Mixture Item Response Models for Inattentive Responding Behavior

Abstract

Inattentive responses can threaten measurement quality, yet they are common in rating- or Likert-scale data. In this study, we proposed a new mixture item response theory model to distinguish inattentive responses from normal responses so that test validity can be ascertained. Simulation studies demonstrated that the parameters of the new model were recovered fairly well using the Bayesian methods implemented in the freeware WinBUGS, and fitting the new model to data that lacked inattentive responses did not result in severely biased parameter estimates. In contrast, ignoring inattentive responses by fitting standard item response theory models to data containing inattentive responses yielded seriously biased parameter estimates and a failure to distinguish inattentive participants from normal participants; the person-fit statistic lz was also unsatisfactory in identifying inattentive responses. Two empirical examples demonstrate the applications of the new model.

Likert-type scales have been commonly used in marketing, psychological and educational research to gauge people's attitude and behaviors, and scores are further compared to understand group differences. In marketing, (cross-country) surveys are administered to understand people's attitudes or shopping habits to target potential customers. Self-reported Likert or rating scale items rely heavily on the accuracy of respondents' reporting. Nichols, Greene, and Schmolck (1989) identified two types of problematic responses: content responsive faking (which refers to fake-good or fake-bad responses to obtain benefits) and content nonresponsivity (responses without regard to item content). When respondents are unable or unwilling to comply with the test demands, they may try to finish questionnaires as quickly as possible, giving careless, haphazard, or random responses. For instance, Johnson (2005) reported that approximately 3.5% of respondents used the same response category throughout without reading the items when responding to an online version of the revised NEO Personality Inventory. Data quality and test validity are thus questionable, because these inattentive responses add noise to the data (Fong, Ho, & Lam, 2010; Huang et al., 2012; Kurtz & Parrish, 2001), particularly in low-stakes questionnaires (Fervaha & Remington, 2013; Huang, Curran, Keeney, Poposki, & DeShon, 2012). It is important to screen out inattentive responses before statistical inferences are made.

Inattentive respondents may respond to questionnaires in several ways (Baumgartner & Steenkamp, 2001; Johnson, 2005; Meade & Craig, 2012). The first type of inattentive response involves skipping questions and leaving answers blank (Johnson, 2005). Such responses can be treated as missing at random and missing values can be imputed to account for them (Rubin, 1976, 1987). The second type is choosing response categories randomly (Glas & Dagohoy, 2007) or according to an arbitrary nonrandom pattern, such as "0, 1, 2, 3, 4, 0, 1, 2, 3, 4, …". The third type is selecting the same response category throughout the whole inventory, regardless of item content. Participants may "agree" with all items (the acquiescent response style; ARS), or "disagree" with all items (the disacquiescent response style; DRS). The fourth type is choosing the middle category (the midpoint response style; MRS). This occurs when participants lack motivation to complete the test accurately, do not have clear opinions, are uncertain about the

item content, or are unwilling to express their true opinions; as a result, their responses do not reflect their true attitudes, and the MRS becomes a nuisance in data analysis.

Recognizing that inattentive responses in Likert or rating scales threaten the reliability and validity of a measure (e.g., Hinkin, 1998), the present study developed a new mixture item response theory (IRT) model to account for inattentive responses. Because the ARS and DRS can be difficult to distinguish from the measured latent trait (De Beuckelaer, Weijters, & Ruttan, 2010; Weijters, Baumgartner, & Schillewaert, 2013), additional information (e.g., a validity scale) is required for judgment but is often not available for inventories. Thus, we mainly focused on random responses and the MRS in this study. The rest of this article is organized as follows. We review existing approaches to detecting inattentive responses and develop a new mixture IRT model that can distinguish inattentive responses from normal responses. We then use simulations to evaluate the performance of the new model and compare its performance with standard IRT models. We apply the new mixture model to two empirical examples to demonstrate the implications and applications of the new approach. Finally, we draw conclusions and give suggestions for future studies.

Existing Approaches to Inattentive Responses

Researchers have proposed several strategies to detect inattentive responses before and after data collection. Some make use of additional survey items to analyze specific behaviors of respondents such as the inconsistency approach, the infrequency approach and the validity-scale approach (Huang & Liu, 2014). Others examine certain patterns or behaviors post-hoc or apply statistical techniques to filter out untrustworthy data (DeSimone, Harms, & DeSimone, 2015), and the most common approaches include the response-pattern approach, the person-fit-index approach, the outlier approach, the response-time approach and mixed item response approach.

The inconsistency approach assumes that inattentive respondents respond to items inconsistently and without reference to item content. Matched item pairs can be used to test whether respondents reply consistently, by comparing the response on one item to the matching one (Huang, Curran, Keeney, Poposki, & DeShon, 2012; McKibben & Silvia, 2015). Studies

have shown that this approach can successfully be applied to identify random responses, but some of the studies also showed mixed results. Huang et al. (2012) further note that this method requires a large number of items, thus it is more complex than other approaches. An additional problem in regard to Likert scales is that some inattentive respondents may slightly change their responses, but keep them around the midpoint of the scale, thus appearing to be more attentive than they actually are (McKibben & Silvia, 2015).

Another way to detect inconsistency is using mixed-format items (including both positively worded [PW] and negatively worded [NW] items). In a mixed-format design, both PW and NW items are assumed to measure the same construct, and respondents with a high score on PW items should have a low score on NW items and vice versa. If a respondent exhibits different patterns (high scores on both PW and NW items, or low scores on both kinds of items), he or she will be identified as an inattentive respondent (Greene, 1978, 1979; Kurtz & Parrish, 2001; Pinsoneault, 1998; Schinka, Kinder, & Kremer, 1997). However, eye-tracking studies (Kamoen, Holleman, Mak, Sanders, & van den Bergh, 2011) and psychometric evidence have indicated that recoded NW items do not merely function as counterparts of PW items. NW items often require more mental resources, and may include a nuisance dimension (Bassili & Scott, 1996; Chessa & Holleman, 2007; Reise, Morizot, & Hays, 2007; Wang, Chen, & Jin, 2015; Woods, 2006).

The infrequency approach is based on the assumption that respondents are likely to provide the same answer to certain items, and if their answers differ from the expectation it could be an indication for random responding (Huang, Bowling, Liu, & Li, 2015; McKibben & Silvia, 2015). Bogus items, which include obvious content and should elicit the same answer from all respondents, could be used to flag possible inattentive respondents (DeSimone, Harms, & DeSimone, 2015). This approach is useful for detecting random answers, but there is a concern in its effectiveness (Maniaci & Rogge, 2014). For instance, some participants might be confused or less motivated after noticing such items, thus affecting the results (McKibben & Silvia, 2015). In addition, aberrant responses might be influenced by faking, and the infrequency approach is

not ideal (Huang et al., 2012). McKibben and Silvia (2015) point out that the infrequency approach could work better with a binary format, such as true/false responses, rather than Likert scales, since the results would be easier to interpret.

The validity-scale approach uses validity scales with the original scales to detect aberrant responses (Arbisi & Ben-Porath, 1995; Archer & Elkins, 1999; Archer, Fontaine, & McCrae, 1998; Baer, Ballenger, Berry, & Wetter, 1997; Schinka et al., 1997; Stein, Graham, & Williams, 1995). Because researchers know the true answers for items on validity scales, if responses are not consistent with the true answers, participants will be classified as inattentive. Examples include the Variable Response Inconsistency Scale and the True Response Consistency Scale in the Minnesota Multiphasic Personality Inventory–Adolescent (MMPI–A; Butcher et al., 1992). Butcher and Rouse (1996) suggested that validity scales should include items assessing cooperation, willingness to share personal information, and degree of response exaggeration. However, empirical studies do not support the use of validity scales (e.g., Nicholson & Hogan, 1990; Piedmont, McCrae, Riemann, & Angleitner, 2000).

The response-pattern or long string approach examines the number or the proportion of items on which a respondent endorses a specific response category (Baumgartner & Steenkamp, 2001, 2006; Chen, Lee, & Stevenson, 1995; Costa & McCrae, 2008; Nichols et al., 1989; Stening & Everett, 1984), such as choosing the “Strongly Agree” option for a number of consecutive items (Costa & McCrae; Huang et al., 2015). The idea behind it is that too many identical answers could indicate a lack of effort or attention. After data collection, researchers can visually screen the data to filter out potential aberrant respondents. For example, participants who skipped the second half of questions or showed a bizarre response pattern (e.g., agreeing on all odd questions and disagreeing on all even questions) are recognized as aberrant respondents. The response-pattern approach is quite straightforward. However, different studies tend to use different arbitrary cut-off points and there is no standard for cut-off points (Huang et al., 2012).

The person-fit-index approach adopts person-fit statistics such as l_z (Drasgow, Levine, & Williams, 1985; Karabatsos, 2003) to detect aberrant responses. De Leeuw and Hox (1994)

found that different questions may induce different kinds of aberrancy and person-fit statistics become uninformative when a person has either the minimum or the maximum possible total score on a test. In addition, the accuracy of the detection rate decreases as the percentage of aberrant-responding persons increases; for example, the detection rate of accuracy is roughly 50%-60% when half of respondents are inattentive (Karabatsos, 2003). A long test, including as many as 65 items, is often needed to yield a detection rate accuracy of 80% (De Leeuw & Hox, 1994; Karabatsos, 2003). Its performance is influenced by several factors (e.g., test design, response style, and model misspecification), and the person-fit approach cannot tell the causes of poor fit. What is more, the person-fit-index approach is somewhat paradoxical. Person-fit statistics require accurate item and person estimates. When data contain many aberrant responses, fitting standard IRT models will yield biased estimates for the item and person parameters, which in turn will invalidate person-fit statistics (Rupp, 2013).

Mahalanobis D approach or outlier analysis compares respondents' scores against the sample means to identify outliers as deviation from the typical response pattern, indicating insufficient effort (Curran, 2015; DeSimone, Harms, & DeSimone, 2015). One limitation of this method is that it is complex and computationally intensive; it requires a number of calculations, of which an average result will be obtained (Meade & Craig, 2012). Since there have not been many studies on this method, its effects are not yet fully understood. Thus, it may not be a good enough means to eliminate inattentive responders yet, but it can be used to flag certain individuals who can then be further examined (Curran, 2015).

The response-time approach uses response time to detect inattentive responses (Wise & DeMars, 2006). Unmotivated respondents usually answer items rashly without considering item content, such that their response time is significantly shorter than that of normal respondents (Huang et al., 2012; Wise & DeMars, 2006). Although response time is strong evidence of response effort, special time-recording equipment is required for this approach (e.g., computerized-based testing), which may not be available in many tests. Curran (2015) argues that this approach as well as the consistency and the infrequency approach are not sufficient for

identifying inattentive respondents. They can be applied as a minimum to remove the worst responders, but it is recommended to follow up with techniques that are more complex.

One limitation applied to most of the above methods is that the results can be affected by missing data. In regard to bogus items, for example, a missing item could indicate an incorrect response, but not necessarily insufficient effort on the part of the respondent. When respondents leave out questions, their response time will be shorter, thus affecting the average response time values. Item pairs could also be rendered useless if responses to one or both of them are missing. Using the Mahalanobis D approach could be problematic in the presence of missing data as well, because the results will be more difficult to analyse (DeSimone, Harms, & DeSimone, 2015).

A mixed item response theory (IRT), a more general form of mixture IRT, has been proposed to differentiate response behavior in personality, attitude and organizational research (e.g., Eid & Rauber, 2000; Eid & Zickar, 2007; Maji-de Meij, Kelderman, & van der Flier, 2008). The mixed IRT combines traditional IRT with latent class analysis (LCA) and does not define respondents' behaviors in advance. It adopts a data-driven framework to identify respondents who differ from the norm, based on their item response functions (Carter, Dalal, Lake, Lin, & Zickar, 2011). When participants use the response scale differently, there will be more than one latent class (Cho, 2007; Maji-de Meij, Kelderman, & van der Flier, 2008). Similar to differential item functioning (DIF) assessment, this approach can be used to model how certain items function differently in different groups, thus providing a better understanding of subgroups. The mixed IRT model provides an alternative to known group membership (e.g., gender or country) by identifying groups along one or more nuisance dimension/s (Cho, 2007) to examine DIF. Therefore, Eid and Zickar (2007) define it as an exploratory DIF technique.

