January 2013

Detecting Aberrant Responding on Unidimensional Pairwise Preference Tests: An Application of based on the Zinnes Griggs Ideal Point IRT Model

Philseok Lee
University of South Florida, philseok@mail.usf.edu

Detecting Aberrant Responding on Unidimensional Pairwise Preference Tests:

An Application of l_z based on the Zinnes Griggs Ideal Point IRT Model

by

Philseok Lee

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Stephen Stark, Ph.D.
Oleksandr S. Chernyshenko, Ph.D.
Michael Coovert, Ph.D.
Robert Dedrick, Ph.D.

Date of Approval:
February 21, 2013

Keywords: Person Fit, Appropriateness Measurement, Item Response Theory,
Noncognitive Assessment, Aberrance, Faking, Random Responding

DEDICATION

I would like to express the deepest appreciation to my advisor, Dr. Stephen Stark, who has continually and convincingly conveyed a spirit of adventure in regard to research. Without his guidance, patience, and persistent help, this thesis would not have been possible. I would also like to thank my committee members, Dr. Oleksandr Chernyshenko, Dr. Michael Coovert, and Dr. Robert Dedrick, for their suggestions and advice throughout the thesis process. I would like to thank my family who has supported me in every way they could.

I dedicate this thesis to the almighty GOD who gave me the strength and patience throughout the entire project. Without him, my life is nothing.

On my bed I remember you; I think of you through the watches of the night. Because you are my help, I sing in the shadow of your wings. My soul clings to you; your right hand upholds me. Psalm 63:6-8

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study investigated the efficacy of the l_z person fit statistic for detecting aberrant responding with unidimensional pairwise preference (UPP) measures, constructed and scored based on the Zinnes-Griggs (ZG, 1974) IRT model, which has been used for a variety of recent noncognitive testing applications. Because UPP measures are used to collect both "self-" and "other-" reports, I explored the capability of l_z to detect two of the most common and potentially detrimental response sets, namely fake good and random responding. The effectiveness of l_z was studied using empirical and theoretical critical values for classification, along with test length, test information, the type of statement parameters, and the percentage of items answered aberrantly (20%, 50%, 100%). We found that l_z was ineffective in detecting fake good responding, with power approaching zero in the 100% aberrance conditions. However, l_z was highly effective in detecting random responding, with power approaching 1.0 in long-test, high information conditions, and there was no diminution in efficacy when using marginal maximum likelihood estimates of statement parameters in place of the true values. Although using empirical critical values for classification provided slightly higher power and more accurate Type I error rates, theoretical critical values, corresponding to a standard normal distribution, provided nearly as good results.

INTRODUCTION

In the fields of psychology and education, there are long histories of research on noncognitive constructs, such as personality, vocational interests, self-efficacy, and values. Measures administered in research settings for developmental and diagnostic purposes were shown early on to predict important outcomes and those successes raised intriguing possibilities about the use of noncognitive tests in the workplace. By the 1950s, however, there were already concerns about the effects of response biases, such as halo error and impression management, on the validities of noncognitive scores for high stakes applications – most notably personnel selection and performance appraisal. Yet, despite these concerns, the need to expand selection testing beyond the cognitive ability realm for predicting a wider variety of job outcomes, the need to derive more accurate information about job performance from employee reviews, and the need for assessments that were quick and easy to administer sparked independent streams of research seeking alternatives to traditional Likert-type formats for noncognitive assessment.

In the 1940s, U.S. military researchers explored the benefits of observer ratings of personality (Connelly & Ones, 2010) as well as forced choice assessment (e.g., Hicks, 1970; Stark, Chernyshenko, Lee, Drasgow, White, & Young, 2011; Waters, 1965; White & Young, 1998) as alternatives to Likert-type self-report measures. Albeit through different mechanisms, both initiatives aimed to reduce social desirability response bias. In the same vein, the critical incident technique (Flanagan, 1954) was designed to improve the accuracy of employee appraisals by focusing the attention of raters on the

key elements of performance, thus reducing the effects of extraneous information. Critical incidents ultimately became the backbone of the behaviorally anchored rating scale (BARS; Smith & Kendall, 1963) appraisal method, which was designed to reduce the leniency, severity, central tendency, and halo errors often associated with Likert-type rating scales. BARS scales order critical incidents along a straight line in terms of effectiveness and require a rater to indicate which incident best characterizes a ratee's typical behavior.

In 1997, Borman, Hanson, Motowidlo, Drasgow, Foster, and Kubisiak proposed a "next-generation" version of BARS, called Computerized Adaptive Rating Scales (CARS; Borman, Buck, Hanson, Motowidlo, Stark, & Drasgow, 2001), which integrated research on observer ratings, forced choice assessment, and modern psychometric theory. Specifically, Borman et al. assessed contextual (i.e., citizenship) performance (Borman & Motowidlo, 1993) using computerized adaptive unidimensional pairwise preference (UPP) measures composed of pairs of statements that represented different levels of employee effectiveness. A rater's task was to choose the statement in each pair that better characterized the behavior of the ratee. By making repeated pairwise preference judgments across items chosen dynamically via computerized adaptive testing (CAT) principles (Stark & Drasgow, 1998; Stark & Chernyshenko, 2011), measurement error was reduced relative to BARS and Likert-type graphical rating scales (Borman et al., 2001).

Since the Borman et al. (2001) study, the suitability of UPP measures has been explored for other organizational applications. For example, Borman and colleagues implemented adaptive UPP measurement in the Navy Computerized Adaptive

Personality Scales assessment (NCAPS; Houston, Borman, Farmer & Bearden, 2005; Underhill, Lords, & Bearden, 2006) and Chernyshenko, Stark, and Williams (2009) used nonadaptive UPP measures to assess dimensions of person-organization fit. Although evidence suggests that UPP scales mitigate the central tendency, leniency, and severity errors that commonly occur when raters evaluate other rating targets (Borman et al., 2001), research is needed to examine their resistance to response biases associated with self-report data, such as socially desirable (fake good) and careless or random responding. Clearly, the validities of self-report applications in personality, person-organization fit, attitude, and values assessment depend on the quality of data collected and the capability to detect response biases or, more generally, aberrant responding that may occur in the absence of external information to verify self-report claims.

This manuscript describes a simulation study that examined the capability to detect two forms of aberrant responding with nonadaptive UPP measures constructed as described by Borman et al. (2001) and Stark et al. (2009). Specifically, this study examined the power and Type I error to detect fake good and random responding using the model-based standardized log likelihood statistic, known as l_z (Drasgow, Levine, & Williams, 1985; Drasgow, Levine, & McLaughlin, 1987). Over the years, l_z has been used to detect aberrant responding in connection with dichotomously and polytomously scored cognitive ability tests, as well as Likert-type noncognitive measures, but there has been little to no research on applications involving forced choice assessments.

l_z is a model-based index that evaluates the standardized log likelihood of a respondent's answer pattern relative to critical values derived from statistical theory or empirical methods. If a respondent's observed l_z is less than the critical value, the

response pattern is classified as aberrant; otherwise the pattern is classified as "normal." In this context, *normal* means that a respondent answered items in accordance with the predictions of an underlying item response theory model, and *aberrant* means that the response pattern was inconsistent with model predictions.

For the UPP applications described above, the model chosen to represent normal responding was the Zinnes and Griggs (ZG, 1974) ideal point IRT model. The ZG model assumes that when a rater is presented with a pair of statements describing different levels of, for example, effectiveness, conscientiousness, or autonomy, the rater carefully considers the statements and chooses the one in each pair that better describes the ratee. In contrast, aberrant responding, such as faking good and random responding, presumes a different psychological process. With UPP assessments, fake good responding implies a rater chooses the more positive or socially desirable statement in a UPP item, regardless of whether it accurately depicts the ratee. In work settings, fake good responding can occur when job applicants want to increase their scores to get hired, when raters wants to give positive impressions of well-liked coworkers in 360 degree appraisals, or when supervisors want to enhance their own reputations for employee development by manipulating the ratings of subordinates under their tutelage. Alternatively, random responding might occur when busy employees are surveyed too frequently without compensation, when respondents don't understand the context or meaning of questionnaire items, or when supervisors have many subordinates to evaluate and are familiar with only a few.

Before delving into details on how normal, faking good, and random responding can be simulated, I will briefly describe the Zinnes-Griggs IRT model, provide some

background on l_z computations and factors affecting detection accuracy, and outline a Monte Carlo study to explore the efficacy of l_z for identifying random and fake good responding with ZG-based UPP tests.

The Zinnes-Griggs IRT Model for UPP Responses

The ZG model assumes that when presented with a pair of statements representing, for example, different levels of employee effectiveness, a rater (e.g., a supervisor) will choose the statement in each pair that better describes the performance of a ratee (e.g., a subordinate). Specifically, the rater will tend to choose the statement in each pair that is closer to the ratee's perceived location on the performance continuum. Following the directions in italics, an example item from Borman et al.'s (2001) UPP assessment is shown below (see the Appendix for additional examples).

> *In each pair that follows, please choose the statement that better describes the employee you are evaluating. Indicate your answer by marking an "x" in the space to the left of that statement.*
>
> _____ 1a. Gathers and then analyzes information from a variety of sources to develop effective and timely solutions to problems.
>
> __x__ 1b. Takes too long to make decisions due to his/her need to gather and analyze more information than necessary.

Formally, if *s* and *t* represent the first and second statements in a performance appraisal item, Zinnes and Griggs (1974) showed that the probability of choosing or preferring statement *s* to statement *t* is given by:

$$P_{st}(\theta) = 1 - \Phi(a_{st}) - \Phi(b_{st}) + 2\Phi(a_{st})\Phi(b_{st}), \text{where} \quad (1)$$

$$a_{st} = (2\theta - \mu_s - \mu_t)/\sqrt{3}\,, \tag{2}$$

$$b_{st} = \mu_s - \mu_t, \text{ and} \tag{3}$$

where θ represents the ratee's perceived location on the performance continuum, μ_s and μ_t represent the perceived locations of the respective effectiveness statements, and $\Phi(a_{st})$ and $\Phi(b_{st})$ are cumulative standard normal density functions evaluated at a_{st} and b_{st}, respectively. Note that each statement is characterized by a single parameter, but to compute UPP item response probabilities, three parameters (μ_s, μ_t, and θ), are needed.

