Using Appropriateness Measurement to Detect Realistic Faking of Personality Tests

Brian Holt
Western Kentucky University
USING APPROPRIATENESS MEASUREMENT TO
DETECT REALISTIC FAKEING OF PERSONALITY TESTS

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts

by
Brian J. Holt

November 8, 2002
USING APPROPRIATENESS MEASUREMENT TO
DETECT REALISTIC FAKING OF PERSONALITY TESTS

Date Recommended: November 8, 2002

Director of Thesis

[Signature]

Dean, Graduate Studies and Research

[Signature] 12/13/02
USING APPROPRIATENESS MEASUREMENT TO DETECT REALISTIC FAKING OF PERSONALITY TESTS

CONTENTS

INTRODUCTION .............................................................................................................1

Faking Personality Tests ...............................................................................................2
Detecting Faking on Personality Measures .................................................................3
Appropriateness Measurement ....................................................................................7
Types of Appropriateness Indices .............................................................................7
Appropriateness Index Research ...............................................................................8
The Present Study .......................................................................................................9

METHOD ....................................................................................................................12

Participants ..................................................................................................................12
Instruments ..................................................................................................................12
Procedure ...................................................................................................................12
Analyses .....................................................................................................................13

RESULTS ....................................................................................................................14

DISCUSSION ...............................................................................................................18

REFERENCES ............................................................................................................20

APPENDIX ................................................................................................................25
Research has shown that personality tests are susceptible to faking and that test takers do indeed take advantage of this vulnerability. This faking creates a problem when organizations use personality tests as screening tools for candidates for employment.

Among the methods available to detect faking, appropriateness measurement (i.e., examining how well a pattern of responses fit item characteristics) has not been thoroughly investigated. The present study examines whether the two most popular appropriateness indices, $Z_3$ and $F_2$, are capable of detecting response distortion among test takers instructed to answer honestly versus fake. The groups demonstrated differences between overall mean scores, but the appropriateness indices did not successfully detect response distortion between the groups.
Introduction

Personality traits are essential individual difference variables for the science and practice of industrial and organizational psychology. Measures of normal personality, most commonly constructs of the five factor model (i.e., extraversion/introversion, agreeableness, conscientiousness, emotional stability, and openness to experience), have been used as selection instruments for decades. Despite the fact that meta-analyses (e.g., Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991) generally have reported only modest relations (e.g., $r = .22$ for Conscientiousness, $r = .28$ for Agreeableness) between tests measuring various constructs of the FFM versus criterion measures of job performance, the popularity of personality tests as selection instruments persists. Although these validity coefficients are substantially lower than those associated with structured interviews or cognitive ability tests (e.g., Huffcutt & Arthur, 1994), personality measures have the benefit of having zero or near zero levels of adverse impact.

Further research concerning the utility of personality testing can be found by exploring the relationship of personality traits and nontraditional measures of job performance. For example, Bernardin (1977) examined absenteeism and turnover using the 16PF and found that conscientiousness and anxiety accounted for most of the variance in both models of organizational withdrawal. A study conducted by White, Nord, Mael, and Young (1993) examined Army enlistees to determine the causes of high dropout rates. They found that enlistees with low emotional stability and high delinquency rates had a higher dropout rate when compared to other enlistees. Thus, personality traits relate to job performance in ways not addressed by performance ratings.
Faking Personality Tests

A major problem with self-report personality tests is their susceptibility to intentional response distortion, or faking. Research has demonstrated that examinees have the ability to substantially distort their scores on an array of personality tests (e.g., Gillis, Rogers, & Dickes, 1990; Krahe, 1989). There is, however, some disagreement on the prevalence and impact of faking in real-world organizational settings.

Hough, Eaton, Dunnette, Kamp, and McCoy (1990) examined test taker faking under varying degrees of motivation to fake. They found that examinees who had no motivation to fake had scores similar to examinees with a motivation to fake, suggesting that little or no faking had occurred. Conversely, Barrick and Mount (1996) reported mean personality score differences between job applicants and incumbents. Although Hough et al. (1990) offered evidence that response distortion does not significantly change personality test validity coefficients, Rosse, Stecher, Miller, and Levin (1998) argued that a comparison of validity coefficients is a poor methodology for the assessment of the impact of faking.

Although the observed (concurrent) validity of the test may not change for the whole sample, its validity for the applicants who are at the top end of the predictor distribution (corresponding to applicants who are most likely to be hired) may approach zero if response distortion occurs primarily among those who receive the highest scores (p. 636).

Their data demonstrated that although correlations with job performance were often similar in faked versus honest response conditions, the actual hiring decisions were very different when applicants faked their answers. In the most extreme case (a selection ratio
of 5%), 88% of the people in the hired sample were hired only because of their extremely high levels of response distortion. Finally, Anderson, Warner, and Spencer (1984) found that almost half of the job applicants for a vast spectrum of positions claimed they had experience with at least one of several imaginary tasks invented by the researchers. In short, although some research suggests that the prevalence and magnitude of personality test faking is minimal, enough contradictory evidence exists that the faking issue cannot be dismissed.