Because the mixed IRT for response styles is exploratory in the sense, the approach simply divides all respondents into the "best" (i.e., most different) set of subpopulations with different types of response behaviors. Researchers must then carefully explain the differences among subpopulations (Draney, Wilson, Glück, & Spiel, 2007) and even need to conduct a series of follow-up studies to investigate correlates among different classes (Zicker et al., 2004). However,

the literature has identified several profound respondents' behaviors in answering Likert-type scales, and users will benefit more from a method that can confirm the definitive behaviors and differentiate people with varied behavioral patterns, than from an exploratory approach.

The existing approaches show their utility under certain conditions. This study stood upon existing knowledge about inattentive responses such as random responses referring to equal probability of endorsing individual response category and moved further by recognizing practical constraints (e.g., no validity scales and no response time records) in most scales as well as the limitations of the mixed IRT. We mainly focused on detection of inattentive responses after data collection and proposed a new mixture IRT model, confirmatory in sense, to account for inattentive responses. Two major research questions were addressed during model development: (a) which respondents provided inattentive responses? and (b) what kinds of inattentive responses did they yield?

Model Formulation

Many IRT models have been developed for Likert or rating-scale items. For example, the generalized partial credit model (GPCM; Muraki, 1992) is defined as:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = \alpha_i(\theta_n - \beta_{ij}), \quad (1)$$

where P_{nij} and $P_{ni(j\text{-}1)}$ are the probabilities of scoring j and $j-1$ on item i for person n, respectively; α_i and β_{ij} are the slope and the jth threshold of item i, respectively; and θ_n is the latent trait level of person n. When fitting the GPCM to data, it is assumed that all participants follow the GPCM when responding to every item (i.e., the GPCM is the true model). In reality, some people may be unmotivated to respond accurately to the test, while some others may exhibit a specific type of response style. Aberrant responses do not reveal respondents' true proficiency levels (e.g., we cannot tell an examinee's proficiency level from item responses if the examinee responds to items randomly). What is more, the item parameter estimates would be seriously biased if there are a large proportion of aberrant responses, which in turn makes the

person parameter estimates biased. As a result, the GPCM is not an appropriate model for evaluating participants' responses, suggesting that a new IRT model is needed.

To meet the demand, we propose the following mixture model for inattentive responses (MMIR), which assumes that there is a set of finite latent classes (g = 1,…, G) and that each latent class captures a specific type of testing behavior, including normal behavior and inattentive behavior. The probability of response vector $\mathbf{x} = (x_1, x_2, \ldots, x_I)^{\mathrm{T}}$ (I = the number of items) is defined as:

$$P(\mathbf{x}) = \sum_{g=1}^{G} \pi_g P(\mathbf{x} \mid g), \tag{2}$$

where π_g is the mixture proportion for latent class g, and $P(\mathbf{x}|g)$ is the conditional probability of $\mathbf{x}$ given class membership g. Specifically, the MMIR considers a normal class (g = 1) and several inattentive classes (g = 2,…, G). The class membership g is assumed to follow a discrete distribution with probability vector $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_G]^{\mathrm{T}}$, so a Dirichlet distribution with G hyperparameters can be used as the prior.

To be general, let the GPCM govern the item responses. The item response function of the MMIR can be then formulated as:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right)_g = \alpha_{ig}(\theta_{ng} - \beta_{ijg}), \tag{3}$$

where α_{ig} is the slope parameter of item i for class g; β_{ijg} is the jth threshold of item i for class g; θ_{ng} is the latent trait level of person n in class g; the other variables are as defined previously. For persons within the normal class (g = 1), the GPCM (Equation 1) governs their item responses. In Bayesian methods, the following prior distributions can be used to estimate the unknown parameters in the GPCM: $\log(\alpha_{i1}) \sim N(0, \sigma_\alpha^2)$, $\beta_{ij1} \sim N(0, \sigma_\beta^2)$, and $\theta_{ng} \sim N(0, \sigma_\theta^2)$. Here, item parameters are treated as fixed effects, and person parameters are random effects (De Boeck, 2008; De Boeck & Wilson, 2004). Often, σ_θ^2 can be constrained at a constant for the purpose of model identification. In this study we use semi-informative priors for α and β parameters: $\sigma_\alpha^2 = 10$ and $\sigma_\beta^2 = 10$.

Apart from the normal class, other latent classes are used to depict various types of inattentive responses. In inattentive responses, the latent trait θ is not involved in item responses and the slope parameter α_{ig} can be fixed as zero, so Equation 3 becomes as follows:

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right)_g = -\beta_{ijg},\ g = 2, ..., G. \quad (4)$$

For example, for a five-category item the probabilities of the categories can be expressed as follows:

$$P_{ni0g} = 1/\Psi,$$
$$P_{ni1g} = \exp(-\beta_{i1g})/\Psi,$$
$$P_{ni2g} = \exp(-\beta_{i1g} - \beta_{i2g})/\Psi,$$
$$P_{ni3g} = \exp(-\beta_{i1g} - \beta_{i2g} - \beta_{i3g})/\Psi,$$
$$P_{ni4g} = \exp(-\beta_{i1g} - \beta_{i2g} - \beta_{i3g} - \beta_{i4g})/\Psi,$$

where Ψ is the summation of all of the numerators for normalization. When a person exhibits random response, the probabilities of endorsing these response categories will theoretically follow a uniform distribution, that is, 20% each. Let $g = 2$ denote random response. For five-category item i, random response implies $P_{n02} = P_{n12} = P_{n22} = P_{n32} = P_{n42} = 0.2$. Operationally, one can constrain the four β_{ij2} parameters in Equation 4 at zero: $\beta_{i12} = \beta_{i22} = \beta_{i32} = \beta_{i42} = 0$.

Other inattentive classes can be operationalized in the same manner. For example, let $g = 3$ denote respondents exhibiting the MRS pattern. One may classify respondents who choose the middle category on 90% (or 80%) of items as MRS respondents, while assuming equal endorsement of the other response categories. Take five-category item i as an example, where j = 0, 1, 2, 3, and 4. If 90% is the threshold indication of the MRS, then $P_{n23} = .90$ and $P_{n03} = P_{n13} = P_{n33} = P_{n43} = .025$. Operationally, β_{ij3} in Equation 4 can be set at $\beta_{i13} = 0$, $\beta_{i23} = -3.58$, $\beta_{i33} = 3.58$, and $\beta_{i43} = 0$. Using other percentages as an indication of the MRS is possible, and the settings for β_{ij3} are straightforward. For example, when 80% is treated as the indication and the other

categories are assumed to be equally endorsed, then $P_{n23} = .80$ and $P_{n03} = P_{n13} = P_{n33} = P_{n43} = .05$. Operationally, β_{ij3} can be set at $\beta_{i13} = 0$, $\beta_{i23} = -2.77$, $\beta_{i33} = 2.77$, and $\beta_{i43} = 0$.

Additional latent classes can be added to the MMIR as long as there is a strong rationale for specifying the corresponding probabilities for all response categories of a specific type of response pattern. The MMIR is "confirmatory" in that (a) the number of latent classes is specified in advance, and (b) the probability of the response categories in each inattentive class is specified in advance. The first requirement is a prerequisite for any mixture model. In practice, one can adopt the Akaike information criterion (AIC), the Bayesian information criterion (BIC), or the deviance information criterion (DIC) to help identify the correct number of latent classes (Cho, Cohen, & Kim, 2014). The second requirement is actually not as inconvenient as it appears. In the classification of inattentive responses, specifications are always somewhat arbitrarily set. For example, Costa and McCrae (2008) studied 983 cooperative participants who were assumed to be highly motived and honest in answering the Revised NEO Personality Inventory. Choosing "neutral" for 10 consecutive items out of 240 items was suggested as the criterion for MRS. Most studies compare differences in the proportions of endorsing the middle response category among groups but do not provide clear cut-offs for practical use (Baumgartner & Steenkamp, 2001; Chen et al., 1995; Hamamura, Heine & Paulhus, 2007; Harzing, Köster & Zhao, 2012; Si & Cullen, 1998). The MMIR has many advantages over traditional approaches to detecting inattentive responses. First, the MMIR is a general model with great flexibility and can account for different types of inattentive responses simultaneously. Classifications of responding behavior are determined by the likelihood of item responses rather than inflexible and subjective criteria for item responses. Second, the MMIR can filter out inattentive respondents and recover the true item and person parameters for direct comparison. Third, the MMIR is free from the logical paradox inherent in person-fit statistics, such as the l_z statistic, due to their reliance on accurate estimates of item and person parameters. In particular, when a person exhibits aberrant responses, the corresponding person measure will be biased, which in turn will result in poor detection of aberrant responses using the biased person measure. When many persons exhibit

aberrant responses, the resulting estimates for the item parameters are also biased, making person-fit statistics even less reliable (Rupp, 2013).

Parameter Estimation

Parameters of the MMIR can be estimated by two major methods. One is the marginal maximum likelihood estimation; the other is the Bayesian approach with the Markov chain Monte Carlo (MCMC) method. We adopted the MCMC method in this study, which is available through free software such as WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007) and JAGS (Plummer, 2012), making the MMIR readily available for practitioners. The common problem of label switching in mixture models, due to an equal likelihood across group labels (Redner & Walker, 1984), does not occur with the MMIR, because the latent group memberships are well determined by constraining the item parameters for different latent classes. Details of the MCMC procedures for the MMIR are given below.

Simulation Study 1

Design

A series of simulations was conducted to evaluate the parameter recovery of the new model. A three-class mixture IRT model (denoted as MMIR-3), including a normal class, a random-responding class, and an MRS class, was used to generate item responses. Three independent variables were manipulated: sample size (500 and 2,000), test length (10 and 20 five-point items), and proportions of the three latent classes. Because detecting aberrant responses requires longer tests (Rupp, 2013), we did not examine short tests in the simulations. For the normal class, the slope parameters were generated from lognormal $(0, 0.3^2)$, and the threshold parameters were set between -3 and 3. The proportions for the three latent classes (normal, random-responding, and MRS) had four levels: $[\pi_1, \pi_2, \pi_3] = [1, 0, 0]$, [0.90, 0.08, 0.02], [0.80, 0.16, 0.04], and [0.70, 0.24, 0.06], suggesting a nil, small, medium, and large magnitude of inattentive responses, respectively. The probability of endorsing the middle category was set at 90%. Note that the normal class always had the largest proportion, which was in line with what is commonly observed in practice (Huang et al., 2012; Johnson, 2005; Woods, 2006). The item

responses for the normal class were generated from the GPCM. After the data were simulated using the MMIR-3, four IRT models were fit to the data, including (a) the GPCM, in which all respondents were treated as normal, (b) the MMIR-3 (i.e., the data-generating model), (c) the MMIR-2NR, in which the normal behavior and the random responding behavior were considered but the MRS behavior was not, and (d) the MMIR-2NM, in which the normal behavior and the MRS behavior were considered but the random-responding behavior was not.

Analysis

Data analyses were conducted in WinBUGS. The distributions given above were set as priors. Our pilot study suggested the use of the first 5,000 iterations for burn-in and the following 5,000 iterations per 10 values for parameter estimates. The convergence of the Markov chain was checked using the Geweke diagnostic method (Geweke, 1992) and found to be satisfactory. The bias and root mean square error (RMSE) of parameter estimates were computed to evaluate parameter recovery, and the BIC was used for model comparison among the four fitted models. The assignment of the membership of latent class was determined by the mode of π_i for each respondent. Due to the computational burden (each replication took about 2 to 32 hours to converge with a personal computer with an Intel Core i7-860 processor and 8G RAM), 20 replications were conducted for each condition. We have increased the replications to 100 in the pilot study and concluded that the results under 20 replications were very stable.

When fitting the GPCM to the MMIR-3 data, the l_z statistic (Drasgow, Levine, & Williams, 1985) was used to detect inattentive responses. The l_z statistic is a standardized likelihood-based person-fit index, and its distribution approximately follows the standard normal distribution when the response string is long. A large negative l_z value indicates an aberrant response pattern; the null hypothesis of no aberrant response pattern will be rejected at the .05 nominal level if the l_z value is smaller than -1.645. It was expected that the l_z statistic would yield inadequate Type I error rates and poor power rates when there was a large proportion of inattentive responses.

The following predictions were examined. First, under the nil condition where only attentive responses were involved (the GPCM was the true model), fitting the four models would

all yield good parameter recovery. Very few persons would be misclassified into the random-responding or MRS class. Second, when there were two types of inattentive responses in the data, only fitting the MMIR-3 would yield a good parameter recovery because it was the true model; fitting the MMIR-2NR would yield a better parameter recovery than fitting the MMIR-2NM, because there was a larger proportion of random responding than MRS in the simulated data; fitting the GPCM would yield the worst parameter recovery, because it neglected both types of inattentive response. Third, the more inattentive responses in the simulated data, the worse the parameter estimates would be for the MMIR-2NR, MMIR-2NM, and GPCM. Fourth, using a larger sample size would be helpful to recover generated parameters. Fifth, the mean BIC across MCMC draws would be able to select the true model correctly.

