JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12286 *Journal of Educational Measurement*
Summer 2021, Vol. 58, No. 2, pp. 262–280

Robust Estimation for Response Time Modeling

Maxwell Hong, Daniella A. Rebouças, and Ying Cheng
University of Notre Dame

Response time has started to play an increasingly important role in educational and psychological testing, which prompts many response time models to be proposed in recent years. However, response time modeling can be adversely impacted by aberrant response behavior. For example, test speededness can cause response time to certain items to deviate from the hypothesized model. In this article, we introduce a robust estimation approach when estimating a respondent's working speed under the log-normal model by down-weighting aberrant response times. A simulation study is carried out to compare the performance of two weighting schemes and a real data example is provided to showcase the use of the new robust estimation method. Limitations and future directions are also discussed.

Introduction

Collecting response times during an assessment is becoming more prominent in educational and psychological research (Lee & Jia, 2014). Response times have much to add to the measurement community, such as improved estimation of item response models (van der Linden et al., 2010), item selection algorithms in computerized adaptive testing (Choe, Kern, & Chang, 2017; Fan, Wang, Chang, & Douglas, 2012; Cheng, Diao, & Behrens, 2017), and detection of aberrant respondents (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; Wang, Xu, & Shang, 2016).

There are several ways to model response times (Van Zandt, 2000; Rouder, Sun, Speckman, Lu, & Zhou, 2003). For instance, some models assume the response times follow a Weibull distribution (Rouder et al., 2003), Gamma distribution (Maris, 1993), Poisson counting process (Ratcliff & Smith, 2004), and others motivated by psychological theory capturing the cognitive process. In recent years, several researchers have utilized the log-normal model in educational measurement (van der Linden, 2006; Thissen, 1983), which has an interpretation analogous to item response theory (IRT). Unfortunately, a specific response time model may not be able to explain the variability of the response times for all respondents of a survey or assessment. This lack of model fit may be due to aberrant behavior such as item preknowledge (cheating), speededness, or a warm-up effect (Sinharay, 2018). If one assumes a parametric model, aberrant response times violate any assumptions of the model when those response times are not taken into account. Parameter estimates will, therefore, be biased unless one takes into consideration the deviating behavior.

In order to overcome this obstacle, one may instead employ robust estimation procedures (Huber & Ronchetti, 2009). Robust estimation has been applied in the general statistical literature, and can overcome violations of model-based assumptions such as heteroscedastic errors in the linear model (Wilcox, 2012). Robust estimation is also one way to effectively improve the estimates of a parametric model when there are outliers or aberrant response patterns in the data. When the population

distribution of a sample is unknown, robust procedures produce more efficient parameter estimates compared to maximum-likelihood (ML) estimates. Although there are many robust estimation approaches, we will limit our discussion to M-estimators, which were proposed by Huber (1967) since they are closely related to ML estimators. M-estimation works by down-weighting data points that are considered extreme cases compared to the rest of observations (Wilcox, 2012; Huber & Ronchetti, 2009). It has been shown to perform well with data that contain outliers (Huber & Ronchetti, 2009; Tyler, 1984; Yuan & Zhong, 2013).

M-estimation has successfully been employed in the general psychometric literature. For example, in the structural equation modeling (SEM) framework, it has been shown to improve the performance of fit indices and relative fit indices (Yuan & Zhong, 2013). Robust estimation has also been shown to improve the ability to produce less biased estimates and better coverage rates of psychometric statistics and indices, such as reliability coefficients α and ω of scores based on psychological scales (Zhang & Yuan, 2016).

In the context of IRT, robust estimation has been employed to address similar concerns (Hong & Cheng, 2019; Schuster & Yuan, 2011; Wainer & Thissen, 2008). M-estimation has been shown to improve latent trait or ability estimates when response vectors are contaminated with careless responses or test speededness (Schuster & Yuan, 2011; Wainer & Thissen, 2008; Mislevy & Bock, 1982). More recently, M-estimation has also been found to improve the ability to recover structural or item parameters using the same general framework (Hong & Cheng, 2019). Employing robust estimation has also been shown to improve the ability to detect deviating behavior compared to no downweighting or data removal (Hong & Cheng, 2019; Kim, Reise, & Bentler, 2018; Sinharay, 2018).

Robust estimation for models beyond those applied to item scores collected during the measurement process has not been previously investigated, which is the aim of the current study. With the advent of technology-enhanced assessment platforms, new information about individuals, such as process data, is becoming more ubiquitous (Bergner & von Daiver, 2018). Data such as eye- and motion-tracking, log data, and response times generated under computerized assessments and game-based learning systems are gaining unprecedented attention in the field of psychometrics. However, similar issues as those observed in traditional assessment settings may occur with new data sources, such as speededness, inattentiveness, and cheating. This problem motivates the current paper's discussion of robust estimation for process data, with a specific emphasis on response time modeling.

We will provide a general framework to estimate participants' working speed parameter based on van der Linden's (2006) log-normal model. More specifically, (ML) and two types of robust estimators, using either Tukey's bisquare/biweight (BW) or Huber (HU) weight functions, will be investigated when estimating the working speed parameter for an individual (Huber, 1964; Schuster & Yuan, 2011). To the authors' knowledge, this is the first time robust estimation has been introduced in the literature for latent variable models with response time data.

In the following sections, we will first introduce a response time model, the log-normal model, by van der Linden (2006). Next, we will review robust statistics and introduce how M-estimation can be used when estimating a participant's working

speed. A simulation study will be presented that demonstrates how robust estimation can improve the parameter estimates of the response time model in terms of bias and efficiency under several conditions. An applied analysis will also be conducted based on a large-scale examination. We will conclude with a discussion of future directions and limitations of the current study.

Response Time Model

van der Linden (2006) introduced how a latent variable model with a log-normal link function can be used to model response times. Given a test with I items, let t_i denote the response time for an examinee on the ith item, where $i = 1, 2, \ldots, I$. Assuming the log response times for an individual are independent, given a participant's latent working speed, τ:

$$\log(t_i)|\tau \sim N\left(\beta_i - \tau, \frac{1}{\alpha_i^2}\right). \tag{1}$$

The average (log) response time for a given item is $\beta_i - \tau$, with a variance of $\frac{1}{\alpha_i^2}$. β_i is the time-intensity parameter and α_i is the discrimination parameter for item i. The item and person parameters can be interpreted in a similar fashion to the two-parameter logistic item response model (Birnbaum, 1968). α_i is analogous to the item discrimination parameter in the 2PL model, and β_i is analogous to the difficulty parameter. The interpretation of the latent working speed τ is analogous to the latent ability in an item response model. We omit any person subscript. For example, a test taker with high working speed is likely to show quicker response times when completing an item compared to test-takers with low working speed. In order to ensure the model is identifiable, τ is assumed to have a mean of 0.