Detecting Faking on Personality Measures

Given the evidence demonstrating that faking occurs on personality measures, organizations need a way to prevent or detect applicant faking. (Note: there are those (e.g., Hogan, Hogan, & Roberts, 1996) who hold the position that personality test faking is a good thing in that those who can successfully fake a test can also successfully moderate their behavior to fit the demands of any situation.) There are three classes of methods available to eliminate or detect faking. The first is to write test items that are difficult to fake. The second is to include validity scales to detect odd responses. The final method is to examine how an examinee responds to the items.

There are two types of items that are difficult to fake: those with verifiable answers and those with ambiguous questions. Becker and Colquitt (1992) found that respondents are less inclined to fake on items for which they perceive their answers can be confirmed. For example, an item such as “Were you a member of any athletic teams in high school?” could be confirmed with little effort. It matters not whether the answer is actually confirmed, only that the respondent thinks his answer will be confirmed. The major problem with constructing tests consisting of only verifiable items as a faking
deterrent is that it limits the type of question that can be asked. The other item format
difficult to fake consists of questions that are ambiguous or less transparent regarding the
construct measured. In other words, if respondents are not sure what the question is trying
to measure, then they will not know the direction in which they should distort their
responses in order to achieve their test taking objectives. One problem with this approach
is that research has shown that subtle or ambiguous items may have lower validity than
more transparent items (e.g., Boone, 1995; Osberg, 1999; Zickar & Ury, 2002).

Much of the research on detecting distortions by measuring odd responses
emanates from the work on the Minnesota Multiphasic Personality Inventory (MMPI)
during the 1940s. The MMPI consists of scales designed to measure abnormal behavior
but also includes scales designed to measure invalid responses (Meehl & Hathaway,
1946). The F scale was designed to detect whether an examinee is malingering (i.e.,
faking bad) and consists of 64 items that measure behaviors and attitudes with very low
frequencies of endorsement in the normal sample. For example, a respondent who
answered false to the statement “Roads without stoplights would be chaotic” would
receive a point on the F scale. The L scale was developed to detect respondents
attempting to appear remarkably socially desirable (i.e., faking good). For example, if the
statement “I research all of the issues in every election” is answered as true, then the test
taker will receive a point on the L scale. Finally, the K scale is used to assess test taker
frankness versus defensiveness and does not directly relate to faking good or bad. The K
scale, however, is used in conjunction with the F scale to produce the F-K malingering
index, which is a popular scale used to detect faking bad.
Many studies have examined the utility of detection scales in the identification of honest versus faked responses (e.g., Bagby, Buis, & Nicholson, 1995; Gillis et al., 1990; Lanyon, 1993). Gillis et al. examined faking on the MMPI using the F – K index with encouraging results. Their results show that 92% of the fakers were correctly identified, whereas only 13% of the non-fakers were misclassified. Lucio, Duran, Graham, and Ben-Porath (2002) examined the extent to which the validity scales of the MMPI-Adolescent could accurately identify individuals who were faking bad compared to individuals who were given standards instructions. They found that when using the F scale there was a 93% detection rate for adolescent girls and a 98% detection rate for adolescent boys. Use of validity scales for faking detection has the unfortunate side effect of lengthening a test. For a test like the MMPI-2, which has 567 items, a substantial reduction in test length could be achieved if the items relating to the validity scales were not needed and, thus, could be removed.

The final method for the detection of faking looks at how test-takers respond. This approach takes one of two forms: examining the amount of time taken to respond to an item and examining whether the pattern of responses is internally consistent given the item characteristics (e.g., difficulty, discrimination).

A response latency is the amount of time an individual takes when responding to an item. It has been hypothesized that respondents who are intentionally faking will take longer to respond, thus causing a greater response latency than if they had answered honestly (Holden, 1995). Holden also hypothesized that job applicants who are lying in order to present themselves positively will endorse some negative characteristics because too many good responses would expose the dissimulation. As a result, applicants who are
faking good will take relatively longer than honest applicants to endorse negative characteristics about themselves. Holden found that respondents in the faking condition had longer response latencies than those instructed to answer honestly. There are several unknowns in the research of response latencies. First, only one study has examined active job seekers and used an inventory specifically designed for personnel selection (Holden, 1995). Second, the value of a response latency index relative to more standard validity measures is unclear. Although, Holden and Hibbs (1995) found that response latencies significantly improved the detection of fakers beyond that associated with a standard validity index, Holden and Kroner (1992) did not find such an increment.