Figure 1 presents three illustrative IRFs for the ZG model computed at θ values ranging from -3 to + 3. Examination of the IRF for each pairwise preference item reveals that the probability of preferring statement *s* to statement *t*, $P_{st}(\theta)$, ranges from near 0 to 1. However, the IRFs differ in slope, because the slope depends on the distance between the statements composing an item: the greater the distance, the steeper the slope.

The item shown in Figure 1a involves statements having location parameters, (μ_s = 2.5, $\mu_t = -1.1$) respectively; the distance between them $(\mu_s - \mu_t)$ is 3.6. Figure 1b presents the IRF for an item involving location parameters (μ_s= 0.3, μ_t= 2.3), with the difference $(\mu_s - \mu_t) = -2.0$. Note also that: 1) the IRF in Figure 1b has a shallower slope than the IRF in Figure 1a because the distance between the respective statements in 1b is smaller; 2) the IRF in Figure 1b is monotonically decreasing, rather than increasing, because $\mu_s < \mu_t$. Finally, Figure 1c shows an IRF involving statements that have the same location parameters, $\mu_s = \mu_t = 0.9$. Because the statements represent equivalent effectiveness levels, each has a 0.5 probability of being selected, regardless of the ratee's performance score, θ.

Stark and Drasgow (1998, 2002) derived item information (I_i) for the ZG model based on Birnbaum's (1968) definition. The result is shown below:

$$I_i(\theta, \mu_{s_i}, \mu_{t_i}) = \frac{\frac{2}{3\pi}\left\{\left[1 - 2\Phi(b_{st_i})\right]e^{-\frac{1}{2}a^2_{st_i}}\right\}^2}{P_{st_i}(\theta)Q_{st_i}(\theta)}, \quad (4)$$

where $Q_{st_i}(\theta) = 1 - P_{st_i}(\theta)$. Using this equation, item information functions (IIFs) were computed for the items having IRFs shown in Figure 1. First, note that the function in Figure 2a is unimodal with a peak occurring halfway between the values of μ_s= 2.5 and $\mu_t = -1.1$. The IIF in Figure 2b has the same general form, but the peak is lower in accordance with the smaller difference between the location parameters, μ_s = 0.3 and μ_t = 2.3. Finally note that the IIF in Figure 2c provides zero information across the entire trait continuum. Because $\mu_s = \mu_t = 0.9$, there is no basis for preferring one statement over another so random responding is expected. In general, item information depends directly on the distance between the statements composing an item, with greater distance being associated with higher information. As was shown by Stark and Drasgow (2002), however, item information nearly attains its maximum when statements are located about two units apart on the typical performance range.

The l_z Person Fit Index

Since the 1960s, many methods have been proposed for detecting aberrant or atypical response patterns (Karabatsos, 2003; Meijer & Sijtsma, 2001). Although early research focused on cognitive ability testing applications, with the primary goal of detecting cheating, answer sheet tampering, and carelessness that could cause spuriously

high or low test scores (Hulin, Drasgow, & Parsons, 1983), applications eventually expanded into the noncognitive realm with the aim of detecting response sets, such as random or patterned responding, untraitedness, and faking (e.g., Drasgow, Levine, & Zickar, 1996; Egberink, Meijer, Veldkamp, Schakel, & Smid, 2010; Ferrando & Chico, 2001; Hendrawan, Glas, & Meijer, 2005; Nering & Meijer, 1998; Reise, 1995; Reise & Due, 1991; Reise & Flannery, 1996; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999; van-Krimpen Stroop & Meijer, 1999; Zickar & Drasgow, 1996; Zickar & Robie, 1999).

Levine and colleagues used the term *appropriateness* indices (Levine & Drasgow, 1982; Levine & Rubin, 1979) in reference to statistics broadly aimed at flagging inconsistencies between observed and expected answer patterns, but today the terms *person fit indices* and *person fit statistics* are common alternatives. Person fit, perhaps more clearly, implies that a psychometric model can merely describe the data better for some examinees than others, and response patterns that are inconsistent with model predictions do not necessarily indicate that anything *inappropriate* occurred during a testing session. By scrutinizing answer patterns having poor person fit statistics and by comparing fit statistics across subgroups, one can identify and remove atypical response patterns for data cleaning, test validation, and personnel selection purposes. One can also generate ideas about why examinees are mischaracterized and perhaps use that information to improve assessments or generalize psychometric models to take those ideas into account.

In general, person fit statistics examine either residuals (i.e., differences between observed and expected responses patterns) or the likelihood of response patterns

assuming a formal model of item responding (Nering & Meijer, 1998). IRT methods generally use the latter. The likelihood of a response pattern is calculated using item and person parameter estimates for a designated item response model, and aberrant or atypical patterns are signaled by low (or, in the log metric, negative) likelihoods. The advantage of IRT methods is that they readily permit the assessment of overall model-data fit unlike classical test theory methods.

One of the most widely used and researched IRT-based person fit statistics is l_z (Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, & Williams; 1985). l_z is popular for cognitive and noncognitive data screening, because it can be readily applied with different IRT models and it is capable of detecting several forms of aberrance (e.g., Bierenbaum, 1985, 1986; Nering, 1996; Nering & Meijer, 1998). For Zinnes-Griggs (1974) UPP model applications, l_z is computed as follows. First, the log likelihood of a rater's response pattern is given by

$$l_0 = \sum_{i=1}^{n} u_i \log P_{st_i}(u_i = 1 \mid \hat{\theta}) + (1 - u_i)\log(1 - P_{st_i}(u_i = 1 \mid \hat{\theta})), \quad (5)$$

where $\hat{\theta}$ is an estimate of θ, the latent trait representing a ratee's trait or performance level, i is an index for items, i =1, …, n, u_i represents an item response coded 1 if statement s is preferred to statement t and 0 otherwise, P_{st_i} is the probability of preferring statement s to statement t in the i^{th} item, and log represents the natural logarithm function. The approximate expectation of this log likelihood is

$$E(l_0) \approx \sum_{i=1}^{n} P_{st_i}(u_i = 1|\hat{\theta}) log P_{st_i}(u_i = 1|\hat{\theta}) + [1 - P_{st_i}(u_i = 1|\hat{\theta}) \log[1 - P_{st_i}(u_i = 1|\hat{\theta})], \quad (6)$$

The approximate variance is

$$Var(l_0) \approx \sum_{i=1}^{n} P_{\pi_i}(u_i = 1 \mid \hat{\theta})[1 - P_{\pi_i}(u_i = 1 \mid \hat{\theta})]\{\log \frac{P_{\pi_i}(u_i = 1 \mid \hat{\theta})}{[1 - P_{\pi_i}(u_i = 1 \mid \hat{\theta})]}\}^2. \quad (7)$$

Finally, the approximately standardized person fit statistic is

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}}. \quad (8)$$

The standardization step is important because it eliminates the dependence of the resulting person fit statistic on test length and θ, which was a concern with the l_0 statistic that was proposed originally by Levine and Rubin (1979).

The use of l_z for ZG UPP data screening thus requires observed item responses and estimates of ZG item parameters (μ_{s_i}, μ_{t_i}) and trait scores ($\hat{\theta}$). By substituting the values into the appropriate equations above, l_z can be computed for each ratee's response pattern and normal versus aberrant classification decisions can be made by comparing each observed l_z to a critical value. When an observed l_z is less than the critical l_z, a response pattern is classified as aberrant because the data are inconsistent with the predictions of the ZG model. Patterns that are highly inconsistent with model predictions will have large negative l_z values (e.g., -2), whereas patterns that are consistent with model predictions will have positive l_z values (e.g., > +1) values on a roughly standard normal scale.

Factors Influencing l_z Efficacy

One of the most widely studied issues associated with l_z is standardization. The original research by Drasgow et al. (1985) as well as subsequent examinations (e.g., Molenaar & Hoijtink, 1990) found that l_z is approximately, but not exactly,

standardized. That is, the empirical distribution of l_z departs somewhat from normality when $\hat{\theta}$ is used instead of θ for the calculations (e.g., Molenaar & Hoijtink, 1990, 1996; Nering, 1997; Snijders, 2001; van-Krimpen Stroop & Meijer, 1999). Therefore, using a lower one-tailed critical value of, say, -1.64 for classification decisions could lead to Type I error rates that differ from the expected (.05) level.

To address this limitation some authors have explored the use of empirical critical values as alternatives to those based on normality assumptions (Stark, Chernyshenko, & Drasgow, 2012; see also Nering, 1997; van-Krimpen-Stroop & Meijer, 1999). Essentially, one must simulate large numbers of normal response patterns based on actual exam or scale characteristics, compute l_z for each pattern, find the l_z value corresponding to the 5th percentile, for example, and use that value as a lower-bound critical value for screening the real response data. Although this method has proven useful in other contexts, such as differential item functioning detection (e.g., Flowers, Oshima, & Raju, 1999; Meade, Lautenschlager, & Johnson, 2007; Seybert & Stark, 2012), research is needed with l_z to determine whether it provides better classification accuracy normal distribution theory critical values, especially when considering the computational complexity it introduces.

Past studies involving applications of l_z and other person fit statistics have identified several other important factors affecting the detection of aberrant response patterns. First, certain types of aberrant responding appear to be easier to detect than others (Drasgow, 1982; Meijer, Molenaar, & Sijtsma, 1994; Rudner, 1983). For example, Levine and Rubin (1979) showed that power rates were consistently higher for aberrant examinees exhibiting spuriously high scoring on cognitive ability tests than for

examinees exhibiting spuriously low scoring; or, in other words, cheating was easier to detect than careless responding. In addition, intuition suggests that faking might be difficult to detect, especially if the degree of distortion is consistent across items. If there are too few apparent inconsistencies between observed and predicted item response probabilities, then it would be virtually impossible to distinguish between truly high or low performance and spuriously high or low performance.