Equivalent to Equation 1, one can write out the model as follows:

$$f(t_i|\tau) = \frac{\alpha_i}{t_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\log t_i - (\beta_i - \tau))]^2\right\}. \tag{2}$$

To obtain an estimate of τ through ML estimation, we take the derivative with respect to τ over i items of the (log) likelihood and set it equal to 0. The likelihood can be written out as:

$$L(\tau) = \prod_{i=1}^{I} L_i(\tau|\alpha_i, \beta_i) = \prod_{i=1}^{I} \frac{\alpha_i}{t_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\log t_i - (\beta_i - \tau))]^2\right\}, \tag{3}$$

and the log-likelihood function as:

$$l(\tau) = \sum_{i=1}^{I} l_i(\tau|\alpha_i, \beta_i) = \sum_{i=1}^{I} \log \alpha_i - \log t_i\sqrt{2\pi} - \frac{1}{2}[\alpha_i(\log t_i - (\beta_i - \tau))]^2. \tag{4}$$

According to van der Linden (2006), the ML estimate of τ (or $\hat{\tau}^{ml}$) is:

$$\hat{\tau}^{ml} = \frac{\sum_i^I \alpha_i^2(\beta_i - \log t_i)}{\sum \alpha_i^2}. \tag{5}$$

Estimates of τ are often used in subsequent analyses, such as item selection algorithms (Choe et al., 2017; Cheng et al., 2017) or investigating test speededness (van der Linden, 2007; Sinharay, 2018). Hence it is very important to ensure its estimation accuracy and precision. Note that when estimating τ through Equation 5, each item has a weight proportional to α_i, or how discriminating the item is. However, each item is equally weighted when summing across each item in the log-likelihood function. Aberrant response behavior such as speededness, inattentiveness, and cheating can influence response times, similar to influencing item responses because of equal contribution when estimating τ. When participants speed through some of the items, those response times deviate from the log-normal model, which will bias the estimate of τ for that test-taker. The bias in τ would impact any subsequent analyses or uses of τ in operational testing. For instance, if a test-taker takes longer periods of time during the beginning of an exam due to an unfamiliar environment, our estimate of τ would be negatively biased compared to if they were already familiar with the testing environment. Therefore, item selection algorithms, such as in Cheng et al. (2017) and Choe et al. (2017), would select inappropriate items later in the test, such as easier items that may require less time. Robust estimation of τ could assuage the impact of earlier warm-up effects and improve test administration. Post hoc analysis, such as diagnosing test speededness, would also improve with robust estimation for response time models. Speedeness is a known nuisance variable where test takers are unable to fully understand a question due to time pressures and can compromise the validity of a test or scale (van der Linden, 2007). Detection methods used to identify test speededness, such as person fit statistics, require good estimates of a user's working speed when analyzing response times (Sinharay, 2018). By using robust methods, one would have a better estimate of τ they could plug into the person fit statistic and have more power to detect said individuals because the deviating behavior is downweighted.

In the next section, we introduce a new robust estimation method for estimating τ using M-estimation, which down-weights the items to which participants respond in an aberrant manner.

Robust Estimation for Response Time Models

A more general approach to estimating τ is to include a weight function in the (log) likelihood. Instead of defining the ML estimator of working speed as the solution to the equation $\sum_i dl_i/d\tau = 0$, one can define weights, $w(r_i)$, as a function of residual, r_i, for item i, and use a weighted likelihood equation,

$$\sum_i^I w(r_i)\left(\frac{dl_i}{d\tau}\right) = 0, \tag{6}$$

to solve for the unknown parameter τ. This leads to another closed-form solution for $\hat{\tau}$:

$$\hat{\tau}^{rml} = \frac{\sum_i w(r_i)\alpha_i^2(\beta_i - \log t_i)}{\sum_i w(r_i)\alpha_i^2}. \tag{7}$$

Equation 7 differs from Equation 5 by including the weights, $w(r_i)$, in both summations in the numerator and denominator. In the robust literature, $w(r_i)$ is determined by two decisions (Huber & Ronchetti, 2009): define the form of the residual and weight function. The residual, r_i, is a measure of which response times are inconsistent under the assumed response model for a given participant. Given a response vector for an individual, under the log-normal model for response times, we can define the residual for a single item similar to Sinharay (2018):

$$r_i = \alpha_i(\log(t_i) - (\beta_i - \hat{\tau})). \tag{8}$$

Other residuals in the context of response time models can be used (van der Linden & van Krimpen-Stoop, 2003). This residual has the same desirable properties as the residual defined for item response models in Schuster and Yuan (2011). The larger the discrepancy between the expected time on the log scale, $\beta_i - \hat{\tau}$, the larger the residual for the observed log-time. Response times with large residuals would contribute less to the estimation of τ. Items that are highly discriminating would also be up-weighted, which is desirable because these items have more information about τ. These two features combined make the residual by Sinharay (2018) a prime candidate to plug into $\hat{\tau}^{rml}$. Note that r_i is conditionally normally distributed given a fixed value of τ. This can be derived following Equation 8.