Results

Tables 1 and 2 summarize the bias and RMSE values when the GPCM, MMIR-3, MMIR-2NR, and MMIR-2NM were fit to the simulated MMIR-3 data with 500 simulees. Under the nil condition, where there were no inattentive responses in the simulated data (i.e., $\pi_1 = 1$, $\pi_2 = \pi_3 = 0$), the GPCM and three mixture models yielded small and comparable biases and RMSE values when the tests had 10 or 20 items. Under the MMIR-3, biases for π_1, π_2, and π_3 were -0.009, 0.007, and 0.002, respectively, when tests had 10 items, and were -0.004, 0.002, and 0.002 when tests had 20 items. Fitting the MMIR-2NR and the MMIR-2NM yielded similar results. As anticipated, fitting the unnecessarily complicated models (the MMIR-3, MMIR-2NR, and MMIR-2NM) to data without inattentive responses did little harm to the parameter estimation.

Under conditions where inattentive responses were involved, fitting the GPCM yielded large bias and RMSE values, and the MMIR-3 yielded a good parameter recovery. For example, when there were 10 items and 10% of 500 respondents were inattentive, the mean RMSE values for the slope and threshold parameters were 0.218 and 0.290 for the GPCM, respectively; 0.130 and 0.213 for the MMIR-3, respectively; 0.131 and 0.218 for the MMIR-2NR, respectively; and 0.209 and 0.280 for the MMIR-2NM, respectively. As anticipated, although the MMIR-2NR,

MMIR-2NM, and GPCM were all inappropriate models, the first two models yielded a better recovery than the GPCM; the MMIR-2NR yielded a slightly better recovery than the MMIR-2NM. Furthermore, as the proportion of inattentive respondents increased, the parameter estimates for the GPCM became worse, suggesting that the more inattentive responses, the worse the parameter estimation for the GPCM. This finding also applied to the MMIR-2NR and MMIR-2NM. It is obvious that the slope parameters were seriously underestimated by the GPCM, suggesting that ignoring inattentive responses would result in underestimating item discrimination. Finally, it was evident that a test of 10 items was sufficient to classify inattentive respondents; increasing the test length to 20 items did not significantly improve the parameter estimation of the two models.

[Tables 1 and 2 about here]

Tables 3 and 4 summarize the bias and RMSE values when 2,000 simulees were generated. Likewise, the MMIR-3, MMIR-2NR, and MMIR-2M recovered item parameters as good as the GPCM under the nil condition in a larger sample size; and only the MMIR-3 can recover item parameters precisely under the three inattentive conditions. For the MMIR-3, as expected, the estimation significantly improved when the sample size increased. Figure 1 shows the bias values against the true values of the slope parameters and the threshold parameters in the GPCM. When the proportion of inattentive respondents was 0%, the GPCM yielded unbiased estimates for the slope parameters and the threshold parameters. When inattentive responses were included in the data, the slope parameters were negatively biased; that is, the GPCM tended to underestimate the slope parameters. The estimation also became less accurate as the proportion of inattentive responses increased. Furthermore, the larger the slope parameter was, the more serious the underestimation. Underestimation in the slope parameters was anticipated, because many respondents exhibited either random response or MRS, so that the items failed to accurately discriminate the levels of the latent trait. When the proportion of inattentive respondents was not zero, the bias was positive for the negative threshold parameters but negative for the positive threshold parameters; that is, the negative threshold parameters were

overestimated, whereas the positive threshold parameters were underestimated. In other words, the scale in the GPCM was shrunken when inattentive responses occurred but were ignored. Due to the sufficiently large sample size and approximately noninformative priors, the magnitude of regression toward the mean of prior for parameter estimates under the MMIR-3 was trivial.

[Tables 3 and 4 and Figure 1 about here]

Table 5 shows the means of the Type I error rates and the power rates for l_z in detecting inattentive responses under the GPCM. In the nil condition, when there were 500 simulees, the means of the Type I error rates were 2.21% and 4.98% for the 10- and 20-item tests, respectively; when there were 2,000 simulees, the means of the Type I error rates were 4.09% and 3.49% for the 10- and 20-item tests, respectively. These Type I errors were slightly smaller than the expected 5% nominal level. This slight conservation was consistent with the literature, because l_z did not exactly follow the standard normal distribution (Nering, 1997). When the data consisted of inattentive responses, the means of the Type I error rates became even more conservative. For example, they were 0.97%, 0.14%, and 0.08% for normal respondents in the 10-item tests taken by 2,000 simulees, when the proportion of inattentive respondents was 10%, 20%, and 30%, respectively. The performance of the l_z statistic to detect inattentive responses was unsatisfactory. For the detection of random responses, the mean of the power rates across conditions was 73%. For the detection of MRS, the mean of the power rates across conditions was 14%. In other words, *lz* had an acceptable power in detecting the random-responding class, but a poor power in detecting the MRS class.

[Table 5 about here]

Consider the effectiveness of the mean BIC in model selection. When the sample size was 500, under the nil condition, in which the GPCM was the true model, the mean BIC preferred the GPCM 14 times and 18 times out of 20 replications in the 10- and 20-item tests, respectively, and preferred the MMIR-2NR 6 times and 2 times in the 10- and 20-item tests, respectively; when the sample size was 2,000, the mean BIC preferred the GPCM 7 times and 18 times in the 10- and 20-item tests, respectively, and preferred the MMIR-2NR 13 times and 2 times in the 10-

and 20-item tests, respectively. Even when the true GPCM was not selected, the differences in parameter estimates between the GPCM and MMIR-2NR were very small. Furthermore, the proportion for the random-responding class under the MMIR-2NR was minimal. Such a small proportion for the random-responding class indicates a preference for the simpler GPCM. When there were inattentive responses (the MMIR-3 was the true model), the mean BIC always preferred the MMIR-3. In summary, the mean BIC was powerful in model selection.

Simulation Study 2

Design

The MMIR-3 considers two types of inattentive responses jointly. The necessity of fitting the MMIR-3 to data with only one type of inattentive response was of interest. In simulation study 2, item responses were generated from either the MMIR-2NR or MMIR-2NM. A simple condition was illustrated, in which 2,000 simulees took a survey with 10 five-point items, and the mixture proportions of normal and inattentive respondents were constantly set at 0.9 and 0.1. The MMIR-3 and the data-generating model were fit to the data.

Analysis

The bias and RMSE for each parameter were computed, and the BIC was used for model comparison. Likewise, 20 replications were carried out. It was anticipated that fitting the MMIR-3 would result in accurate parameter estimates, as the case of fitting a more complicated mixture model to the GPCM data in simulation study 1. Because the MMIR-3 includes the MMIR-2NR and MMIR-2NM as some models and one mixture proportion parameter is estimated additionally, the efficiency of the BIC to select the simpler models was tested.

Results

Table 6 summarizes the bias and RMSE values when the MMIR-3 was fit to data. The MMIR-3 yielded an accurate recovery of the item parameters as well as the mixture proportions of latent classes. For example, when data was generated from the MMIR-2NR, the bias for the mixture proportion of the normal and random-responding classes was -0.002 and 0.002, and the estimated mixture proportion of the inexistent MRS class was 0.001. In terms of the performance

of the BIC, the MMIR-2NR was favored 15 times and the MMIR-2NM was favored 11 times, which again supports that fitting the MMIR-3 did little harm and yielded almost identical results to the simpler models.

[Table 6 about here]

Simulation Study 3

Design

In the MMIR, the latent distribution of the normal class is constrained to follow a normal distribution, which is aligned with the settings in most IRT software, such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), ConQuest (Adams, Wu, & Wilson, 2015), and PARSCALE (Muraki & Bock, 1997). The normal distribution might be inappropriate to real situations, however. We conducted this simulation to evaluate the applicability of the MMIR when the latent distribution was incorrectly specified as a normal distribution.

A total of 2,000 simulees answering responding to 10 five-point items were generated. The mixture proportions for the normal, random-responding, and MRS classes were 0.8, 0.16, and 0.04, respectively. Particularly, the intended-to-be-measured latent trait was generated from either a skewed or a bimodal distribution, as illustrated in Figure 2. Then the MMIR-3 was fit to the generated data by constraining the latent trait as a standard normal distribution. Likewise, 20 replications were conducted.

[Figure 2 about here]

Analysis

The correct rate of classification was computed as the dependent variable. Because the referencing distribution was no longer normally distributed, it would influence the scale for item parameters, making that the estimates cannot be directly compared to the true values. Nevertheless, the classification of normal and inattentive respondents was based on the likelihood of response pattern and not affected by the scale of item parameters. It was anticipated that the MMIR-3 would still perform well in detecting inattentive respondents.

Results

When the latent distribution was skewed, the correct classification rates for the normal, random-responding, and MRS classes across replications were 99.5%, 94.6%, and 99.9%, respectively. When the latent distribution was bimodal, the correct classification rates for the normal, random-responding, and MRS classes across replications were 99.7%, 95.9%, and 100%, respectively. In brief, the anticipation was confirmed.

Example 1: The Transport Customer Survey

The objective of the 2011 Transport Customer Survey (TCS) (Bureau of Transport Statistics, 2011) was to provide a robust measure of customers' satisfaction with transport services for New South Wales policy, planning, coordination, community engagement, and contract management. The TCS covered three transport services, including train, bus, and ferry. In this study, we used the train service data, which consisted of 6,183 participants' responses to 25 five-point items (Very dissatisfied = 0, Dissatisfied = 1, Neither = 2, Satisfied = 3, and Very satisfied = 4). The GPCM, MMIR-2NR, MMIR-2NM, and MMIR-3 were fit to the data using WinBUGS. The mean BIC values of the GPCM and MMIR-3 were 309,178 and 297,842, respectively (the mean BIC values of the MMIR-2NR and MMIR-2NM were between 297,842 and 309,178), suggesting that the MMIR-3 had a better fit among the four models. The estimated proportions of the three latent classes (normal, random-responding, and MRS) were 92%, 6%, and 2%, respectively. In other words, 8% of respondents were classified as inattentive respondents. Each respondent was then assigned to one of the three latent classes according to the mode of the posterior probabilities for the three latent classes. To check the consistency of classification, we computed the means and standard deviations (in parenthesis) of the posterior probability for each latent class; they were 0.99 (0.04), 0.93 (0.12), and 0.98 (0.08), respectively, suggesting that the respondents were classified with a very high degree of confidence.

The test reliability of θ was .95 in the GPCM and .87 in the MMIR-3. The difference in the reliability estimates might be because the GPCM treated all participants as normal respondents and ignored the possibility of inattentive responses in the TCS. A comparison on the standard errors of the θ estimates between the GPCM and MMIR-3 for the three latent classes revealed

that the GPCM tended to yield smaller standard errors, particularly for those respondents who were classified as inattentive in the MMIR-3 (Figure 3). Compared to the MMIR-3, the GPCM tended to yield smaller slope parameter estimates and shrunken threshold parameter estimates (Figure 4), which was consistent with the findings in the simulation studies in that the GPCM tended to underestimate the slope parameters and yield shrunken threshold parameter estimates.

[Figures 3 and 4 about here]

Table 7 summarizes the distributions of l_z for the normal, random-responding, and MRS latent classes when the GPCM was fit. Among the 6,163 respondents, 83.4% of them had l_z statistics < -1.645 and were thus identified as inattentive. More specifically, among those respondents that were classified into the normal, random-responding, and MRS latent classes by the MMIR-3, 11.3%, 100% and 3.0% of them were identified as inattentive by l_z, respectively. It suggested that the MMIR-3 and *lz* were in 89% and 100% agreements on the response patterns classified as normal and random responding, respectively, but they showed strong a substantial disagreement (only 3%) in MRS detection.

[Table 7 about here]

Table 8 lists the raw responses of 11 customers and their θ estimates under the MMIR-3 and GPCM. The MMIR-3 clearly classified the first nine cases into three groups: cases 1-3 as the normal class; cases 4-6 as the random-responding class; cases 7 and 8 as the random-responding class. The θ estimates for inattentive respondents (i.e., cases 4-8) under the MMIR-3 were close to zero and the standard errors were close to unity, suggesting that the point estimates and standard errors were mainly attributed to the prior (the standard normal distribution). For cases 7 and 8, they were identified as the normal cases by l_z but as the inattentive class by the MMIR-3. Furthermore, the responses patterns of cases 9-11 were too atypical to assign the cases to any latent class.

