Behavior Research Methods (2019) 51:573–588
https://doi.org/10.3758/s13428-018-1150-4

Robust maximum marginal likelihood (RMML) estimation for item response theory models

Maxwell R. Hong[1] · Ying Cheng[1]

Published online: 22 October 2018

Abstract

Self-report data are common in psychological and survey research. Unfortunately, many of these samples are plagued with careless responses, due to unmotivated participants. The purpose of this study was to propose and evaluate a robust estimation method to detect careless or unmotivated responders, while leveraging item response theory (IRT) person-fit statistics. First, we outlined a general framework for robust estimation specific for IRT models. Subsequently, we conducted a simulation study covering multiple conditions in order to evaluate the performance of the proposed method. Ultimately, we showed that robust maximum marginal likelihood (RMML) estimation significantly improves detection rates for careless responders and reduces bias in item parameters across conditions. Furthermore, we applied our method to a real data set, to illustrate the utility of the proposed method. Our findings suggest that robust estimation coupled with person-fit statistics offers a powerful procedure to identify careless respondents for further review and to provide more accurate item parameter estimates in the presence of careless responses.

Keywords Item response theory · Careless responses · Person fit · Robust estimation · Robust maximum marginal likelihood

With recent technological advances, researchers have access to large samples that are rapidly collected via online platforms, such as Amazon Mechanical Turk (Sakaluk, 2016). These new rapid data collection methods improve one's ability to efficiently collect large samples, which in turn improves the ability to detect an effect in a more precise way. However, crowdsourcing data collection poses a trade-off when we consider "noisy data," since researchers have little control over the data collection environment (Pauszek, Sztybel, & Gibson, 2017). Noisy data, or data with large amounts of measurement error, hinder the researchers' ability to detect any true effect due to bad quality data (Loken & Gelman, 2017; Maniaci & Rogge, 2014). Oftentimes, measurement error can arise from aberrant responses, in which participants respond independently of item-level content (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Meade & Craig, 2012). The range of inattentive responses in one sample can vary, where researchers have reported proportions of compromised samples from 3.8% up to 50.0% (Berry et al., 1992; Johnson, 2005).

Different types of aberrant responses in the psychological literature have been studied extensively. For instance, aberrant responses can manifest themselves as a form of careless or insufficient-effort response (Curran, 2016; Meade & Craig, 2012). One manifestation of carelessness is random responses, in which a participant endorses every response option available within an item with equal probability—that is, the probability of endorsing each response option follows a discrete uniform distribution (Beach, 1989). However, it is very difficult for a participant to repeatedly produce truly discrete uniform responses across multiple items (Meade & Craig, 2012). Instead, participants may still answer independent of the item-level content by consecutively endorsing, for example, "strongly agree" within a set of responses (Huang et al., 2012). Therefore, in this study we differentiate between two types of careless responses: those in which participants respond *randomly*, versus *systematic* carelessness. In contrast to random responses, systematically carelessness participants respond in a nonrandom fashion, such as by choosing the same response category over multiple items.

Careless responses can lead to many serious consequences in research, such as biased correlations and attenuated or inflated scale reliability (Attali, 2005; Clark, Gironda, & Young, 2003; Wise & DeMars, 2009). Reducing the effect of careless responses is no trivial problem to solve. Some researchers

✉ Ying Cheng
ycheng4@nd.edu

[1] Department of Psychology, University of Notre Dame, Notre Dame, IN, USA

have proposed ways to reduce the harmful effects of careless responses by removing response patterns associated with self-reported insufficient efforts or implausible responses on bogus items (Maniaci & Rogge, 2014; Meade & Craig, 2012). However, these approaches require planning on the researcher's part, to include bogus items or self-reports on response effort. For instance, some researches work with precollected data, and thus may not have the luxury of such preplanned methods for outlier detection. In addition, the efficacy of such approaches is questionable, at best (Meade & Craig, 2012).

Some researchers control for aberrant responses by directly modeling these types of complex processes (Shao, Li, & Cheng, 2016; Wang, Xu, & Shang, 2018; Yamamoto & Everson, 2003). In the psychological literature, others consider careless responses a "special case" in a more general modeling framework of response styles (Böckenholt, 2017; Falk & Cai, 2016; Wetzel & Carstensen, 2017). Although these methods are attractive, directly modeling carelessness or low motivation can be challenging. First, researchers have to assume they know the aberrant mechanism underlying the data. Many different types of carelessness have already been identified, and correctly choosing what type is in a data set would be challenging, because the researcher would never know (Meade & Craig, 2012). Moreover, the general goal when modeling response styles is to gain a better understanding of the cognitive process of item responding. Modeling carelessness or response styles is certainly an interesting line of inquiry, but it is qualitatively different from considering carelessness as a nuisance factor.

Our proposed method belongs to another category—that is, post-hoc procedures to identify careless responses and minimize the impact of careless responders. These methods require no prior planning on the researcher's part, thus providing a much more flexible solution. To this end, researchers have proposed many procedures of this nature, such as the popular Mahalanobis distance (MD) method for outlier detection (Curran, 2016). Unfortunately, the computation of MD is prohibitively intensive for long surveys (Meade & Craig, 2012) in which careless responses are more likely to occur, because each item is essentially one dimension in the calculation. Latent-variable modeling that allows for dimension reduction and model-based outlier detection therefore become desirable (Yuan, Fung, & Reise, 2004b).

Person-fit statistics based on item response theory (IRT) is such an example. Given a chosen IRT model, outliers, or outlying response patterns, can be identified by comparing person-fit statistics to established cutoffs. This approach has been shown in the literature as a viable way to classify careless respondents (Niessen, Meijer, & Tendeiro, 2016). Person-fit statistics have a rich place in the education literature on detecting aberrant responses (e.g., Karabatsos, 2003; Meijer & Sijtsma, 2001), and they have also begun to emerge in psychological research when evaluating scores from psychological scales. For instance, person-fit statistics have been used to refine scales for health outcomes, anxiety and depression, mindfulness, personality, and attitudes (Chien, Shao, & Kuo, 2017; Edelen & Reeve, 2007; Ferrando, 2004; Pallant & Tennant, 2007; Van Dam, Earleywine, & Borders, 2010). Beyond identifying person misfit, person-fit statistics have been used in moderated multiple-regression analysis to evaluate validity estimates (McGrath, Mitchell, Kim, & Hough, 2010). Others have shown that person-fit scores add information when calculating criterion scores in a validation sample (Schmitt, Cortina, & Whitney, 1993). Person-fit statistics also provide more nuance when identifying participant subtypes, such as for screening people at high risk for suicide who would remain undetected otherwise (Conrad et al., 2010).

When carelessness is viewed as a nuisance variable, detecting careless responses is crucial for the research process. Carelessness can lead to unreliable measures, underestimated effect size, and attenuated correlations between the predictor and outcome variables, which can contribute to replication failure (Stanley & Spence, 2014). Recently, a special issue in the *Journal of Experimental Social Psychology* aimed to identify rigorous and replicable methods for replication studies in social psychology. The very first article in the issue highlights the importance of detecting carelessly invalid responses in survey data, for which person-fit statistics are a clear contender (Curran, 2016). As a field, we must first identify powerful methods that can correctly classify careless responders in data.

Most, if not all, person-fit statistics require estimated parameters based on the data. One would normally estimate an IRT model from the full sample, which would include both normal and careless responders. Researchers can then use the estimated item and person parameters to calculate person-fit statistics to classify careless responders. However, the existence of a nonignorable portion of outliers could bias the model parameter estimates (Kim & Moses, 2016; Meijer & Sijtsma, 2001; Oshima, 1994) to the extent of "masking" outliers—that is, reducing the probability of detecting the outliers. This phenomenon is well known in the literature of outlier detection as the "masking effect" (Yuan et al., 2004b; Yuan & Zhong, 2008).

Removing careless respondents can potentially overcome the "masking effect" in outlier detection and some researchers have even proposed to do so iteratively to achieve the best results (Cheng & Patton, 2014). However, removing data can be a very controversial practice. In addition, removing outliers could potentially remove participants who exhibit only partial carelessness—that is, only some of his or her responses are affected by carelessness. By removing all of a participant's responses, we may lose useful information. Given these considerations, we propose a flexible robust estimation framework based on IRT to improve item parameter estimation, which in turn improves the detection of careless respondents. This method is flexible in that different weighting mechanisms can be used in the same general framework. In this study the weights are derived from the normalized *p* value of Sinharay's (2016a) person-fit statistic l_z^*, but it does not necessarily have to be so. There is a large and diverse set of person-fit statistics to choose from. Moreover, our parameter estimation procedure is demonstrated through the simple unidimensional IRT model case. However, the framework is by no means limited to the unidimensional case. It can also work with multidimensional IRT models or be applied to each individual subscale in a test battery (Felt, Castaneda, Tiemensma, & Depaoli, 2017).

The rest of our article will be organized as follows. We will provide a review of the literature pertaining to careless responses and person-fit statistics. Then, we will describe a robust procedure to estimate IRT models given response data contaminated by careless responses. By obtaining more accurate IRT model parameters we will demonstrate that we will be able to achieve better classification rates (of careless vs. noncareless response patterns) via a simulation study. Finally, we apply our methodology to the Feelings Scale in the National Longitudinal Study of Adolescent Health (AddHealth). We will conclude with a discussion and future directions.