A second factor affecting the power of person fit statistics is the proportion of items answered aberrantly. Several studies involving dominance IRT models have shown that higher proportions of aberrant responding are associated with higher detection rates (Drasgow et al., 1987; Karabatsos, 2003; Levine & Rubin, 1979). However, for some types of aberrance such as cheating or faking good, this relationship may actually be curvilinear. For example, with really high proportions of aberrant item responding, it would be difficult to separate cheaters from high ability examinees because both would be expected to answer most items correctly.

Test composition has also been found to influence detection rates. Given the same type and relative proportions of aberrant responding, detection rates are consistently higher with longer tests (Emons, Sijtsma, & Meijer, 2004; Karabatsos, 2003; Nering & Meijer, 1998; Reise & Due, 1991), perhaps because trait scores are more accurately estimated and there are more opportunities to observe inconsistencies with model predictions. Second, higher power and lower Type I error are typically observed with tests having more discriminating items (Emons et al, 2004; Meijer, 1997; Meijer, Molenaar, & Sijtsma; 1994) and more variation in item extremity (Reise, 1995). This makes intuitive sense because higher discrimination leads to higher test information and,

thus, more accurate trait estimation. And, variations in extremity highlight inconsistencies between predicted and observed responses given one's estimated trait score.

Finally, some recent studies have examined the effects of parameter estimation error on the power and Type I error of person fit indexes. In accordance with the statistical principle of consistency, large samples are always desirable for item/statement parameter estimation. The more these parameter estimates differ from their true values, the more error there will be in the estimated trait scores and, thus, the lower the power to detect aberrance. Fortunately, person fit research with single-statement measures has shown only small detrimental effects for parameter estimation error on power (Hendrawan, Glas, & Meijer, 2005; St-Onge, Valois, Adbous, & Germain, 2009). However, research is needed to see whether this finding generalizes to ZG-based UPP measures calibrated via marginal maximum likelihood estimation (Stark & Drasgow, 2002).

METHOD

Study Design

This research investigated the power and Type I error rates for l_z aberrance detection using a Monte Carlo simulation involving four primary factors: 1) UPP test length (10 items, 20 items), 2) percent of items answered aberrantly (20%, 50%, 100%), 3) response style (normal, random, fake good), and 4) test information (medium, high). In addition, to examine how UPP statement parameter estimation accuracy in a pretesting scenario would affect subsequent operational l_z screening decisions, classification accuracy was studied with empirical critical values using true and marginal maximum likelihood (MML) statement parameter estimates, based on samples of 1000 and 500 examinees, respectively (TRUE, MML1000, MML500), as well as critical values based on normal distribution theory. For this aspect of the research, lower one-tailed critical values corresponding to nominal alpha levels of .01, .05, .10, and .20 were used. Table 1 presents the study design.

Test Characteristics

In preparation for this simulation, four tests were created to satisfy the test information and test length considerations mentioned above. First, in accordance with the recommendation by Stark and Drasgow (2002), a 10-item high information UPP test was assembled by pairing statement parameters that differed by about 2.5 units along different parts of the trait continuum. The result was a test information function that had

an amplitude of approximately 5 near $\theta = 0.8$, as shown in Figure 3b. Next, a 10-item medium information test was created by paring statement parameters that differed by about 1.5 units along different parts of the trait continuum. The resulting test information function had an amplitude of about 3.5 near $\theta = 0.8$, as shown in Figure 3a. Finally, 20-item medium and high information tests were created by adding replicas of the respective 10-item tests. The resulting test information functions are shown in Figures 3c and 3d. Table 2 shows the "true" parameters for the 20-item medium and high information tests.

Data Generation and l_z Analyses

Power and Type I error rates for l_z classification decisions were computed over 100 replications in each of the 28 experimental conditions shown in Table 1. *Power* is defined as a "hit" or correct detection of an aberrant response pattern, whereas *Type I error* represents a "false alarm" or incorrect classification of a normal response pattern as aberrant. A C++ program was developed to perform the following sequence of steps for data generation and analysis in the simulation study.

1. 1,000 trait scores (thetas) were obtained by sampling from a standard normal distribution. These "true" thetas were used in conjunction with the true item parameters, shown in Table 2, to simulate UPP responses to the four tests having information functions shown in Figure 3.
2. "Normal" responses to each item of each test were simulated by computing the probability of preferring statement *s* to statement *t* in item *i* given a simulee's true trait score (see Equation 1) and comparing the value to a random uniform number. Specifically, if $P_{st_i}(\theta)$ was greater than the random number, the response was

scored as 1; otherwise the response was scored as 0. (The data generated in this step corresponds to the assumption that in applied settings there are uncontaminated pretest data available for IRT estimation before operational l_z screening.)

3. Three sets of statement parameters for each of the four tests were used to investigate the effects of MML estimation error on l_z classification accuracy. Specifically the response data generated in Step 2 were calibrated using the ZG MML computer program (Stark & Drasgow, 2002) using the full sample of 1000 simulees (MML1000) and a randomly selected subsample of 500 (MML500). The TRUE statement parameters served as a baseline for comparison.
4. Three sets of l_z values were computed for each simulee using the true and estimated ZG statement parameters from Step 3 in conjunction with the respective expected a posteriori (EAP) trait score estimates provided by the ZG_EAP computer program (Stark, 2006).
5. Lower one-tailed *empirical* critical values for "operational" l_z classification decisions were obtained by sorting the respective sets of observed l_z values from Step 4 in ascending order and identifying the values corresponding to the 1st, 5th, 10th, and 20th percentiles. *Theoretical* critical values for those same percentiles under normality assumptions were obtained by using an inverse normal probability table.
6. New response data reflecting varying degrees of aberrance (0%, 20%, 50%, 100%) were generated to investigate l_z power and Type I error under an operational testing scenario. For the 0% (no aberrance or normal) conditions, 1,000 new trait

scores were sampled from a standard normal distribution and used to generate UPP responses to each of the four tests using the TRUE statement parameters. For the 20% and 50% aberrance conditions, item responses from those same data sets were randomly designated for replacement with fake good or random responses. In the 100% condition, all of the responses were replaced with fake good or random responses. *Random responses* were generated by sampling a random number from a uniform distribution and comparing it to 0.5. If the result exceeded 0.5, the response was scored as 1; otherwise the response was scored as 0. *Fake good* responses were simulated by adding 1.5 to a simulee's trait score when computing $P_{st_i}(\theta)$ for the designated items. If the result exceeded a randomly sampled uniform number, then the response was coded 1; otherwise 0. Simulating faking in this way has been referred to as the *theta-shift* method (Zickar, 2000; Zickar & Drasgow, 1996; Zickar & Robie, 1999). Note that the 0% conditions were used to investigate Type I error for l_z classification, while the 20%, 50% and 100% conditions were used to examine power.

7. As in Step 4, three sets of l_z values were computed for each response pattern generated in the previous step using the true and MML statement parameter estimates and the resulting EAP trait scores. Each response pattern was then classified as normal or aberrant using each of the observed l_z values in conjunction with each of the empirical and theoretical critical values from Step 5. If the observed l_z was less than the critical l_z, then the response pattern was classified as aberrant; otherwise the response pattern was classified as normal under the respective conditions.

8. Type I error was computed for each of the critical values by calculating the proportion of response patterns in the 0% conditions that were misclassified as aberrant. Power was computed in the 20%, 50%, and 100% conditions by computing the proportion of response patterns that were correctly identified as aberrant.
9. Steps 1 through 8 were repeated until 100 replications were performed. Upon completion of the replications, overall power and Type I error were calculated for each experimental condition by averaging the findings from Step 8 over replications.

Hypotheses

Based on theoretical assumptions and previous l_z research, the following hypotheses were formulated.

1. Power will be higher for detecting random responding than fake good responding.
2. Power to detect random responding will increase as a function of test length, information, and the percent of aberrant items. The highest power will be observed in the 20-item, high information conditions compared to the 10-item medium information conditions.
3. Power to detect fake good responding will be low overall. It will be near zero in the 100% aberrance conditions because it will be impossible to distinguish between responding based on true and inflated (spuriously higher) trait score estimates. Slightly higher power will be observed in the 20% and 50% aberrance

conditions where there will be at least some inconsistencies between expected and observed response probabilities on faked and non-faked items.

4. Power will be highest and Type I error lowest overall when the TRUE statement parameters are used for l_z computations. Higher power and Type I error rates will be observed with MML1000 statement parameter estimates than MML500 statement parameter estimates.
5. Higher power and lower Type I error will be observed when using empirical l_z critical values in place of theoretical critical values for classification decisions, where the theoretical critical values for the 1st, 5th, 10th, and 20th percentiles under standard normal assumptions are -2.33, -1.64, -1.28, and -0.84, respectively.

Power and Type I error results were tabulated and ANOVA was used to test for the statistical significance of main effects and interactions involving up to three variables. Omega-square (ω^2) was used to examine these effect sizes, with values of .01, .06, and .14 representing small, medium, and large effects, respectively (Cohen, 1998). To address some specific hypotheses, a few planned comparisons were also performed.

To provide a visual illustration of l_z efficacy for detecting random and fake good responding, I also computed receiver operating characteristic (ROC) curves, which portray power (hits) as a function of Type I error (false alarms). Good performance is indicated by ROC curves that rise sharply to a level well above a diagonal line of reference corresponding to equal proportions.

The ROCs were computed as follows. First, the l_z values for the samples of normal and aberrant examinees in each condition were sorted and the minimum and

maximum l_z values were identified. Then cut scores (t) for classification decisions were obtained by starting with the lowest value and moving to the highest in increments of 0.1. For each cut score, I computed the proportion of normal examinees that would be misclassified as aberrant at that cut score (x(t)) and the proportion of aberrant examinees who would be correctly classified as aberrant (y(t)). These (x(t), y(t)) data points were used to plot power as a function of Type I error.

RESULTS

Tables 3 through 5 show the average Type I error and power rates across the 100 replications in each simulation condition. In particular, Table 3 presents detailed results for Type I error under conditions of test length (10 and 20 items), test information (medium and high information), type of statement parameters (TRUE, MML1000, MML500), and type of critical values (empirical, theoretical).