Second, one needs to specify a weight function $w(\cdot)$, which can be selected from a standard pool of functions; see table 11-1 in Hoaglin, Mosteller, and Tukey (2006). Two common weight functions used in this study are the so-called Huber (HU) and Tukey's Biweight (BW) or bisquare functions. The general form of the BW weight function is:

$$w(r_i) = \begin{cases} [1 - (r_i/B)^2]^2, & \text{for } |r_i| < B \\ 0, & \text{for } |r_i| \geq B. \end{cases} \tag{9}$$

The tuning constant, B, determines which residuals are down-weighted when estimating τ. Smaller values of B lead to more observations with a weight of 0. For a given r_i, smaller B leads to smaller weight. Previous research in item response models have found $B = 4$ to be an adequate cut-off (Schuster & Yuan, 2011). For $B = 4$, any response with a residual that is larger than 4 will be removed entirely when estimating τ. Another weight function is HU-type weights. The formula for HU-type weights is:

$$w(r_i) = \begin{cases} 1, & \text{for } |r_i| \leq H \\ H/|r_i|, & \text{for } |r_i| > H. \end{cases} \tag{10}$$

Similar to the BWs, large values of H will lead to less down-weighting and small values lead to more down-weighting of the observations. Previous literature have found that when $H = 1$, the BW function appears to work well (Schuster & Yuan, 2011). In order to get a sense of how the different weight functions perform, Figure 1 plots the different weights and values of B and H against the residuals for both HU and BW functions. The BW function decreases immediately when the residual deviates from 0, whereas the HU weights down-weight responses for any residual less than or equal to H. Note that the HU weight function asymptotically approaches 0 as the residual approaches positive or negative infinity.

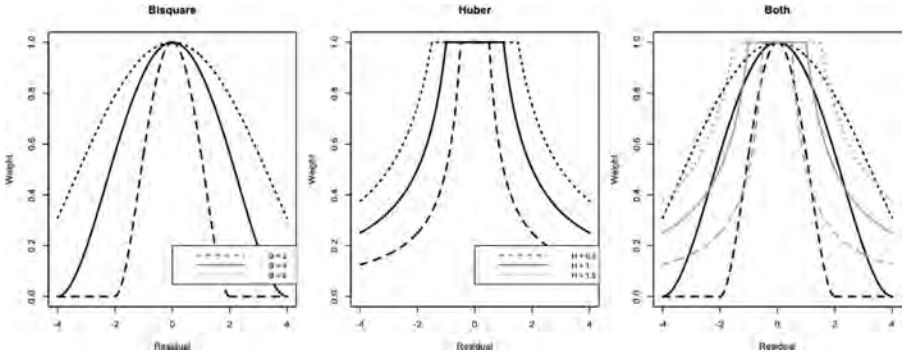

Figure 1. Comparison of bisquare and Huber-type weight functions with different tuning parameters.

Algorithm 1 outlines the robust estimation algorithm to compute $\hat{\tau}^{rml}$. For a single response time vector, one first calculates $\hat{\tau}^{ml}$, which is also the first estimate of $\hat{\tau}^{rml}$. Next, one plugs in the current estimate of $\hat{\tau}^{rml}$ into Equation 8 to obtain the residuals for each item i. One then transforms the residuals calculated in the previous step using a suitable weight function, such as Equation 9 or 10. Finally, one calculates a new $\hat{\tau}^{rml}$, which is also $\hat{\tau}_{m+1}$ or the new current estimate $\hat{\tau}^{rml}$, in the algorithm. We repeat the steps starting from calculating the residuals using the new estimate $\hat{\tau}^{rml}$ until some convergence criteria is met. The algorithm needs to be iterated because the residuals depend on the current estimate of τ. Two criteria are the total number of iterations, $m = 1 \ldots M$, and the absolute difference in adjacent estimates of $\hat{\tau}$, or $|\hat{\tau}_{(m+1)} - \hat{\tau}_{(m)}|$. If the total number of iterations reach M or if the absolute difference between adjacent iterations of $\hat{\tau}$ drops below some threshold K, the algorithm is terminated and the most recent estimate is used as $\hat{\tau}^{rml}$. In the current study, we fixed M to be 30 and K to be .0001. However, it is rare for the algorithm to reach M. In the following sections, we will illustrate how $\hat{\tau}^{rml}$ compares to $\hat{\tau}^{ml}$ in a simulation study.

Algorithm 1: Robust estimation for τ based on van der Linden (2006)'s log normal model

1: $\hat{\tau}_1 = \frac{\sum_i^I \alpha_i^2(\beta_i - \log t_i)}{\sum \alpha_i^2}$
2: Set m = 1
3: **while** $m < M$ and $|\hat{\tau}_{(m+1)} - \hat{\tau}_{(m)}| < K$ **do**
4: $\quad r_{(m+1)i} = \alpha_i[\log t_i - (\beta_i - \hat{\tau}_m)]$
5: $\quad$ Compute $w(r_{(m+1)i})$ based on Equations 9 or 10
6: $\quad \hat{\tau}_{(m+1)} = \frac{\sum_i^I w(r_{(m+1)i})\alpha_i^2(\beta_i - \log t_i)}{\sum_i^I w(r_{(m+1)i})\alpha_i^2}$
7: $\quad$ m = m + 1
8: **end while**
9: **return** $\hat{\tau}^{rml} = \hat{\tau}_{(m+1)}$

Simulation

A simulation study was conducted to compare three estimators: ML and robust estimators using either BW or HU weight functions. Response times were generated with 30, 50, or 100 items under van der Linden's (2006) log-normal model. Data were simulated such that $\beta \sim N(3.8, .25)$ and $\alpha \sim U(1.75, 3.25)$, which mirrors previous simulation studies with response times (Patton, 2014). These choices generate response times that are in the scale of seconds (instead of minutes or hours) and are congruent with the expected amount of time necessary to complete an item. Working speed τ was generated at fixed values between -2.0 and 2.0 at intervals of .5.

We generated three types of aberrancy of response time that mirror real testing scenarios. We followed work done by van der Linden & van Krimpen-Stoop (2003):

$$\log(t_i)|\tau \sim N\left(\beta_i - \tau + \delta, \frac{1}{\alpha_i^2}\right), \tag{11}$$

where δ introduces a shift due to aberrant behavior. We varied δ from -2 to 2 by increments of .5. When $\delta = 0$, Equation 11 reduces to Equation 1 where there is no aberrant behavior. Larger absolute values of δ correspond to more deviant behavior. When $\delta > 0$, the aberrant behavior mimics test speededness. Speededness occurs when participants take shorter time due to fatigue or time pressure (van der Linden, 2009). When $\delta < 0$, the aberrant behavior mimics warm-up effect. A warm-up effect occurs when participants take a longer time for items in the beginning of a test, which could be due to unfamiliarity with the testing environment (van der Linden, 2009). In order to avoid computational errors, each observed response time was constrained to never be lower than 1 second. We varied the number of aberrant response times to be either 0%, 10%, 20%, or 30% of a test.

