Research Report

The Impact of Aberrant Responses and Detection in Forced-Choice Noncognitive Assessment

ETS RR–18-32

Sooyeon Kim
Tim Moses

December 2018

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Impact of Aberrant Responses and Detection in Forced-Choice Noncognitive Assessment

Sooyeon Kim[1] & Tim Moses[2]

1 Educational Testing Service, Princeton, NJ
2 College Board, New York, NY

The purpose of this study is to assess the impact of aberrant responses on the estimation accuracy in forced-choice format assessments. To that end, a wide range of aberrant response behaviors (e.g., fake, random, or mechanical responses) affecting upward of 20%–30% of the responses was manipulated under the multi-unidimensional pairwise preference (MUPP) model with forced-choice format. Under each aberrance condition, we also computed the *lz* statistic (a common person-fit index) to determine how well this index can accurately detect aberrant responses without erroneously misidentifying honest examinees. Some aberrant behaviors (e.g., limited responses) were more problematic than others (e.g., omitting) in terms of true trait level recovery. The *lz* statistic associated with each dimension of the MUPP form led to higher detection rates than did the overall *lz* statistic. When the dimensional *lz* approach was used, however, many honest examinees were misclassified as outliers. The detection rates of *lz* differed as a function of aberrant response types.

Keywords Noncognitive assessment; forced-choice format; multi-unidimensional pairwise preference; aberrant response; *lz* statistic

doi:10.1002/ets2.12222

Noncognitive test scores are intended to measure aspects of personality in both educational and occupational settings. Often it is assumed that noncognitive personality measures can be useful predictors of educational outcomes (Richardson, Abraham, & Bond, 2012) and task performance in the workplace (Barrick & Mount, 1991). Noncognitive assessments are rapidly gathering popularity in many organizational settings as a selection tool for identifying qualified job candidates whose behavioral skills fit a specific position (e.g., ACT's WorkKeys, ETS's *WorkFORCE*® assessments, or SHL[Saville & Holdsowrth Ltd]'s occupational personality questionnaire). Personality traits have both a strong theoretical foundation and empirical support of their importance in predicting workplace outcomes (for details, see Naemi, Seybert, Robbins, & Kyllonen, 2014). Even so, use of personality assessment in high-stakes contexts has been limited mainly owing to potential response bias, which may distort test scores (see Heggestad, Morrison, Reeve, & McCloy, 2006; McGrath, Mitchell, Kim, & Hough, 2010).

The most traditional way of collecting responses to personality items is to ask examinees to evaluate a single item at a time (called a *single-stimulus format*), independently of other items. With the single-stimulus format, examinees make absolute judgments about every item. Because they can easily discern more desirable answers under this circumstance, examinees can endorse desirable statements rather than answering honestly. Such socially desirable responding is a common problem confounding the results in noncognitive tests with a single-stimulus format (see Heggestad et al., 2006).

The forced-choice format has been proposed as an effective format that can make noncognitive tests considerably less susceptible to faking. Fake-resistant items can be created by pairing statements similar in desirability but representing different dimensions (called *multi-unidimensional pairs*) or the same dimension (called *unidimensional pairs*).[1] Thus those pairings constitute a fake-resistant test. With the forced-choice format, examinees have to evaluate several statements (more than one) presented at the same time to indicate which statement describes their personal attributes best. Even if the examinees evaluate either all or none of the statements favorably, they are forced to choose a single statement that is the closest to them. The examinees engage in comparative judgments, assessing relative properties of the statements. Presenting statements in the same set such that neither is clearly a more desirable choice may prevent the examinees from endorsing desirable statements, and thus their socially desirable responding can be reduced (Christiansen, Burns, & Montgomery, 2005; Jackson, Wroblewski, & Ashton, 2000; Martin, Bowen, & Hunt, 2002; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). Furthermore, forced-choice formats remove rater effects, such as leniency (uniformly high

Corresponding author: S. Kim, E-mail: skim@ets.org

judgments across all the stimuli) or severity (uniformly low judgments across all the stimuli), which often occur with single-stimulus format tests (Cheung & Chan, 2002).

Multi-Unidimensional Pairwise Preference

Although the forced-choice format has potential advantages, scoring the comparative responses arising from forced-choice items is not straightforward because of their ipsative nature (i.e., normative comparisons that are within individuals rather than between them). Under the classical scoring framework, where scores would be obtained as sums or percentages of choices, these ipsative scores are suitable for intraindividual assessment but not for interindividual assessment (Brown, 2016). In an attempt to circumvent the ipsativity problem, various item response theory (IRT) models have been developed to infer proper measurements from forced-choice data.[2]

Stark et al. (2005) proposed the multi-unidimensional pairwise preference (MUPP) model to infer measurement of individual differences from the forced-choice format assessment. The MUPP model assumes that when presented with a pair of statements representing the same or different constructs (i.e., dimensions or traits), an examinee evaluates his or her agreement with each statement independently based on his or her trait level (i.e., two independent unidimensional comparisons). In the MUPP model, each individual statement is described by a single dimension, but the test as a whole is composed of multiple dimensions. In practice, the MUPP approach was adopted by the U.S. Army for use in selecting and classifying recruits into job categories (e.g., Tailored Adaptive Personality Assessment System [TAPAS]; Drasgow et al., 2012).

Stark et al. (2005) described the general MUPP model as shown in Equation (1). Under the MUPP model, the probability of endorsing one statement over another is modeled as the joint probability of choosing one statement and not choosing the other statement in the pair. The probability of an examinee preferring statement s to statement t in an item i is given by

$$P_i\left(u_{is}=1\right)=P_{(s>t)i}\left(\theta_{ds},\theta_{dt}\right)=\frac{P_s(1)\,P_t(0)}{P_s(1)\,P_t(0)+P_s(0)\,P_t(1)}, \tag{1}$$

where i is the index for each pairwise preference item ($i=1, 2, \ldots, N$); s and t are indices for two statements in an item i; ($u_{is}=1$) is the response code for denoting that the statement s and its trait are chosen over the statement t and its trait in an item i; d is the dimension associated with a given statement ($d=1, 2, \ldots, D$); θ_{ds} and θ_{dt} are latent trait values for an examinee on the two statements' dimensions, ds and dt, respectively; $P_s(1)$ and $P_t(1)$ are the probabilities of endorsing statements s and t, respectively; and $P_s(0)$ and $P_t(0)$ are the probabilities of not endorsing statements s and t, respectively.