As can be seen in Table 3, the Type I error rates for the empirical critical values matched perfectly with the respective nominal alpha levels. Specifically, Type I errors of .01, .05, .10, and .20 were found for the nominal alphas of .01, .05, .10, and .20, respectively, regardless of test length, test information, and the type of statement parameters. In contrast, with the exception of the .01 nominal alpha level, the theoretical critical values resulted in consistently lower than expected Type I errors and the negative bias increased as the nominal alpha increased from .05 to .20. Importantly, however, there were no marked differences in Type I error as a function of the type of statement parameters, test information, or test length.

Table 4 presents the power results for random response detection using empirical and theoretical l_z critical values. Examination of the conditions within Table 4 revealed several interesting patterns. First, with the exception of the .01 nominal alpha, power was slightly higher when using the empirical l_z critical values, which is consistent with the findings of lower than expected Type I error for the theoretical critical values in Table 3. Second, power to detect random responding increased somewhat with test length and test

information, and there was a sharp improvement in overall power as the percentage of aberrant items increased. Importantly, these results clearly show that there was ample power to detect 100% random responding with informative tests, regardless of the type of critical value or statement parameters used for the l_z computations.

Table 5 shows the power results for detecting faking good, which was operationalized as a consistent upward shift in trait scores on items that were designated as aberrant. As can be seen in the table, power to detect faking was poor in every case. In what were optimal conditions for detecting random responding (20 items, high information, 100% aberrance), power for detecting faking was only .16 with empirical critical values and a nominal alpha of .20, and the results were even worse with stricter alphas. Neither test length nor test information had a beneficial effect on power, nor did the use of true statement parameters nor empirical critical values. The only interesting finding is that power was lowest, as expected, in the respective 100% aberrance conditions due to the inability to distinguish an "across-the-board" faker from a truly high-trait responder.

To buttress the interpretation of the power results in Tables 3 through 5 and address the specific hypotheses that were proposed above, an ANOVA and planned comparisons were conducted. Table 6 shows the ANOVA results for main effects and interactions that accounted for at least 1% of the variance in power. All of the factors manipulated were statistically significant ($p < .05$), with the largest effect observed for the type of aberrance. Power was markedly higher for detecting random responding than fake good responding ($p < .0001$; $\omega^2 = .478$), which supported Hypothesis 1.

Note also that the ANOVA results showed a significant and large interaction effect (ω^2 = .231) between response style and percentage of aberrance. This interaction is portrayed graphically in Figure 4.

Hypothesis 2 was also supported. It stated that power to detect random responding would increase with test length ($p < .0001$; ω^2 =.012), test information ($p < .0001$; ω^2 =.017), and the percentage of aberrant items ($p < .0001$; ω^2 = .122). As expected, the highest power was observed in the 20-item, high information conditions and the lowest power was found in the 10-item, medium information conditions.

Hypothesis 3 stated that power to detect fake good responding would be low overall, which was confirmed by the results in Table 5. It also stated that power would be near zero in the 100% aberrance conditions and slightly higher in the 20% and 50% aberrance conditions. Individual planned comparisons supported that finding ($p < .0001$) and a trend analysis based on orthogonal polynomials revealed a statistically significant quadratic effect ($F = 18.67$; $p < .001$).

Hypothesis 4 proposed that power would be highest and Type I error lowest overall when using the TRUE statement parameters for l_z computations, and better power and Type I error rates would be observed with MML1000 statement parameter estimates in comparison with MML500. Although this main effect was statistically significant ($p < .05$), there were no noteworthy differences in power across conditions and the effect size was extremely small ($\omega^2 < .001$). Similarly, the Type I error rates were identical across types of statement parameters in the empirical conditions and only negligibly different in the theoretical critical value conditions.

Hypothesis 5 stated that higher power and lower Type I error would be obtained when using empirical critical values in place of theoretical critical values for aberrance detection. This hypothesis was not clearly supported. Although there was a statistically significant increase in power associated with using empirical critical values for aberrance detection ($p < .05$; $\omega^2 = .003$), that can be attributed to the surprisingly lower than expected Type I error rates for the theoretical critical values shown in Table 3.

Finally, panels (a) through (d) of Figure 5 present ROC curves illustrating the efficacy of l_z for detecting random responding. The sharply rising, nearly right-angle shapes of the 100% random responding ROCs indicate nearly ideal performance – near perfect power with low Type I error. And although power dropped relative to Type I error in the 50% and 20% conditions, the ROC curves were still well above the reference line, indicating solid performance.

In striking contrast, the inability of l_z to detect fake good responding in this simulation is demonstrated by the ROC curves in panels (a) through (d) of Figure 6. Consistent with expectations, the ROC curves for the 100% conditions either straddled or were below the reference line representing equal proportions of hits and false alarms. Slightly better and quite similar performance was observed for the 50% and 20% fake good conditions, but the slowly rising, relatively flat ROCs indicated generally poor performance.

DISCUSSION

The primary goal of this study was to investigate the efficacy of the l_z person fit statistic for detecting aberrant responding with UPP measures constructed and scored based on the Zinnes-Griggs (1974) IRT model, which has proven useful for a variety of noncognitive testing applications in organizational settings (e.g., Borman et al., 2001; Chernyshenko et al., 2009; Houston, Borman, Farmer and Bearden, 2005; Underhill, Lords, & Bearden, 2006). Because UPP measures are now being used to collect both "self-" and "other-" reports, we explored the capability of l_z to detect two of the most common and potentially detrimental response sets, namely fake good and random responding. The effectiveness of l_z was studied using empirical as well as theoretical critical values for classification, along with test length, test information, the type of statement parameters, and the percentage of items answered aberrantly (20%, 50%, 100%).

In short, l_z was ineffective for detecting fake good responding, with power approaching zero in the 100% aberrance conditions. However, l_z was highly effective for detecting random responding, with power approaching 1.0 in the long-test, high information conditions, and there was no diminution in efficacy when using MML estimates of statement parameters in place of the true values. Furthermore, although using empirical critical values for classification provided slightly higher power, theoretical critical values, corresponding to a standard normal distribution provided nearly as good results.

Finding that faking good is difficult to detect is not surprising. If a respondent fakes on a large proportion of items, there will be few apparent inconsistencies in the response pattern, making it difficult to distinguish a spuriously high from a truly high trait score. Similarly, if just a small percentage of items are faked, the likelihood of the response pattern would be very similar to that of a normal responder, which would also reduce hit rates. These results are consistent with the optimal appropriateness measurement findings and conclusions of Zickar and Drasgow (1996), who examined fake good response detection with Likert-type personality scales in an experiment involving coached and ad-lib faking conditions.

Another interesting and important finding was that using MML statement parameter estimates based on samples of 500 yielded power and Type I error rates that were nearly identical to the true parameter values. This is consistent with the findings of small effects in research involving single-statement IRT models (Hendrawan, Glas, & Meijer, 2005; St-Onge, Valois, Adbous, & Germain, 2009). It is also good news for practitioners because the true parameters are never known and, for obvious reasons, pretest samples of 500 and smaller are preferred. In the future, it would be interesting to explore whether subject matter expert (SME) ratings of statement location would be as effective for l_z aberrance detection as the MML500 estimates, given that recent research with a ZG-based computer adaptive test showed little differences between trait scores based on true statement parameters, MML estimates, and SME ratings of statement location (Stark, Chernyshenko, & Guenole, 2011).

Finally, although previous l_z research has raised concern about the use of critical values based on normality assumptions (Nering, 1995; Reise, 1995; van Krimpen-Stoop

& Meijer, 1999), we found that a computationally intensive method of obtaining empirical critical values provided relatively small improvements. In this study, using theoretical critical values (i.e., those corresponding to a standard normal distribution) resulted in lower than expected Type I error rates, which, in turn, reduced power somewhat. In practice, using empirical critical values should enhance classification accuracy, but a reasonable simple alternative might be just to choose a slightly higher theoretical critical value for flagging examinees.

Limitations and Suggestions for Future Research

This study has some limitations that can be addressed in future research. First, it would be interesting to compare the performance of l_z with the performance of model-based detection methods (i.e., optimal appropriateness measurement, OAM; Levine & Drasgow, 1988), which postulate different models for aberrance. Second, it might be beneficial to compare l_z efficacy with empirical and theoretical critical values using trait scores sampled from a negatively skewed distribution, which might better reflect the distribution of job performance scores among experienced incumbents. Finally, it might be interesting to investigate how l_z can be adapted for use with more complex forced choice formats, such as multidimensional pairs or tetrads, and whether the findings for the key variables examined here will generalize.

Implications for Organizations

This research clearly demonstrated that l_z can be an effective method for detecting some forms of aberrant responding with noncognitive measures. It is highly

effective for detecting random responding, which may occur if incumbents become unmotivated because they are surveyed too frequently, if managers are inattentive in evaluating subordinates' performance, or employees or applicants are informed that measures are being administered for "research only" purposes. Flagging unmotivated respondents can be helpful to organizations in identifying courses of action that will increase engagement, such as counseling, incentives or, conversely, sanctions. Organizations should also explore whether simply being flagged as "aberrant" predicts important components of job performance.

This simulation also showed that l_z cannot be recommended for detecting faking good at this time. Although faking good remains a preeminent concern with noncognitive testing, particularly in selection environments, this research indicates that organizations should continue to actively explore other methods for detecting and preventing faking, such as social desirability scales, tracking response latencies, warnings, and multidimensional forced choice formats. Moreover, even if statistically effective faking detection methods are eventually developed, organizations will still have to grapple with what to do with the *individuals* who are flagged. Disqualifying them from an application or promotion process with the looming possibility that a flag is a false positive could have important legal ramifications that would eradicate anticipated utility gains. With that in mind, allowing a retest or conducting a follow-up diagnostic interview might be a more judicious next step in the review process.