In total, we had 9 fixed levels of τ, 3 estimation techniques (ML, HU, BW), 18 types of aberrant behavior (values of δ), 4 levels of aberrancy (0%, 10%, 20%, or 30% items), and 3 test lengths (30, 50, and 100). We also varied the parameter B to be 2, 4, and 6 and H to be 0.5, 1, and 1.5 to examine the effect of the tuning parameters. The tuning parameters were based on values used when evaluating item response models (Mislevy & Bock, 1982; Wainer & Thissen, 2008). We replicated each condition 1,000 times and evaluated each condition in terms of bias, $\frac{1}{1{,}000}\sum(\hat{\tau} - \tau)$, and MSE, $\frac{1}{1{,}000}\sum(\hat{\tau} - \tau)^2$. Example code for the estimation process can be found in the supplementary material.

Simulation Results

Due to the large number of simulation conditions, results from a subset of the conditions are presented. Other simulation conditions can be obtained by contacting the authors. We only present the simulation results with tuning parameters $B = 4$ and $H = 1$. This is due to the fact that using these tuning parameters resulted, on average, the smallest MSE across simulation conditions. The tuning parameter appears to be only sensitive when using the BW function when there are no aberrant responses. When there is warm-up response in the data for a 30 item test, when $\tau = 0$, the

Table 1
*Bias using Maximum Likelihood (ML), Biweight (*B = *4), and Huber (*H = *1) Estimation of Working Speed When There Is No Contamination in the Data*

		τ								
ML	Items	−2	−1.5	−1	−0.5	0	.5	1	1.5	2
	30	−.000	.001	.001	.001	−.000	.000	.001	−.002	.001
	50	−.001	−.001	.000	.001	−.001	−.001	.001	.001	.001
	100	.000	.000	−.001	−.001	.001	.001	−.001	−.000	−.000
Huber	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	−.001	.001	.001	.001	−.001	.000	.000	−.001	.001
	50	.002	.001	.000	−.001	−.001	−.001	−.003	.001	.000
	100	.002	−.001	.000	.000	.000	.000	−.001	−.001	.000
Tukey	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	−.001	.001	.001	.001	−.001	.000	.001	−.001	.001
	50	−.001	.000	.000	.001	−.001	−.001	.001	.001	.001
	100	.001	.001	−.001	−.001	.001	−.000	−.001	−.000	−.001

MSE is .020, .020, .024 when H is .5, 1, or 1.5, respectively, and the MSE is .018, .013, .020 when B is 2, 4, or 6, respectively. When there is aberrant behavior in the data, fixing $B = 4$ and $H = 1$ appears to work the best. However, it is important to note there are only small differences when we change the tuning parameter when comparing MSE.

Table 1 presents the average bias across 1,000 replications when there is no contamination in the data, or $\delta = 0$. Across all simulation conditions, both the robust and ML estimation provide unbiased estimates of τ. Table 2 presents the MSE across 1,000 replications when there is no contamination in the data. We find that there is little difference between ML and robust estimation when we consider the MSE across all simulation conditions when there is no contamination in the data. When there was a longer scale, the MSE decreased from .005 to .002 when the test length was 30 or 100, respectively. These findings suggest that when there is no contamination in the data, one does not need to worry about using either ML or robust estimation because they produce similar results.

When analyzing simulation results with contaminated data, increasing δ or the proportion of aberrant responses led to more bias and larger MSE across simulation conditions. Moreover, MLE had larger bias and MSE with increasing δ or the proportion of aberrant responses. Therefore, we averaged across these factors. Interested readers can contact the authors for detailed information.

Table 3 presents the bias using ML and both robust estimation methods of working speed when there are warm-up response times. There were little differences between test length. On average, using either HU or Tukey weights led to less biased estimates compared to ML estimation. Between the two robust estimation methods, Tukey

Table 2
*MSE Using Maximum Likelihood (ML), Biweight (*B = *4), and Huber (*H = *1) Estimation of Working Speed When There Is No Contamination in the Data*

		τ								
ML	Items	−2	−1.5	−1	−0.5	0	.5	1	1.5	2
	30	.005	.005	.005	.005	.005	.005	.005	.005	.005
	50	.003	.003	.003	.003	.003	.003	.003	.003	.003
	100	.002	.002	.002	.002	.002	.002	.002	.002	.002
Huber	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.005	.006	.006	.006	.005	.005	.006	.006	.006
	50	.004	.004	.004	.003	.004	.003	.003	.004	.003
	100	.002	.002	.002	.002	.002	.002	.002	.002	.002
Tukey	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.005	.006	.006	.006	.005	.005	.006	.006	.006
	50	.003	.003	.004	.003	.004	.003	.003	.004	.003
	100	.002	.002	.002	.002	.002	.002	.002	.002	.002

weights tend to provide less biased estimates on average. For instance, when $\tau = 0$ for a 30-item test, ML estimation bias was −.345. Robust estimation using HU weights bias was −.213 and Tukey weights bias was −.113. This finding provides some evidence that using Tukey weights are superior to HU weights and both robust estimation procedures are superior to ML. This is most likely due to features of the weight functions themselves. Tukey weights down-weight response times more so than HU weights conditioning on different values of the residual. Therefore, Tukey weights should provide less biased estimates than HU weights.

Table 4 presents the MSE using ML and both robust estimation methods of working speed when there are warm-up response times. Longer tests in general lead to smaller MSE. On average, using either HU or Tukey weights led to smaller MSE estimates. However, Tukey weights tend to provide smaller MSE on average. For instance, when $\tau = 0$ for a 30-item test, ML estimation MSE was .161. Robust estimation using HU weights MSE was .063 and Tukey weights MSE was .029.