[Table 8 about here]

Figure 5 summarizes the response profiles of the 25 items for the normal (5a), random-responding (5b), and MRS (5c) latent classes. The response profiles of the 25 items were

sorted according to the mean of the four thresholds in the normal class. In the normal class, there was a large percentage of Dissatisfied, Neither, and Satisfied responses, and a small percentage of Very Dissatisfied and Very Satisfied responses. In the random-responding class, the distribution of the five categories appeared relatively uniform for most items. In the MRS class, the middle category had a percentage as high as approximately 90%. To sum up, the MMIR-3 had a better fit than the GPCM, and a small proportion of respondents exhibited random responses or MRS based on the MMIR-3 suggestions.

We further investigated whether the classification under the MMIR-3 was related to respondents' gender, age, and the frequency of traveling via multinomial logistic regression. Consequently, the estimates of the pseudo R^2 were approximately zero (around .01), suggesting that the performance of inattentive responding was unrelated to respondents' gender, age, and the frequency of traveling.

[Figure 5 about here]

Example 2: The Egypt and Saudi Arabia Survey

The Egypt and Saudi Arabia Survey (ESAS) assessed the values, general opinions, and sociopolitical and cultural attitudes of youths in Egypt and Saudi Arabia (Moaddel, Karabenick, & Thornton, 2010). A representative sample of 928 youths answered the Egyptian survey, and a representative sample of 954 youths responded to the Saudi Arabian survey. A subscale in the ESAS measured the extent to which a respondent believed and trusted different information sources, including parents, teachers at school or university, secular teachers, friends, religious leaders, media, satellite TV, and the Internet. This subscale had 35 five-point items (0 = Do not rely, 1 = Generally do not rely, 2 = Sometimes reply and sometimes not, 3 = Generally rely, and 4 = Rely completely). The GPCM and the MMIR-3 were fit to the subscale using WinBUGS. To identify the model, the latent distribution for the Saudi Arabian sample was assumed to follow the standard normal distribution, whereas the mean and variance of the latent distribution for the Egyptian sample were freely estimated.

The mean BIC values of the GPCM and the MMIR-3 were 178,432 and 173,136, respectively, suggesting that the MMIR-3 had a better fit (we also fit the MMIR-2NR and the MMIR-2NM, and their mean BIC values were larger than that of the MMIR-3). In the MMIR-3, the estimated proportions of the normal, random-responding, and MRS latent classes were 64%, 34%, and 2%, respectively, for the Egyptian sample and 83%, 14%, and 3% for the Saudi Arabian sample. The Egyptian sample had a smaller proportion of normal respondents and a larger proportion of random-responding respondents than the Saudi Arabian sample. The means and standard deviations (in parenthesis) of the posterior probability for the three latent classes were .98 (0.08), .93 (0.13), and .98 (0.06), indicating a very high degree of confidence in the classification.

Consistent with the findings of the first example, the GPCM yielded smaller standard errors for the θ estimates and slope parameter estimates, as compared to the MMIR-3. The test reliability was .91 for the Egyptian sample and .85 for the Saudi Arabian sample under the GPCM, and .58 for the Egyptian sample and .71 for the Saudi Arabian sample under the MMIR-3. Figure 6 shows the slope and threshold parameter estimates under the GPCM against those under the MMIR-3. The threshold parameter estimates obtained from these two models were very different, which was because of a large proportion of inattentive respondents (approximately 26.3%) in the full dataset of the two countries.

[Figure 6 about here]

We investigated the difference in receptiveness of information between Saudi Arabians and Egyptians and found Saudi Arabians exhibited a higher receptiveness than the Egyptian by 1.54 (Cohen's $D = 1.50$) under the GPCM but 1.74 (Cohen's $D = 1.58$) under the MMIR-3. The discrepancy was because inattentive respondents were not excluded in the GPCM. As shown in Figure 7, the estimates for inattentive respondents under the GPCM were very diverse, but those under the MMIR-3 were centralized to the grand means of the two samples (around 0 for Saudi Arabians and -1.7 for Egyptians). Apparently, including inattentive respondents would attenuate the difference and effect size between the two samples.

[Figure 7 about here]

Conclusion and Discussion

In real-life surveys, it is common that some respondents do not respond to questions conscientiously and thus yield inattentive responses. Inattentive responses will reduce the quality of measurements, if they are not properly accounted for. Existing methods provide several solutions to inattentive responses, but they suffer from practical constraints, such as the need for recording response time, adding additional items, and lack of agreement in setting cut-points. This study recognizes the obstacles and resolves these problems by developing a set of mixture IRT models applicable for common data collection. The newly developed MMIR can classify different kinds of inattentive responses effectively by specifying response probabilities for each inattentive latent class and is easy for further extensions.

In this study, we focused on two types of inattentive response: random responses and MRS. The parameters in the MMIR were estimated using a Bayesian approach with MCMC methods. A series of simulations were conducted to evaluate parameter recovery of the MMIR and to analyze the consequences when inattentive responses were ignored by fitting standard IRT models. The results confirmed the utility of the proposed model, showing evidence of satisfactory parameter recovery. In addition, fitting the MMIR to data that did not include any inattentive responses did little harm to the parameter estimation; however, fitting standard IRT models to data containing inattentive responses yielded biased parameter estimates and inflated estimates of test reliability.

Once inattentive respondents are identified using the MMIR, they can be excluded from the entire dataset, so that the resulting parameter estimates will not be biased and the test reliability will be accurately estimated. The present study also examined conventional approaches to detecting inattentive responses by fitting the GPCM and adopting l_z. The findings suggested that fitting the GPCM to data with inattentive responses underestimated the slope parameters and shrank the threshold parameters, and that l_z failed to distinguish normal from inattentive responses. As expected, the mean EBIC statistic was useful for model comparison.

In addition to the simulation studies, two empirical examples were discussed. Example 1 contained a relative small percentage of inattentive respondents; Example 2 contained two samples and a larger proportion of inattentive respondents. Both examples suggested that the MMIR was a better fit than the GPCM. Consistent with findings in the simulations, ignoring inattentive responses in these empirical examples by fitting the GPCM yielded inaccurate parameter estimates and inflated estimates of test reliability.

Although this study focused on two types of inattentive responses, the MMIR is general enough to also account for other kinds of inattentive responses. For example, we further examined the ESAS dataset and found that 43 respondents often chose the two extreme categories among the 35 items (extreme response styles). When necessary, the MMIR can be easily extended to four latent classes to detect extreme response styles, together with the existing three latent classes. Parameterization in the MMIR can be adjusted in accordance with practical needs. In practice, some practitioners may prefer even numbers of response categories to push MRS respondents toward one end or the other of a dimension (Bradburn, Sudman, & Wansink, 2004). In such cases, the middle option is no longer available, so that the MMIR-3 and MMIR-2NM become inapplicable and the MMIR-2NR is more appropriate. Furthermore, the latent distribution of the normal class in the MMIR is assumed to be normal. One can release such constraint by choosing an item and fixing the slope and one of the threshold parameters to constants (e.g., $\alpha_{11} = 1$ and $\beta_{111} = 0$), so that a non-normal distribution function (such as student's *t*-distribution or logistic distribution) can be applied.

Screening out aberrant responses (e.g., faking) is important, and can improve measurement quality. However, it comes at a price: mistakenly classifying normal respondents as inattentive respondents may have unintended consequences. Finally, yet importantly, simply analyzing response patterns may not be sufficient for the identification of inattentive responses. Information such as response time or video recording, if available, is often very helpful. Future studies can be conducted to incorporate other kinds of inattentive responses and evaluate the performance of the MMIR in these complicated situations.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2015). *ACER ConQuest: Generalised Item Response Modelling Software* [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

Arbisi, P. A., & Ben-Porath, Y. S. (1995). An MMPI-2 infrequent response scale for use with psychopathological populations: The Infrequency-Psychopathology Scale. *Psychological Assessment, 7,* 424-431. doi: 10.1037/1040-3590.7.4.424

Archer, R. P., & Elkins, D. E. (1999). Identification of random responding on the MMPI-A. *Journal of Personality Assessment, 73,* 407-421. doi: 10.1207/S15327752JPA7303_8

Archer, R. P., Fontaine, J., & McCrae, R. R. (1998). Effects of two MMPI-2 validity scales on basic scale relations to external criteria. *Journal of Personality Assessment, 70,* 87-102. doi: 10.1207/s15327752jpa7001_6

Baer, R. A., Ballenger, J., Berry, D. T. R., & Wetter, M. W. (1997). Detection of random responding on the MMPI-A. *Journal of Personality Assessment, 68,* 139-151. doi: 10.1207/s15327752jpa6801_11

Bassili, J. N., & Scott, S. B. (1996). Response latency as a signal to question problems in survey research. *The Public Opinion Quarterly, 60,* 390-399. doi: 10.2307/2749743

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38,* 143-156. doi: 10.1509/jmkr.38.2.143.18840

Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). Response biases in marketing research. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research: Uses, misuses, and future advances* (pp. 95-110). Thousand Oaks, CA: Sage.

Bishop, G. R., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly, 50,* 240-250. doi: 10.1086/268978

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design - For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco, CA: John Wiley & Sons.

Bureau of Transport Statistics (2011). *2011 Transport Customer Survey Train, Bus and Ferry*. New South Wales Government.

Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47,* 87-111. doi: 10.1146/annurev.psych.47.1.87

Butcher, J., Williams, C., Graham, J., Archer, R. P., Tellegen, A., Ben-Porath, Y., & Kaemmer, B. (1992). *Manual for the administration, scoring, and interpretation of the MMPI–A.* Minneapolis: University of Minnesota Press.

Carter, N. T., Dalal, D. K., Lin, B. C., Lake, C., & Zickar, M. J. (2011). Using mixed-model item response theory to ask questions concerning scale usage: An illustration using the Job Descriptive Index. *Organizational Research Methods 14,* 116-146. doi: 10.1177/1094428110363309

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6,* 170-175. doi: 10.1111/j.1467-9280.1995.tb00327.x

Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology, 21,* 203-225. doi: 10.1002/acp.1337

Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling: A Multidisciplinary Journal, 21,* 375-395. doi: 10.1080/10705511.2014.915371

Costa, P. T., & McCrae, R. R. (2008). The revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment: Personality measurement and testing* (pp. 179-198). London: Sage.

Curran, P. G. (2015). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4-19. doi: 10.1016/j.jesp.2015.07.006

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity, 44*, 761-775. doi: 10.1007/s11135-009-9225-z

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559. doi: 10.1007/S11336-008-9092-X

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Leeuw, E. D., & Hox, J. J. (1994). Are inconsistent respondents consistently inconsistent? A study of several nonparametric person fit indices. In J.J. Hox & W. Jansen (Eds)., *Measurement problems in the social sciences* (pp. 67-87). Amsterdam: SISWO.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*, 171-181. doi: 10.1002/job.1962

Draney, K., Wilson, M., Glück, J., & Spiel, C. (2007). Mixture models in developmental context. In G. R. Hancock & K. M. Samuelsen (Eds)., *Advances in Latent Variable Mixture Models* (pp. 199-216). Charlotte, NC: Information Age Publishing.

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

Eid, M. & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*, 20-30. doi: 10.1027//1015-5759.16.1.20

Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. Von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

Fervaha, G., & Remington, G. (2013). Invalid responding in questionnaire-based research: Implications for the study of schizotypy. *Psychological Assessment, 25*, 1355-1360. doi: 10.1037/a0033520

Fong, D. Y., Ho, S. Y., & Lam, T. H. (2010). Evaluation of internal reliability in the presence of inconsistent responses. *Health and Quality of Life Outcomes, 8,* 27. doi: 10.1186/1477-7525-8-27

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*, 159-180. doi: 10.1007/s11336-003-1081-5

Greene, R. L. (1978). An empirically derived MMPI carelessness scale. *Journal of Clinical Psychology, 34*, 407-410. doi: 10.1002/1097-4679(197804)34:2<407::aid-jclp2270340231>3.0.co;2-a

Greene, R. L. (1979). Response consistency on the MMPI: The TR index. *Journal of Personality Assessment, 43,* 69-77. doi: 10.1207/s15327752jpa4301_10

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2007). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44*, 932-942. doi: 10.1016/j.paid.2007.10.034

Harzing, A.-W., Brown, M., Köster, K., & Zhao, S. M. (2012). Response style differences in cross-national research. *Management International Review, 52*, 341-363. doi: 10.1007/s11575-011-0111-2

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*, 104-121. doi: 10.1177/109442819800100106

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30,* 299-311. doi: 10.1007/s10869-014-9357-6

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. doi: 10.1007/s10869-011-9231-8

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

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & Van Den Bergh, H. (2011). Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Processes, 48*, 355-385. doi: 10.1080/0163853X.2011.578910

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298. doi: 10.1207/s15324818ame1604_2

Kopf, J., Zeileis, A., & Strobl, C. (2015b). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. Educational and Psychological Measurement, 75, 22-56.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment, 76*, 315-332. doi: 10.1207/s15327752jpa7602_12

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. doi: 10.1016/j.jrp.2013.09.008

McKibben, W. B., & Silvia, P. J. (2015). Evaluating the distorting effects of inattentive responding and social desirability on self-report scales in creativity and the arts. *The Journal of Creative Behavior*. doi:10.1002/jocb.86

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

Moaddel, M., Karabenick, S. A., & Thornton, A. (2010). Youth, emotional energy, and political violence: The cases of Egypt and Saudi Arabia Survey, 2005. Inter-university Consortium for Political and Social Research (ICPSR) [distributor].