TABLE 1. Simulation Study Design

Cell#	Response Style	Test Length(# Items)	Test Information	% Aberrant Items
1	Normal	10	Medium	0
2	Normal	10	High	0
3	Normal	20	Medium	0
4	Normal	20	High	0
5	Random	10	Medium	20
6	Random	10	Medium	50
7	Random	10	Medium	100
8	Random	10	High	20
9	Random	10	High	50
10	Random	10	High	100
11	Random	20	Medium	20
12	Random	20	Medium	50
13	Random	20	Medium	100
14	Random	20	High	20
15	Random	20	High	50
16	Random	20	High	100
17	Fake Good	10	Medium	20
18	Fake Good	10	Medium	50
19	Fake Good	10	Medium	100
20	Fake Good	10	High	20
21	Fake Good	10	High	50
22	Fake Good	10	High	100
23	Fake Good	20	Medium	20
24	Fake Good	20	Medium	50
25	Fake Good	20	Medium	100
26	Fake Good	20	High	20
27	Fake Good	20	High	50
28	Fake Good	20	High	100

TABLE 2. True Statement Parameters for the 20-Item Medium and High Information Tests

	Medium Information				High Information			
Item	μ_s	μ_t	$\lvert\mu_s-\mu_t\rvert$	$(\mu_s+\mu_t)/2$	μ_s	μ_t	$\lvert\mu_s-\mu_t\rvert$	$(\mu_s+\mu_t)/2$
1	0.63	2.12	1.49	1.37	0.13	2.62	2.49	1.37
2	0.01	1.37	1.36	0.69	-0.49	1.87	2.36	0.69
3	-0.72	0.82	1.54	0.05	-1.22	1.32	2.54	0.05
4	1.83	0.29	1.54	1.06	2.33	-0.21	2.54	1.06
5	-2.84	-1.26	1.58	-2.05	-3.34	-0.76	2.58	-2.05
6	-1.24	0.26	1.49	-0.49	-1.74	0.76	2.49	-0.49
7	1.65	0.12	1.53	0.89	2.15	-0.38	2.53	0.89
8	0.63	-0.82	1.45	-0.10	1.13	-1.32	2.45	-0.10
9	-0.85	-2.39	1.54	-1.62	-0.35	-2.89	2.54	-1.62
10	2.89	1.39	1.50	2.14	3.39	0.89	2.50	2.14
Mean	0.20	0.19	1.50	0.20	0.20	0.19	2.50	0.20
11	0.63	2.12	1.49	1.37	0.13	2.62	2.49	1.37
12	0.01	1.37	1.36	0.69	-0.49	1.87	2.36	0.69
13	-0.72	0.82	1.54	0.05	-1.22	1.32	2.54	0.05
14	1.83	0.29	1.54	1.06	2.33	-0.21	2.54	1.06
15	-2.84	-1.26	1.58	-2.05	-3.34	-0.76	2.58	-2.05
16	-1.24	0.26	1.49	-0.49	-1.74	0.76	2.49	-0.49
17	1.65	0.12	1.53	0.89	2.15	-0.38	2.53	0.89
18	0.63	-0.82	1.45	-0.10	1.13	-1.32	2.45	-0.10
19	-0.85	-2.39	1.54	-1.62	-0.35	-2.89	2.54	-1.62
20	2.89	1.39	1.50	2.14	3.39	0.89	2.50	2.14
Mean	0.20	0.19	1.50	0.20	0.20	0.19	2.50	0.20

**Note.* Means in the last row are for the full 20 item tests.

TABLE 3. Type I Error Rates for Empirically and Theoretically Driven Critical Values

			Empirical Critical Values				Theoretical Critical Values			
			Nominal Alpha				Nominal Alpha			
Length	Information	Parameter	.01	.05	.10	.20	.01	.05	.10	.20
10	Medium	TRUE	.01	.05	.10	.20	.01	.04	.07	.13
		MML1000	.01	.05	.10	.20	.02	.04	.07	.13
		MML500	.01	.05	.10	.20	.01	.04	.08	.13
	High	TRUE	.01	.05	.10	.20	.02	.04	.06	.11
		MML1000	.01	.05	.10	.20	.02	.04	.06	.10
		MML500	.01	.05	.10	.20	.01	.03	.05	.09
20	Medium	TRUE	.01	.05	.10	.20	.01	.04	.08	.14
		MML1000	.01	.05	.10	.20	.01	.04	.07	.14
		MML500	.01	.05	.10	.20	.01	.04	.07	.13
	High	TRUE	.01	.05	.10	.20	.02	.04	.07	.13
		MML1000	.01	.05	.10	.20	.01	.04	.06	.12
		MML500	.01	.05	.10	.20	.01	.03	.06	.11

TABLE 4. Power Rates of Random Responding for Empirically and Theoretically Driven Critical Values

				Empirical Critical Values				Theoretical Critical Values			
				Nominal Alpha				Nominal Alpha			
Length	Information	Aberrancy	Parameter	.01	.05	.10	.20	.01	.05	.10	.20
10	Medium	20%	TRUE	.03	.13	.23	.38	.04	.11	.16	.26
			MML1000	.03	.13	.23	.38	.04	.11	.17	.26
			MML500	.03	.13	.23	.38	.04	.11	.16	.26
		50%	TRUE	.15	.40	.53	.69	.19	.37	.47	.58
			MML1000	.16	.40	.54	.69	.20	.37	.47	.59
			MML500	.16	.40	.54	.69	.21	.37	.47	.59
		100%	TRUE	.47	.71	.81	.89	.52	.69	.77	.84
			MML500	.47	.71	.81	.89	.52	.68	.76	.84
			MML1000	.47	.71	.81	.89	.52	.69	.77	.84
	High	20%	TRUE	.08	.22	.34	.51	.11	.20	.29	.36
			MML1000	.08	.23	.35	.51	.09	.18	.25	.34
			MML500	.08	.24	.35	.51	.08	.17	.24	.33
		50%	TRUE	.43	.62	.71	.80	.52	.60	.66	.73
			MML1000	.44	.63	.71	.81	.50	.59	.64	.72
			MML500	.43	.63	.71	.80	.48	.58	.63	.71
		100%	TRUE	.76	.89	.92	.96	.82	.88	.90	.93
			MML1000	.77	.89	.92	.96	.80	.87	.90	.92
			MML500	.76	.89	.92	.96	.78	.86	.89	.92
20	Medium	20%	TRUE	.05	.17	.27	.43	.06	.14	.22	.34
			MML1000	.05	.17	.27	.43	.06	.14	.22	.33
			MML500	.05	.17	.27	.43	.06	.14	.21	.33
		50%	TRUE	.37	.63	.75	.86	.40	.60	.70	.81
			MML1000	.37	.63	.75	.86	.39	.59	.69	.80
			MML500	.37	.63	.75	.86	.39	.59	.69	.79
		100%	TRUE	.81	.93	.96	.98	.83	.92	.95	.97
			MML1000	.81	.93	.96	.98	.82	.91	.95	.97
			MML500	.81	.93	.96	.98	.82	.91	.95	.97
	High	20%	TRUE	.12	.27	.38	.55	.15	.25	.33	.44
			MML1000	.12	.28	.40	.56	.14	.23	.31	.42
			MML500	.12	.28	.40	.56	.13	.22	.29	.41
		50%	TRUE	.70	.86	.91	.95	.75	.85	.89	.93
			MML1000	.71	.86	.91	.95	.74	.84	.88	.92
			MML500	.71	.86	.91	.95	.72	.82	.87	.92
		100%	TRUE	.96	.99	.99	1.00	.97	.99	.99	1.00
			MML1000	.96	.99	.99	1.00	.97	.98	.99	.99
			MML500	.96	.99	.99	1.00	.96	.98	.99	.99

TABLE 5. Power Rates of Fake Good Responding for Empirically and Theoretically Driven Critical Values

				Empirical Critical Values				Theoretical Critical Values			
				Nominal Alpha				Nominal Alpha			
Length	Information	Aberrancy	Parameter	.01	.05	.10	.20	.01	.05	.10	.20
10	Medium	20%	TRUE	.02	.10	.19	.33	.03	.09	.13	.23
			MML1000	.02	.10	.19	.33	.03	.09	.13	.23
			MML500	.02	.10	.19	.33	.03	.09	.13	.23
		50%	TRUE	.03	.11	.19	.32	.04	.10	.14	.23
			MML1000	.03	.11	.20	.34	.04	.10	.15	.23
			MML500	.03	.11	.19	.34	.04	.09	.15	.23
		100%	TRUE	.01	.05	.09	.18	.01	.04	.07	.12
			MML1000	.01	.05	.10	.20	.01	.04	.07	.12
			MML500	.01	.05	.10	.19	.02	.04	.07	.12
	High	20%	TRUE	.04	.16	.25	.45	.06	.15	.22	.27
			MML1000	.04	.17	.26	.45	.05	.13	.19	.25
			MML500	.04	.17	.26	.44	.05	.12	.18	.25
		50%	TRUE	.05	.17	.31	.44	.08	.15	.20	.33
			MML1000	.05	.17	.29	.44	.07	.13	.19	.30
			MML500	.06	.17	.29	.43	.06	.12	.18	.28
		100%	TRUE	.01	.04	.08	.16	.02	.03	.05	.09
			MML1000	.01	.04	.08	.16	.01	.03	.04	.07
			MML500	.01	.04	.08	.16	.01	.03	.04	.07
20	Medium	20%	TRUE	.03	.11	.20	.34	.04	.10	.16	.26
			MML1000	.03	.12	.20	.34	.04	.10	.16	.26
			MML500	.03	.11	.20	.34	.04	.10	.15	.25
		50%	TRUE	.03	.11	.19	.33	.03	.09	.15	.25
			MML1000	.03	.11	.19	.33	.03	.09	.14	.24
			MML500	.03	.11	.19	.33	.03	.09	.14	.24
		100%	TRUE	.01	.04	.09	.18	.01	.04	.06	.12
			MML1000	.01	.05	.09	.18	.01	.04	.07	.12
			MML500	.01	.05	.10	.19	.01	.04	.06	.12
	High	20%	TRUE	.05	.17	.29	.45	.07	.15	.22	.32
			MML1000	.06	.18	.29	.45	.07	.14	.20	.31
			MML500	.06	.18	.29	.45	.06	.13	.19	.29
		50%	TRUE	.04	.15	.26	.41	.06	.13	.20	.30
			MML1000	.05	.16	.27	.41	.06	.12	.18	.29
			MML500	.05	.16	.26	.41	.05	.11	.17	.27
		100%	TRUE	.01	.04	.08	.16	.01	.03	.06	.10
			MML1000	.01	.04	.08	.15	.01	.03	.05	.09
			MML500	.01	.04	.08	.15	.01	.02	.04	.08

TABLE 6. Main Effects and Interactions for Studied Variables on Power Rates

Source	df_B	F	ω^2
Response style (R)	1	77698.7	0.48
Percentages of aberrance (A)	2	9945.61	0.12
Nominal alpha	3	469.65	0.09
Test information	1	284.17	0.02
Test length (L)	1	1961.02	0.01
Type of critical values	1	528.7	0.00
Type of statement parameters	2	3.19	0.00
R*A	2	18474.4	0.23
R*L	1	1859.85	0.01

**Note.* All effects shown were significant at $p < .05$. Only interaction effects that accounted for at least 1% of the variance in power are included. ω^2=proportion of variance accounted for by the independent variables. df_B = degrees of freedom between; for all effects; degrees of freedom within = 392.