Table 5 presents the bias using ML and both robust estimation methods of working speed when there are speeded response times in the data. One important note is that the bias observed for the speeded response times has the opposite affect compared to the warm-up effect. For instance, when $\tau = 0$ for a 30-item test, ML estimation bias was −.345 when there is warm-up effect compared to .346 when there was test speededness. We find similar patterns across different test lengths and values of τ. On average, using either HU or Tukey weights led to less biased estimates. However, Tukey weights tend to provide less biased estimates on average. For instance, when $\tau = 0$ for a 30-item test, ML estimation bias was .346. Robust estimation using HU weights bias was .213 and Tukey weights bias was .113. This finding provides

Table 3
*Bias Using Maximum Likelihood (ML), Biweight (*B = *4), and Huber (*H = *1) Estimation of Working Speed When There Are Warm-Up Responses in the Data*

		τ								
ML	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	−.380	−.366	−.397	−.398	−.345	−.390	−.385	−.369	−.385
	50	−.373	−.367	−.376	−.385	−.364	−.377	−.387	−.380	−.378
	100	−.376	−.375	−.377	−.379	−.367	−.390	−.377	−.374	−.377
Huber	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	−.228	−.228	−.251	−.252	−.213	−.243	−.239	−.230	−.243
	50	−.238	−.232	−.232	−.242	−.228	−.236	−.240	−.233	−.234
	100	−.238	−.234	−.236	−.235	−.230	−.240	−.236	−.233	−.237
Tukey	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	−.114	−.112	−.121	−.121	−.113	−.117	−.117	−.113	−.120
	50	−.116	−.120	−.116	−.115	−.114	−.118	−.118	−.116	−.116
	100	−.114	−.115	−.114	−.111	−.117	−.110	−.113	−.113	−.118

some evidence that using Tukey weights are superior to HU weights and both robust estimation procedures are superior to ML when there are speeded response times. This finding corroborates when there are warm-up response times. Regardless of the underlying mechanism generating the data in the context of this simulation study, Tukey weights provide the least biased estimates of working speed across different test lengths and values of τ.

Table 6 presents the MSE using ML and both robust estimation methods of working speed when there are speeded response times in the data. Longer tests lead to smaller MSE. Moreover, both warm-up and speeded response times provide similar trends when we compare MSE. On average, using either Huber or Tukey weights led to smaller MSE estimates. However, Tukey weights tend to provide smaller MSE on average. For instance, when $\tau = 0$ for a 30-item test, ML estimation MSE was .162. Robust estimation using Huber weights MSE was .063 and Tukey weights MSE was .029. Putting it all together, both data generation mechanisms imply that Tukey weights provide the smallest MSE and least biased estimates across different simulation conditions.

Our simulations results are consistent with the robust literature. $\hat{\tau}^{rml}$ was less biased or comparable to $\hat{\tau}^{ml}$ across simulation conditions when data were contaminated with speededness or warm-up effect. In the majority of the simulation conditions, $\hat{\tau}^{rml}$ achieved smaller MSE compared to using traditional ML estimation. Moreover, our simulations suggest that using the Tukey-weights would be preferred.

This finding contradicts findings from Schuster and Yuan, (2011). Their simulation results suggest that using Huber weights are preferred, at least in the context of item response models. An important distinction between Schuster and Yuan (2011) and

Table 4
*MSE Using Maximum Likelihood (ML), Biweight (*B = *4), and Huber (*H = *1) Estimation of Working Speed When There Are Warm-Up Responses in the Data*

		τ								
ML	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.194	.181	.205	.221	.161	.203	.194	.185	.201
	50	.192	.176	.180	.198	.177	.185	.194	.185	.188
	100	.187	.181	.187	.185	.175	.196	.184	.182	.185
Huber	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.071	.073	.087	.093	.063	.081	.079	.074	.084
	50	.080	.071	.070	.080	.070	.073	.075	.070	.072
	100	.075	.069	.072	.070	.067	.072	.072	.070	.072
Tukey	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.032	.031	.038	.041	.029	.034	.035	.032	.034
	50	.029	.029	.028	.029	.027	.029	.030	.029	.029
	100	.024	.024	.024	.022	.025	.022	.024	.024	.025

this study is that we focus on response time models. The data generation scheme to generate aberrancy is different for response times. For instance, aberrant behavior on a test can result in a string of incorrect answers where a user has 50% chance of a correct response for an item with two categories. Aberrant behavior for response times can manifest itself differently as demonstrated in the simulation studies and what is observed in practice.

Furthermore, our simulations focus on different models that have different link functions compared to (Schuster & Yuan, 2011). Convergence issues for item response models can arise when there is a significant amount of aberrant behavior. This is due to the fact that downweighing a significant amount of item responses can lead to unstable parameter estimates when using a Newton Raphson algorithm when estimating latent ability. Response time models do not suffer from extreme loss of information, at least in the simulation conditions considered in this study. One does not need to use a Newton Raphson algorithm to estimate τ because there is a closed-form solution for both the robust and ML estimation. We only need to iterate to update the residual for response time data. Therefore, using either Huber or Tukey weights does not suffer from convergence issues for response time models.

An Application to Real Data

In this section, we compare all three estimators for τ with data obtained from a large-scale test administered to a midwestern region of the United States. The test was administered as a low-stakes computerized exam to assess how third-grade students perform on a language/arts domain compared to other schools in the state.

Table 5
*Bias Using Maximum Likelihood (ML), Biweight (*B = *4), and Huber (*H = *1) Estimation of Working Speed When There Are Speeded Responses in the Data*

		τ								
ML	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.380	.365	.400	.396	.346	.391	.385	.367	.387
	50	.375	.367	.376	.384	.363	.376	.386	.381	.379
	100	.377	.374	.377	.378	.367	.391	.375	.374	.377
Huber	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.227	.229	.253	.250	.213	.243	.238	.229	.246
	50	.240	.232	.231	.241	.226	.235	.238	.236	.235
	100	.239	.233	.235	.234	.231	.241	.234	.233	.237
Tukey	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.112	.113	.122	.120	.113	.116	.118	.113	.123
	50	.118	.119	.114	.113	.111	.117	.117	.118	.117
	100	.116	.115	.113	.111	.118	.111	.111	.113	.118

Motivation, warm-up or fatigue effects may be present due to the nature and novelty of the assessment. The data included 49,163 response times from students on 32 items from the language arts subject area. Only a subset of the data was used.

The response time data were recorded in seconds. Data were then cleaned to include only complete response times and response times with total response time of at least 1 second per item (minimum total of 32 seconds). Two analyses were performed. First, item (and person) parameters of the log-normal response time model were obtained for each item through Bayesian estimation, as described in van der Linden (2006), on a randomly selected subsample of $N = 5,000$. Details on how this procedure was implemented can be found in the Appendix. Although working speed was also estimated, only the mean of the posterior distributions of the item parameters, α and β, were used. Those were taken as true item parameter estimates in subsequent analyses and are presented in Table 7.