The assumption with the MUPP model is that each statement is evaluated separately, and then a choice for one statement over the other is made based on the degree of agreement between each statement and its corresponding trait level. The more agreeable the statement is, the more likely it is to be selected. It is also assumed that the probabilities for the individual statements, $P_s(1)$ and $P_t(1)$, reflect IRT models involving parameters for each statement and trait values for an examinee, θ_{ds} and θ_{dt}. Stark et al. (2005) linked the probability of endorsing the statement to a personality trait using each statement's three parameter estimates derived from the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000). The three parameter estimates are discrimination (α), statement location (δ), and threshold (τ). The GGUM parameter estimates for all of the statements in the forced-choice assessment are assumed to be available at the time of the forced-choice scoring (usually obtained from an earlier study where statements were administered and scored individually in Likert formats).

Under the GGUM (often called the *ideal-point model*), it is assumed that a respondent estimates the distance between his or her location and the statement's location on the underlying trait continuum. The probability of endorsing a statement increases as the distance between the person and statement locations decreases. The probability of agreement is highest when $(\theta-\delta)=0$, and it decreases in both directions, resulting in a single-peaked, bell-shaped item response function. The rate of decrease in the probability of agreement depends on the joint relationship between the item discrimination and item threshold parameters (see more details in Roberts et al., 2000). The following equations indicate the probabilities of agreement [$P_s(1)$] and disagreement [$P_s(0)$] associated with the statement s at the latent trait value of an examinee on the dimension s. The same equations hold for the statement t:

$$P_S(1)=\frac{\exp\left(\alpha_s\left[\left(\theta_{ds}-\delta_s\right)-\tau_s\right]\right)+\exp\left(\alpha_s\left[2\left(\theta_{ds}-\delta_s\right)\right]-\tau_s\right)}{\gamma_s} \tag{2}$$

$$P_S(0) = \frac{1 + \exp\left(\alpha_s\left[3\left(\theta_{ds} - \delta_s\right)\right]\right)}{\gamma_s}, \tag{3}$$

where $\gamma_s = 1 + \exp(\alpha_s[3(\theta_{ds} - \delta_s)]) + \exp(\alpha_s[(\theta_{ds} - \delta_s) - \tau_s]) + \exp(\alpha_s[2(\theta_{ds} - \delta_s)] - \tau_s)$, θ_{ds} is an examinee's trait level on the dimension s, α_s is discrimination (i.e., slope), δ_s is statement location, and τ_s is threshold.

Patterns of response choices to all forced-choice items on the noncognitive assessment are assumed to be locally independent. The resulting likelihood of a given response pattern and trait values reflects the fit of the forced-choice item models to the actual responses multiplied by an assumed prior distribution for all D traits, $P(\theta_s, \theta_t, \ldots, \theta_D)$:

$$\begin{aligned} L\left(u_{1,}u_2, \ldots, u_N, \theta_s, \theta_t, \ldots, \theta_D\right) &= \left[\Pi_{i=1}^{i=N} p_i^{ui}\left(1-p\right)^{1-ui}\right] p\left(\theta_s, \theta_t, \ldots, \theta_D\right) \\ &= L\left(u_{1,}u_2, \ldots, u_N | \theta_s, \theta_t, \ldots, \theta_D\right) p\left(\theta_s, \theta_t, \ldots, \theta_D\right). \end{aligned} \tag{4}$$

The Stark et al. (2005) approach involves finding trait values that maximize or obtain the mode of the likelihood given in Equation (4), $L(u_{1,}u_2, \ldots, u_N, \theta_s, \theta_t, \ldots, \theta_D)$, which is the same approach as the Bayesian maximum (mode) a posteriori (MAP) estimation.

Aberrant Response Behaviors

The MUPP model with forced-choice format is presumed to be more resistant to faking. By the same token, however, this format can lead to different aberrances, such as random or mechanical responses. For example, if the distance between the examinee's trait and the location (difficulty) of the first statement is almost identical to the distance between the examinee's trait and the location of the second statement, the statements may be perceived as fairly similar (no clear preference), and nearly random responding will occur. Owing to such uneasy decisions, the examinees often describe their overall testing experience associated with the forced-choice format as unpleasant. Perhaps this negativity deteriorates the examinees' motivation, leading to careless responses. Even under this format, some examinees may use tricks to present themselves in the best possible manner, particularly when there are strong incentives in the testing situation to score high. The following example item has two equally desirable statements representing different personality dimensions. The first statement (*s*) represents a high level of adjustment; the second one (*t*) represents a high-level of cooperation. If applicants have to compete to get an open position for customer service representative based on their trait scores, they may be inclined toward the second statement because of its plausible content relevance to that particular job position and not because of its closeness to their trait levels.

Statement *s*: I work well under pressure.

Statement *t*: I enjoy working for others, even strangers.

Scores will be valid when examinees have little motivation to distort their answers. In the noncognitive assessment with forced-choice format, examinees' responses could be classified into the following three types: (a) responses reflecting examinees' true traits; (b) responses made through anxiety, randomness, carelessness, lack of motivation, or distraction; and (c) responses made through intentional faking. The latter two types of aberrant responses might cause error in trait estimation because they would not reflect the examinees' actual trait and assumed item response patterns given their true trait levels. Thus the aberrant responses may artificially inflate (or deflate) trait estimates.

Meijer (1996) described various factors that can cause an examinee's score on a test to be spuriously high or spuriously low, particularly in the cognitive assessment setting. Those factors are cheating, careless responses, guessing, or random responses. Some factors may occur even in the noncognitive testing context. Most studies for aberrance and person-fit statistics as a method to detect aberrant examinees have been conducted under the cognitive testing context (see Kim & Moses, 2016). Such research rarely exists under the noncognitive context using a forced-choice item format. Recently, Lee, Stark, and Chernyshenko (2014) assessed the efficacy of a common person-fit statistic (i.e., *lz*) for detecting aberrant responding with unidimensional pairwise preference (UPP) measures in a situation where aberrance response patterns (e.g., fake good, random [careless]) were present. In their simulation, detecting fake-good (socially desirable) responding was much harder than detecting random responding. Table 1 presents six aberrant response types that can occur with a forced-choice format in noncognitive assessments. It also includes a description of each aberrant response and the direction of estimation bias due to aberrance.