Method

Carelessness is one type of content-independent behavior—that is, participants responding independently of an item's content. As we discussed earlier, carelessness under that definition can manifest itself in several ways. For instance, for a *K*-point Likert-scale item, participants can endorse any of the *K* response categories with equal probability. This would constitute a random response. Meanwhile, carelessness may also manifest itself as having invariant responses to multiple items consecutively on the survey. This type of response behavior ties into our definition of content-independent responses, yet the careless response itself is not necessarily random.

In this study we focused on carelessness that does create additional randomness in item responses, which is less straightforward to detect (Meade & Craig, 2012). The most common, or "off-the-shelf," statistic for the detection of outliers is the well-known Mahalanobis distance,

$$D(\boldsymbol{u}_i) = \sqrt{\left(\boldsymbol{u}_i - \overline{\boldsymbol{u}}\right)' \boldsymbol{S}_u^{-1} \left(\boldsymbol{u}_i - \overline{\boldsymbol{u}}\right)},$$

where the MD for the response vector $\boldsymbol{u}_i$, $D(\boldsymbol{u}_i)$, is evaluated on the basis of a general distance metric between a response vector and average item scores, for person *i* (Curran, 2016). Assuming that $\boldsymbol{u}_i$ is multivariate normally distributed with average item scores $\overline{\boldsymbol{u}}$ and the covariance matrix $\boldsymbol{S}_u$, the null distribution for $D(\boldsymbol{u}_i)$ should follow a central chi-square distribution with the degrees of freedom equal to the number of items. Participants whose $D(\boldsymbol{u}_i)$ is larger than a critical value from the central chi-square distribution are considered aberrant and can be flagged at a nominal Type I error level. Unfortunately, the computation of MD is prohibitively intensive for long surveys (Meade & Craig, 2012). Methods that incorporate scale dimension reduction therefore become more attractive.

For many questionnaires that have items based on a Likert rating scale, participants rate an item on a fixed number of response options such as from *Strongly Disagree* to *Strongly Agree*. Item responses are typically coded as values one up to the total number of response options. Arguably, the most widely used IRT model for such item response data is the graded response model (GRM; Samejima, 1969). Under a GRM, each item subsumes $K - 1$ boundary functions specified between item steps for *K* response options. The GRM represents the probability of responding either above or below the boundary as a function of latent ability, θ. The probability of endorsing the score category *k* or above can be characterized by a simple two-parameter logistic model (Birnbaum, 1968):

$$P_{jk}^*(\theta) = \frac{1}{1 + \exp\left[-Da_j\left(\theta - b_{jk}\right)\right]}. \tag{1}$$

Note that *D* is a scaling constant often fixed to 1.7, so that item parameters are approximately put on the same scale as in a normal ogive model. Here, b_{jk} is the location parameter for the boundary that separates the (k – 1)th and *k*th response categories of item *j*, and a_j is the discrimination parameter for item *j* for all boundary functions within item *j*. The probability of endorsing response option *k* can be expressed by taking the difference between adjacent boundary functions:

$$P_{jk}(\theta) = P_{jk}^*(\theta) - P_{j(k+1)}^*(\theta). \tag{2}$$

The probabilities of responding below the first and below the highest option are set to 0.0 and 1.0, respectively. This model subsumes the widely used two-parameter logistic (2PL) model for dichotomously scored items—that is, K = 2. It further subsumes the one-parameter logistic (1PL) model, which sets all the a_j to be equal across items.

For content-independent responses, the GRM would not be able to capture them. Person-fit statistics provide researchers a statistical method to decide whether or not a participant's response vector is aberrant, in the sense that it cannot be sufficiently explained by the IRT model. Careless responses, as one type of aberrant response, can therefore possibly be detected by person-fit statistics. A commonly used person-fit statistic is the standardized log-likelihood person-fit index l_z (Drasgow, Levine, & Williams, 1985; Van Krimpen-Stoop &

Meijer, 2002). Denote the item parameter matrix as γ, where each column $\gamma_j = (a_j, b_{j1}, b_{j2}, \ldots, b_{j(K-1)})'$. The likelihood function for the ith respondent's response vector $\boldsymbol{u}_i$ can be written as

$$L_i(\boldsymbol{u}_i|\theta;\gamma) = \prod_{j=1}^{J}\prod_{k=1}^{K} P_{ijk}(\theta)^{\delta_{ijk}}, \tag{3}$$

where δ_{ijk} is an indicator function that the ith respondent endorses the kth response category of the jth item:

$$\delta_{ijk} = \begin{cases} 1, \text{if } u_{ij} = k, \\ 0, \text{otherwise} \end{cases}, \tag{4}$$

where u_{ij} is the response of the ith respondent to the jth item. The standardized person-fit statistic l_z for the ith respondent takes the following form:

$$l_{zi} = \frac{l_i - E(l_i)}{\sqrt{\mathrm{Var}(l_i)}}, \tag{5}$$

where l_i is the logarithm of the likelihood function of the GRM for person i, $E(l_i)$ is the expectation of l_i, and $\mathrm{Var}(l_i)$ is the corresponding variance. Respondents who respond carelessly should exhibit large negative residuals of the likelihood—that is, a large negative l_{zi}. The standardized likelihood function should asymptotically follow a standard normal distribution under the null hypothesis of person-model fit. It is therefore common practice to flag response patterns with $l_{zi} < -1.64$, which corresponds to a nominal Type I error of 5%.

A well-known issue to this strategy is that the computation of the l_z statistic requires the true latent trait value. In practice, this value is never known, so it is estimated on the basis of the item responses. Estimated latent trait estimates would shrink the variance of l_z, so the statistic would not follow a standard normal distribution. In consequence, the Type I error rate has been shown to be over-conservative. To overcome this issue, researchers have proposed corrections to the original l_z statistic, such as Snijders's (2001) l^*_{z-d} for dichotomous items, and Sinharay's (2016a) correction for polytomous items, l^*_{z-p}. In this study we will consider l^*_{z-p} because we are dealing with polytomous items, or l^*_z for short. Appendix A includes the formulas required to compute l^*_z.

Using the proposed correction addresses the uncertainty introduced by the latent trait estimate. However, the person-fit statistic still relies on estimated item parameter, which are biased due to careless responders in the sample. The magnitude of the bias would depend on the manifestation of carelessness and proportion of careless responses. It is noteworthy that the biased item parameter estimates would also lead to bias in latent trait estimates (Oshima, 1994). Together the biases in the structural and person parameters will distort the l^*_z, and hence the "masking effect" may result.

In the literature, some have suggested that responses should be removed once they are identified as aberrant (e.g., Curran, 2016; Niessen, Meijer, & Tendeiro, 2016). However, data removal can be controversial. Such controversy has been well documented in the context of the detection and treatment of outliers (e.g., Orr, Sackett, & Dubois, 1991). Furthermore, complete removal of a participant's response vector may not be ideal when a careless responder exhibits partial carelessness—that is, he or she responds carelessly to some but not all items. Partial carelessness can manifest in many forms, but the most commonly observed form is called "back random responding" (BRR), meaning that careless responses occur in the later part of a questionnaire, due to cognitive fatigue or boredom or other reasons (Clark et al., 2003). Full removal would cause loss of information when a participant exhibits partial carelessness.

We propose a robust estimation procedure that addresses the masking effect while overcoming the two limitations mentioned. This was done by weighting response patterns according to their fit in the process of item parameter estimation; that is, misfitting patterns were be down-weighted. Note that misfitting patterns were not removed anywhere in the analysis. Meanwhile, partial carelessness was taken into account because a response pattern with a higher proportion of careless responses would be expected to result in a smaller weight, whereas a response pattern with a smaller proportion of careless responses should receive a higher weight. A normal response pattern would be expected to receive close to full weight. Next, we introduce the details of the robust estimation procedure.

Robust estimation of GRM through robust maximum marginal likelihood

Robust estimation has been previously examined in other measurement models, such as confirmatory factor analysis (Yuan & Zhong, 2008, 2013). It has also been previously applied in different applications of IRT, such as scale equating and latent trait estimation (e.g., Bejar & Wingersky, 1981; Linn, Levine, Hastings, & Wardrop, 1980; Schuster & Yuan, 2011; Stocking & Lord, 1983). However, these applications in IRT have concerned themselves with robust estimation of incidental parameters—that is, the latent traits of individuals—instead of with robust estimation of structural parameters—that is, the item parameters.

There are at least two reasons why we chose to focus on robust estimation of the item parameters. First of all, without accurate item parameter estimates, it would be difficult to accurately detect response patterns affected by carelessness, due to the masking effect.

Previous studies have investigated the utility of robust estimation for the latent trait estimator when conducting outlier detection with person fit (Sinharay, 2016b). Researchers found that using robust estimation led to better detection rates for aberrant responders than for nonrobust estimates, at the price of inflated Type I error rates. The inflation became negligible for very long tests. We therefore expect to be able to improve the detection of carelessness with robust estimation of item parameters under certain conditions. Second, obtaining accurate item parameters by itself is important. Item parameters are often treated as if they are known and without error when used in many IRT applications, such as latent trait estimation (see Baker & Kim, 2004; Cheng & Yuan, 2010), scale linking and equating (van der Linden & Barrett, 2016), and computerized adaptive testing (Patton, Cheng, Yuan, & Diao, 2013). If they are not estimated accurately, using item parameter estimates in those applications can cause many undesirable consequences (Cheng, Liu, & Behrens, 2015; Patz & Junker, 1999; Tsutakawa & Johnson, 1990).