In the second analysis, the working speed parameter was estimated on a separate sample of $N = 1,000$ using the ML and robust estimation with either Huber or BW functions, assuming known item parameters from the first subsample. In Figure 2, the boxplots show the sampling distribution of $\hat{\tau}$ for each estimator. Estimates of τ obtained using ML were binned into six intervals for graphical purposes. Standard errors (SEs) for $\hat{\tau}$ were estimated by calculating the standard deviation of $\hat{\tau}$ for each estimator within each of the six intervals, which had a sample size of 6, 125, 455, 281, 79, and 54 in increasing order of $\hat{\tau}$, respectively. Note that intervals with fewer individuals are less reliable compared to other cells.

Based on our empirical analysis, SEs are larger for extreme values of the working speed parameter using any type of estimation method. We assume it is due to less

Table 6
*MSE Using Maximum Likelihood (ML), Biweight (*B = *4), and Huber (*H = *1) Estimation of Working Speed When There Are Speeded Responses in the Data*

		τ								
ML	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.194	.181	.207	.219	.162	.203	.195	.184	.202
	50	.193	.177	.180	.198	.176	.184	.194	.186	.188
	100	.187	.180	.187	.185	.176	.197	.183	.182	.184
Huber	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.071	.073	.088	.091	.063	.081	.079	.074	.084
	50	.080	.071	.069	.079	.069	.072	.074	.071	.073
	100	.075	.069	.071	.070	.068	.073	.072	.070	.072
Tukey	Items	−2	−1.5	−1	−.5	0	.5	1	1.5	2
	30	.031	.031	.038	.042	.029	.034	.035	.032	.035
	50	.029	.028	.027	.028	.026	.028	.030	.029	.029
	100	.024	.024	.024	.022	.026	.022	.024	.024	.025

information for individuals when $\tau > .5$. In other words, we are unable to measure these individuals well given the current item features. For instance, the average time for individuals with a $\tau = 0$ is approximately 20 seconds. On the other hand, the average time for individuals who have a $\tau = 1$ is approximately 7.38 seconds. Our item bank may not be able to differentiate between individuals that take a shorter amount time. Moreover, the robust estimators had larger standard errors compared to using the ML. This is not surprising, as there can be a bias-variance trade-off when using robust estimators (Carroll & Pederson, 1993). This finding has also been reported when using robust estimation with item response models (Schuster & Yuan, 2011).

Figure 3 plots the values of $\hat{\tau}^{ml}$ and $\hat{\tau}^{rml}$ using either weight functions. In most cases, we see that the robust estimators produce similar estimates of τ when using ML, suggesting that there are not many aberrant cases in the dataset. Given the data come from a well-established large-scale operational testing program, this is not a surprise because quality control measures have been taken to minimize warm-up or speededness. Nonetheless, there are a small number of individuals with smaller estimates or larger estimates of $\hat{\tau}^{rml}$ compared to $\hat{\tau}^{ml}$. Both robust estimators tend to have similar estimates of τ. In order to probe whether the small number of individuals with larger or smaller $\hat{\tau}^{rml}$ were possibly affected by warm-up or speededness effects, we further investigate each case individually.

Warm-up respondents were those whose median total response times (on the first seven items) were larger than the 80th percentile, and whose overall $\hat{\tau}_{bisquare} - \hat{\tau}_{MLE} > .1$. Only four respondents met this criteria. In Figure 4, median response times (in seconds) for each item are plotted for the warm-up group compared to

Table 7
Estimates of the Item Parameters

Item	α	β	Item	α	β
1	1.45	3.17	17	1.68	2.79
2	1.52	3.46	18	1.71	3.57
3	1.75	3.39	29	2.08	3.26
4	1.25	3.22	20	1.72	3.30
5	1.91	3.66	21	1.75	3.10
6	2.05	3.34	22	1.75	3.41
7	2.14	3.52	23	1.10	2.89
8	2.11	3.42	24	1.24	3.39
9	1.83	3.32	25	1.51	3.48
10	1.79	3.54	26	1.92	3.32
11	1.80	3.43	27	1.20	3.40
12	1.28	3.59	28	1.67	3.39
13	1.94	3.40	29	1.67	3.27
14	2.20	3.59	30	1.69	3.42
15	2.13	3.34	31	1.95	3.16
16	1.33	3.26	32	1.57	3.29

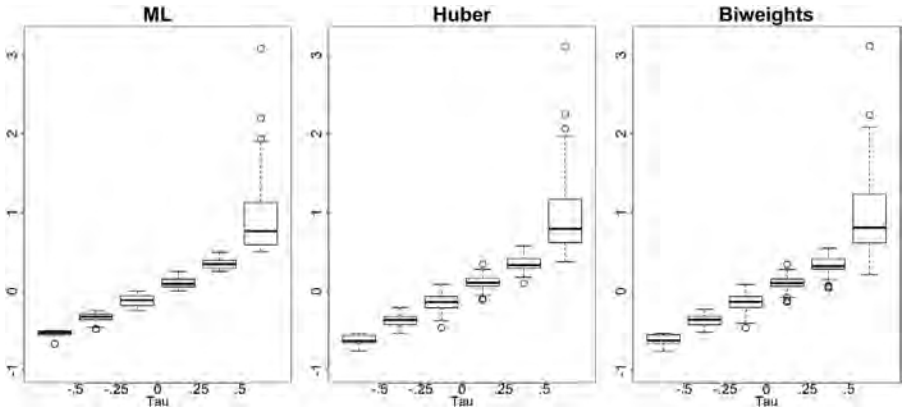

Figure 2. Comparison of sampling distributions of $\hat{\tau}$ using ML and robust methods with Huber or biweight functions.