Table 1 Aberrant Response Types Under the Multi-Unidimensional Pairwise Preference Model

Aberrance type	Description	Trait estimates artificially	How to generate aberrant examinees
Random responding	Examinees randomly choose one statement from a pair. Random selection can occur owing to indecisive preference between two statements.	Increased/ decreased	Impute random responses (i.e., no functional relationship with examinees' trait levels) to some statement pairs whose location levels between the two statements in a pair are similar (~22%–23% [28–30 pairs] of the test) to represent examinees who blindly respond.
Mechanical responding	Systematic patterns appear in a sequence of option choices. Examinees frequently choose a first option (s–s), a second option (t–t), or a mix (s–t/t–s).[a]	Increased/ decreased	Impute one of the following systematic patterns on some portions of the test (~30% of the pairs): 1 First statement ($s-s$) 2 Second statement ($t-t$) 3 Mix of ($s-t$) or ($t-s$)
Limited responding	In the multidimensional pairs, a particular dimension is rarely chosen, leading to estimation problems for that trait owing to lack of data.	Increased/ decreased	Among the 13 dimensions, a particular 3 dimensions are rarely chosen in the multidimensional pairs.
Fake responding	Socially desirable statements are preferred in a pair.	Increased	Faking will take place in all 26 (20%) unidimensional pairs. Under the unidimensional pairs, select statements whose location parameters are higher than their counterparts. Faking may also occur under certain multi-unidimensional pairs. Select statements whose social desirability values are higher than their counterparts' in a situation where the location difference between the two statements in a pair is smaller than 1.00.
Peculiar group	Examinees with different languages/cultural backgrounds show atypical response patterns (potential differential item functioning).	Increased/ decreased	Generate responses of peculiar examinees using fake location parameters. For the four dimensions, a positive constant (e.g., 1.5) is added to the statements' location parameters. For other four dimensions, a negative constant (e.g., −1.5) is added to the statements' location parameters.
Omitting	In the paper-and-pencil testing, some pairs are neglected because of their sensitivity or negativity.	Increased/ decreased	Impose missing responses to some statement pairs whose location levels between the two statements are similar (~22%–23% [28–30 pairs] of the test).

[a] s = first statement in an item pair; t = second statement in an item pair.

Purpose

The existence of aberrant responses can cause error in trait estimations because they may not reflect the examinees' actual trait levels. In reality, some level of aberrance is unavoidable. Given that circumstance, the question is how to identify examinees' unusual patterns of responses that would lead to an inaccurate estimate of their underlying trait levels. This study has two purposes. One is to assess the impact of aberrant responses on the examinees' trait levels as a function of aberrance types under the MUPP model with a forced-choice item format. Another is to find out how well common person-fit measures (e.g., *lz* statistic) distinguish aberrant responses from normal responses without penalizing honest examinees. It is assumed that the effectiveness of the *lz* statistic may differ as a function of aberrance types. For those purposes, we used simulated data to manipulate six aberrance types explained in Table 1 in a systematic manner. We imposed a certain type of aberrance when generating examinees' responses under the MUPP model with a forced-choice item format. Because both the simulated examinees' true trait levels and the item parameter estimates were available in this simulation, we were able to investigate the impact of aberrance on the trait estimations and the effectiveness of *lz* in detecting aberrant examinees.

Method

Forced-Choice Item Form Assembly

We assembled a linear form measuring 13 latent traits/dimensions under the MUPP model. The number of statements per dimension was 20, and thus the number of total statements on the form was 260. Because an item consists of two statements in the MUPP model, there were 130 items (i.e., statement pairs). To make the study form realistic, we mimicked an existing form assembled from the statement pool currently in use. Each statement in the statement pool has three parameter estimates derived from GGUM (discrimination, location, and threshold), along with its social desirability value.[3] Those statement characteristics were mimicked to the 260 statements used in this study. We treated those estimates as each statement's true parameters.

Two statements, when paired in an item, can be selected either as same dimension (unidimensional pair) or different dimensions (multi-unidimensional pair). A certain proportion of unidimensional pairs per dimension is considered necessary to identify the metric of trait scores. Stark, Chernyshenko, Drasgow, and White (2012) recommended that it would be better to minimize the proportion of unidimensional pairs because unidimensional pairs are less resistant to faking than multidimensional pairs. The linear MUPP form targeted in this study was a mix of 104 multidimensional pairs and 26 unidimensional pairs (20%).[4] An essential key to assemble a MUPP format item is to pair statements similar not only in social desirability but also in the trait location as a method to minimize faking.[5] In practice, the decisions for form specifications (e.g., number of dimensions, number of statements per dimension) and pairing constraints (e.g., maximum allowable difference in social desirability between two statements in an item, minimum location difference between two statements in a unidimensional pair) depend on several factors, such as capacity of the statement pool, testing population, or score usage. We took into account the operational practice to make the study form realistic.

Simulating Observed Scores for the Multi-Unidimensional Pairwise Preference Model

The first step in the simulations was to generate true traits for 1,000 examinees, $\theta_{d=1}, \theta_{d=2}, \ldots, \theta_{D=13}$. Those traits were generated as normally distributed variables, for one condition with population intercorrelations of .00 and for another condition with population intercorrelations of .50.[6] The true traits were rounded to a single decimal place, which was helpful for accuracy evaluations (described later). Then, using the estimated GGUM parameters for all paired statements, true MUPP probabilities of endorsing the first of the two statements were computed (Equation (1)). From these true probabilities, observed endorsements of the first of the two statements were generated by comparing the true probabilities to a uniformly distributed random number between 0 and 1; that is, when the true probability was greater than the uniformly distributed number, the observed choice for the first statement was $u_{is} = 1$. This process was used to simulate normal (no aberrance) responding.

Table 1 presents descriptions of each of the six aberrant response types and the method used to generate each aberrant response. Aberrant responding was implemented by introducing variations to the simulation model, such as by selecting statements for random or omitted responding, changing some statement parameters prior to simulating responses (e.g., peculiar group), or using true response probabilities that differed from the MUPP model. We adopted the method used by Karabatsos (2003) and Kim and Moses (2016) to generate aberrant examinees under the cognitive testing context, but the level of aberrance was rather arbitrary.

Scoring for the Multi-Unidimensional Pairwise Preference Model

The Stark et al. (2005) approach involves finding trait values that maximize or obtain the mode of the likelihood given in Equation (3), $L(u_{1,}u_2, \ldots, u_N, \theta_s, \theta_t, \ldots, \theta_D)$. The recommended approach for MAP scoring is to use a numerical algorithm such as the DFPMIN numerical recipes subroutine (Press, Teukolsky, Vetterling, & Flannery, 2007). This routine implements D-dimensional minimization of $-L(u_{1,}u_2, \ldots, u_N, \theta_s, \theta_t, \ldots, \theta_D)$ using a Broyden–Fletcher–Goldfarb–Shanno algorithm (implemented in SAS as the NLPQN subroutine; SAS Institute, 2010). An important aspect of this minimization is the assumption that the prior distribution, $P(\theta_s, \theta_t, \ldots, \theta_D)$, reflects traits that are distributed as standard normal and independent variables. The traits estimated from this approach, $\widehat{\theta}_{d=1}, \widehat{\theta}_{d=2}, \ldots, \widehat{\theta}_{D=13}$, are treated as scores. The resulting trait score estimates are traditionally reported on a standard normal distribution scale and typically range from −3.00 to 3.00.

Evaluations of Scoring Accuracy

To evaluate scoring accuracy for the simulations, examinees' estimated traits from the MUPP scoring, $\widehat{\theta}_d$, are compared to the true values, θ_d . This comparison is described in terms of four accuracy measures computed for each of the 13 true and estimated traits in each simulated condition of interest (i.e., aberrant response types, intercorrelations of the true traits).

