	0.307	0.280	0.279
SD	0.093	0.076	0.076
SE θ_2			
Mean	0.252	0.242	0.241
SD	0.082	0.071	0.070
SE θ_3			
Mean	0.178	0.171	0.171
SD	0.079	0.074	0.074

variations such as different and changing test-taking behaviors (e.g., Wang and Xu, 2015; Molenaar et al., 2018; Wang et al., 2018b,c). However, the usage of item-level RT information has rarely been explored in health measurement.

This study systematically investigated the application of RTs for improving measurement precision of the target latent traits and the estimation precision of the item parameters. The bivariate joint model discussed in Molenaar et al. (2015) was applied and expanded in two respects: (1) a multivariate θ was considered in the measurement model for responses, and this θ vector was correlated with the latent speed through the cross-relation term; (2) an interviewer covariate was entered into the model to explain the variability of the observed RTs. Patient-reported outcomes obtained from personal interview surveys are widely used in health services research studies (Clancy and Collins, 2010), especially when conducting such surveys among older adults or patients with severe symptoms like the sample used in the present research. Thus, the observed RTs might be contaminated by the interviewer's reaction speed and, hence, the interviewer variable should be included in the model.

Several approaches to including the interviewer variable were explored. Results indicated that Model 2, which is a hierarchical model, consistently best fit the data. In this model, the interviewer's effect on the observed RTs is mediated through patients' latent speed. It is more parsimonious than Model 1 in which the interviewer's effect could differ for different items. Indeed, Model 2 also makes more intuitive sense because the interviewer effect reflects the different interviewers' response styles (i.e., fast or slow responders) that could be considered as the latent speed of an interviewer; hence, it should not change from item to item.

Results from the data analysis revealed that (1) adding response time information did not affect the item parameter estimates and their standard errors significantly; and (2) adding response time information helped reduce the standard error of patients' multidimensional latent trait estimates, but adding interviewer as a covariate did not result in further improvement, although the interviewer effect was significant. Regarding the first point, it is not surprising because Ranger (2013) has proven that the amount of (Fisher) information RTs provide to θ cannot be $> \frac{\rho^2}{1-\rho^2}$ (i.e., an upper bound) regardless of test length and RT distributions. A simple explanation is that RTs only contribute to θ via τ due to the hierarchical structure in van der Linden (2007), and hence the maximum information RTs provide is when τ is "observed," resulting in the information upper bound. As a result, the collateral information provided by RT will be useful when test length is short, but its role diminishes in longer tests when information accrued through responses is already high. That said, it is still worth pointing out that the role of speed as a self-contained construct might be useful for psychological and health assessment. It might be particularly promising to investigate the additional validity of the assessment by including speed in the prediction of external criteria.

An immediate implication for the follow-up adaptive design of the AM-PAC is that RT does not need to be included in interim θ estimation (i.e., selecting items during assessment delivery), but it could be used to improve the final θ estimates. Moreover, to further improve the time efficiency of adaptive testing, the maximum information per time unit (Fan et al., 2012) or its simplified version (Cheng et al., 2017) could be applied. In this case, the interviewer effect could be ignored when estimating an individual patient's speed, as long as item time parameters are provided. This is pragmatically sound because it is likely that different interviewers will be used for adaptive testing data collection in some measurement environments.

Due to the positive skewness of the RT distribution, typical log-transformations were used (van der Linden, 2007; Wang and Xu, 2015; Qian et al., 2016), and the raw RT data was cleaned by trimming the extremely short and long observations. However, recent research by Marmolejo-Ramos et al. (2015a) suggested that the Box-Cox transformation outperformed the elimination methods in normalizing positively skewed data. Vélez et al. (2015) proposed a new approach to estimate the parameter λ in the Box-Cox transformation. In cases in which the

TABLE 7 | Final Pearson correlation parameter estimates for the three models from two calibration stages.