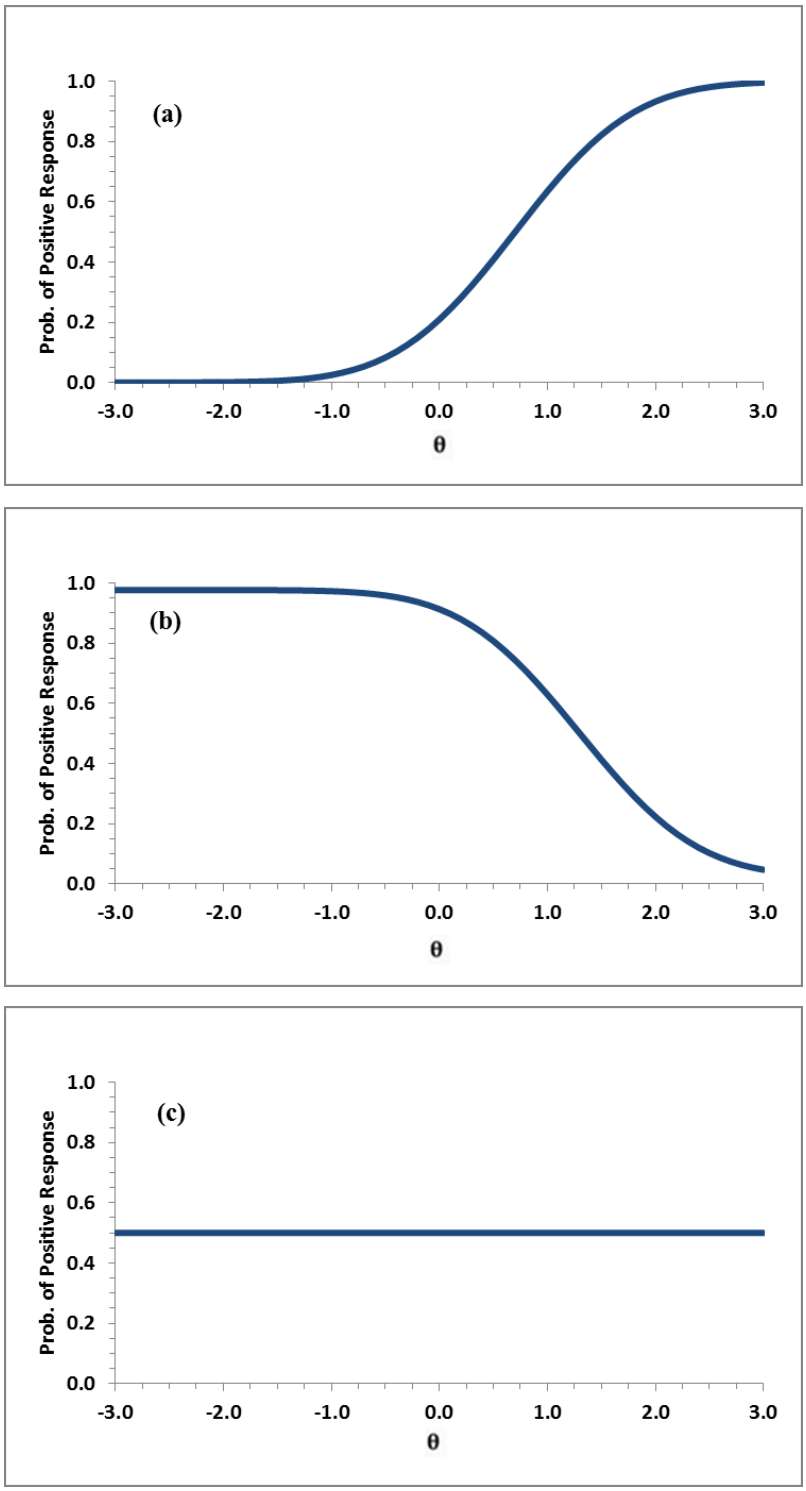

FIGURE 1. Illustrative Item Response Functions (IRFs) for the Zinnes and Griggs Model: (a) the item involves statements having location parameter $\boldsymbol{\mu_s}$= 2.5 and $\boldsymbol{\mu_t}$= -1.1, (b) the item involves statements having location parameters, $\boldsymbol{\mu_s}$= 0.3 and $\boldsymbol{\mu_t}$= 2.3, (c) the item involves statement having location parameters, $\boldsymbol{\mu_s}$=$\boldsymbol{\mu_t}$= 0.9.