The Pearson correlation of the estimated and true traits was computed for all 1,000 simulated examinees. The following four (essentially three) accuracy measures exploited the rounded population traits, which usually resulted in more than one examinee with the same population trait and allowed for the consideration of accuracy conditional on population trait value. The average bias was computed as the difference in average estimated traits $\left(\overline{\widehat{\theta_{dj}}}\right)$ and average true traits $\left(\overline{\theta_d}\right)$, $\sum_j P\left(\theta_{dj}\right)\left(\overline{\widehat{\theta_{dj}}} - \theta_{dj}\right)\Big|\theta_{dj} = \overline{\widehat{\theta}_d} - \overline{\theta_d}$, where d indicates the dimension and j indicates the rounded population trait level. The average squared bias was also computed as $\sum_j P\left(\theta_{dj}\right)\left(\overline{\widehat{\theta_{dj}}} - \theta_{dj}\right)^2\Big|\theta_{dj}$ to prevent negative bias at one trait level from canceling out positive bias at another. The average variance was computed as $\sum_j P\left(\theta_{dj}\right)\sigma^2\left(\widehat{\theta_{dj}}\right)\Big|\theta_{dj}$. Average total error was computed as $\sum_j P\left(\theta_{dj}\right)\left(\overline{\widehat{\theta_{dj}}} - \theta_{dj}\right)^2\Big|\theta_{dj} + \sum_j P\left(\theta_{dj}\right)\alpha^2\left(\widehat{\theta_{dj}}\right)\Big|\theta_{dj}$.

For the reported results, all five accuracy measures were computed for each of the 13 dimensions (traits) and then averaged. For each accuracy measure, its standard deviation, minimum, and maximum were also calculated over the 13 values associated with the 13 dimensions to determine whether the estimation results were comparable over the 13 dimensions.

The *lz* Statistic for Forced-Choice Response Patterns and Multi-Unidimensional Pairwise Preference Scores

Person-fit indices generally measure the extent to which an examinee's observed pattern of response deviates from the response pattern expected from an examinee with a certain trait level. One of the statistics used to detect aberrant response patterns is the *lz* statistic (Drasgow, Levine, & Williams, 1985). The *lz* statistic is a common person-fit measure to evaluate the fit of response pattern to a particular test model. This index has been used to identify respondents with seemingly idiosyncratic response patterns on cognitive ability tests as well as Likert-type noncognitive measures. Recently, Lee et al. (2014) assessed the effectiveness of the *lz* statistic on forced-choice item assessments (UPP) under various simulation conditions.

The *lz* statistic is based on comparing the logarithm of an examinee's observed likelihood, $\ln[L(u_{1,}u_2, \ldots, u_N | \theta_s, \theta_t, \ldots, \theta_D)]$ computed as in Equation (4), to the logarithm of his or her expected likelihood with the estimated traits and an assumption that the response model is correct, $E\left[\ln(L)\right] = \sum_{i=1}^{N}\left[P_i \ln\left(P_i\right) + \left(1 - P_i\right)\ln\left(1 - P_i\right)\right]$. When computing the *lz* statistic, the differences in observed and expected log likelihoods are standardized:

$$\frac{\ln(L) - E\left[\ln(L)\right]}{\sqrt{\sum_{i=1}^{N}\left\{\frac{P_i(1-P_i)}{\ln\left[P_i/(1-P_i)\right]}\right\}^2}} = \frac{\ln(L) - E\left[\ln(L)\right]}{\sigma\left[\ln(L)\right]}. \tag{5}$$

The *lz* statistic quantifies the difference between an examinee's observed pattern of item responses to responses expected on the basis of that person's standing in the latent trait and a set of statement parameters as specified by the GGUM model.

The *lz* statistic is a model-based index that evaluates the standardized log likelihood of an examinee's answer pattern relative to cutoff threshold values derived from either statistical theory or empirical methods (Lee et al., 2014). If an examinee's observed *lz* is lower than a cutoff threshold value, the response pattern will be classified as aberrant because the lower *lz* indicates that the response pattern is inconsistent with model prediction. In reality, there is no clear rule for setting a cutoff value of *lz*. The choice of cutoff threshold values can be rather arbitrary and may depend on the purpose of assessments (e.g., screening, certifying) and their specific constraints, such as test length, number of dimensions, or test mode. If it is critical to minimize false positives (i.e., flagging honest examinees; Type I error) while flagging the most extreme of the aberrant responses, using a large negative *lz* value (e.g., −2.0) would be appropriate.

In this study, we used two cutoff threshold values to differentiate aberrant examinees from the normal ones. One threshold of *lz* was −1.64, which is associated with Type I error rates of .05 level on a roughly standard normal scale.[7] Another threshold was the value corresponding to the 5th percentile of the *lz* statistic obtained under the normal (no aberrance) condition. The threshold value from the normal condition was −0.267 under the correlation level of .00 and − 0.250 under the correlation level of .50. For convenience, the former value (−1.64) was called the theoretical threshold and the latter one (either −0.267 or − 0.250) was named as the empirical threshold.

Use of the *lz* statistic computed over all responses (from a total of 130 statement pairs) is a common practice. In this study, however, we computed not only the overall *lz* statistic (conventional) but also the dimensional *lz* statistic related to each of the 13 dimensions so as to assess whether the use of dimensional *lz* information may enhance the detection rate of aberrant examinees. To identify aberrant examinees, the overall *lz* statistic was compared to the two types of thresholds (theoretical and empirical), whereas the dimensional *lz* was compared to the theoretical *lz* threshold only to alleviate Type I error rates.

Markov Chain

A Markov chain is a type of Markov process that has particular types of response patterns. We computed a Markov chain particularly for the mechanical response condition. In reality, various patterns of mechanical responses can occur under the MUPP model with the forced-choice format. As presented in Table 1, we imposed four typical mechanical patterns (e.g., consistently choosing the first statement [*s*] or consistently choosing the second statement [*t*]) on the simulation. The Markov chain values were calculated using Equation (6):

$$\text{Markov chain} = \frac{\left(ss - \text{avgPairCount}\right)^2 + \left(st - \text{avgPairCount}\right)^2 + \left(ts - \text{avgPairCount}\right)^2 + \left(tt - \text{avgPairCount}\right)^2}{\text{avgPairCount}}, \tag{6}$$

where *s* is the first statement in an item pair, *t* is the second statement in an item pair, *ss* is *ss*_pattern_count, *st* is *st*_pattern_count, *ts* is *ts*_pattern_count, *tt* is *tt*_pattern_count, and avgPairCount = (*ss* + *st* + *ts* + *tt*)/4.