Stage and Model	$\rho_{\theta_1\theta_2}$	$\rho_{\theta_1\theta_3}$	$\rho_{\theta_2\theta_3}$	ρ_1	ρ_2	ρ_3	γ_p
STAGE 1: CONCURRENT CALIBRATION (BATCH 2 TO 4)							
MGRM	0.624	0.468	0.846	–	–	–	–
Model 0	0.625	0.488	0.839	0.425	0.458	0.418	–
Model 2	0.628	0.492	0.840	0.583	0.629	0.578	(1.150, 1.053, 0.725, −0.911)
STAGE 2: FIXED PARAMETER CALIBRATION							
MGRM	0.702	0.545	0.869	–	–	–	–
Model 0	0.707	0.584	0.881	0.400	0.433	0.457	–
Model 2	0.706	0.583	0.880	0.527	0.577	0.605	(1.063, −0.286, 1.315, −0.135, 1.046)

log-transformation is insufficient, the Box-Cox transformation could be a viable alternative. In the present study, the extremely long and short RTs were trimmed because those RTs were considered as outliers. On the other hand, when there is lack of information on the outliers, Ueda's method could be used to automatically detect discordant outliers (Marmolejo-Ramos et al., 2015b). Because observed RTs could exhibit different skewed distributions, a careful decision needs to be made with respect to dealing with outliers, data transformation, and using the mean vs. the median, for making valid inferences (Rousselet and Wilcox, 2018). When the median is used, then quantile regression instead of a mean-based linear model should be considered instead.

AUTHOR CONTRIBUTIONS

CW contributed to constructing the ideas, explaining the results, and drafting the paper. DW contributed to the constructing the ideas and editing the draft. SS contributed to conducting the data cleaning, model fitting/analysis, and constructing tables/figures.

ACKNOWLEDGMENTS

This research was supported by the Eunice Kennedy Shriver National Institutes of Child Health and Human Development of the National Institutes of Health under Award Number R01HD079439 to the Mayo Clinic in Rochester, Minnesota through a subcontract to the University of Minnesota.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00051/full#supplementary-material

REFERENCES

Anstey, K. J., Wood, J., Lord, S., and Walker, J. G. (2005). Cognitive, sensory and physical factors enabling driving safety in older adults. *Clin. Psychol. Rev.* 25, 45–65. doi: 10.1016/j.cpr.2004.07.008

Bassili, J. N. (1996). "The how and why of response latency measurement in telephone surveys," in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, eds N. Schwarz and S. Sudman (San Francisco, CA: Jossey-Bass), 319–346.

Braver, T. S., and Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neurosci. Biobehav. Rev.* 26, 809–817. doi: 10.1016/S0149-7634(02)00067-2

Bridgeman, B., and Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *J. Educ. Measur.* 41, 137–148. doi: 10.1111/j.1745-3984.2004.tb01111.x

Cai, L. (2013). *flexMIRT® Version 2: Flexible Multilevel Multidimensional Item Analysis and Test Scoring [Computer Software]*. Chapel Hill, NC: Vector Psychometric Group.

Cheng, Y., Diao, Q., and Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behav. Res. Methods* 49, 502–512. doi: 10.3758/s13428-016-0712-6

Clancy, C., and Collins, F. S. (2010). Patient-centered outcomes research institute: the intersection of science and health care. *Sci. Trans. Med.* 2, 37cm18-37cm18. doi: 10.1126/scitranslmed.3001235

Cox, B., Blaxter, M., Buckle, A., Fenner, N., Golding, J., Gore, M., et al. (1987). *The Health and Lifestyle Survey. Preliminary Report of a Nationwide Survey of the Physical and Mental Health, Attitudes and Lifestyle of a Random Sample of 9,003 British Adults*. London: Health Promotion Research Trust.

Der, G., and Deary, I. J. (2006). Age and sex differences in reaction time in adulthood: results from the United Kingdom Health and Lifestyle Survey. *Psychol. Aging* 21, 62. doi: 10.1037/0882-7974.21.1.62

Diao, Q., and van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Appl. Psychol. Meas.* 35, 398–409. doi: 10.1177/0146621610392211

Fan, Z., Wang, C., Chang, H.-H., and Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *J. Educ. Behav. Stat.* 37, 655–670. doi: 10.3102/1076998611422912

Fletcher, R. B., and Hattie, J. A. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychol. Sport Exerc.* 5, 423–446. doi: 10.1016/S1469-0292(03)00036-0

Fraley, R. C., Waller, N. G., and Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *J. Pers. Soc. Psychol.* 78, 350. doi: 10.1037/0022-3514.78.2.350

Galesic, M., and Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opin. Q.* 73, 349–360. doi: 10.1093/poq/nfp031

Gauggel, S., Rieger, M., and Feghoff, T. (2004). Inhibition of ongoing responses in patients with Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* 75, 539–544.