Robust procedures typically involve assigning a weight to each individual case "according to its distance from the center of the majority of the data" (Yuan & Zhang, 2012). For our purposes when estimating item parameters, robust estimation entails assigning a weight to each respondent when applying the common maximum marginal likelihood estimation method. More specifically, it entails assigning a weight to each respondent when computing the overall marginal likelihood for the GRM model (Eqs. 1 and 2), using the expectation maximization (EM) algorithm proposed in Bock and Aitkin (1981). Assuming that $\theta \sim f(\theta)$, the marginal probability of observing the item response vector $\boldsymbol{u}_i$ can be written as

$$P(\boldsymbol{u}_i|\gamma) = \int L_i\Big(\theta|\boldsymbol{u}_i, \gamma\Big) f(\theta) d\theta, \quad (6)$$

where $f(\theta)$ is often assumed to be the PDF of the standard normal distribution. The overall marginal likelihood of observing the entire response data matrix is therefore

$$P(\boldsymbol{u}|\gamma) = \prod_{i=1}^{N} P(\boldsymbol{u}_i|\gamma). \quad (7)$$

The MML procedure seeks to computationally find the $\hat{\gamma}$ that maximizes $P(\boldsymbol{u}|\gamma)$ after the data are observed, but the observed data are regarded as a random sample from a population. The persons are considered a random effect, whereas "the items are still considered fixed" (de Ayala, 2009, p. 69). This leads to maximizing over the log-likelihood of the response data in MML:

$$\log P(\boldsymbol{u}) = \sum_{i=1}^{N} \log P(\boldsymbol{u}_i|\gamma). \quad (8)$$

Estimation based on Eq. 8 implicitly assumes an equal weight for each respondent. We can modify Eq. 8 by adding a weight, w_i, in order to weight individuals differentially given the plausibility of their response patterns:

$$\log P(\boldsymbol{u}) = \sum_{i=1}^{N} w_i \log P(\boldsymbol{u}_i|\gamma). \quad (9)$$

The EM algorithm proceeds in the same manner as usual, with just a modified log-likelihood function (Baker & Kim, 2004, pp. 285–291). Weighting the likelihood function based on groups of individuals is similar to the so-called *pesudolikelihood approach* for complex survey data (Thomas & Cyr, 2002). Note that if we remove participants from the data set, we are essentially fixing their weights to be 0.

By incorporating different weighting mechanisms into the likelihood function, we can devise different robust estimation procedures. Weighting the likelihood function by some weight function follows the general form of M-estimation in the robust literature for general linear models (Carroll & Pederson, 1993; Künsch, Stefanski, & Carroll, 1989). For an accessible introduction to the properties of general robust estimation, we refer readers to chapter 2 in Wilcox (2016). In general, robust estimation requires two parts: a residual, r, and weight function, $w(.)$. The residual requires one to specify which observations will be considered inconsistent. Fortunately, l_z^* functions as a standardized residual that we used in this study. Using a standardized residual is similar to the approach used in Yang and Yuan (2016). Next, we needed to choose a weight function. One possibility was to choose common weight functions used in the robust statistical literature, such as Huber weights or Tukey biweights (Beaton & Tukey, 1974; Huber, 1964). However, prior work on a 2PL model showed that these common weights in the robust literature have undesirable properties for IRT model estimation (Cheng & Patton, 2014). Huber-type weights have been recommended in some previous research (e.g., Yuan, Bentler, & Chan, 2004a; Yuan & Zhang, 2012) on robust estimation of SEM when the majority of the data are normally distributed but some moderate outlying cases exist (not sitting too far away from the center). In our case, the Huber-type weights might not work well, because with item response data (essentially ordinal data) we can hardly expect the majority of the data to be normally distributed. Meanwhile, the type of "outliers" expected would not lead to fatter tails than a normal distribution, but would result from model misspecification: for the inattentive respondents, their responses would no longer follow the usual IRT model. Hence, weights based on multivariate t distributions were not suitable.

Given the wide use of person-fit statistics in detecting response aberrance, the previous success of robust estimation of the latent trait estimator when conducting outlier detection with person fit (Sinharay, 2016b), and the superiority of l_z^* over l_z, we proposed a weighting scheme based on l_z^*. Note that l_z^* asymptotically follows a standard normal distribution; therefore, we proposed to set

$$w_i = \frac{\Phi(l_{zi}^*)}{\sum_i \Phi(l_{zi}^*)}, \tag{10}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, and l_{zi}^* is the l_z^* statistic computed for the ith respondent. This essentially sets w_i equal to the normalized p value of the l_z^* statistic under a one-sided hypothesis-testing framework. The following discussion will be based on such a weight, although this will not prevent other researchers from choosing a different person-fit statistic or weight function.

By setting the weight to be equal to the normalized p value of the l_z^* statistic, we expected to give smaller weights to aberrant responses, such as careless responses, while maintaining larger weights for normal response patterns. We provide a brief illustration to show the utility of the proposed weight function. Item responses were simulated from a GRM with six response categories. The discrimination parameter was sampled from a uniform distribution with a lower bound of .5 and upper bound of 2. The location parameters were also sampled from a uniform distribution with a lower bound of – 2 and upper bound of 2. To avoid the confounding effect of sparse categories, we resampled location parameters if the difference between adjacent categories was smaller than .3, a similar design to those in previous studies (Jiang, Wang, & Weiss, 2016). We generated 1,000 latent traits drawn from a standard normal distribution for a 50-item test. However, 100 participants gave random responses to 25 of the items, randomly chosen from the test. Carelessness was simulated by generating responses from the discrete uniform distribution U[0, 5]. These participants were considered the aberrant responders. Data generation and analysis were conducted in R (R Development Core Team, 2017).

In Fig. 1, the left panel plots a respondent's l_z^* statistic against the true latent ability. Note that if participants respond aberrantly—in this case, for 50% of a survey—their l_z^* statistic is more negative than the normal responder's counterpart. The line drawn in the graph shows when participants are typically flagged in studies, at a cutoff value of – 1.64, given that the asymptotic distribution of l_z^* is standard normal. The right panel shows the corresponding weights if one integrates the normal distribution from $-\infty$ to q, where q is the l_z^* statistic, as compared to the corresponding l_z^* statistic. This integrand is the associated p value for each l_z^* statistic. Note that the weights in this figure are rescaled from Eq. 9 by multiplying each weight by N, so that the weights sum to the total sample size. We can observe that aberrant responders are now given a smaller weight than normal responders. Again, the rationale behind applying a gradient weight instead of data removal (which is essentially assigning a weight of 0 for aberrant response patterns) is to take into account the severity, or proportion of careless responses within a single participant. If we plugged in these weights from our given sample into Eq. 9, we conjectured that our robust procedure would produce less-biased parameter estimates than the unweighted (or equal-weight) estimation procedure. As compared to the method of complete data removal, we expected our method to yield smaller standard errors, because we would still glean valuable information from partial valid responses and did not decrease our sample size, neither of which would be true for data removal.

Our robust method can then be described as a two-stage approach to estimating a given IRT model. First, we estimate the item parameters from the full data, which would include equal weights for all participants. Next, we calculate the person-fit statistic, l_z^*, for each participant, given the current estimated item parameters from the first step. Finally, we use normalized l_z^* 's p value as the weights to plug into Eq. 9 and reestimate the model. Smaller p values would indicate a greater severity of response aberrance, and the corresponding response pattern would be assigned a smaller weight. Such a weighting mechanism would allow us to form a gradient differentiating between partial and complete carelessness. The item parameter estimates should be less biased, which would in turn improve the detection of careless responses. A summary for our RMML estimation procedure is provided in Algorithm 1.

Note that this general algorithm can be adapted in multiple ways. For example, one can devise different weighting mechanisms and incorporate them in Step 2. In addition, Steps 2–4 can be iterative. That is, person-fit statistics, weights, and item parameters can be iteratively updated until some convergence criteria are met. Here we only consider one iteration; that is, the item parameters are updated only once. If the end goal is only to obtain item parameters, then one may remove Step 5.

To evaluate the performance of the proposed robust procedure, we conducted a simulation study by generating a data set containing both regular and careless responses. The accuracy and precision of parameter estimations were examined. The power and Type I error of the proposed procedure in detecting carelessness were also investigated. It is important to note that our method should accommodate various forms of aberrant response patterns. Carelessness, given its prevalence in low-stakes testing, is one example we are most concerned with.