As in the *lz* statistic, there is no clear rule to set a cutoff value of the Markov chain value so as to differentiate mechanical response patterns from normal response patterns. Using the simulated data under the normal (no aberrance) condition, we found the value corresponding to the 99th percentile of the Markov chain values associated with 1,000 normal examinees. Then we used this value, which is 13, as a cutoff threshold to identify mechanical responses. For simplicity, we used the same threshold of 13 in both the trait correlation levels of .00 and .50, because the difference between the two levels was very minor.

Table 2 Descriptives of Accuracy Measures Associated With Aberrance Types, Correlations = 0.00

	Aberrance type						
Deviance measure[a]	Normal[b]	Random	Mechanical	Limited	Fake	Peculiar	Omitting
Correlation							
Mean	0.854	0.69	0.693	0.622	0.753	0.797	0.806
SD	0.027	0.13	0.099	0.245	0.086	0.06	0.056
Min.	0.809	0.441	0.502	−0.009	0.611	0.683	0.685
Max.	0.893	0.836	0.839	0.832	0.842	0.886	0.872
Bias							
Mean	−0.002	−0.066	−0.071	−0.076	0.151	0.004	0.003
SD	0.017	0.156	0.161	0.395	0.293	0.789	0.021
Min.	−0.03	−0.4	−0.368	−0.756	−0.295	−1.015	−0.032
Max.	0.031	0.178	0.275	0.582	0.737	1.041	0.031
Squared bias							
Mean	0.307	0.567	0.572	0.746	0.636	0.802	0.376
SD	0.029	0.128	0.059	0.237	0.175	0.39	0.056
Min.	0.254	0.421	0.49	0.456	0.393	0.296	0.278
Max.	0.387	0.828	0.71	1.177	0.945	1.186	0.452
Variance							
Mean	0.416	0.451	0.471	0.36	0.362	0.397	0.45
SD	0.041	0.064	0.128	0.059	0.09	0.039	0.056
Min.	0.353	0.357	0.29	0.233	0.272	0.338	0.364
Max.	0.489	0.589	0.717	0.46	0.537	0.476	0.56
MSE							
Mean	0.518	0.726	0.747	0.838	0.74	0.918	0.587
SD	0.041	0.136	0.103	0.206	0.16	0.329	0.074
Min.	0.454	0.573	0.614	0.583	0.577	0.468	0.497
Max.	0.582	1.016	0.93	1.213	1.086	1.233	0.72

[a] $r = .00$. [b] Normal = no aberrance.

Results

Accuracy Measures

Table 2 presents the summary statistics of the five accuracy measures calculated over the 13 dimensions (traits) for each aberrance type under the trait intercorrelation level of zero. Table 3 presents the same type of information under the trait intercorrelation level of .50. For each of the five accuracy measures, four descriptive statistics (mean, standard deviation, min., and max.) were computed over the 13 values associated with the 13 dimensions.[8] Because the MUPP model assumes no relationships among the multiple traits/dimensions, the estimation results derived from the zero correlation level were slightly more accurate than the results derived from the .50 correlation level. Under the no aberrance (normal) condition, the correlation between estimate and true trait scores, averaged over the 13 dimensions, was .85 under the zero correlation level and .80 under the .50 correlation level. The range of correlations over the 13 dimensions was much wider (.70–.87) under the .50 level than under the zero level (.81–.89). Consequently, the averaged mean squared error (.52) in the zero level was smaller than the mean squared error (.60) in the .50 level.

Figure 1 graphically presents the correlations between the estimated and true trait scores as a function of aberrance types and trait intercorrelation levels. The relative accuracy patterns among the six aberrance types were fairly consistent regardless of the intercorrelation levels among the 13 traits. The difference between the zero and .50 levels was almost negligible in both random and limited-response conditions, however, indicating some degree of interaction between aberrance response types and trait intercorrelation levels. As displayed in Figure 2, a similar trend occurred with respect to the overall estimation error (mean squared error). The .50 correlation level produced larger mean squared errors than did the zero correlation level in most conditions, except for random and limited responding conditions. Even so, we focused on the estimation results from the six aberrant response types derived from the zero intercorrelation level for simplicity.

As shown in Table 2 and Figures 1 and 2, estimation accuracy was severely deteriorated owing to the aberrant responses. This trend was particularly clear under the limited-response case. As expected, three dimensions rarely chosen by simulation design yielded very low correlations (−.01, .33, and .47) and large overall errors (1.05, 1.21, and 1.17).[9] The mean

Table 3 Descriptives of Accuracy Measures Associated With Aberrance Types: Correlations = 0.5

	Aberrance type						
Deviance measure[a]	Normal[b]	Random	Mechanical	Limited	Fake	Peculiar	Omitting
Correlation							
Mean	0.798	0.689	0.638	0.61	0.692	0.719	0.732
SD	0.045	0.128	0.109	0.151	0.111	0.081	0.097
Min.	0.699	0.427	0.469	0.232	0.495	0.584	0.523
Max.	0.867	0.846	0.812	0.765	0.834	0.849	0.861
Bias							
Mean	−0.003	−0.077	−0.059	−0.081	0.156	−0.004	−0.001
SD	0.016	0.15	0.154	0.392	0.31	0.816	0.026
Min.	−0.028	−0.364	−0.295	−0.764	−0.278	−1.065	−0.036
Max.	0.019	0.136	0.301	0.572	0.827	1.06	0.059
Squared bias							
Mean	0.405	0.581	0.629	0.778	0.7	0.886	0.47
SD	0.045	0.12	0.049	0.18	0.171	0.38	0.078
Min.	0.324	0.45	0.567	0.552	0.499	0.368	0.361
Max.	0.504	0.824	0.712	1.126	1.053	1.245	0.599
Variance							
Mean	0.441	0.44	0.474	0.372	0.371	0.417	0.48
SD	0.046	0.064	0.125	0.064	0.095	0.05	0.074
Min.	0.355	0.337	0.294	0.246	0.276	0.347	0.374
Max.	0.542	0.6	0.73	0.47	0.566	0.514	0.637
MSE							
Mean	0.599	0.73	0.793	0.869	0.798	0.999	0.672
SD	0.058	0.13	0.102	0.155	0.167	0.325	0.105
Min.	0.51	0.569	0.66	0.671	0.639	0.531	0.52
Max.	0.7	1.02	0.981	1.163	1.184	1.295	0.862

[a] $r = .5$. [b] Normal = no aberrance.