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. doi: 10.1177/014662169201600206

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data* [Computer software]. Chicago, IL: Scientific Software International.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127. doi: 10.1177/01466216970212002

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: rationale, development, and empirical trials. *Journal of Clinical Psychology, 45*, 239-250. doi: 10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

Nicholson, R. A., & Hogan, R. (1990). The construct validity of social desirability. *American Psychologist, 45*, 290-292. doi: 10.1037/0003-066x.45.2.290

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609. doi: 10.1037/0022-3514.46.3.598

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78*, 582-593. doi: 10.1037/0022-3514.78.3.582

Pinsoneault, T. B. (1998). A variable response inconsistency scale and a true response Inconsistency scale for the Jesness Inventory. *Psychological Assessment, 10*, 21-32. doi: 10.1037/1040-3590.10.1.21

Plummer, M. (2012). JAGS version 3.3.0 user manual. Retrieved from http://mcmc-jags.sourceforge.net/

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review, 26,* 195-239. doi: 10.1137/1026034

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16,* 19-31. doi: 10.2307/40212571

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55,* 3-38.

Schinka, J. A., Kinder, B. N., & Kremer, T. (1997). Research validity scales for the NEO-PI-R: Development and initial validation. *Journal of Personality Assessment, 68,* 127-138. doi: 10.1207/s15327752jpa6801_10

Si, S. X., & Cullen, J. B. (1998). Response categories and potential cultural bias: Effects of an explicit middle point in cross-cultural surveys. *International Journal of Organizational Analysis, 6,* 218-230. doi: 10.1108/eb028885

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2007). WinBUGS version 1.4.3. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs

Stein, L. A. R., Graham, J. R., & Williams, C. L. (1995). Detecting fake-bad MMPI-A profiles. *Journal of Personality Assessment, 65,* 415-427. doi: 10.1207/s15327752jpa6503_3

Stening, B. W., & Everett, J. E. (1984). Response styles in a cross-cultural managerial study. *The Journal of Social Psychology, 122,* 151-156. doi: 10.1080/00224545.1984.9713475

Wang, W.-C., Chen, H.-F., & Jin, K.-Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement, 75,* 157-178. doi: 10.1177/0013164414528209

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18,* 320-334. doi: 10.1037/a0032121

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38. doi: 10.1111/j.1745-3984.2006.00002.x

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*, 186-191. doi: 10.1007/s10862-005-9004-7

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiplegroup IRT analysis and test maintenance for binary items* [Computer software]. Chicago: Scientific Software International.

Appendix: WinBUGS code for the MMIR-3

```
# N is the number of respondents;
# I is the number of five-point Likert-type items;
# r is the data matrix with N rows and I columns;
# theta is the measured latent trait;
# a is the slope parameter;
# b is the threshold parameter;
# and class is the group membership.

model {
    for (n in 1:N) {
        class[n] ~ dcat(pi[])
        theta[n] ~ dnorm(0, 1)

        for (i in 1:I) {
            Q[n,i,1] <- 1
            Q[n,i,2] <- exp(a[i,class[n]]*1*theta[n] - b[i,1,class[n]])
            Q[n,i,3] <- exp(a[i,class[n]]*2*theta[n] - b[i,1,class[n]] -
b[i,2,class[n]])
            Q[n,i,4] <- exp(a[i,class[n]]*3*theta[n] - b[i,1,class[n]] -
b[i,2,class[n]] - b[i,3,class[n]])
            Q[n,i,5] <- exp(a[i,class[n]]*4*theta[n] - b[i,1,class[n]] -
b[i,2,class[n]] - b[i,3,class[n]] - b[i,4,class[n]])

            denom[n,i] <- sum(Q[n,i,])
            PP[n,i,1] <- Q[n,i,1]/denom[n,i]
            PP[n,i,2] <- Q[n,i,2]/denom[n,i]
            PP[n,i,3] <- Q[n,i,3]/denom[n,i]
            PP[n,i,4] <- Q[n,i,4]/denom[n,i]
            PP[n,i,5] <- Q[n,i,5]/denom[n,i]
            r[n,i] ~ dcat(PP[n,i,])
        }
    }

    # Priors

    alpha <- c(1,1,1)
    pi[1:3] ~ ddirch(alpha[1:3])
```

```
    for (i in 1:I) {
        # Class 1: normal class

        a[i,1] ~ dlnorm(0, 0.1)
        b[i,1,1] ~ dnorm(0, 0.1)
        b[i,2,1] ~ dnorm(0, 0.1)
        b[i,3,1] ~ dnorm(0, 0.1)
        b[i,4,1] ~ dnorm(0, 0.1)

        # Class 2: random-responding class
        a[i,2] <- 0
        b[i,1,2] <- 0
        b[i,2,2] <- 0
        b[i,3,2] <- 0
        b[i,4,2] <- 0

        # Class 3: MRS class
        a[i,3] <- 0
        b[i,1,3] <- 0
        b[i,2,3] <- -3.58
        b[i,3,3] <- 3.58
        b[i,4,3] <- 0
    }
}
```

Table 1. Summary statistics for the parameter estimates under the GPCM and MMIR-3 (N = 500)

	GPCM								MMIR-3							
	Nil		10%		20%		30%		Nil		10%		20%		30%	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
10 *Items*																
Slope																
Max	0.049	0.145	0.019	0.437	-0.076	0.838	-0.116	0.799	0.097	0.161	0.169	0.227	0.088	0.209	0.152	0.275
Min	-0.008	0.061	-0.434	0.073	-0.832	0.106	-0.795	0.133	0.000	0.063	0.003	0.059	-0.017	0.087	-0.003	0.080
Mean	0.028	0.102	-0.178	0.218	-0.333	0.351	-0.395	0.406	0.039	0.108	0.044	0.130	0.019	0.128	0.063	0.150
Threshold																
Max	0.116	0.430	0.532	1.152	1.911	2.168	1.710	1.726	0.122	0.476	0.109	0.403	0.179	0.762	0.201	0.552
Min	-0.099	0.099	-1.121	0.116	-2.141	0.128	-1.225	0.109	-0.196	0.100	-0.230	0.116	-0.133	0.140	-0.164	0.094
Mean	0.010	0.176	-0.019	0.290	0.043	0.507	0.093	0.469	0.003	0.181	-0.002	0.213	0.006	0.270	0.000	0.255
Mixture Proportion																
π_1	–	–	–	–	–	–	–	–	-0.009	0.010	-0.008	0.010	-0.006	0.009	-0.011	0.016
π_2	–	–	–	–	–	–	–	–	0.007	0.008	0.006	0.008	0.004	0.008	0.010	0.015
π_3	–	–	–	–	–	–	–	–	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.002
20 *Items*																
Slope																
Max	0.064	0.210	-0.048	0.654	-0.041	0.592	-0.156	0.665	0.067	0.214	0.060	0.185	0.064	0.218	0.108	0.188
Min	-0.017	0.055	-0.652	0.079	-0.590	0.062	-0.662	0.163	-0.018	0.056	-0.027	0.055	-0.014	0.058	-0.037	0.061
Mean	0.023	0.103	-0.174	0.199	-0.226	0.247	-0.381	0.389	0.024	0.105	0.009	0.094	0.022	0.122	0.016	0.122
Threshold																
Max	0.128	0.455	0.771	1.583	1.794	1.809	1.671	1.813	0.125	0.454	0.180	0.458	0.250	0.654	0.080	0.755
Min	-0.079	0.088	-1.554	0.095	-0.813	0.089	-1.796	0.084	-0.082	0.089	-0.223	0.091	-0.101	0.119	-0.230	0.121
Mean	0.006	0.200	-0.042	0.307	-0.004	0.359	-0.077	0.483	0.004	0.200	-0.011	0.202	0.004	0.226	-0.012	0.243
Mixture Proportion																
π_1	–	–	–	–	–	–	–	–	-0.004	0.004	-0.004	0.005	-0.004	0.006	-0.005	0.007
π_2	–	–	–	–	–	–	–	–	0.002	0.002	0.002	0.004	0.003	0.005	0.004	0.005
π_3	–	–	–	–	–	–	–	–	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002

Note. – = not applicable.

Table 2. Summary statistics for the parameter estimates under the MMIR-2NR and MMIR-2NM (N = 500)

	MMIR-2NR								MMIR-2NM							
	Nil		10%		20%		30%		Nil		10%		20%		30%	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
10 *Items*																
Slope																
Max	0.096	0.164	0.196	0.248	0.095	0.209	0.168	0.290	0.052	0.147	-0.002	0.407	-0.086	0.809	-0.123	0.776
Min	-0.001	0.063	-0.029	0.063	-0.008	0.090	-0.209	0.086	-0.008	0.061	-0.403	0.077	-0.804	0.112	-0.772	0.139
Mean	0.040	0.109	0.036	0.131	0.020	0.130	-0.001	0.158	0.030	0.103	-0.173	0.209	-0.329	0.344	-0.401	0.411
Threshold																
Max	0.127	0.482	0.134	0.407	0.179	0.822	0.730	0.774	0.111	0.438	0.529	1.078	1.871	2.134	1.703	1.719
Min	-0.190	0.102	-0.169	0.119	-0.178	0.136	-0.528	0.113	-0.104	0.098	-1.045	0.117	-2.105	0.125	-1.181	0.115
Mean	0.003	0.182	-0.006	0.218	0.002	0.276	0.009	0.332	0.008	0.176	-0.019	0.280	0.043	0.491	0.093	0.453
Mixture Proportion																
π_1	-0.007	0.008	0.004	0.009	-0.006	0.010	0.045	0.048	-0.002	0.002	0.078	0.078	0.157	0.157	0.238	0.238
π_2	0.007	0.008	0.016	0.018	0.046	0.047	0.015	0.023	–	–	–	–	–	–	–	–
π_3	–	–	–	–	–	–	–	–	0.002	0.002	0.002	0.003	0.003	0.003	0.002	0.002
20 *Items*																
Slope																
Max	0.067	0.206	0.065	0.190	0.120	0.230	0.105	0.181	0.065	0.210	-0.056	0.614	-0.053	0.565	-0.161	0.655
Min	-0.016	0.056	-0.021	0.054	-0.146	0.062	-0.037	0.061	-0.017	0.056	-0.611	0.084	-0.563	0.072	-0.652	0.167
Mean	0.024	0.105	0.009	0.094	0.016	0.135	0.012	0.123	0.023	0.103	-0.170	0.194	-0.233	0.252	-0.386	0.394
Threshold																
Max	0.132	0.454	0.152	0.447	0.457	0.765	0.109	0.744	0.127	0.453	0.745	1.491	1.764	1.781	1.636	1.785
Min	-0.081	0.089	-0.226	0.092	-0.276	0.115	-0.263	0.124	-0.080	0.087	-1.459	0.093	-0.809	0.093	-1.766	0.090
Mean	0.005	0.200	-0.011	0.202	0.022	0.266	-0.014	0.250	0.006	0.200	-0.040	0.298	-0.006	0.352	-0.076	0.474
Mixture Proportion																
π_1	-0.002	0.002	-0.001	0.003	0.027	0.030	0.000	0.010	-0.002	0.002	0.078	0.078	0.158	0.158	0.238	0.238
π_2	0.002	0.002	0.021	0.022	0.013	0.017	0.060	0.060	–	–	–	–	–	–	–	–
π_3	–	–	–	–	–	–	–	–	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002

Note. – = not applicable.