Simulation study

Item and person responses were simulated in the same way as in the data generation scheme previously described, with minor changes. For accurate parameter recovery of the GRM,

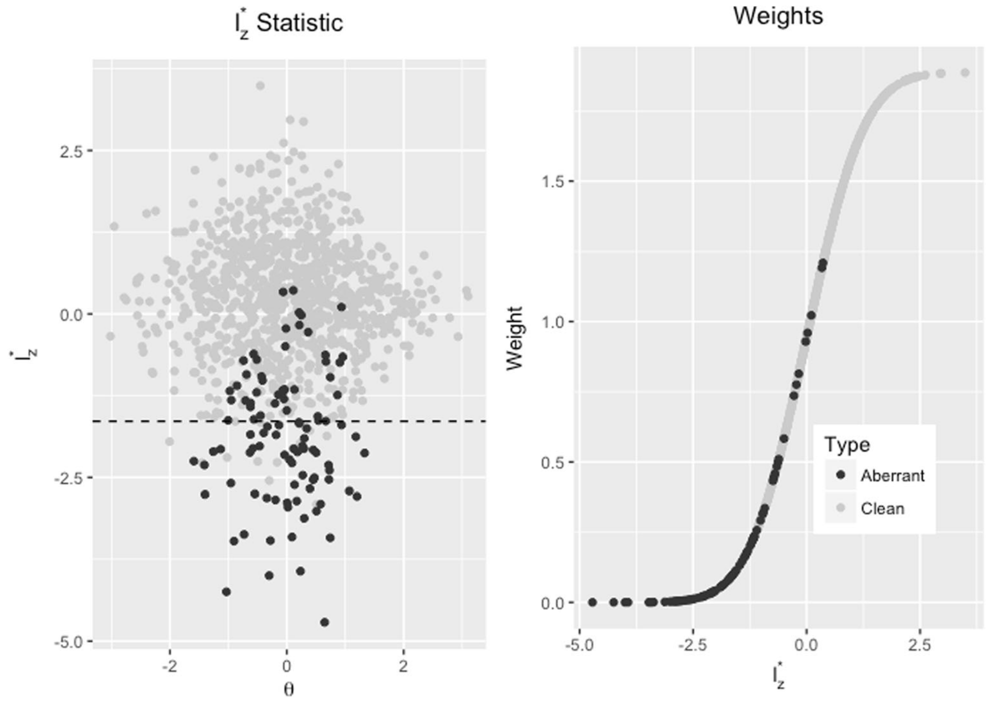

Fig. 1 l_z^* statistics and weight values for aberrant and clean responses. Note that the weights are rescaled so they sum to the total sample size

previous literature had suggested that the minimum required sample size ranges from 375 to 750, depending on the length of the scale (de Ayala, 1994) as well as the number of response categories per item. This is true when the responses follow the GRM. With misfitting response patterns in the data, it is unclear what would be an adequate sample size. Hence, we included two sample sizes in our study, 500 or 3,000, with the former being within the recommended sample size range, and the latter being way more than an adequate size in traditional settings.

Meade and Craig (2012) suggested that the efficacy of detection methods for carelessness depends strongly on the nature of the response data. Following their study, we simulated both *random* and *midpoint* carelessness. Random carelessness was simulated by generating responses from the discrete uniform distribution U[0, 5]. Each response option had an equal chance of endorsement, 1/6. Midpoint carelessness was simulated in which participants were more likely to select the middle categories, in our case Categories 2 and 3. This was done by drawing from a binomial distribution in which the number of categories was equal to five and the probability of success was fixed to .5. In this way, the expected probability of endorsing 0 or 5 was roughly 3%, that of endorsing 1 or 4 was around 16%, and that of endorsing 2 or 3 was around 31%.

It is also important to consider the prevalence and severity of careless responses in the response data, because a large amount of careless responses can cause the "masking effect." *Prevalence* refers to the proportion of participants who respond carelessly to some or all of the items. *Severity* refers to the proportion of items affected by carelessness within a response vector. Together, they indicate the proportion of responses in an $N \times J$ data matrix affected by carelessness. In this study, prevalence in the total sample was set at either 10% or 30%, with the former being consistent with previous research on prevalence rates, which reported prevalence rates of 5%–20% (Curran, Kotrba, & Denison, 2010) and 11% (Meade & Craig, 2012); we chose the latter rate as a more extreme rate to challenge our proposed procedure. With respect to severity, in past research people had self-reported responding carelessly to 50% or more of items on long surveys that involved multiple subscales (Baer, Ballenger, Berru, & Wetter, 1997; Berry et al., 1992). Given that our study focused on shorter, unidimensional measures, we simulated the severity rate to be either 30% or 50%. In other words, for a careless participant, we simulated 30% or 50% of his or her responses to be careless, distributed intermittently throughout the survey. We also varied the test length to be either 30 or 50.

In sum, we had a total of 32 conditions for data generation, with 100 replications for each condition and five factors fully crossed (2 test lengths × 2 sample sizes × 2 types of careless responses × 2 prevalence rates × 2 severity rates). For each data generation condition, we compared the performance of three procedures that involved the use of l_z^* in terms of the detection of careless respondents and quality of item parameter estimates. The first procedure (denoted as *full-sample* l_z^*) used the full sample to both estimate the GRM and identify participants who aberrantly responded, using l_z^* based on the estimated parameters. Participants whose l_z^* was smaller than – 1.64 were flagged as aberrant. The second procedure (denoted *data removal*) involved data removal—that is, participants whose l_z^* was smaller than – 1.64 were removed. Then the model was reestimated and participants' l_z^* values were updated with the new parameter estimates. Data removal implies that during model reestimation, each unflagged participant receives a weight of 1 and each flagged participant receives a weight of 0, based on Eq. 9. Those with an updated l_z^* smaller than – 1.64 were flagged as aberrant (data removal). The third procedure (denoted *robust estimation*) was our RMML estimation procedure detailed in the previous sections. All programming was done using the R language. For the RMML procedure, the number of quadrature points during the E step was fixed to 60. Maximization over the log-likelihood was done using a Newton–Raphson procedure during the M step. Convergence was established if the log-likelihood between adjacent iterations did not change beyond .0001.

We evaluated the power and Type I errors of the three procedures in identifying careless respondents. Power was defined as the proportion of flagged participants over the total number of participants affected by carelessness. Type I error was defined as the proportion of participants unaffected by carelessness who were flagged. For ability estimation, we used maximum likelihood estimation procedures. When participants endorsed 0 for all items or obtained a full score, we fixed their ability estimates to be – 3 and 3, respectively. We compared these three procedures in power and Type I error against two scenarios: (a) the baseline scenario, in which item parameters were known and l_z^* was used to flag participants on the basis of the true parameter values; and (b) when MD was used for outlier detection on all the data. This was the same procedure as the full-sample l_z^* procedure, except that MD was used for flagging participants instead of l_z^*. The motivation for including the full-sample MD conditions was the ubiquity of MD in the outlier detection literature. The known item parameters condition served as an ideal case that would never hold in practice.

We evaluated the average bias of the location parameter estimates, $\sum_{j=1}^{J}\left(\widehat{b_{jk}}-b_{jk}\right)/J$, and the root mean square error (RMSE), $\left(\sum_{j=1}^{J}\left(\widehat{b_{jk}}-b_{jk}\right)^{2}/J\right)^{-.5}$, for $k = 1, 2, \ldots, (K-1)$. For the discrimination parameter, we evaluated the parameter using conditional bias and RMSE, where we first binned the true discrimination parameters into six equally spaced intervals between .5 and 2, and then evaluated the parameter bias and RMSE. The average parameter bias and RMSE were also evaluated across only the three estimation procedures using l_z^*. Bias, RMSE, and classification accuracy were averaged across the 100 replications.

Results

Preliminary results showed little difference between the sample sizes of 500 and 3,000 and the test lengths. We therefore only present results for the sample size of 500 and a test length of 50. Tables 1 and 2 show the Type I error, and average power for the respective conditions.

The full-sample condition using either l_z^* or MD is the common practice in reality. Both the data removal method and the robust estimation method try to take inaccurate item parameter estimates into account when detecting careless responses. Between the two methods, data removal and robust estimation, the former is a special case of the latter in which the flagged participants are given a weight of 0. The known parameter condition is an ideal condition. Evidently, power does not appear to be high even in the ideal situation, especially when the prevalence rate is low. This is consistent with the existing literature, which has reported traditional methods with power well below .5 in classifying random responders (Huang et al., 2012; Marjanovic, Holden, Struthers, Cribbie, & Greenglass, 2015), even when multiple indices are employed, as was suggested by Meade and Craig (2012). Our simulation conditions particularly challenged the detection methods, because we simulated partial carelessness (in contrast to careless responses on every item) and midpoint carelessness, both of which have been shown in previous studies to reduce power (Cheng & Patton, 2014; Meade & Craig, 2012). In Table 2, we also observe that the power is uniformly higher for detecting random careless responses than for midpoint carelessness.

As compared to the known item parameter condition, the full-sample condition leads to an evident drop of power when using either l_z^* or MD. The data removal method does help improve the power slightly. Data removal and the full-sample conditions result in conservative Type I error rates. On the other hand, the robust estimation procedure increased power across conditions as compared to the baseline and other estimation methods. However, we also find an inflated Type I error in the robust condition as compared to all other methods. Upon discussion, we did not evaluate procedures based on MD any further, due to its poor performance in detecting simulated careless respondents as compared to methods that utilized l_z^*.

Figures 2 and 3 present the empirical bias and RMSE for the item location and discrimination parameters for the three procedures that involve l_z^* (full-sample l_z^*, data removal, and robust estimation). For item location parameters, a clear trend emerges when we look across groups. The largest and smallest

Table 1 Type I errors for aberrant and normal response patterns

Aberrant Type	Sev.	Prev.	Full-Sample l_z^*	Data Removal	Robust Estimation	Known Item Parameters	Full-Sample MD
Random	10%	30%	.025	.036	.115	.036	.026
Random	10%	50%	.018	.034	.111	.035	.021
Random	30%	30%	.011	.019	.078	.035	.011
Random	30%	50%	.005	.014	.067	.034	.007
Midpoint	10%	30%	.030	.039	.123	.037	.038
Midpoint	10%	50%	.028	.039	.120	.038	.039
Midpoint	30%	30%	.021	.029	.098	.036	.036
Midpoint	30%	50%	.019	.028	.096	.035	.043

Sev. = severity, Prev. = prevalence. The columns l_z^*, MD, and Known Item Parameters are based on the full data set

location parameters produce the most biased estimates across replications. Data removal reduces the bias in many scenarios. However, robust estimation either outperforms or operates at the same level as the data removal procedure. A different pattern emerges when we look at RMSE. Data removal reduces the RMSE when compared to the full estimation. The robust estimation procedure has a slightly larger RMSE than data removal across conditions, and it even exceeds the RMSE for the full-data sample occasionally. This perhaps could explain the inflated Type I error rates for the robust estimation procedure.