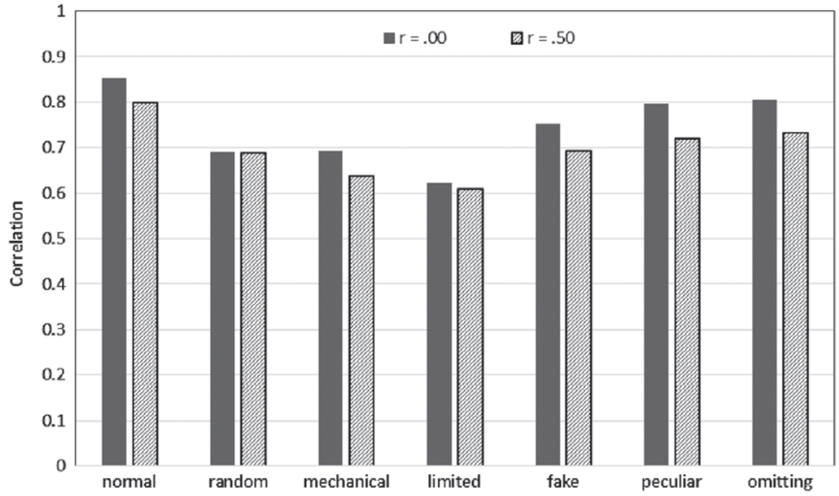

Figure 1 Correlations between the estimated and true trait scores under normal and six aberrance types.

squared errors associated with those dimensions were twice as large as their mean squared errors under the normal response condition (.51, .46, and .54). The random, mechanical, fake responses also produced low correlations and large estimation errors compared to the normal response condition. For most dimensions of the MUPP form used in this study, the statements' location values were highly correlated to their social desirability values ($r = .88 - .96$; a single dimension has $r = .80$). Consequently, the faking response spuriously increased the trait estimated values, leading to positive bias.[10]

Among the six aberrance types, the peculiar group responses led to the largest overall estimation error, whereas the omitting responses led to the least estimation error. Both aberrance types, however, showed similar correlations between

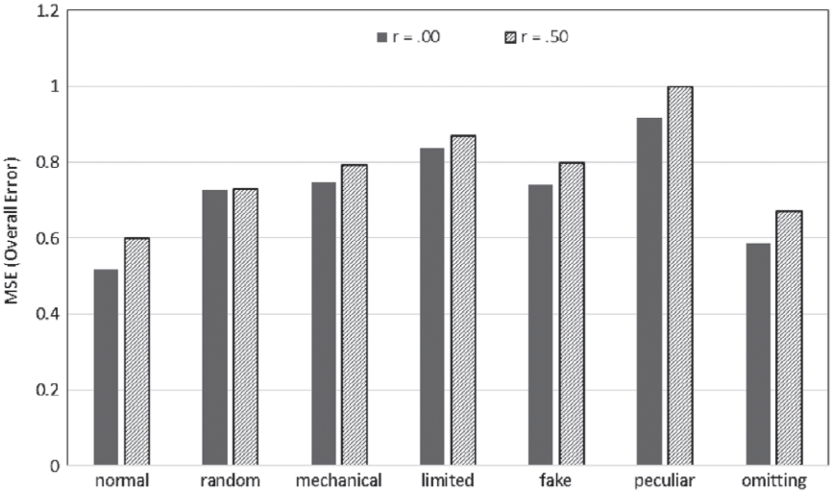

Figure 2 Overall estimation error (mean squared error) under normal and six aberrance types.

the estimated and true trait scores (.80 vs. .81). As the aberrance associated with the peculiar group response condition was due to the statements' incorrect parameter estimates rather than examinees' dishonest behaviors, this aberrance clearly increased the overall estimation error but did not decrease the correlation. Of particular interest in the present simulation is the comparison between random responding and no responding (omitting). The reason was that we imposed either random responding or no responding on the same set of items whose statement pairs were equally desirable (or undesirable) to reflect an indecisive preference between two statements. Interestingly, random responding reduced estimation accuracy more than did omitting.

The *lz* Statistic and Markov Chain

Table 4 presents the detection results based on either the overall *lz* or dimensional *lz* threshold under each of the six aberrance type conditions. When the overall *lz* of −1.64 (theoretical threshold) was applied as a cutoff threshold to detect aberrant response patterns, very few honest examinees were flagged as outliers in the no aberrance condition, indicating negligible low misclassification (Type I error) rates. This overall *lz* index, however, did not perform effectively in detecting aberrant response patterns when aberrations were present. Except for the limited-response condition (43%), the detection rates of the overall *lz* index were much lower than 30%. When the theoretical threshold was applied to detect aberrance responses, the effectiveness of the overall *lz* was questionable under all aberrance conditions. When the empirical threshold was applied to detect aberrant responses, however, the detection rates of the overall *lz* greatly increased in the four aberrance conditions, except for the peculiar and omitting conditions. As expected, approximately 5% of honest examinees were misclassified as outliers under the normal condition. The trends associated with the overall *lz* statistic were consistent regardless of the correlation levels.

As an alternate method to improve the detection rates, examinees were identified as outliers in cases where any dimensional *lz* (among the 13 dimensional *lz* values) was lower than −1.64 (theoretical threshold). The use of dimensional *lz* identified a larger number of aberrant examinees, with approximately 18%–21% of honest examinees being misclassified under this approach. The detection rates were over 96% in both the limited-response and peculiar group conditions when the dimensional *lz* approach was used. Additionally, the dimensional *lz* approach yielded over a 65% detection rate in the random, mechanical, and fake response conditions. A few outliers were identified under the no-response (omit) condition.

Table 5 presents the descriptives of the Markov chain values obtained from the mechanical response condition in both correlation levels. As presented in Tables 5, 77% of aberrant responses were flagged as outliers when the Markov chain value of 13 was applied as a threshold to detect them. As explained in Table 1, we manipulated four types of mechanical response patterns in this simulation, with each mechanical type occurring in one-fourth of the examinees, respectively. It appeared that the effectiveness of the Markov chain index would depend on the mechanical response patterns. The Markov chain used in this study is capable of detecting certain systematic response behaviors (e.g., consistently choosing

Table 4 Flagging Rates of the Overall and Dimensional *lz* Index Under Each of the Aberrance Types (%)

			Aberrance type						
Correlation	*lz*	Threshold	Normal[a]	Random	Mechanical	Limited	Fake	Peculiar	Omitting
.00	Overall	Theoretical	0.1	11.7	21.0	42.9	6.6	0.0	0.0
		Empirical	5.0	55.7	75.4	88.4	60.1	1.0	0.0
	Dimensional	Theoretical	17.6	68.7	76.3	97.8	79.6	96.7	1.2
.50	Overall	Theoretical	0.2	10.6	23.2	38.8	6.6	0.0	0.0
		Empirical	5.0	55.2	77.5	89.8	58.3	1.2	0.0
	Dimensional	Theoretical	20.9	67.6	77.1	98.0	78.5	97.1	1.8

Note. The theoretical threshold value was −1.64 in both correlation conditions. The empirical threshold value derived from the no aberrance condition was −0.267 in the correlation = 0.00 condition and − 0.250 in the correlation = 0.50 condition. Because the empirical threshold value associated with the dimensional *lz* cannot be well defined, the result from the theoretical threshold was presented for the dimensional *lz*. Examinees whose *lz* values were lower than the threshold values were flagged as outliers.