Table 3. Summary statistics for the parameter estimates under the GPCM and MMIR-3 (N = 2,000)

	GPCM								MMIR-3							
	Nil		10%		20%		30%		Nil		10%		20%		30%	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
10 *Items*																
Slope																
Max	0.020	0.115	-0.064	0.396	-0.099	0.943	-0.228	0.871	0.020	0.113	0.044	0.108	0.042	0.187	0.040	0.186
Min	-0.019	0.036	-0.393	0.089	-0.941	0.107	-0.866	0.229	-0.018	0.035	-0.018	0.046	-0.002	0.054	-0.020	0.042
Mean	0.005	0.065	-0.195	0.203	-0.395	0.400	-0.461	0.463	0.006	0.065	0.012	0.068	0.019	0.083	0.010	0.087
Threshold																
Max	0.059	0.303	0.709	0.727	1.184	1.822	0.681	1.205	0.063	0.302	0.082	0.226	0.040	0.223	0.097	0.261
Min	-0.028	0.050	-0.463	0.053	-1.818	0.046	-1.200	0.073	-0.024	0.050	-0.063	0.059	-0.079	0.054	-0.057	0.049
Mean	0.008	0.121	0.029	0.232	-0.003	0.434	-0.139	0.372	0.009	0.120	0.002	0.114	-0.009	0.120	0.000	0.135
Mixture Proportion																
π_1	–	–	–	–	–	–	–	–	-0.001	0.001	-0.003	0.005	-0.002	0.005	-0.002	0.002
π_2	–	–	–	–	–	–	–	–	0.001	0.001	0.003	0.005	0.001	0.005	0.001	0.002
π_3	–	–	–	–	–	–	–	–	0.000	0.000	0.001	0.001	0.000	0.001	0.000	0.001
20 *Items*																
Slope																
Max	0.054	0.128	-0.058	0.675	-0.101	0.809	-0.157	1.108	0.055	0.128	0.063	0.116	0.057	0.114	0.079	0.158
Min	-0.025	0.035	-0.673	0.063	-0.808	0.104	-1.108	0.159	-0.027	0.035	-0.008	0.033	-0.005	0.035	-0.030	0.040
Mean	0.000	0.059	-0.190	0.198	-0.316	0.321	-0.490	0.492	0.000	0.059	0.018	0.063	0.020	0.067	0.010	0.078
Threshold																
Max	0.060	0.258	1.026	1.200	2.221	2.226	3.106	3.109	0.064	0.259	0.136	0.268	0.067	0.361	0.140	0.529
Min	-0.062	0.039	-1.191	0.054	-1.945	0.058	-2.674	0.065	-0.065	0.039	-0.049	0.058	-0.116	0.061	-0.096	0.052
Mean	0.001	0.109	0.005	0.236	0.089	0.386	0.129	0.581	0.001	0.109	0.005	0.106	-0.005	0.113	0.008	0.135
Mixture Proportion																
π_1	–	–	–	–	–	–	–	–	-0.001	0.001	-0.002	0.002	-0.002	0.003	-0.001	0.001
π_2	–	–	–	–	–	–	–	–	0.001	0.001	0.001	0.002	0.001	0.003	0.001	0.001
π_3	–	–	–	–	–	–	–	–	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.000

Note. – = not applicable.

Table 4. Summary statistics for the parameter estimates under the MMIR-2NR and MMIR-2NM (N = 2,000)

	MMIR-2NR								MMIR-2NM							
	Nil		10%		20%		30%		Nil		10%		20%		30%	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
10 *Items*																
Slope																
Max	0.019	0.114	0.037	0.106	0.058	0.192	0.128	0.188	0.021	0.113	-0.077	0.376	-0.111	0.926	-0.246	0.892
Min	-0.016	0.035	-0.011	0.048	-0.010	0.058	-0.121	0.043	-0.018	0.036	-0.372	0.094	-0.925	0.117	-0.887	0.247
Mean	0.006	0.064	0.005	0.066	0.019	0.085	-0.004	0.105	0.005	0.064	-0.189	0.197	-0.392	0.396	-0.473	0.475
Threshold																
Max	0.064	0.298	0.114	0.214	0.080	0.263	0.612	0.625	0.060	0.300	0.573	0.588	1.142	1.786	0.688	1.197
Min	-0.027	0.051	-0.103	0.060	-0.107	0.065	-0.568	0.048	-0.025	0.050	-0.441	0.054	-1.782	0.062	-1.193	0.060
Mean	0.010	0.111	0.009	0.125	-0.011	0.130	0.009	0.294	0.010	0.111	0.040	0.207	-0.141	0.411	-0.117	0.368
Mixture Proportion																
π_1	-0.001	0.001	0.008	0.009	-0.001	0.007	0.058	0.058	0.000	0.000	0.079	0.079	0.159	0.159	0.239	0.239
π_2	0.001	0.001	0.012	0.013	0.041	0.042	0.002	0.008	–	–	–	–	–	–	–	–
π_3	–	–	–	–	–	–	–	–	0.000	0.000	0.001	0.001	0.001	0.002	0.001	0.001
20 *Items*																
Slope																
Max	0.054	0.035	0.091	0.034	0.062	0.037	0.084	0.039	0.054	0.035	-0.064	0.071	-0.110	0.113	-0.168	0.169
Min	-0.026	0.035	-0.033	0.034	-0.027	0.037	-0.027	0.039	-0.025	0.035	-0.630	0.071	-0.798	0.113	-1.098	0.169
Mean	0.000	0.059	0.015	0.064	0.021	0.068	0.013	0.079	-0.001	0.059	-0.190	0.198	-0.322	0.326	-0.498	0.500
Threshold																
Max	0.060	0.259	0.093	0.269	0.066	0.350	0.147	0.534	0.061	0.259	0.984	1.134	2.191	2.197	3.081	3.084
Min	-0.059	0.039	-0.123	0.057	-0.138	0.060	-0.092	0.051	-0.063	0.038	-1.124	0.056	-1.888	0.067	-2.656	0.058
Mean	0.000	0.109	0.004	0.110	-0.005	0.115	0.009	0.136	0.001	0.109	0.007	0.228	0.088	0.378	0.127	0.565
Mixture Proportion																
π_1	-0.001	0.001	0.005	0.005	0.000	0.003	-0.002	0.002	0.000	0.000	0.079	0.079	0.160	0.160	0.239	0.239
π_2	0.001	0.001	0.015	0.016	0.040	0.040	0.062	0.062	–	–	–	–	–	–	–	–
π_3	–	–	–	–	–	–	–	–	0.000	0.000	0.001	0.001	0.000	0.001	0.001	0.001

Note. – = not applicable.

Table 5. Mean Type I error rates and power rates (%) of the l_z statistic in detecting aberrant responses with the GPCM

	Proportion of inattentive responses			
Latent class	Nil	10%	20%	30%
500 simulee, 10 items				
Normal	*2.21*	*0.84*	*0.16*	*0.03*
Random	–	75.00	69.75	44.75
MRS	–	22.50	12.50	0.00
500 simulee, 20 items				
Normal	*4.98*	*0.57*	*0.16*	*0.01*
Random	–	*95.38*	*81.44*	*79.92*
MRS	–	*79.00*	*0.75*	*0.50*
2,000 simulee, 10 items				
Normal	*4.09*	*0.97*	*0.14*	*0.08*
Random	–	75.63	66.14	37.19
MRS	–	19.50	4.88	0.08
2,000 simulee, 20 items				
Normal	*3.49*	*0.47*	*0.06*	*0.02*
Random	–	92.56	84.53	76.24
MRS	–	23.00	1.69	0.08

Note. The italic numbers denote Type I error rates and the remaining numbers denote power rates; Class 1 = normal class, Class 2 = random-responding class, and Class 3 = midpoint class.

Table 6. Summary statistics for the parameter estimates under the MMIR-3 when data was generated from the MMIR-2NR and MMIR-2NM

	Data generating model			
	MMIR-2NR		MMIR-2NM	
	Bias	RMSE	Bias	RMSE
Slope				
Max	0.036	0.121	0.037	0.093
Min	-0.010	0.036	-0.021	0.044
Mean	0.014	0.066	0.005	0.065
Threshold				
Max	0.084	0.296	0.060	0.168
Min	-0.046	0.049	-0.073	0.060
Mean	0.004	0.107	0.008	0.097
Mixture proportion				
π_1	-0.002	0.005	-0.001	0.002
π_2	0.002	0.004	0.001	0.001
π_3	0.001	0.001	0.000	0.001

Table 7. Classification of latent classes and l_z statistic in Example 1

Latent class	*N*	Mean	SD	Minimum	Maximum	l_z < -1.645
Normal	5,702	0.41	1.60	-7.75	3.22	11.3%
Random	380	-5.54	2.42	-20.08	-2.45	100.0%
MRS	101	-0.25	0.93	-6.87	0.78	3.0%
Total	6,163	0.03	2.18	-20.08	3.22	16.6%

Table 8. Customers' response patterns, θ estimates and posteriori probabilities of classifications under the MMIR3, and θ estimates and l_z under the GPCM in Example 1

		MMIR-3				GPCM	
No.	Response Pattern	θ estimate	Normal (%)	Random (%)	MRS (%)	θ estimate	l_z
	Good classifications						
1	14244344414344443423424444	0.00 (0.22)	**100**	0	0	0.03 (0.22)	0.44
2	34443444444444444433444434	0.00 (0.21)	**100**	0	0	0.03 (0.22)	1.40
3	52344414454343345344454454	0.00 (0.22)	**100**	0	0	0.04 (0.23)	-1.51
4	41434121441454544334554555	0.00 (0.99)	2	**98**	0	-0.29 (0.20)	-4.01
5	54143454515425544443514544	0.00 (0.98)	0	**100**	0	0.09 (0.24)	-5.47
6	13544314243332245455454355	0.00 (0.98)	5	**95**	0	-0.27 (0.19)	-3.85
7	33334343334333343343343333	0.00 (1.04)	1	0	**99**	-0.65 (0.17)	-0.07
8	33333333333333333333333333	0.00 (1.03)	0	0	**100**	-0.87 (0.17)	-0.16
	Poor classifications						
9	44235444454415543532544455	0.08 (0.69)	**52**	48	0	0.21 (0.23)	-3.78
10	52344542445345445234453 53	0.11 (0.81)	41	**59**	0	0.07 (0.23)	-3.74
11	43335333442353453534333 33	-0.04 (1.01)	1	39	**60**	-0.44 (0.18)	-3.72

Note. Items are sorted according to the mean of the four thresholds in the normal class. Values in parenthesis are standard errors.

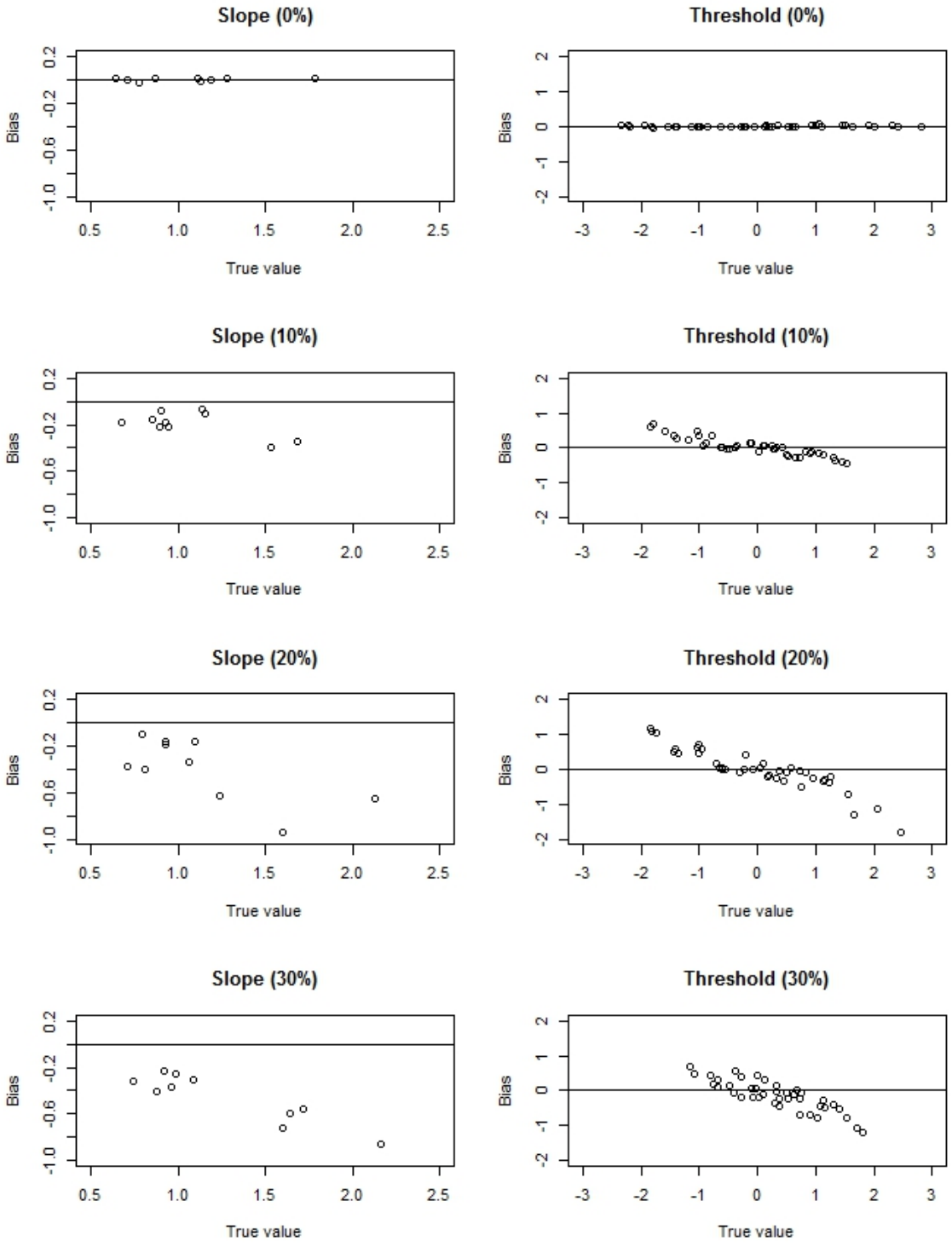

Figure 1. Bias for slope and threshold parameters under the GPCM when sample size was 2,000 simulees and test length was 10 items