For the conditional bias and RMSE of the discrimination parameters, the full-sample approach creates a (negative) bias for the discrimination parameters; the magnitude of the bias increases when the true discrimination parameter increases. The bias and RMSE also increase with greater severity and prevalence of careless responses. Data removal improves the estimation for larger discrimination values when we look at both RMSE and bias. However, robust estimation outperforms both full-sample estimation and data removal across the majority of the conditions, whether we look at RMSE or bias.

Taken together, the results from the simulation study suggest that the robust estimation method improves item parameter estimation in the presence of careless responses, even when the prevalence and severity rates are high. It also improves the power for detecting careless responses. However, the improvement in power seems to come at the cost of an inflated Type I error rate. This suggests that we need to be careful with the next steps once we flag participants. How to handle suspicious responses is an evolving practice, and our goal is not to provide strict guidelines for practitioners. For instance, Allalouf, Gutentag, and Baumer (2017) recommended that following statistical quality control procedures, a human review of suspicious cases should be conducted. Other precautionary steps may be taken, as outlined in Wainer (2014), for suspicious responses. We argue that robust estimation's high power makes it a powerful tool for data-driven methods when identifying suspicious responses. We strongly caution against data removal on the sole basis of results from statistical analysis in testing.

To further exemplify our key findings, we built upon our simulation study with an empirical analysis to illustrate the robust estimation procedure.

Empirical analysis

We applied both the data removal and robust procedures to data from the National Longitudinal Study of Adolescent

Table 2 Power for aberrant and normal response patterns

Aberrant Type	Sev.	Prev.	Full-Sample l_z^*	Data Removal	Robust Estimation	Known Item Parameters	Full-Sample MD
Random	10%	30%	.332	.389	.626	.408	.189
Random	10%	50%	.607	.700	.857	.729	.354
Random	30%	30%	.223	.293	.539	.415	.117
Random	30%	50%	.390	.562	.787	.726	.183
Midpoint	10%	30%	.169	.201	.405	.213	.022
Midpoint	10%	50%	.278	.342	.570	.381	.015
Midpoint	30%	30%	.120	.150	.340	.216	.016
Midpoint	30%	50%	.158	.208	.422	.388	.014

Sev. = severity, Prev. = prevalence. The columns l_z^*, MD, and Known Item Parameters are based on the full data set

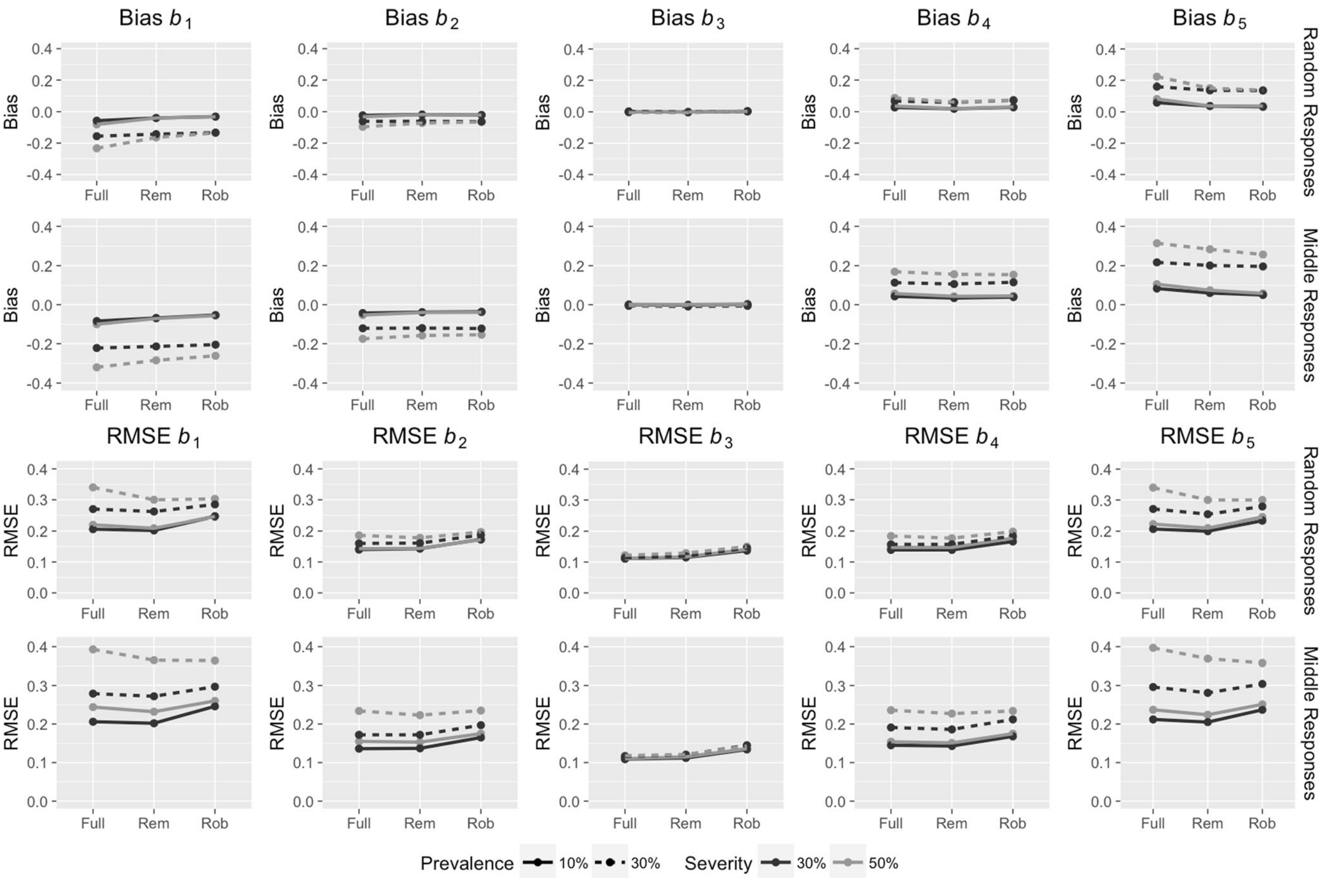

Fig. 2 Bias and RMSE for the location parameters for the three estimation procedures based on l_z^* (Full = full data, Rem = data removal; Rob = robust estimation). The top two rows show estimated bias, and the bottom two rows show RMSE. Solid and dashed lines represent different prevalence rates, and gray and black colors represent different severity rates

Health (AddHealth), which is an education- and health-related study of adolescents in grades 7–12 (Harris & Udry, 2010). We restricted our analysis to the base year, which was an in-home questionnaire administered during 1994–1995. We analyzed the Feelings scale, which collects information about the respondents' current emotional state. The scale contains 19 items to which participants responded on a Likert scale with four response categories, describing how often each of the statements applied to the past week (e.g., "You were bothered by things that usually don't bother you"). However, participants rarely endorsed Category 4, "most of the time or all of the time." Therefore, we collapsed this category with Response Category 3. This approach mirrors similar studies using this scale, in which a unidimensional IRT model was fit to the data and found to have adequate fit (Edelen & Reeve, 2007).

The original sample contains 6,504 respondents, but respondents with any missing data in the Feelings scale were removed, leaving 6,457 respondents. Using this final sample, item parameter estimates were obtained without cleansing (i.e., using all 6,457 participants), with data removal based on l_z^*, and with the robust estimation method. We did not utilize cleaning based on MD, due to its suboptimal performance relative to the procedures that used l_z^* in our simulation conditions. We evaluated all three methods on the basis of classification of respondents and differences among the resulting item parameter estimates and their standard errors. Without any cleansing, 7.6% of the sample was flagged. With data removal, 11.18% of the sample was flagged, and with robust estimation, 20.59% of the sample was flagged. This is consistent with our simulation studies, in which using the entire data set resulted in reduced power and an overly conservative Type I error rate. According to the simulation study, using data removal increases power, and here we saw that data removal flagged 3.5% more participants than without any cleansing. Robust estimation flagged an additional 13% of the total sample, as compared to using the nonweighted full sample.