[a] Normal = no aberrance.

Table 5 Descriptives of Markov Chain Values Under the Mechanical Response Condition

	$r = 0.00$		$r = 0.50$	
Correlation	No aberrance	Mechanical	No aberrance	Mechanical
$(s-s/s-t/t-s/t-t)$[a]				
Mean	2.69	25.66	2.53	25.72
SD	2.77	15.05	2.54	14.80
Min.	0.02	3.68	0.02	2.32
Max.	20.74	106.50	21.73	83.06
Flag = Y (%)	1.10	76.80	0.70	77.40
$(s-s)$ or $(t-t)$[b]				
Mean	2.59	37.37	2.57	37.47
SD	2.59	12.23	2.65	11.40
Min.	0.02	8.71	0.02	13.30
Max.	17.51	106.50	21.73	83.06
Flag = Y (%)	0.80	99.80	0.80	100.00
$(s-t)$ or $(t-s)$[b]				
Mean	2.77	13.95	2.50	13.97
SD	2.93	5.39	2.43	5.64
Min.	0.02	3.68	0.02	2.32
Max.	20.74	33.82	16.27	35.19
Flag = Y (%)	1.40	53.80	0.60	54.80

Note. Cutoff criterion for flagging = 13. Examinees whose Markov value is higher than 13 are flagged as outliers.

[a] $N = 1{,}000$. [b] $n = 500$.

the first statement [s] or consistently choosing the second statement [t]), resulting in the detection of almost 100% of aberrant response patterns. Yet, its effectiveness was severely limited in detecting certain systematic response patterns (e.g., $s-t-s-t-s-t-s-t$ or $s-s-t-t-s-s-t-t-s-s-t-t$), which resulted in the detection of only 54% of aberrant response patterns.

Discussion

It is essential to provide examinees with fair and accurate scores. In reality, however, the aberrant responses that can cause an examinee's score on a test to be spuriously high or spuriously low are often unavoidable. Although the MUPP model with forced-choice format is fake resistant, this format can lead to other types of aberrance, such as random responses or careless responses (similar to mechanical responses), in actual testing settings.

The purpose of this study was to assess the impact on estimation accuracy of aberrant responses that can occur in a real testing setting. Using simulated data, we imposed six aberrant response types on the examinees' responses and assessed their impact on the estimation results by comparing to the results from the no aberrance condition. In a situation where aberrance clearly existed, trait score recovery was assessed using the correlations between true and estimated trait values as well as accuracy measures. We also assessed the effectiveness of *lz* (person-fit index) in detecting outliers at low rates of misclassification of honest examinees.

Aberrance types are not equally problematic. Some aberrance types, such as limited response and peculiar group response, were much more serious than others, resulting in larger overall errors compared to the no aberrance condition. Even so, the overall *lz* using the empirical threshold detected approximately 90% of the aberrant responses at low rates of false positives (5% of Type I error) under the limited-response condition. This trend did not appear in the peculiar group response condition, leading to very low detection rates of the overall *lz* index. Any potential impact due to poor parameter estimation must be a real concern, because testing programs are not completely free from this matter unless they use a large representative sample for item (statement) calibration. This issue can be further complicated particularly for the noncognitive assessments, because item parameter estimates for all of the statements in the forced-choice assessments are usually obtained from an earlier study (e.g., item pretesting and calibration) where the statements are administered and scored individually in single-stimulus formats (e.g., Likert scale), which are prone to faking. Perhaps the statements' parameter estimates are not free from common biases associated with single-stimulus format responses (e.g., faking or rater effects, such as severity or leniency).

Many noncognitive assessments in practice are computerized adaptive tests (CATs), and thus they do not allow skipping any items under the MUPP model. Under the CAT framework, examinees must choose a statement over another to move to the next item. The real examinees often describe their overall testing experiences associated with the forced-choice format as unpleasant mainly due to uneasy decisions (e.g., forced choice between equally unlikeable statements). Under this condition, examinees tend to choose one of the statements randomly. In this simulation, we manipulated two cases, random response and omitted response, where indecisive preferences between two statements were likely to occur due to similar location (i.e., equally desirable/undesirable). In this study, the random response deteriorated the estimation results more than did the omitted (no) response. Not only is omitting less problematic but it is also one of the easiest aberrant response types to detect in practice. These two features raise the question of whether it is always best practice to force examinees to choose from statement pairs without exception. It is worthwhile to investigate this matter under the CAT framework to reach a concrete conclusion.

With a forced-choice item format, statements should not confuse respondents. The GGUM employed to estimate statement parameters is often called an ideal-point model, and there are some issues related to item writing under the GGUM framework. One major concern is that good intermediate items located in the middle of the trait continuum are difficult to write in practice. The ambiguity applied to mid-level items may lead to a considerable increase in the likelihood of responding randomly. The investigation of aberrant responses with a forced-choice format including statements whose parameters were estimated using an ideal-point model is important in practice.

It is unrealistic to assume that all the examinees' responses will be free from any type of aberrance. It is generally assumed that the impact of aberrance on the estimation accuracy would be minimal as long as a sufficient number of high-quality statements are available with which to estimate the examinees' trait levels. Even so, this trend will no longer be true unless the degree of aberrance is moderate. Relatively less attention has been placed on identifying examinee characteristics that may lead to inaccurate trait scores under the noncognitive assessment context. Many testing programs remove extremely unmotivated examinees from all the operational analyses based on a certain criterion (e.g., no response for most items). In this study, the *lz* statistic and Markov chain value (to detect mechanical responses) did not perform well in detecting aberrant examinees. Setting an appropriate threshold level is not always clear in practice. The detection measure needs to be expanded further to capture various types of aberrant examinees.

This study assessed the impact of aberrant responses on the estimation accuracy and the effectiveness of *lz* in detecting aberrant examinees under each of the six aberrant conditions. To make the present simulation manageable, we limited the study design on the basis of some information from a particular testing program, and we did not incorporate the following topics in the process of simulation.

One limitation is related to the manner of manipulation of examinees' responses. We manipulated the six aberrant responses that may likely occur in a real testing setting. By design, certain aberrant responses were rather unrealistic in severity, while others looked more realistic. Although it appeared that certain response types looked more problematic than others in terms of the estimation error, the trend will vary depending on the degree of manipulation imposed on each of the aberrance types. Because we simulated a single level of aberrance for each of the six aberrant types, however, it is uncertain whether the same trend would appear for either higher or lower aberrance than the one imposed for this simulation. In addition, this simulation did not offer the information associated with potential interaction effects due to a mix of multiple aberrance types, because we simulated a single aberrance type for each condition.