Holden, R. R., and Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychol. Assess.* 4, 170. doi: 10.1037/1040-3590.4.2.170

Hsieh, C.-A., von Eye, A. A., and Maier, K. S. (2010). Using a multivariate multilevel polytomous item response theory model to study parallel processes of change: the dynamic association between adolescents' social isolation and engagement with delinquent peers in the National Youth Survey. *Multivariate Behav. Res.* 45, 508–552. doi: 10.1080/00273171.2010.483387

Huang, M. E., Cifu, D. X., and Keyser-Marcus, L. (2000). Functional outcomes in patients with brain tumor after inpatient rehabilitation: comparison with traumatic brain injury. *Am. J. Phys. Med. Rehabil.* 79, 327–335. doi: 10.1097/00002060-200007000-00003

Hultsch, D. F., MacDonald, S. W., and Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *J. Gerontol.* 57, 101–115. doi: 10.1093/geronb/57.2.P101

Jiang, S., and Wang, C. (2019). "The different effects of collapsing categories on the graded response model and the generalized partial credit model," in *Paper to be Presented at the 2019 AERA Meeting* (Toronto, ON).

Jiang, S., Wang, C., and Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Front. Psychol.* 7:109. doi: 10.3389/fpsyg.2016.00109

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *J. Educ. Meas.* 43, 355–381. doi: 10.1111/j.1745-3984.2006.00021.x

Kolen, M. J., and Brennan, R. L. (2004). Test *Equating, Scaling, and Linking*. New York, NY: Springer-Verlag.

Loeys, T., Rosseel, Y., and Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika* 76, 487–503. doi: 10.1007/s11336-011-9211-y

Marmolejo-Ramos, F., Cousineau, D., Benites, L., and Maehara, R. (2015a). On the efficacy of procedures to normalize Ex-Gaussian distributions. *Front. Psychol.* 5:1548. doi: 10.3389/fpsyg.2014.01548

Marmolejo-Ramos, F., and González-Burgos, J. (2013). A power comparison of various tests of univariate normality on ex-Gaussian distributions. *Methodology* 9, 137–149. doi: 10.1027/1614-2241/a000059

Marmolejo-Ramos, F., Vélez, J., and Romão, X. (2015b). Automatic detection of discordant outliers via the Ueda's method. *J. Stat. Distrib. Appl.* 2: 8. doi: 10.1186/s40488-015-0031-y

Molenaar, D., Bolsinova, M., and Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *Br. J. Math. Stat. Psychol.* 71, 205–228. doi: 10.1111/bmsp.12117

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *Br. J. Math. Stat. Psychol.* 68, 197–219. doi: 10.1111/bmsp.12042

Muthén, L. K., and Muthén, B. O. (1998-2015). *Mplus User's Guide 7th Edn*. Los Angeles, CA: Muthén & Muthén

O'Dell, M. W., Barr, K., Spanier, D., and Warnick, R. E. (1998). Functional outcome of inpatient rehabilitation in persons with brain tumors. *Arch. Phys. Med. Rehabil.* 79, 1530–1534. doi: 10.1016/S0003-9993(98)90414-2

Osmon, D. C., Kazakov, D., Santos, O., and Kassel, M. (2018). Non-Gaussian distribuitonal analyses of reaction time: improvements that increase efficacy of RT tasks for describing cognitive processes. *Neuropsychol. Rev.* 28, 359–376. doi: 10.1007/s11065-018-9382-8

Pearson, K. (1924). *The Life Letters and Labours of Francis Galton: Volume II. Researches of Middle Life*. Cambridge: Cambridge University Press.