Item parameter estimates based on the full sample, data removal, and robust cleansed samples, with their standard errors, are presented in Table 3. Data removal resulted in little change or in slightly larger discrimination parameter estimates. For Item 1, the full sample yielded an estimate of 1.42, whereas data removal yielded 1.48, and the robust

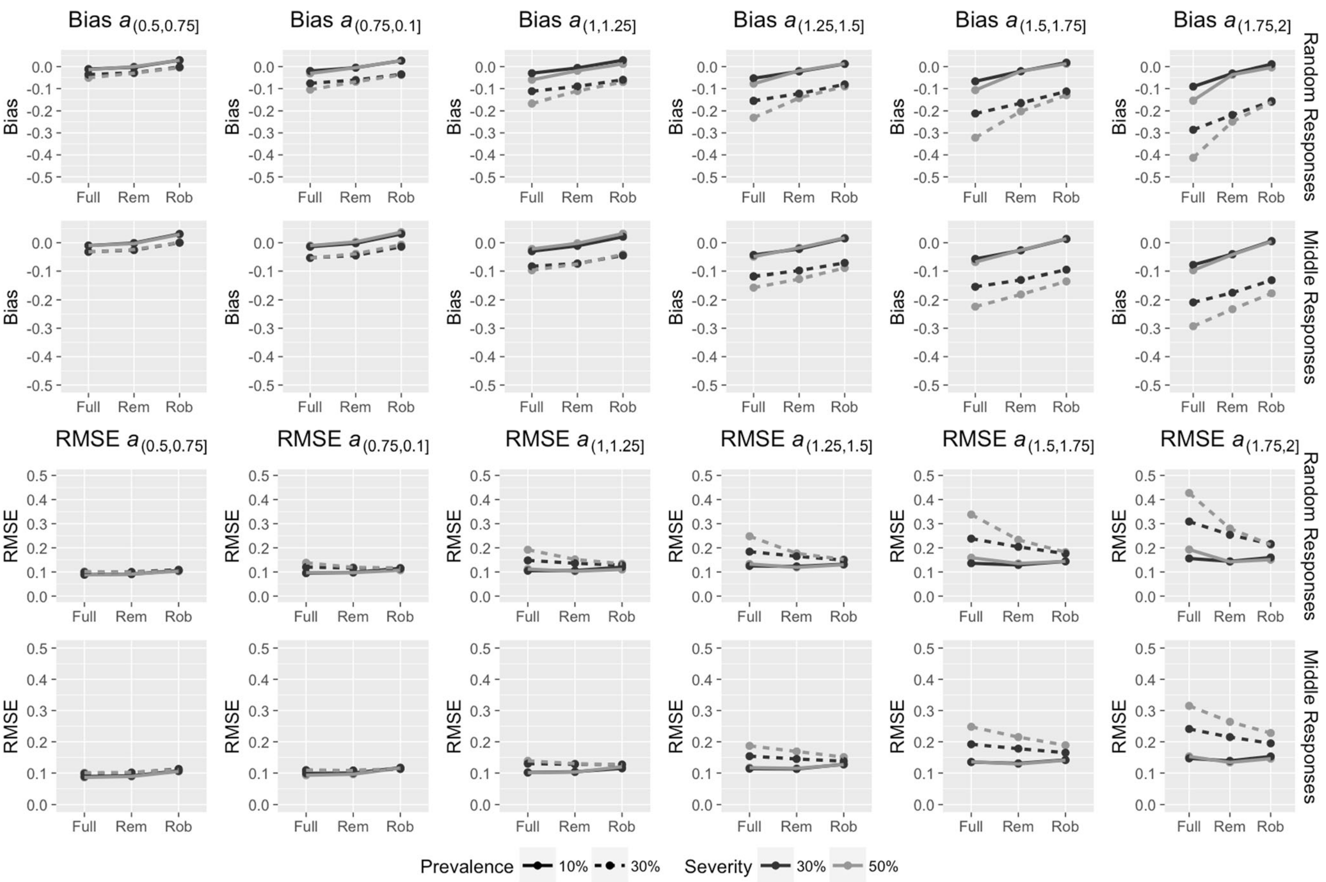

Fig. 3 Conditional bias and RMSE for the discrimination parameters for three estimation procedures based on l_z^* (Full = full data, Rem = data removal; Rob = robust estimation). The top two rows are estimated bias and the bottom two rows are RMSE. Solid and dashed lines represent different prevalence rates and gray and black colors represent different severity rates

estimation procedure produced 1.67. Larger discrimination estimates were observed in the majority of the other items, as well. Moreover, the effect of careless responses on the *b* parameter estimates varied across items. Some parameter estimates increased or decreased, depending on the estimation procedure. Potentially, different kinds of aberrant responses might be attributed to different amounts of bias for the location parameter, which can explain why there are different manifestations of item parameter bias. It would be difficult to generalize the overall effect of careless responses specifically for the location parameter. All standard error estimates for the item parameters were also comparable, which could be attributed to the large sample size.

Discussion

This article has proposed a robust estimation framework to improve item parameter estimation in presence of aberrant responses. We provide a general method that could leverage person-fit statistics, such as l_z^*, to detect careless responders and down-weight them in the estimation process. This procedure was compared to what is common in practice, the unweighted full-data estimation procedure. We also applied a data cleansing procedure, which is essentially a special case of the robust estimation procedure, when one fixes the weights to be zero for flagged participants. We compared all three procedures that involve the use of l_z^* using a simulation study in which different careless response behaviors, percentages of severity and prevalence, sample sizes, and item bank sizes were considered. In general, the impact of careless responses was regulated mostly by the severity and prevalence or by the number of careless responses. Test length and sample size made minimal differences when considering the impact of careless responses. Moreover, the type of careless responses had similar impacts across conditions, with varying magnitudes.

The results from our simulation study suggest that using the entire data set when fitting an IRT model severely biases both the location and discrimination parameters, at least in the context with the GRM polytomous model. Furthermore, we found that the biased item parameter estimates significantly reduces the ability of detecting outliers with the person-fit statistic, l_z^*. This finding is consistent with previous research,

Table 3 Item parameter estimates for the Feelings scale

Item	a_{Full}	a_{Remove}	a_{Robust}	$b_{1\text{–Full}}$	$b_{1\text{–Remove}}$	$b_{1\text{–Robust}}$	$b_{2\text{–Full}}$	$b_{2\text{–Remove}}$	$b_{2\text{–Robust}}$
1	1.42 (0.04)	1.48 (0.05)	1.67 (0.05)	0.40 (0.03)	0.41 (0.03)	0.47 (0.02)	2.27 (0.06)	2.39 (0.06)	2.53 (0.06)
2	1.04 (0.04)	1.08 (0.04)	1.25 (0.04)	0.69 (0.03)	0.71 (0.04)	0.79 (0.03)	2.67 (0.09)	2.79 (0.09)	2.82 (0.08)
3	2.20 (0.06)	2.31 (0.07)	2.58 (0.08)	0.72 (0.02)	0.74 (0.02)	0.79 (0.02)	1.8 (0.04)	1.89 (0.04)	1.99 (0.04)
4	0.86 (0.03)	0.94 (0.03)	1.04 (0.03)	– 0.76 (0.04)	– 0.79 (0.04)	– 0.85 (0.04)	1.01 (0.05)	0.92 (0.04)	0.82 (0.04)
5	1.24 (0.04)	1.27 (0.04)	1.41 (0.04)	– 0.40 (0.03)	– 0.40 (0.03)	– 0.42 (0.03)	1.64 (0.05)	1.70 (0.05)	1.82 (0.05)
6	2.74 (0.08)	2.80 (0.08)	3.08 (0.09)	0.34 (0.02)	0.34 (0.02)	0.35 (0.02)	1.54 (0.03)	1.61 (0.03)	1.73 (0.03)
7	1.01 (0.03)	1.04 (0.04)	1.16 (0.04)	– 0.38 (0.03)	– 0.40 (0.03)	– 0.40 (0.03)	2.28 (0.07)	2.38 (0.08)	2.50 (0.07)
8	0.78 (0.03)	0.85 (0.03)	0.95 (0.03)	– 1.16 (0.05)	– 1.19 (0.05)	– 1.31 (0.05)	0.89 (0.05)	0.79 (0.04)	0.63 (0.04)
9	2.13 (0.07)	2.33 (0.08)	2.62 (0.09)	1.29 (0.03)	1.33 (0.03)	1.40 (0.03)	2.35 (0.06)	2.50 (0.06)	2.67 (0.06)
10	1.25 (0.04)	1.28 (0.05)	1.52 (0.05)	1.00 (0.04)	1.03 (0.04)	1.07 (0.03)	3.20 (0.10)	3.57 (0.12)	3.62 (0.11)
11	1.28 (0.04)	1.36 (0.04)	1.51 (0.04)	– 0.56 (0.03)	– 0.60 (0.03)	– 0.72 (0.03)	1.29 (0.04)	1.26 (0.04)	1.25 (0.03)
12	0.86 (0.03)	0.89 (0.03)	1.04 (0.04)	0.33 (0.03)	0.36 (0.04)	0.41 (0.03)	2.90 (0.10)	3.10 (0.11)	3.13 (0.10)
13	1.96 (0.06)	2.02 (0.06)	2.25 (0.06)	0.47 (0.02)	0.47 (0.02)	0.50 (0.02)	1.86 (0.04)	1.98 (0.05)	2.11 (0.04)
14	0.97 (0.04)	0.98 (0.04)	1.19 (0.04)	0.83 (0.04)	0.87 (0.04)	0.94 (0.04)	3.38 (0.12)	3.79 (0.14)	3.82 (0.13)
15	1.37 (0.04)	1.46 (0.04)	1.60 (0.04)	– 0.07 (0.02)	– 0.09 (0.02)	– 0.15 (0.02)	1.33 (0.04)	1.31 (0.04)	1.34 (0.03)
16	2.27 (0.06)	2.40 (0.07)	2.61 (0.07)	0.07 (0.02)	0.05 (0.02)	0.03 (0.02)	1.88 (0.04)	1.98 (0.04)	2.13 (0.04)
17	1.38 (0.04)	1.39 (0.05)	1.63 (0.05)	0.62 (0.03)	0.65 (0.03)	0.73 (0.03)	2.55 (0.07)	2.82 (0.08)	2.95 (0.08)
18	0.96 (0.03)	0.97 (0.04)	1.07 (0.04)	– 0.10 (0.03)	– 0.12 (0.03)	– 0.12 (0.03)	2.85 (0.09)	3.03 (0.10)	3.18 (0.10)
19	2.14 (0.08)	2.43 (0.09)	2.71 (0.11)	1.55 (0.04)	1.58 (0.04)	1.68 (0.03)	2.46 (0.06)	2.56 (0.06)	2.75 (0.06)

Standard error estimates are in parentheses

in which the masking effect reduced the ability of finding outliers due to biased structural parameter estimates (Yuan & Zhong, 2008). To assuage the impact of item parameter bias, our goal was to compare two procedures that could potentially improve both item parameter estimates and detection rates: a data removal process and a more general robust estimation procedure. Both procedures improved item calibration; that is, they resulted in less biased item parameter estimates. In turn, both procedures were able to better detect outliers, due to a reduction in the masking effect.