We manipulated the aberrance responses based only on the location levels of statements in an item (statement pair) under the random, fake, and omit response conditions. Under those conditions, aberrance was imposed on approximately 25% of items having equally desirable (or undesirable) statements. According to the assumption of the MUPP model, the distance between a statement's location and an examinee's trait level will be more crucial than the location difference between two statements in a pair when the examinees choose one statement over the other. For simplicity, however, we did not take into account a factor such as the difference between statements' location levels and examinees' trait levels in this study. Owing to the limited study design, a definitive conclusion of the more problematic aberrance types cannot be clearly determined from this simulation. Additional investigation is necessary to assess the impact of the percentage of aberrant responses on the trait estimation in a solid manner.

Another limitation is related to the linear MUPP form assembly. In this simulation, we used a linear MUPP form, which contained well-fitting statement pairs from the test specification perspective, to generate all examinees' responses in all aberrant response conditions. Owing to the use of a single MUPP form, any investigation related to outlier detections under each of the six aberrant response conditions was straightforward in this application. Even so, it would be worthwhile to investigate the test constraints required to pair two statements in an item (e.g., maximum allowable location difference in multi-unidimensional pairs, maximum allowable social desirability difference in unidimensional pairs) to determine whether using MUPP forms with less constraint would confirm the current findings. The impact of limited responding, which is directly related to the test design constraints, can be further investigated when various types of MUPP forms are available.

In practice, many operational noncognitive assessments follow the CAT. By design, the impact of aberrance response on the trait estimation would be more critical under CAT than under the linear form case, because the selection of item pairs must depend on the provisional trait estimates derived from the examinee's precedent statement choice. It is premature to generalize the current findings to the CAT context. Despite some limitations, we think that the present findings will contribute to the noncognitive assessment literature. Future research geared at addressing the limitations listed herein may further expand the area of noncognitive assessment literature. A comparison with real data is also worthy of future research.

Notes

[1] Stark, Chernyshenko, and Drasgow (2005) recommend creating a small portion of unidimensional pairs by pairing statements that are similar in desirability but have different location parameters so as to identify the latent trait metric and permit interindividual comparisons.

[2] In this report, we do not mention all the IRT models because this is beyond the scope of the study. See Brown (2016) for the specific references for those IRT models. We focused solely on the MUPP model used in this study.

[3] Each statement's social desirability information can be collected through two strategies. The ideal strategy is to collect data from at least 50 individuals from the target population under "fake-good" instructions, with the individuals endorsing the statements on a 1 (*strongly disagree*) to 4 (*strongly agree*) scale. The average of the responses to each can be used as the social desirability estimate of the statement. A second approach is to use subject matter experts to rate the desirability of each statement. These experts could be instructed to indicate how desirable or undesirable it would be to tell someone the statement described him or her accurately on a 1 (*highly undesirable*) to 4 (*highly desirable*) scale. The average of the responses to each can be used as the social desirability estimate of the statement. See the TAPAS report (Drasgow et al., 2012, pp. 41–47, 63–64) for the specific references related to getting the social desirability values. The TAPAS was designed to support army selection and classification decisions upon noncognitive characteristics.

[4] We assembled three forms, differing the proportion of unidimensional pairs (20%, 30%, and 50%) to assess their impact on the estimation.

[5] This may not be case for the unidimensional pairs. The major use of unidimensional pairs is to identify the metric of trait scores. Therefore it is desirable to pair statements whose social desirability levels are similar but whose location levels are dissimilar.

[6] The choice of correlations was made based on a simulation study conducted by Drasgow et al. (2012, pp. 73–77).

[7] If the empirical distribution of *lz* departs from normality, using the cutoff threshold of −1.64 could lead to Type I error rates that differ from the expected level of .05 (Lee et al., 2014).

8 We conducted exactly the same simulation and analyses using two other MUPP forms having different pairing combinations: One had 39 unidimensional pairs (30%), and another had 65 unidimensional pairs (50%). The 20% form produced slightly better estimation results than the 30% and 50% forms did. Even so, the estimation results from the three linear MUPP forms were generally comparable. We can provide the results derived from the 30% and 50% forms on request.

9 For simplicity, we did not present specific values associated with each of the 13 dimensions in the table and figures. However, we can provide details on request.

10 The magnitude of positive bias was greater when the proportion of unidimensional pairs increased.

References

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1 – 26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika, 81*, 135 – 160. https://doi.org/10.1007/s11336-014-9434-9

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9*, 55 – 77. https://doi.org/10.1207/S15328007SEM0901_4

Christiansen, N., Burns, G., & Montgomery, G. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance, 18*, 267 – 307. https://doi.org/10.1207/s15327043hup1803_4

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

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support army selection and classification decisions* (Technical Report No. 1311). Fort Belvoir, VA: /Army Research Institute for the Behavioral and Social Sciences.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessment of personality for selection: Evaluation issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9 – 24. https://doi.org/10.1037/0021-9010.91.1.9

Jackson, D., Wroblewski, V., & Ashton, M. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371 – 388. https://doi.org/10.1207/S15327043HUP1304_3

Karabatsos, G. (2003). Comparing the atypical 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). *Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing* (GRE Board Report No. 16-03). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12111

Lee, P., Stark, S., & Chernyshenko, O. S. (2014). Detecting aberrant responding on unidimensional pairwise preference tests: An application of *lz* based on the Zinnes – Griggs ideal point IRT model. *Applied Psychological Measurement, 38*, 391 – 403. https://doi.org/10.1177/0146621614526636

Martin, B. A., Bowen, C.-C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247 – 256. https://doi.org/10.1177/0146621614526636

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

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9,* 3 – 8. https://doi.org/10.1207/s15324818ame0901_2

Naemi, B., Seybert, J., Robbins, S., & Kyllonen, P. (2014). *Examining the WorkFORCE Asessment for Job Fit and core capabilities of FACETS (Research Report No. RR-14-32)*. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12040

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.). New York, NY: Cambridge University Press.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353 – 387. https://doi.org/10.1037/a0026838

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3 – 32. https://doi.org/10.1177/01466216000241001

SAS Institute. (2010). *SAS/IML 9.22 user's guide*. Cary, NC: Author.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*, 184 – 203. https://doi.org/10.1177/0146621604273988

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*, 463–487. https://doi.org/10.1177/1094428112444611

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*, 175–199. https://doi.org/10.1207/s15327043hup1903_1

Suggested citation:

Kim, S., & Moses, T. (2018). *The impact of aberrant responses and detection in forced-choice noncognitive assessment* (Research Report No. RR-18-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12222

Action Editor: Donald Powers

Reviewers: Hongwen Guo and Jacob Seybert

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/