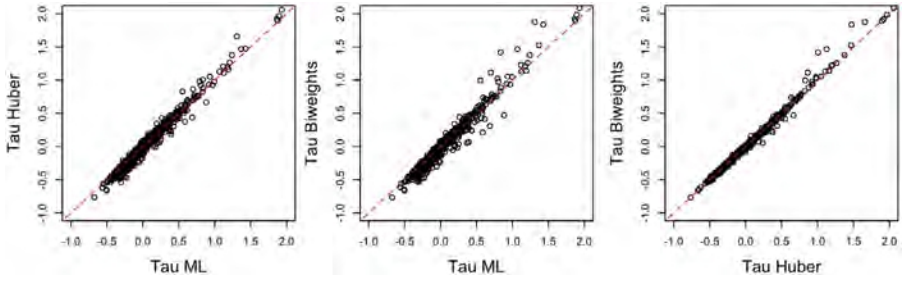

Figure 3. Comparison of ML, bisquare, and Huber-type estimates of τ. [Color figure can be viewed at wileyonlinelibrary.com]

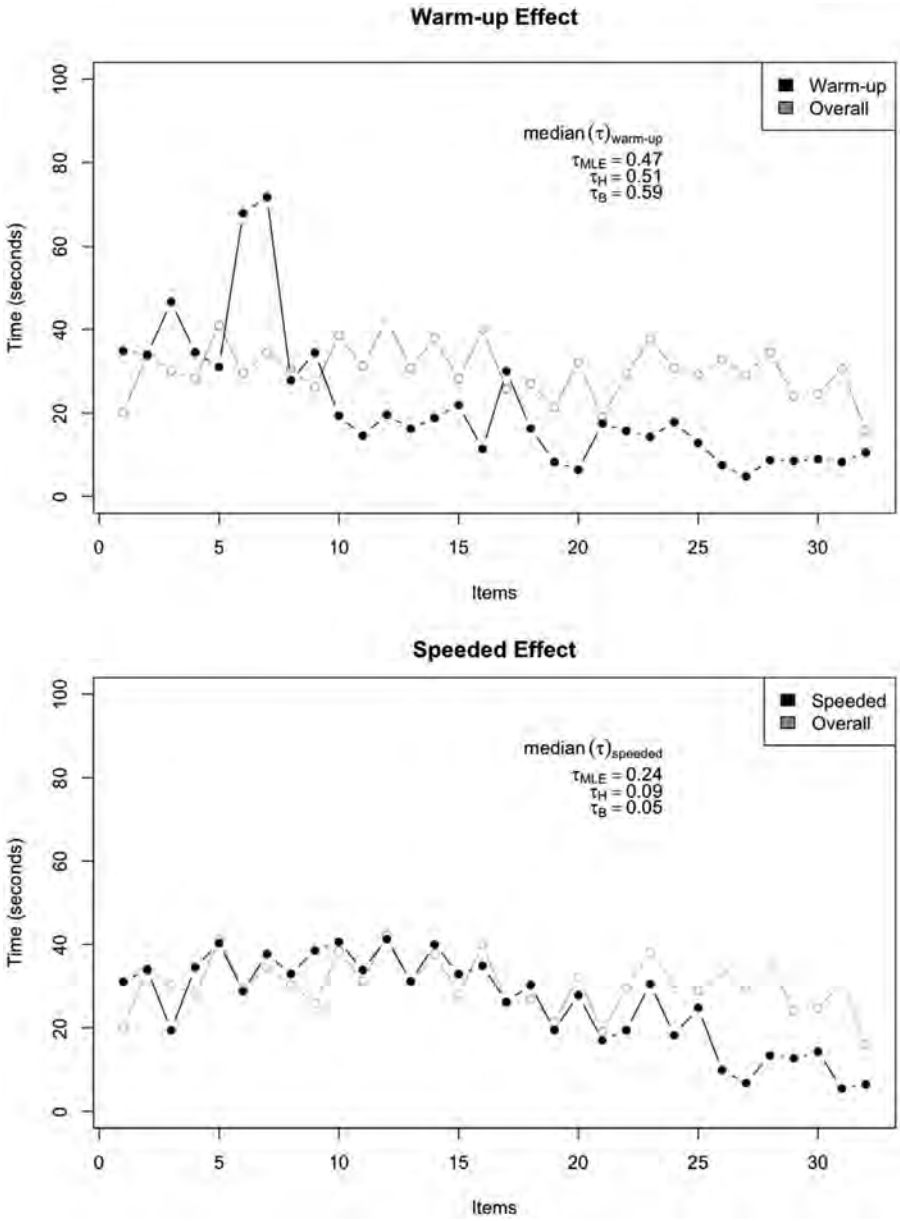

Figure 4. Median response times of warm-up/speeded respondents.

the overall median response time without those warm-up response times. Longer response times in the warm-up group are associated with underestimating τ when using the ML compared to the robust approach. Median estimates of τ were .47, .51, and .59 when using ML, Huber weights, and bisquare weights, respectively. Thus,

these trends parallel results from our simulation studies, where ML tended to be biased downwards when there was a warm-up effect.

Speeded respondents were identified by selecting those whose median total response times (on the last 7 items) were smaller than the 20th percentile, and whose overall $\hat{\tau}_{bisquare} - \hat{\tau}_{MLE} < -.1$. A total of 20 respondents met this criteria. Similarly, in Figure 4, median response times (in seconds) are plotted for the speeded group versus the overall sample without those speeded response times. Smaller response times than overall in the speeded group are associated with overestimating τ when using the ML versus when applying a robust approach. Median estimates of τ were .24, .09, and .05 when using ML, Huber weights, and bisquare weights, respectively. Thus, similarly to our simulation results, the robust estimates correct the upward bias in τ caused by the speeded response times.

Discussion

Statistical models are always approximations of the true model. With real data, a useful model may not be able to capture fully explain the response process for each individual. In other words, some outliers or deviating individuals may exist and do not conform to the assumed model. In such a case, using ML estimates will produce biased estimates. Our simulation and applied analysis suggest that parameter estimates can be improved using M-estimation in the context of response time modeling.

This study proposes robust estimators based on van der Linden's (2006) response time model. The robust estimators were studied under a variety of conditions and evaluated in a simulation study and real data example. The proposed method was shown to reduce bias when estimating the working speed parameter when there is either test speededness or warm-up effect in the data. However, there is a trade-off between bias and efficiency. This study extends previous work in the robust estimation framework designed for item responses to response times (Hong & Cheng, 2019; Schuster & Yuan, 2011). Moreover, our paper suggests that including BW functions for robust estimators would be appropriate for response time modeling with a tuning parameter set to $B = 4$. This finding is different compared to Schuster and Yuan (2011). Schuster and Yuan (2011) demonstrated how BWs had some convergence issues. In our study, we did not encounter the same problem with response time models.