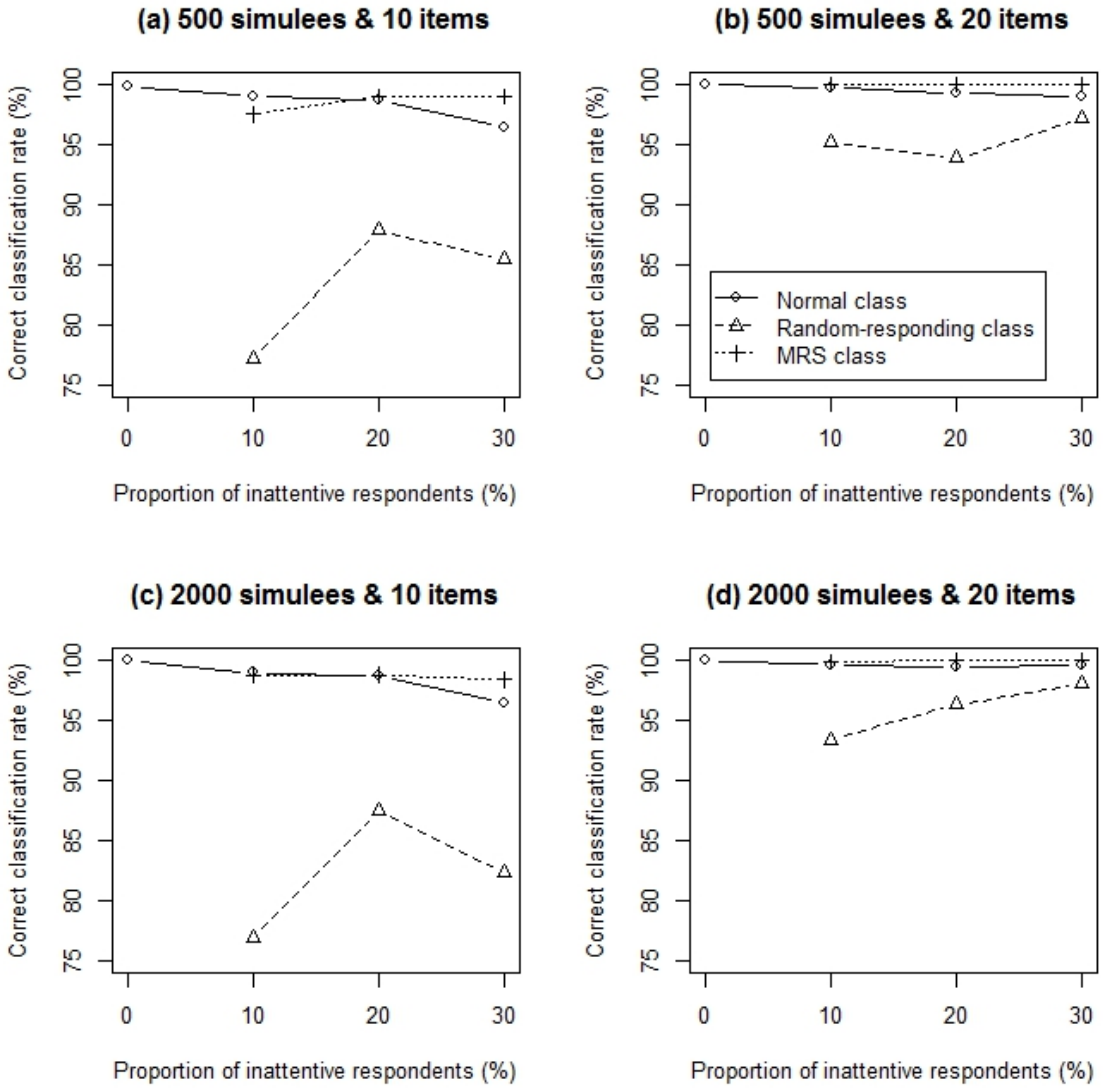

Figure 2. Correct classification rates of the three latent classes in the MMIR-3.

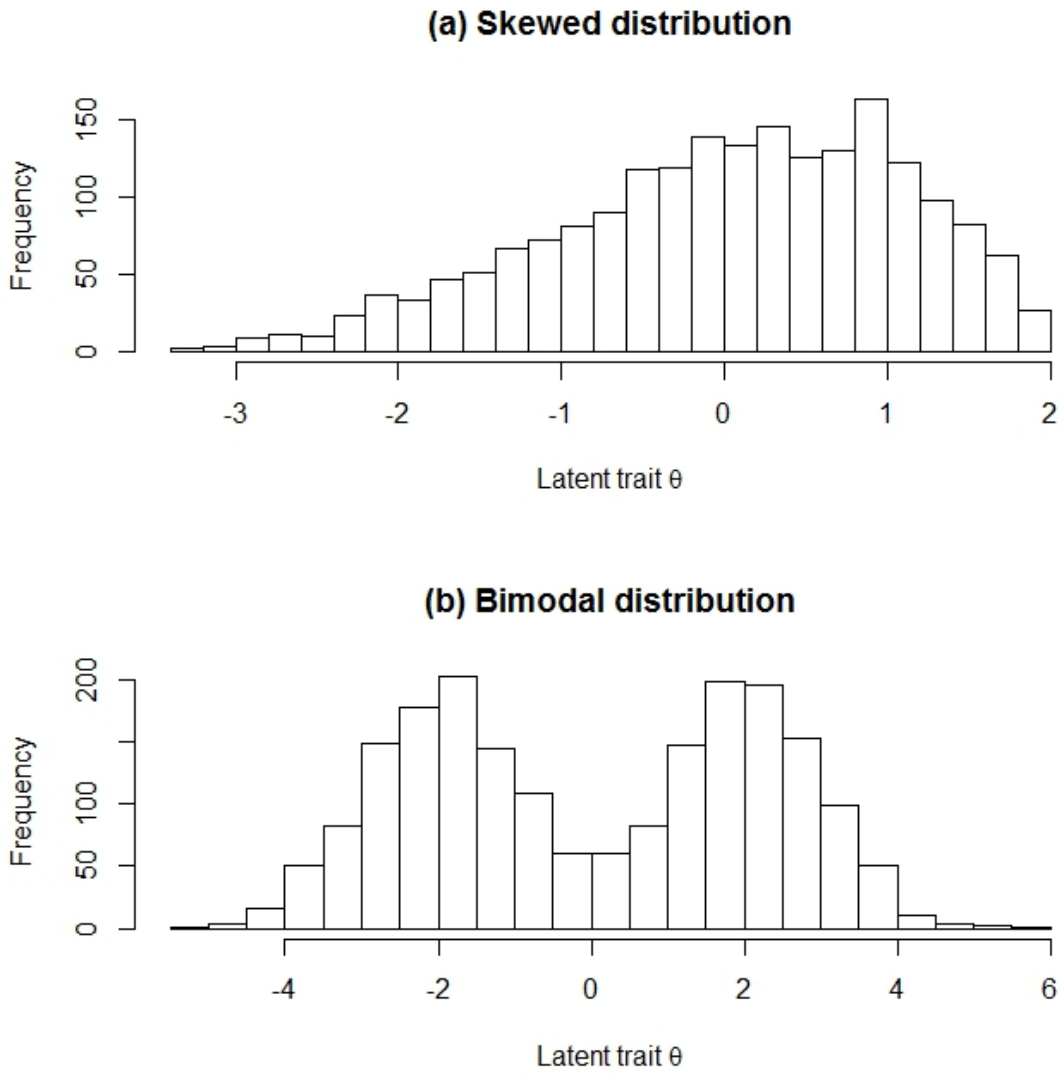

Figure 3. Non-normal distributions generated in Simulation 3

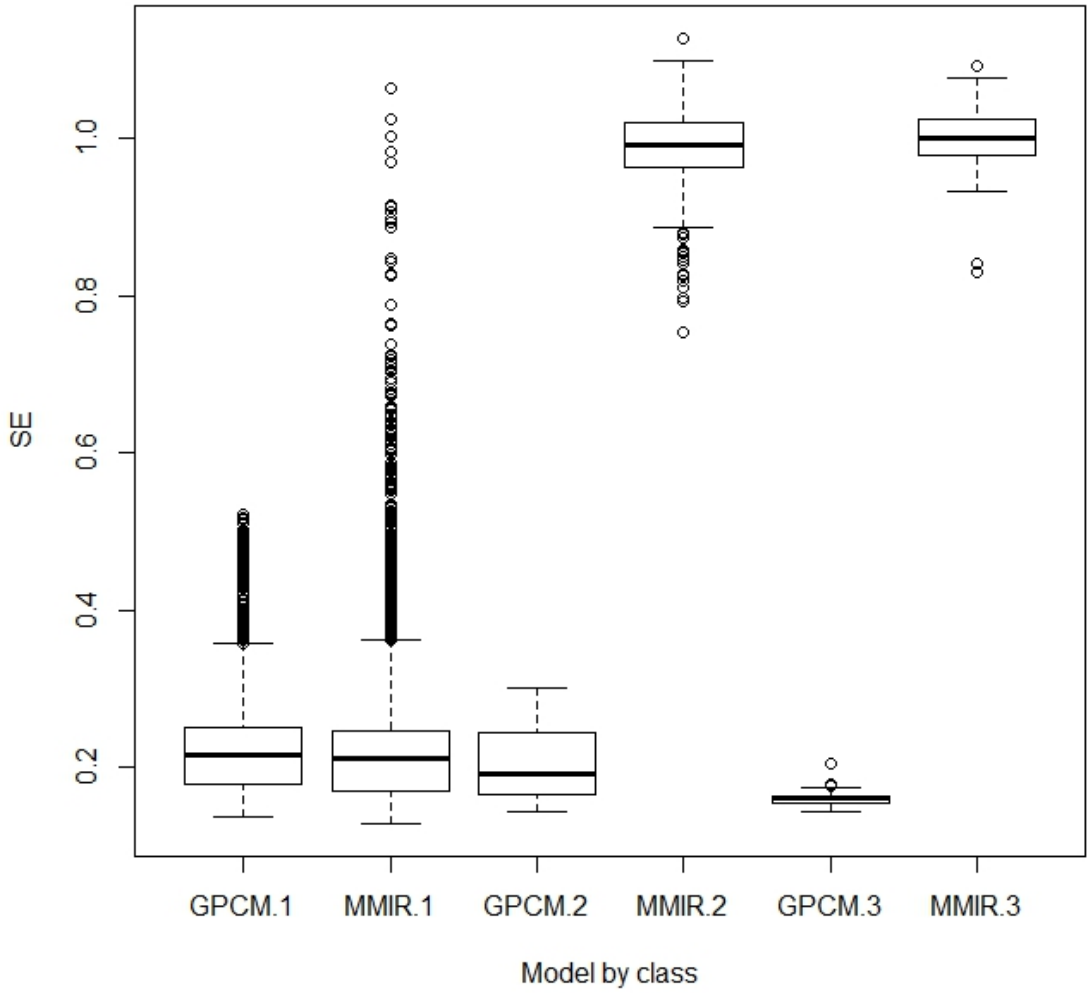

Figure 4. Standard errors for θ estimates under the GPCM and MMIR-3
Note. MMIR.1, MMIR.2, and MMIR.3 denote the normal class, the random-responding class, and the MRS class yielded from the MMIR-3, respectively. GPCM.1, GPCM.2, and GPCM.3 denote the normal class, the random-responding class, and the MRS class yielded from the GPCM, respectively.