Pilkonis, P. A., Kim, Y., Yu, L., and Morse, J. Q. (2014). Adult Attachment Ratings (AAR): an item response theory analysis. *J. Pers. Assess.* 96, 417–425. doi: 10.1080/00223891.2013.832261

Qian, H., Staniewska, D., Reckase, M., and Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educ. Meas.* 35, 38–47. doi: 10.1111/emip.12102

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika* 78, 538–544. doi: 10.1007/s11336-013-9324-6

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., and Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika* 68, 589–606. doi: 10.1007/BF02295614

Rousselet, G. A., and Wilcox, R. R. (2018). *Reaction Times and Other Skewed Distributions: Problems With the Mean and the Median*. Available online at: https://www.biorxiv.org/content/biorxiv/early/2018/08/10/383935.full.pdf.

Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Appl. Stat.* 31, 115–124. doi: 10.2307/2347973

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr. Suppl.* 34(4 Pt. 2), 100. doi: 10.1007/BF03372160

Siem, F. M. (1996). The use of response latencies to self-report personality measures. *Military Psychol.* 8, 15–27. doi: 10.1207/s15327876mp0801_2

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* 19, 279–281. doi: 10.1214/aoms/1177730256

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

van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *J. Educ. Behav. Stat.* 34, 378–394. doi: 10.3102/1076998609332107

Van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

van der Linden, W. J., Klein Entink, R. H., and Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Appl. Psychol. Meas.* 34, 327–347. doi: 10.1177/0146621609349800

Van Der Linden, W. J., Scrams, D. J., and Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Appl. Psychol. Meas.* 23, 195–210. doi: 10.1177/01466219922031329

Vélez, J. I., Correa, J. C., and Marmolejo-Ramos, F. (2015). A new approach to the Box–Cox transformation. *Front. Appl. Math. Stat.* 1:12. doi: 10.3389/fams.2015.00012

Verbruggen, F., Chambers, C. D., and Logan, G. D. (2013). Fictitious inhibitory differences: how skewness and slowing distort the estimation of stopping latencies. *Psychol. Sci.* 24, 352–362. doi: 10.1177/0956797612457390

Wang, C., Chang, H. H., and Douglas, J. A. (2013a). The linear transformation model with frailties for the analysis of item response times. *Br. J. Math. Stat. Psychol.* 66, 144–168. doi: 10.1111/j.2044-8317.2012.02045.x

Wang, C., Fan, Z., Chang, H.-H., and Douglas, J. A. (2013b). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J. Educ. Behav. Stat.* 38, 381–417. doi: 10.3102/1076998612461831

Wang, C., Kohli, N., and Henn, L. (2016). A second-order longitudinal model for binary outcomes: item response theory versus structural equation modeling. *Struct. Equat. Model.* 23, 455–465. doi: 10.1080/10705511.2015.1096744

Wang, C., Su, S., and Weiss, D. J. (2018a). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate Behav. Res.* 53, 403–418. doi: 10.1080/00273171.2018.1455572

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Xu, G., and Shang, Z. (2018b). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Wang, C., Xu, G., Shang, Z., and Kuncel, N. (2018c). Detecting aberrant behavior and item preknowledge: a comparison of mixture modeling method and residual method. *J. Educ. Behav. Stat.* 43, 469–501. doi: 10.3102/1076998618767123

Yost, K., Jette, A., Wang, C., Weiss, D., and Cheville, A. (2018). "Computerized adaptive testing to direct delivery of hospital-based rehabilitation: item-bank generation and data collection," in *Poster Prestend at the 25th International Society for Quality of Life Research* (Dublin).

Zagorsek, H., Stough, S. J., and Jaklic, M. (2006). Analysis of the reliability of the leadership practices inventory in the item response theory framework. *Int. J. Select. Assess.* 14, 180–191. doi: 10.1111/j.1468-2389.2006.00343.x

Zickar, M. J., and Robie, C. (1999). Modeling faking good on personality items: an item-level analysis. *J. Appl. Psychol.* 84, 551. doi: 10.1037/0021-9010.84.4.551

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.