There are several future directions not addressed in the current study. First, we did not consider item order. Aberrant responses can easily manifest themselves in different portions of the assessment Shao, Li, and Cheng (2016). Second, we only proposed robust estimators for the working speed parameter, τ. A limitation of the current study is that we assume there is a small amount of aberrant responses so that the item parameters are close enough to the true values. In practice, this may or may not hold. One could attempt to robustly estimate structural parameters as done in Hong and Cheng (2019) or attempt to perform a cleansing method as described in Patton et al. (2019). Third, robust estimators have been shown to improve other psychometric analysis such as identifying aberrant responders (Sinharay, 2016). One

could further investigate the utility of robust estimators for response time models in this context.

In conclusion, our study provides a framework for researchers who want to analyze response time data with robust methodology. We also provide an R function for applied researchers in the supplementary material.

Appendix A: Estimation of the Log-Normal Model Using JAGS

Bayesian estimation of the log-normal model was implemented in JAGS through the R package `rjags`. To set the scale of the working speed parameter, we chose a normal prior centered around 0 for τ and a multivariate normal prior for the item parameters. Additionally, the variance of τ was a hyperparameter with prior distribution $\Gamma^{-1}(2, 2)$, where the inverse-gamma distribution is a common conjugate prior for variance parameters. The mean and variance of the multivariate normal prior were hyperparameters with hyperprior distributions set as a multivariate normal and inverse-Wishart distribution, respectively. The prior distribution for the mean had mean $\mu_{\alpha,\beta} = (1, 3)$ and variance–covariance matrix equal to the variance–covariance matrix of the distribution for the item parameters. The inverse-Wishart distribution had a variance–covariance matrix

$$\Sigma = \begin{bmatrix} .2 & 0 \\ 0 & .2 \end{bmatrix}$$

and $k = 2$. The chosen priors were set as to represent expected mean values of α and β and as to approximate to the expected variance for such parameters.

References

Bergner, Y., & von Daiver, A. A. (2018). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, (pp. 397–472).

Carroll, R. J., & Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(3). https://doi.org/10.1111/j.2517-6161.1993.tb01934.x.

Cheng, Y., Diao, Q., & Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, *49*. https://doi.org/10.3758/s13428-016-0712-6.

Choe, E. M., Kern, J. L., & Chang, H.-H. (2017). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*(2), 1–14. https://doi.org/10.3102/1076998617723642.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37* (5), 655–670. https://doi.org/10.3102/1076998611422912.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (2006). *Understanding robust and exploratory data analysis*. New York: Wiley-Interscience. https://doi.org/10.2307/2988240.

Hong, M., & Cheng, Y. (2019). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods*, *51*, 573–588.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Retrieved from https://doi.org/citeulike-article-id:2607115.

Huber, P., & Ronchetti, E. M. (2009). *Robust statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc., (2nd ed.). https://doi.org/10.1002/9780470434697.

Kim, D. S., Reise, S. P., & Bentler, P. M. (2018). Identifying aberrant data in structural equation models with IRLS-ADF. *Structural Equation Modeling*, *25*(3), 343–358. https://doi.org/10.1080/10705511.2017.1379881.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(1). https://doi.org/10.1186/s40536-014-0008-1.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6). https://doi.org/10.3102/1076998614559412.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*. https://doi.org/10.1007/BF02294651.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, *42*(3). https://doi.org/10.1177/001316448204200302.

Patton, J. (2014). *Some consequences of response time model misspecification in educational measurement*. (Unpublished doctoral dissertation).

Patton, J., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, *44*(3). https://doi.org/10.3102/1076998618825116.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2). https://doi.org/10.1037/0033-295X.111.2.333.

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

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, *36*(6), 720–735. https://doi.org/10.3102/1076998610396890.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*(4), 1118–1141. https://doi.org/10.1007/s11336-015-9476-7.

Sinharay, S. (2016). The choice of the ability estimate with asymptotically correct standardized person-fit statistics. *British Journal of Mathematical and Statistical Psychology*, *69*(2), 175–193. https://doi.org/10.1111/bmsp.12067.

Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. *Journal of Educational Measurement*, *55*(4), 457–476. https://doi.org/10.1111/jedm.12188.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press. 179–203.

Tyler, D. E. (1984). Robustness and efficiency properties of scatter matrices. *Biometrika*, *70*(2). https://doi.org/10.1093/biomet/71.3.656-a.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2). 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). https://doi.org/10.1007/s11336-006-1478-z.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3). https://doi.org/10.1111/j.1745-3984.2009.00080.x.

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5). https://doi.org/10.1177/0146621609349800.

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68* 251–265. https://doi.org/10.1007/BF02294800.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*. https://doi.org/10.3758/BF03214357.

Wainer, H., & Thissen, D. (2008). Estimating ability with the Wrong model. *Journal of Educational Statistics*, *12*(2). https://doi.org/10.3102/10769986012004339.

Wang, C., Xu, G., & Shang, Z. (2016). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*, 1–32. https://doi.org/10.1007/s11336-016-9525-x.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Cambridge, MA: Academic Press.

Yuan, K. H., & Zhong, X. (2013). Robustness of fit indices to outliers and leverage observations in structural equation modeling. *Psychological Methods*, *18*(2), 121–136. https://doi.org/10.1037/a0031604.

Zhang, Z., & Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: methods and software. *Educational and Psychological Measurement*, *76*(3). https://doi.org/10.1177/0013164415594658.

Authors

MAXWELL HONG is a graduate student, Psychology Department, 390 Corbett Family Hall, University of Notre Dame, Notre Dame, IN 46556; maxwell.hong@gmail.com. His primary research interests include psychometric methods and statistical learning.

DANIELLA A. REBOUÇAS is a graduate student, Psychology Department, 390 Corbett Family Hall, University of Notre Dame, Notre Dame, IN 46556; drebouca@nd.edu. Her primary research interests include psychometric methods and process data.

YING CHENG is a professor, Psychology Department, 390 Corbett Family Hall, University of Notre Dame, Notre Dame, IN 46556; ycheng4@nd.edu. Her primary research interests include psychological and educational measurement.