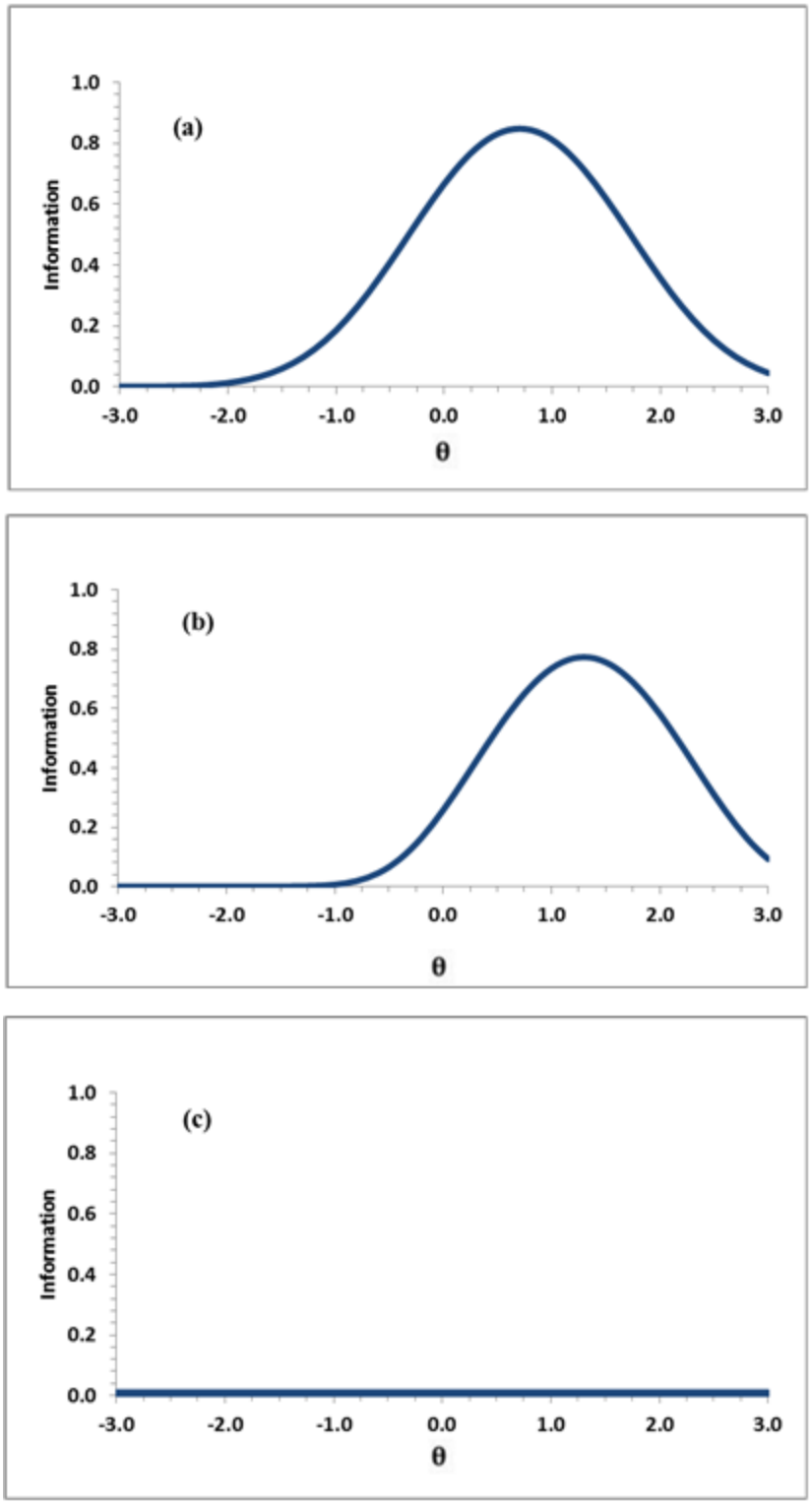

FIGURE 2. ZG Item Information Functions for the Items Appearing in FIGURE 1

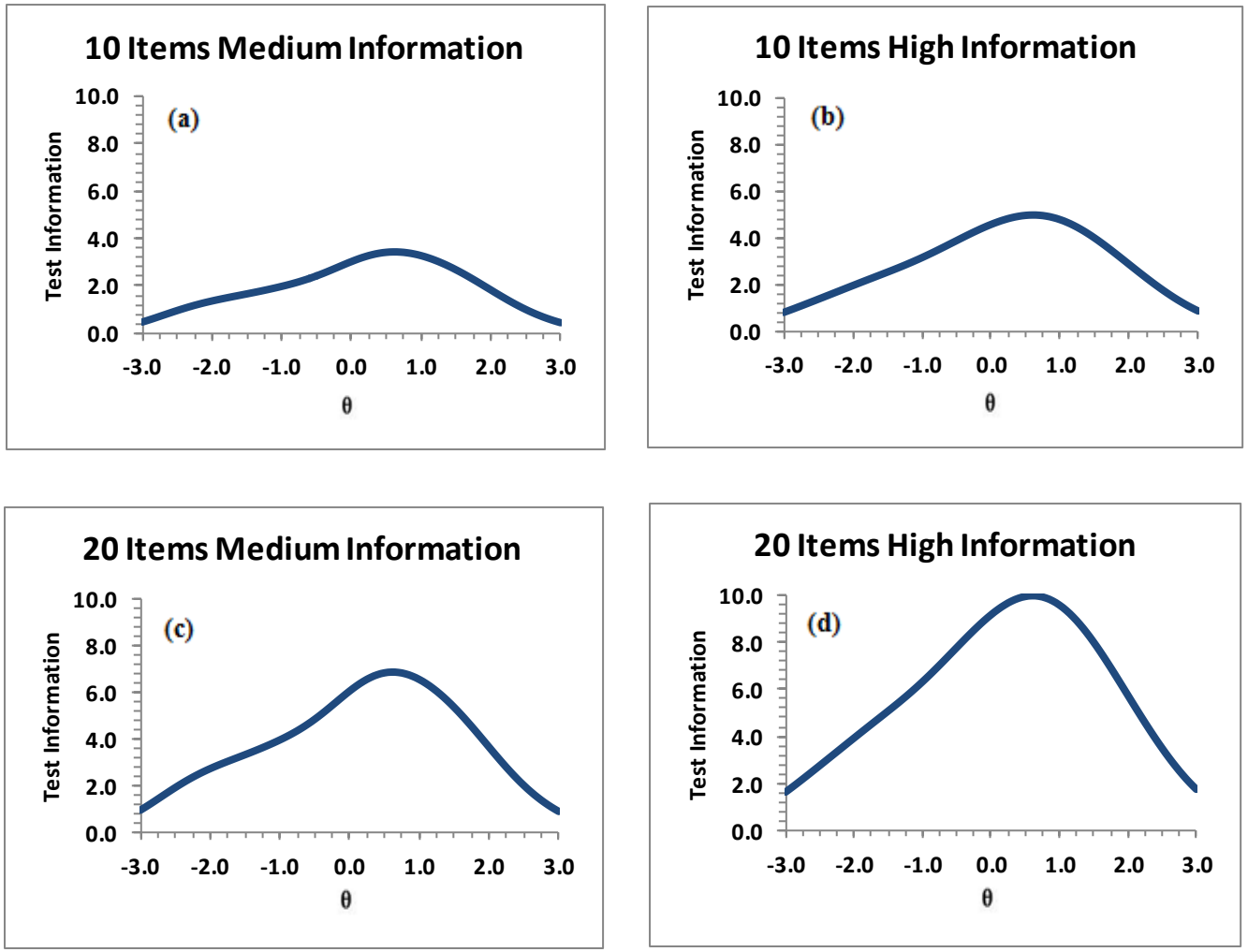

FIGURE 3. Test Information Functions for the l_z Monte Carlo Simulation

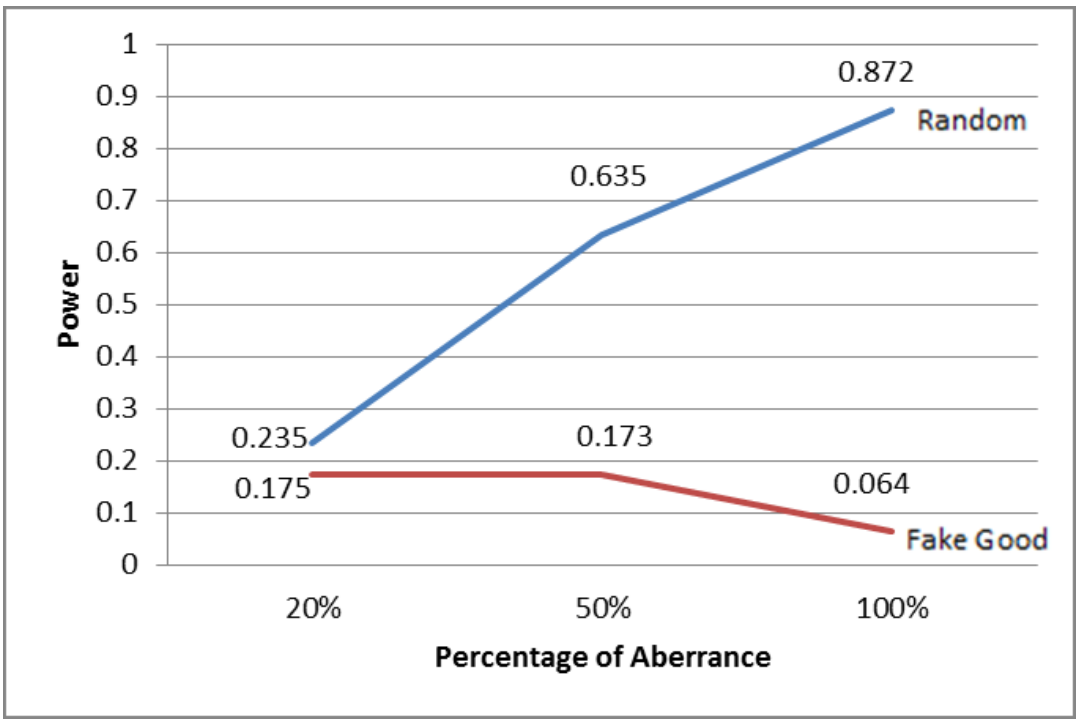

FIGURE 4. Interaction of Response Style and Percentage of Aberrance

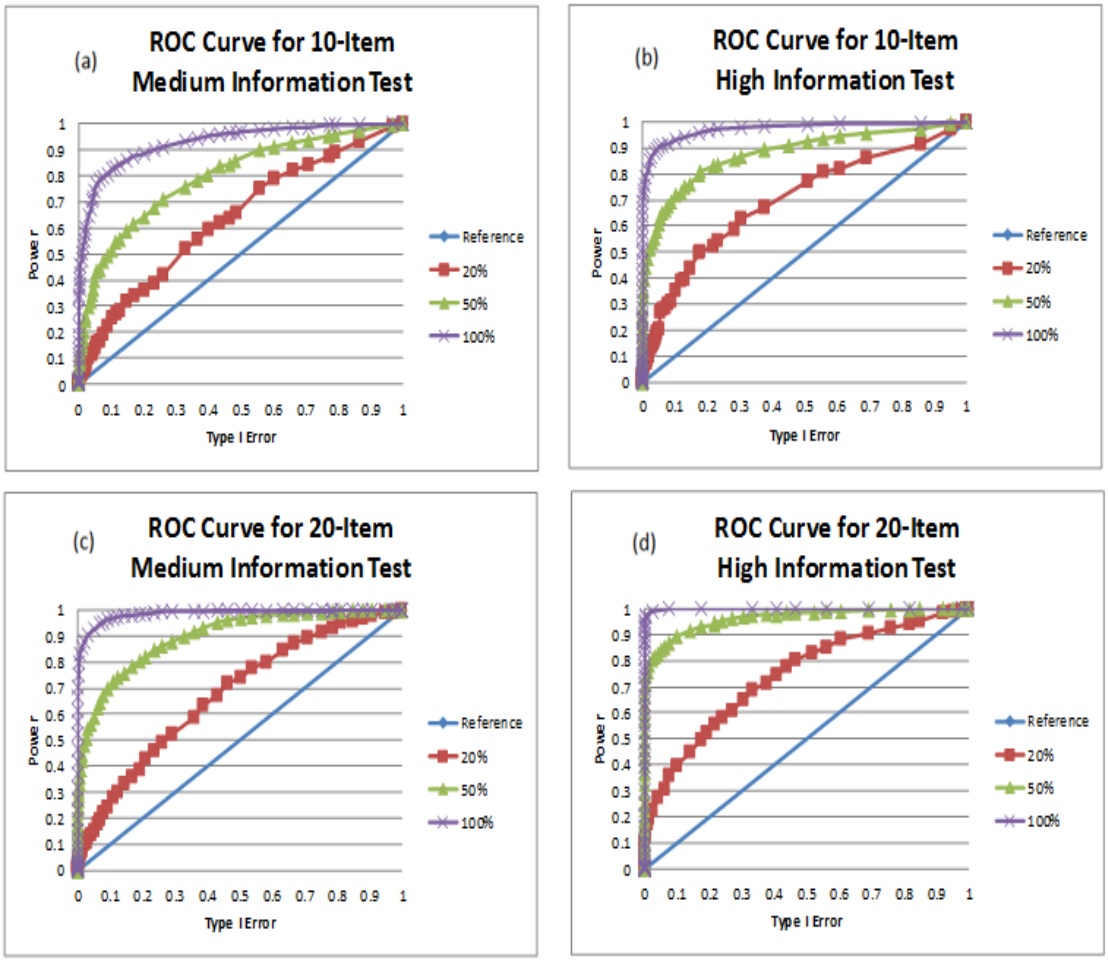

FIGURE 5. Receiver Operating Characteristic (ROC) Curve for Detecting Random Responding