In spite of the encouraging results, there are several limitations and room for improvement concerning our robust procedure. One confounding element for our weight function is that we require an estimate of latent ability. For a finite test length, l_z^* may not follow a normal distribution. Sinharay (2016b) and Snijders (2001) found that l_z^* may have smaller variances for extreme values of latent ability. This would, in turn, influence our weight function. Second, aberrant responses contribute their own form of bias when estimating latent trait estimates (Oshima, 1994). To correct this form of bias in the latent trait estimates, robust estimators for latent ability can certainly be used (e.g., Schuster & Yuan, 2011). Coupling robust estimators for latent ability and item parameters could be studied in the future. The performance of our method may be evaluated under more simulation conditions. For instance, the tendency to respond carelessly can also be correlated with latent trait ability (Falk & Cai, 2016), or carelessness may lead to responses that are not completely independent of the item content (Shao & Cheng, 2017; Yu & Cheng, 2017). These scenarios are not covered in this article but certainly warrant attention in future studies.

Moreover, robust estimators of an IRT model can have a breakdown point. A breakdown point for a robust estimator corresponds to the proportion of aberrant observations an estimator can accommodate before giving a severely biased or inconsistent estimate. For an IRT model, the breakdown point for robust estimation would become most evident concerning the discrimination parameter estimate. The robust estimator for slope parameter estimates could explode after several iterations of down weighting outliers (Croux, Flandre, & Haesbroeck, 2002). This occurs when it is easy to distinguish between high- and low-ability examinees or when a vertical line can separate when a person would or would not endorse an item. This would cause discrimination parameter estimate to approach infinity. Outliers can be thought of a "safeguard" to minimize this effect.

An unforeseen side effect of the robust estimation procedure was inflated RMSEs as compared to the two other estimation procedures, especially in the largest location parameters in absolute terms. A trade-off between bias and efficiency is not unusual, in which a less biased estimate using robust estimation might be worth the loss of efficiency as compared to nonrobust estimation (Carroll & Pederson, 1993). Moreover, larger RMSEs could potentially explain why our Type I error rates for outlier detection were significantly larger in the case of the robust estimation procedures. A similar finding when using robust estimators for latent ability further validates this argument (Sinharay, 2016b). Perhaps coupling our robust estimators with a resampling procedure, such as the bootstrap, could build an empirical null distribution for the l_z^* statistic (Efron & Tibshirani, 1994). The bootstrap procedure would retain the nice properties of unbiased item parameter estimates and could potentially reduce the Type I error rates to a nominal level.

These limitations aside, both other researchers and applied psychometricians should find our study meaningful. First, our method provides a general framework for robust estimation. Our study provides the initial steps that outline a general methodology, but they can be improved upon or modified. For instance, our choice of utilizing the l_z^* statistic was based on its popularity, familiarity, and simplicity of use. Researchers could potentially choose other weighting procedures, such as using the MD (Meade & Craig, 2012). Other person-fit statistics can also be used, including nonparametric person-fit statistics such as Gnorm (Emons, 2008).

Our overall approach in this article was done by maximizing a so-called *psuedolikelihood function*, which directly maps onto the literature on estimating IRT models with complex survey designs (Thomas & Cyr, 2002) and robust generalized linear model (Carroll & Pederson, 1993; Künsch et al., 1989). However, other robust methods might be employed, such as those in the general structural equation modeling framework, by iteratively reweighted least squares (Yuan & Zhong, 2008).

In addition, our study could be applied to other measurement situations, such as using IRT models suited for educational testing scenario as opposed to a psychological battery. In that scenario, other common IRT models, such as the 2PL or 3PL for dichotomous items, might be more appropriate. Because GRM subsumes 2PL and 1PL, our RMML procedure applies directly to those models. With proper adaptation, the procedure can be used with 3PL, as well. Other considerations could be to understand our method in the context of more complex IRT models that are common in the practice of psychology, such as multidimensional IRT models. Our method can address this situation in two ways. First, one could apply the robust estimation procedure to the subscales in a psychological battery. This approach would certainly be the easiest to perform computationally, and it would provide a rich picture of scales on which a participant does not perform at his or her true ability level. Our method could also be framed using a multidimensional IRT model, in which one would just add a weighting procedure to the likelihood when estimating a multidimensional IRT model.

Our method is also easy to adopt in practice. Example code to calculate conduct robust estimation is provided in Appendix B using R. Moreover, one could also use other R packages to calculate person-fit statistics, such as PerFit, or could create their own weights to employ in mirt, which has a weight function argument that can be used to implement our method (Chalmers, 2012; Tendiero, 2015).

Finally, our study contributes to the literature more broadly in psychology and other sciences that have large sources of measurement error (Maxwell, Lau, & Howard, 2015). More specifically, we addressed issues that arise when identifying powerful statistics to detect outliers in a data set that could potentially confound replication studies (Loken & Gelman, 2017). Detecting outliers is an important step psychologists should take during data preprocessing; our method provides one improvement on preexisting methods.

Appendix A

Suppose $P_{jk}(\theta)$ follow the graded response model for $j = 1, \ldots, J$ items and $k = 1, \ldots, K$ response options. Its first-order derivative is denoted as $P_{jk}^{'}(\theta)$. The derivation and notation in Sinharay (2016a), $l_z^*\left(\hat{\theta}\right)$ can be expressed as

$$l_z^*\left(\hat{\theta}\right) = \frac{\left[T\left(\hat{\theta}\right) + c_J^{'}\left(\hat{\theta}\right)s_0\left(\hat{\theta}\right)\right]}{\left[\sqrt{J}\zeta_J\left(\hat{\theta}\right)\right]},$$

where

$$T\left(\hat{\theta}\right) = \textstyle\sum_{j=1}^{J}\sum_{k=1}^{K}\left[\delta_{jk} - P_{jk}\left(\hat{\theta}\right)\right]\left[\tilde{w}_{jk}\left(\hat{\theta}\right)\right],$$

$$\delta_{jk} = \begin{cases} 1, \text{if } u_j = k \\ 0, \text{otherwise}, \end{cases}$$

$$c_J^{'}\left(\hat{\theta}\right) = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}P_{jk}^{'}\left(\hat{\theta}\right)\log P_{jk}\left(\hat{\theta}\right)}{\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{P_{jk}^{'}\left(\hat{\theta}\right)^2}{P_{jk}\left(\hat{\theta}\right)}}.$$

For $s_0\left(\hat{\theta}\right)$, it depends on the estimation method of θ. For example, $s_0\left(\hat{\theta}\right) = 0$ for ML estimate of θ. In addition,

$$\zeta_J^{\,2}\left(\hat{\theta}\right) = \frac{1}{J}\textstyle\sum_{j=1}^{J}\mathbf{v}_j^{'}\left(\hat{\theta}\right)\mathbf{D}_j\left(\hat{\theta}\right)\mathbf{v}_j\left(\hat{\theta}\right),$$

where $\mathbf{D}_j\left(\hat{\theta}\right)$ is the variance–covariance matrix of $(\delta_{j1}, \delta_{j2}, \ldots, \delta_{jK})^{'}$:

$$\mathbf{D}_{\boldsymbol{j}}\left(\hat{\theta}\right) = \begin{pmatrix} P_{j0}\left(\hat{\theta}\right)\left(1-P_{j0}\left(\hat{\theta}\right)\right) & -P_{j0}\left(\hat{\theta}\right)P_{j1}\left(\hat{\theta}\right) & \dots & -P_{j0}\left(\hat{\theta}\right)P_{jK}\left(\hat{\theta}\right) \\ -P_{j1}\left(\hat{\theta}\right)P_{j0}\left(\hat{\theta}\right) & P_{j1}\left(\hat{\theta}\right)\left(1-P_{j1}\left(\hat{\theta}\right)\right) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ -P_{jK}\left(\hat{\theta}\right)P_{j0}\left(\hat{\theta}\right) & -P_{jK}\left(\hat{\theta}\right)P_{j1}\left(\hat{\theta}\right) & \dots & P_{jK}\left(\hat{\theta}\right)\left(1-P_{jK}\left(\hat{\theta}\right)\right) \end{pmatrix},$$

and

$$\mathbf{v}_{\boldsymbol{j}}^{'}\left(\hat{\theta}\right)=\left(\tilde{w}_{j1}\left(\hat{\theta}\right),\tilde{w}_{j2}\left(\hat{\theta}\right),\dots,\tilde{w}_{jK}\left(\hat{\theta}\right)\right),\text{ where }\tilde{w}_{jk}\left(\hat{\theta}\right)$$

$$=\log\left(P_{jk}\left(\hat{\theta}\right)\right)-c_{J}^{'}\left(\hat{\theta}\right)\frac{P_{jk}^{'}\left(\hat{\theta}\right)}{P_{jk}\left(\hat{\theta}\right)}.$$

Appendix B

```
# Load package and example data set
    require(mirt)
    data(Science)
    # Initial Model estimation
    mod <- mirt(Science, 1, optimizer = 'NR')
    # Calculating person fit
    per.fit <- personfit(mod, method = 'ML')$Zh
    # Calculating weight
    weight <- pnorm(per.fit)*nrow(Science)/
sum(pnorm(per.fit))
    # Robust estimation
    robust.mod <- mirt(Science, 1, survey.weights = weight,
optimizer = 'NR')
```

References

Allalouf, A., Gutentag, T., & Baumer, M. (2017). Quality control for scoring tests administered in continuous mode: An NCME instructional module. *Educational Measurement: Issues and Practice*, *36*, 58–68. https://doi.org/10.1111/emip.12140

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, *29*, 357–368. https://doi.org/10.1177/0146621605276676

Baer, R. A., Ballenger, J., Berru, D., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment*, *68*, 139–151.