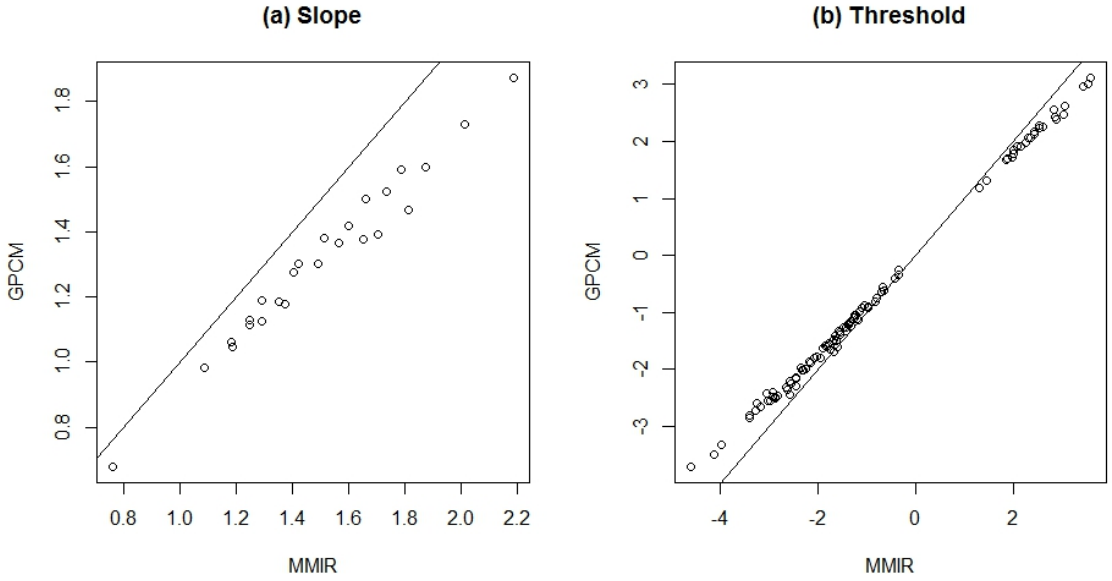

Figure 5. Slope and threshold parameter estimates under the GPCM and the MMIR-3 in Example 1

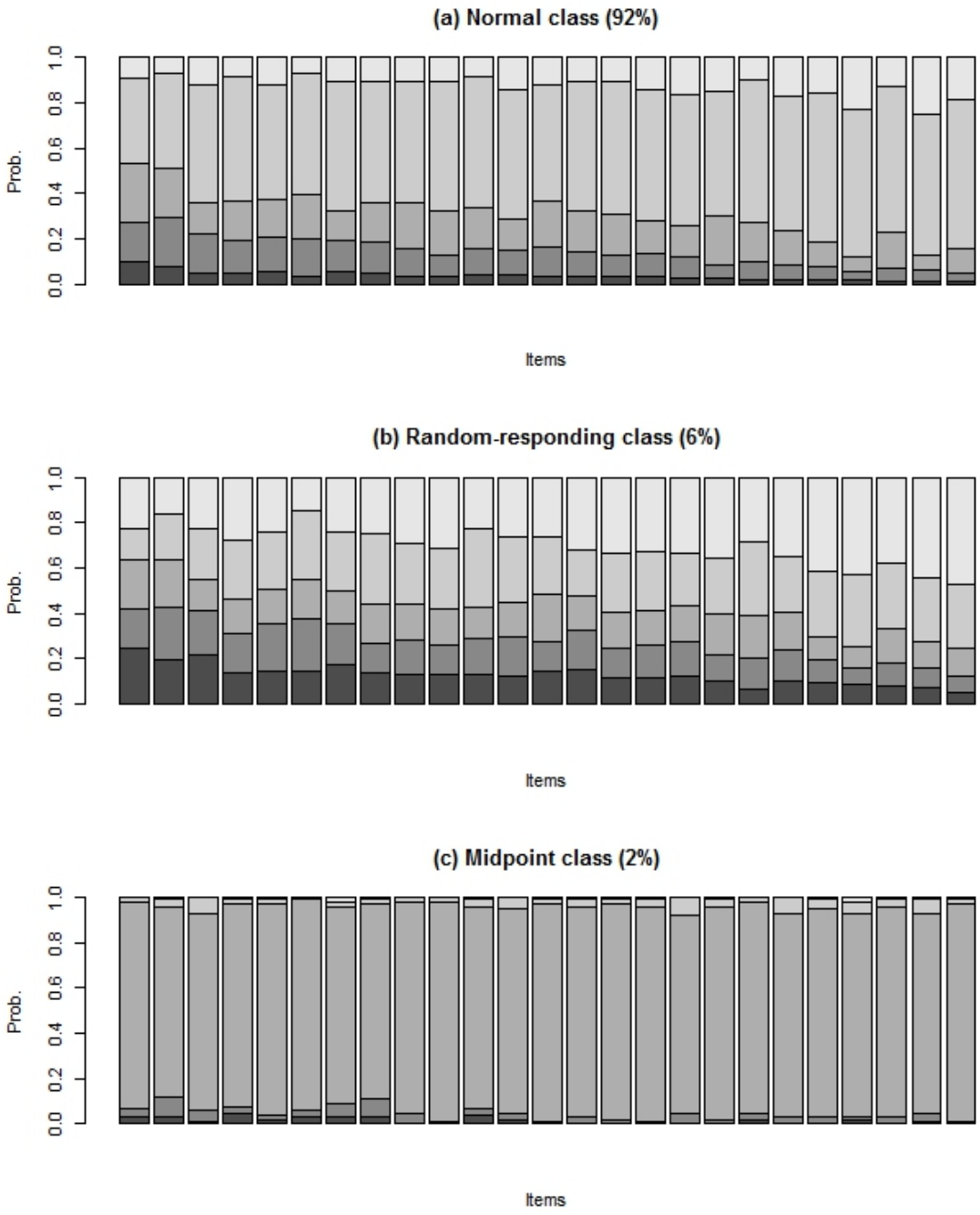

Figure 6. Response profiles of the three latent classes in Example 1
Note. The grey bars from dark to bright represent the response categories from very dissatisfied (1) to very satisfied (5).

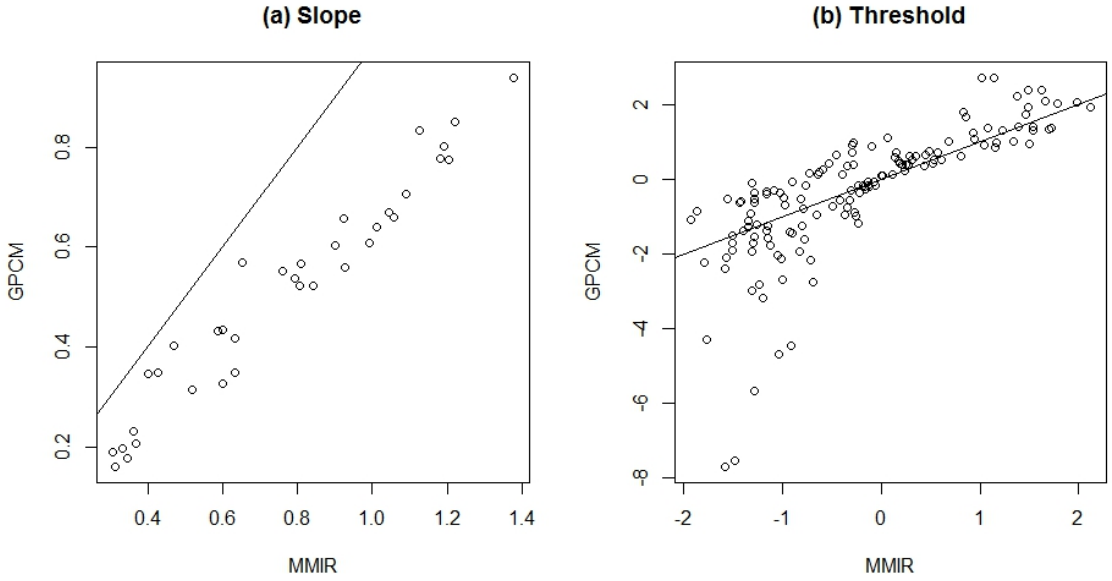

Figure 7. Slope and threshold parameter estimates under the GPCM and the MMIR-3 in Example 2

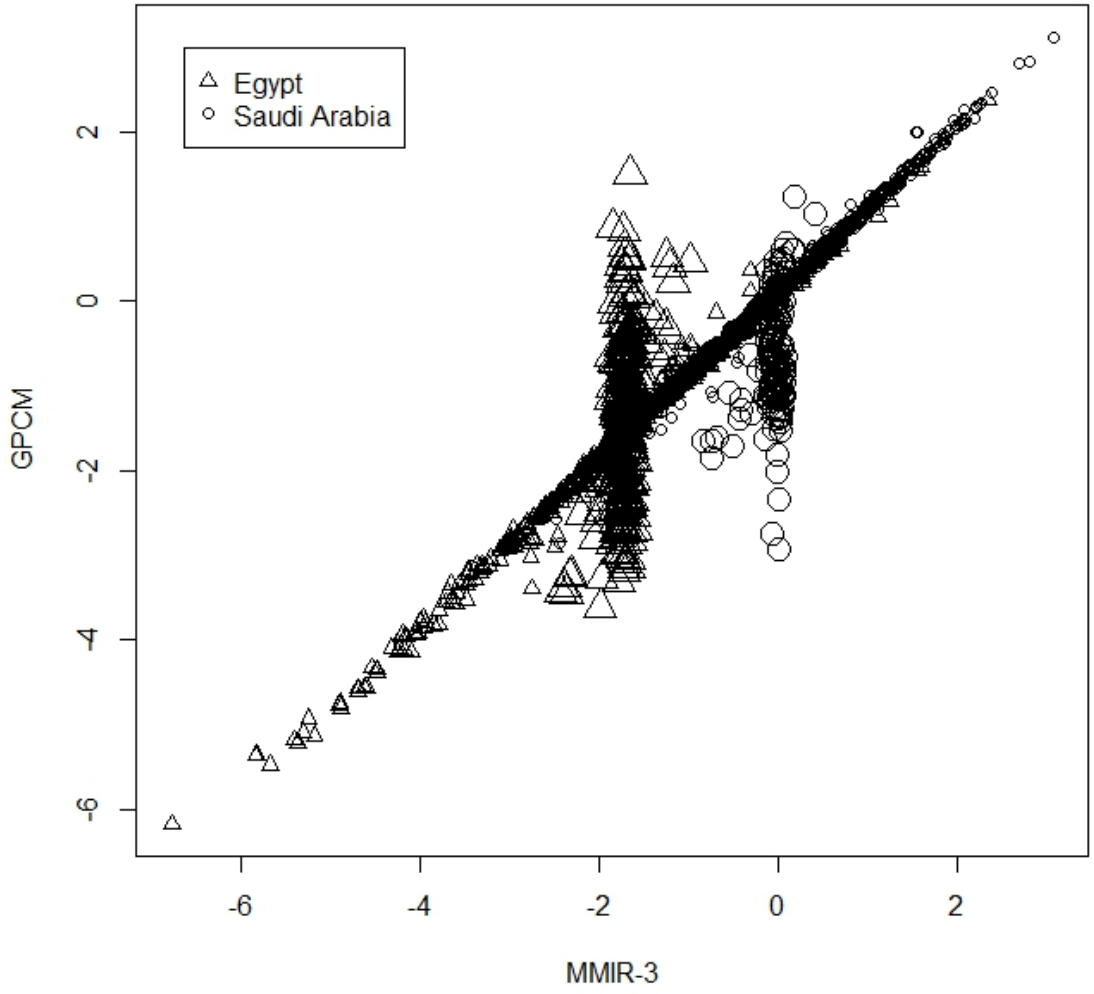

Figure 8. Person estimates under the GPCM and the MMIR-3 in Example 2
Note. Larger symbols refer to inattentive respondents.