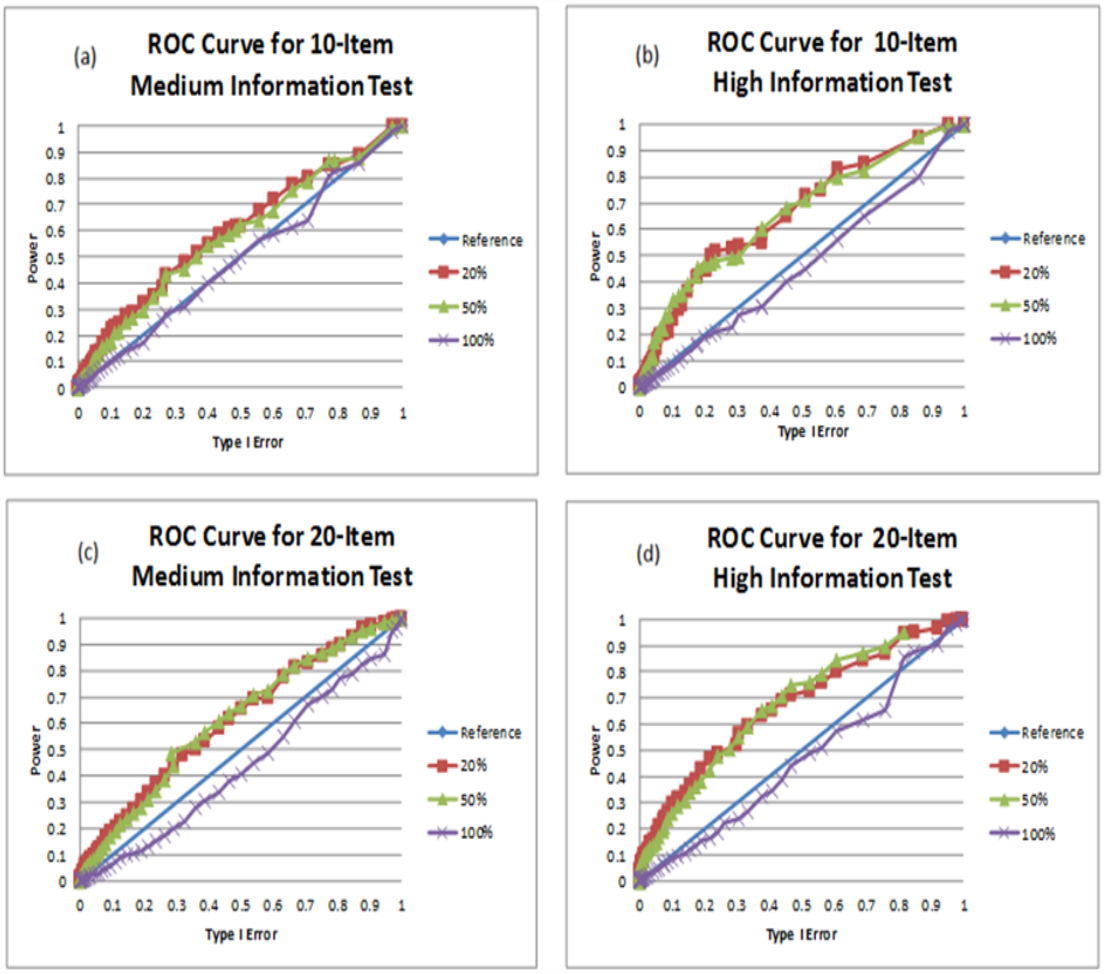

FIGURE 6. Receiver Operating Characteristic (ROC) Curve for Detecting Fake Good Responding

REFERENCES

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15,* 113-141.

Borman, W. C., Buck, D., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86,* 965-973.

Borman, W. C. & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt and W. C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 71–98). San Francisco: Jossey-Bass.

Chernyshenko, O. S., Stark, S., & Williams, A. (2009). Latent trait theory approach to measuring person-organization fit: Conceptual rationale and empirical evaluation. *International Journal of Testing, 9*, 358–380.

Connelly, B.S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity, *Psychological Bulletin, 136*, 1092–1122.

Drasgow, F. (1982). Choice of test models for appropriateness measurement. Applied *Psychological Measurement, 6*, 297–308.

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.

Drasgow, F.,Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59–79.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47–64.

Egberink, J. L., Meijer, R. R., Veldkamp, B. P., Schakel, L., & Smid, N. G. (2010). Detection of aberrant item score patterns in computerized adaptive testing: An empirical example using the CUSUM. *Personality and Individual Differences, 48,* 921–925.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person response function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1–35.

Ferrando, P. J., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection. *Educational and Psychological Measurement, 61,* 997–1012.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327–359.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*, 309–326.

Harnisch, D. L., & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory (pp. 104–122). In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, Canada: Kluwer.

Hendrawan, I., Glas, C., & Meijer, R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29*, 26–44.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184.

Houston, J. S., Borman, W. C., Farmer, W. L., & Bearden, R. M. (2005). *Development of the Enlisted Computer Adaptive Personality Scales (ENCAPS) for the United States Navy, Phase 2 (Institute Report No. 503).* Minneapolis, MN: Personnel Decisions Research Institute.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement.* Homewood, IL: Dow Jones Irwin.

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

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–289.

Levine, M. V., & Drasgow, F. (1998). Optimal Appropriateness Measurement. *Psychometrika, 53*, 161-176.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42–56.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31*, 430–455.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21,* 99–113.

Meijer, R. R. (1998). Consistency of test behavior and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71*, 147–160.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18,* 111–120.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8*, 261–272.

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

Molenaar, I.W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75–106.

Molenaar, I.W., & Hoijtink, H. (1996). Person-fit and the Rasch model, with an application to knowledge of logical quantors. *Applied Measurement in Education, 9*, 27–45.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121–129.

Nering, M. L. (1996). The effects of person misfit in computerized adaptive testing (Doctoral dissertation, University of Minnesota, Minneapolis, 1996). *Dissertation Abstracts International*, 57, 04B.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115–127.

Nering, M. L. & Meijer, R. R. (1998). A comparison of the person response function and the l_z person-fit statistic. *Applied Psychological Measurement, 22*, 53–69.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213–229.

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217–226.

Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit measurement of typical performance applications. *Applied Measurement in Education, 9*, 9-26.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20*, 207–219.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? Human Resource Management Review, 16, 155–180.

Schmitt, N., Chan, D., Sacco, J., McFarland, L., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41–53.

Seybert, J. M., & Stark, S. (2012). Iterative Linking with the Differential Functioning of items and Test (DFIT) Method: Comparison of Testwide and Item Parameter Replication (IPR) Critical Values. *Applied Psychological Measurement, 36*, 495-515.

Smith, P., & Kendall, L. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155.

Snijders, T.A.B. (2001). Asymptotic distribution of person fit statistics with estimated person parameters. *Psychometrika, 66,* 331–342.

Stark, S. (2006). ZG-EAP: A computer program for EAP scoring with the Zinnes-Griggs model. Unpublished manuscript. University of South Florida.

Stark, S., & Chernyshenko, O.S. (2011). Computerized adaptive testing with the Zinnes and Griggs pairwise preference ideal point model. *International Journal of Testing, 11*, 231–247.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (July, 2012). *Development of a person-fit index for multidimensional pairwise preference tests.* Invited presentation at the 8th conference of the International Test Commission. Amsterdam, NE.

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.

Stark, S., Chernyshenko, O.S., Lee, W.C., Drasgow, F., White, L.A., & Young, M.C. (2011). Optimizing prediction of attrition with the U.S. Army's Assessment of Individual Motivation (AIM). *Military Psychology, 23*, 180–201.

Stark, S. & Drasgow, F. (April, 1998). *Application of an item response theory ideal point model to computer adaptive assessment of job performance.* Paper presented at the 13th annual conference for the Society of Industrial and Organizational Psychology. Dallas, TX.

Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26*, 208–227.

St-Onge, C., Valois, P., Adbous, B., & Germain, S., (2009) A Monte Carlo study of the effect of item characteristic curve estimation on the accuracy of three person-fit statistics, *Applied Psychological Measurement, 33*, 307–324.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95–110.

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models (pp. 83–108). In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Underhill, C.M., Lords, A.O., & Bearden, R.M. (2006, October). *Fake resistance of a forced choice paired-comparisons personality measure.* Annual meeting of the International Military Testing Association. Kingston, Canada.

van Krimpen-Stroop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327–345.

Waters, L. K. (1965). A note on the "fakability" of forced-choice scales. *Personnel* Psychology, 18, 187–191.

White, L.A. & Young, M.C. (1998, August). *Development and validation of the Assessment of Individual motivation (AIM).* Paper presented at the Annual Meeting of the American Psychological Association, San Francisco.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71–87.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item level analysis. *Journal of Applied Psychology, 84*, 551–563.

APPENDIX

10 Pairs of CARS UPP Items

In each pair that follows, please choose the statement that better describes the employee you are evaluating.

1a. Approaches work with a strong sense of urgency (e.g., constantly pushes self and others for positive results, has a strong tendency to take action).

1b. In most cases, takes the initiative to complete tasks on or ahead of time.

2a. Is at times overly reactive rather than proactive, but generally produces a reasonably effective product.

2b. Usually seeks help when a work-related problem occurs; is hesitant to initiate action that results in moving forward on important tasks.

3a. Completes own tasks with some initiative, but requires a fair amount of oversight to achieve acceptable standards for most tasks or missions.

3b. Tasks are always completed after the established deadline despite considerable prompting by supervisors.

4a. Generally will complete assigned tasks on time with occasional oversight from own immediate supervisor.

4b. Routinely demonstrates a good ability to complete all assigned tasks by initiating early actions that provide momentum toward task completion.

5a. May gather insufficient or irrelevant data, inaccurately assess available resources, or develop inadequate plans relative to completing work/assignments.

5b. Aims at providing balanced analyses when situations require the integration of input from a variety of sources.

6a. When analyzing data or a problem, effectively identifies the important pieces of information to help solve problems or accomplish different tasks.

6b. Fully understands and analyzes relatively straightforward tasks, and can sometimes provide analyses of more complex tasks.

7a. Understands the issues surrounding most problems or situations, but cannot always apply that knowledge to construct the best possible solution.

7b. Conducts analyses those are usually helpful for decision making.

8a. Resists new directions, priorities, or objectives, but respects the chain of command sufficiently to help implement those changes.

8b. Most of the time effectively adapts to changing situations, but is not as good at adapting in highly ambiguous or uncertain conditions.

9a. Has some difficulty in an environment where changing unit goals create uncertainty, but is able to adjust reasonably well to change and convey new goals to subordinates to meet objectives.

9b. Is sometimes unsure how to help subordinates cope with periods of change and transition.

10a. Is comfortable working with diverse groups of individuals in a broad range of situations and settings.

10b. Thrives in a dynamic work environment, and is always open to new ideas and methods for accomplishing goals.