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques (2nd). New York: Marcel Dekker.

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied*, *123*, 101–103. https://doi.org/10.1080/00223980.1989.10542966

Beaton, A. E., & Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, *16*, 147–185. https://doi.org/10.1080/00401706.1974.10489171

Bejar, I., & Wingersky, M.S. (1981). An application of item response theory to equating the Test of Standard Written English (College Board Resport No. 81-8, ETS No. 81-35). Princeton: Educational Testing Service.

Berry, D., Wetter, M., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*, 340–345. https://doi.org/10.1037/1040-3590.4.3.340

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*.

Bock, R. D., & Aitkin, M. (1981). EM solution of the marginal likelihood equations. *Psychometrika*, *46*, 443–459.

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*, 69–83. https://doi.org/10.1037/met0000106

Carroll, R. J., & Pederson, S. (1993). On robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B*, *84*, 693–706.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cheng, Y., Liu, C., & Behrens, J. (2015). Standard error of ability estimates and the classification accuracy and consistency of binary decisions. *Psychometrika*, *80*, 645–664. https://doi.org/10.1007/s11336-014-9407-z

Cheng, Y., & Patton, J. M. (2014). Detection and treatment of careless responses in survey data. Poster presented at the annual convention of the Association for Psychological Science, San Francisco.

Cheng, Y., & Yuan, K. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*, 280–291. https://doi.org/10.1007/s11336-009-9144-x

Chien, T. W., Shao, Y., & Kuo, S. C. (2017). Development of a Microsoft Excel tool for one-parameter Rasch model of continuous items: An application to a safety attitude survey. *BMC Medical Research Methodology*, *17*. https://doi.org/10.1186/s12874-016-0276-2

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and Personality Assessment Inventory validity indices. *Psychological Assessment*, *15*, 223–234. https://doi.org/10.1037/1040-3590.15.2.223

Conrad, K. J., Bezruczko, N., Chan, Y. F., Riley, B., Diamond, G., & Dennis, M. L. (2010). Screening for atypical suicide risk with person fit statistics among people presenting to alcohol and other drug treatment. *Drug and Alcohol Dependence*, *106*, 92–100. https://doi.org/10.1016/j.drugalcdep.2009.07.023

Croux, C., Flandre, C., & Haesbroeck, G. (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters*, *60*, 377–386.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Curran, P. G., Kotrba, L., & Denison, D. (2010). Careless responding in surveys: Applying traditional techniques to organizational settings. Poster presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta.

Van Dam, N. T., Earleywine, M., & Borders, A. (2010). Measuring mindfulness? An Item Response Theory analysis of the Mindful Attention Awareness Scale. *Personality and Individual Differences*, *49*, 805–810. https://doi.org/10.1016/j.paid.2010.07.020

de Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*, 155–170. https://doi.org/10.1177/014662169401800205

de Ayala, R. J. (2009). The theory and practice of item response theory. New York: Guilford Press.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*, 5–18. https://doi.org/10.1007/s11136-007-9198-0

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. New York: Chapman & Hall.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*, 224–247. https://doi.org/10.1177/0146621607302479

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*, 328–347. https://doi.org/10.1037/met0000059

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, *8*, 863. https://doi.org/10.3389/fpsyg.2017.00863

Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement*, *28*, 126–140. https://doi.org/10.1177/0146621603260917

Harris, K. M., & Udry, J. R. (2010). *National* Longitudinal Study of Adolescent Health (Add Health), 1994–2008: Core files [restricted use] (Technical report). Ann Arbor: Inter-University Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR27021.v11

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99–114. https://doi.org/10.1007/s10869-011-9231-8

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

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

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*, 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298. https://doi.org/10.1207/S15324818AME1604_2

Kim, S., & Moses, T. (2016). ETS GRE® board research report investigating robustness of item response theory proficiency estimators to two-stage multistage testing. Princeton: Educational Testing Service.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement*, *26*, 164–180. https://doi.org/10.1177/01421602026002004

Künsch, H. R., Stefanski, L. A., & Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, *84*, 460–466. https://doi.org/10.1080/01621459.1989.10478791

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). An investigation of item bias in a test of reading comprehension (Technical Report No. 163). Urbana: University of Illinois, Center for the Study of Reading.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585. https://doi.org/10.1126/science.aal3618

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79–83. https://doi.org/10.1016/j.paid.2014.08.021

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. https://doi.org/10.1037/a0039400

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*, 450–470. https://doi.org/10.1037/a0019216

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455. https://doi.org/10.1037/a0028085

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135. https://doi.org/10.1177/01466210122031957

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Orr, J. M., Sackett, P. R., & Dubois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology*, *44*, 473–486. https://doi.org/10.1111/j.1744-6570.1991.tb02401.x

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*, 200–219. https://doi.org/10.1111/j.1745-3984.1994.tb00443.x

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*, 1–18. https://doi.org/10.1348/014466506X96931

Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement*, *37*, 24–40.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Pauszek, J. R., Sztybel, P., & Gibson, B. S. (2017). Evaluating Amazon's Mechanical Turk for psychological research on the symbolic control of attention. *Behavior Research Methods*, *49*, 1969–1983. https://doi.org/10.3758/s13428-016-0847-5

R Development Core Team. (2017). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to The New Statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology*, *66*, 47–54. https://doi.org/10.1016/j.jesp.2015.09.013

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*. https://doi.org/10.1007/BF02290599

Schmitt, N., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement*, *17*, 143–150. https://doi.org/10.1177/014662169301700204

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

Shao, C., & Cheng, Y. (2017). Detection of test speededness using change-point analysis with response time data. Paper presented at the Annual Meeting of National Council for Measurement in Education, San Antonio.

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

Sinharay, S. (2016a). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, *81*, 992–1013. https://doi.org/10.1007/s11336-015-9465-x

Sinharay, S. (2016b). The choice of the ability estimate with asymptotically correct standardized person-fit statistics. *British Journal of*

Mathematical and Statistical Psychology, *69*, 175–193. https://doi.org/10.1111/bmsp.12067

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342. https://doi.org/10.1007/BF02294437

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications. *Perspectives on Psychological Science*, *9*, 305–318. https://doi.org/10.1177/1745691614528518

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201–210. https://doi.org/10.1177/014662168300700208

Tendeiro, J. N. (2015). Perfit (R package version 1.4) [Computer software]. Available from http://cran.r-project.org/web/packages/PerFit/index.html.

Thomas, D. R., & Cyr, A. (2002). Applying item response theory methods to complex survey data. In *Proceedings of the SSC Annual Meeting, Survey Methods section* (pp. 17–26). Ottawa: Statistical Society of Canada.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, *55*, 371–390.

van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika*, *81*, 650–673. https://doi.org/10.1007/s11336-015-9469-6

Wainer, H. (2014). Cheating: Some ways to detect it badly. In N. M. Kingston & A. K. Clark (Eds.), Test fraud: Statistical detection and methodology (pp. 8–20). New York: Taylor & Francis.

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

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, *33*, 352–364. https://doi.org/10.1027/1015-5759/a000291

Wilcox, R. R. (2016) Introduction to robust estimation and hypothesis testing (4th). San Diego: Academic Press.

Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient: A note on Attali's reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, *33*, 488–490. https://doi.org/10.1177/0146621607304655

Yamamoto, K., & Everson, H. (2003). Estimating the effects of test length and test time on parameter estimation using the hybrid model. *ETS Research Report Series*, *1995*, 277–298. https://doi.org/10.1002/j.2333-8504.1995.tb01637.x

Yang, M., & Yuan, K.-H. (2016). Robust methods for moderation analysis with a two-level regression model. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2016.1235965

Yu, X., & Cheng, Y. (2017). Using change point analysis to detect inattentiveness in polytomous survey response data. Paper presented at the 2017 Conference on Test Security, Madison.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004a). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*. https://doi.org/10.1007/BF02295644

Yuan, K.-H., Fung, W. K., & Reise, S. P. (2004b). Three Mahalanobis distances and their role in assessing unidimensionality. *British Journal of Mathematical and Statistical Psychology*, *57*, 151–165. https://doi.org/10.1348/000711004849231

Yuan, K.-H., & Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika*, *77*, 803–826. https://doi.org/10.007/s11336=012-9282-4

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology*, *38*, 329–368. https://doi.org/10.1111/j.1467-9531.2008.00198.x

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