The use of response latencies has three practical limitations. First, the test must be computer administered. Second, differences in item lengths must be standardized. Finally, differences between test takers in terms of reading speed must be standardized. The alternative method for examining how test takers respond, by quantifying the internal consistency of their responses, is based upon the principles of item response theory (IRT).

IRT is a theory of measurement that relates the probability of making a correct response to an item to characteristics of the item and the ability of the test taker. The three-parameter IRT model defines each test item's difficulty (b parameter), discrimination (a parameter), and pseudo-guessing (c parameter). Ideally, a general purpose test designed for use with a diverse population should be composed of items with low c-parameters, high a-parameters, and a range of b parameters. IRT can be applied to both ability tests and personality tests. When applied to personality tests, difficulty (b-parameter) does not mean hard or easy, instead it refers to an item that is likely to be
answered correctly be people with high versus low levels of the trait. Moreover, IRT can be applied to items with dichotomous or polytomous scoring schemes. Finally, IRT can be used to score tests in a more sophisticated manner than traditional number right scoring. As a test taker answers some items correctly and others incorrectly, her ability (called theta and symbolized as \( \theta \)) can be estimated. For example, a test taker who answers all of the easy items correctly and most of the hard items correctly would have a high ability estimate. Conversely, a test taker who misses all of the hard items and some of the easy items would have a low ability estimate.

**Appropriateness Measurement**

Scoring a test becomes more complicated when a given test taker has a pattern of responses that includes missing most of the easy items and very few of the harder items. IRT can be used to examine how well a pattern of responses fits the item characteristics. For example a person who gets 50% of the items correct by answering all of the easy items correctly and missing all of the hard items has a pattern of responses that fits the items better than a person who get 50% of the items correct by answering half of the easy items correctly and half of the hard items correctly. This process of examining the fit of a pattern of responses is called appropriateness measurement. The application of appropriateness measurement to personality testing and faking detection is clear: patterns of responses that do not fit the item characteristics well may be indicative of intentional response distortion by the test taker.

**Types of Appropriateness Indices**

Two appropriateness indices, known as \( Z_3 \) and \( F_2 \), have been proposed by researchers in the attempt to quantify profile invalidity. The \( Z_3 \), sometimes referred to as
standardized \( l_o \), has been examined by Drasgow and associates (e.g., Drasgow, Levine, & Williams, 1985). The \( Z_3 \) index is determined by the height of the likelihood function standardized in an attempt to control for the fact that log likelihood values are not all equally likely at all theta locations. The height (at its maximum) of the likelihood function produced by a given pattern of responses should determine whether distortion has occurred. More specifically, smaller maxima, and thus lower \( Z_3 \) values, are seen as indicative of profiles that are less likely, presumably due to the presence of distorted or internally inconsistent responses (Burnkrant & Harvey, 2000). The more serious the inconsistencies, or the more numerous in the profile, the lower the likelihood would be for the response profile.

Rudner (1983) described the \( F_2 \) as an index that assesses fit by determining the overall degree of deviation from the expected response summarized across items. With the \( F_2 \), a number is derived from the test’s item parameters and from that number it is determined the acceptable amount of deviation possible for honest response patterns. A larger deviation of the score would indicate a more inappropriate response profile.

**Appropriateness Index Research**

Drasgow et al. (1985) found that when aberrant responses are introduced into an otherwise normal set of data the \( Z_3 \) index has a detection rate of 91%. Rudner (1983) examined \( F_2 \) and found varying levels of success: in the best case, the \( F_2 \) statistic correctly identified 75% of those cases with high levels of aberrant responses designed to raise one’s score (faking good). By contrast, at best \( F_2 \) correctly identified only 35% of those cases containing high amounts of responses designed to lower one’s score (faking bad).
Burnkrant and Harvey (2000) examined the Z3 and F2 using the Myers-Briggs Type Indicator (MBTI). The researchers used real data from the administration of the MBTI by a large nonprofit research organization as well as simulation data employing a Monte Carlo methodology similar to previous research of appropriateness measurement (e.g., Drasgow, Levine, & McLaughlin, 1987; Hulin, Drasgow, & Parsons, 1983). The real data group was not manipulated in any way but rather was used as a comparison group to Monte Carlo simulations. Overall, they found encouraging results in terms of high hit rates and low false-positives, suggesting that aberrant responding can be detected in certain situations. However, their analyses also indicate that relatively good rates of detection were found only for moderate-to-large degrees of aberrant responding. In other words, the detection rates for lower levels of aberrant responding (e.g., 0% to 20%) were quite low. These variable detection rates present a problem because the magnitude of most faked responses is not known in applied situations. If large numbers of applicants fake at modest levels, then very few will be successfully identified.

The Present Study

The present study examines the ability of appropriateness indices to detect faking using two types of faking manipulation. To date, neither the Z3 or the F2 have been studied using data from real test takers instructed to fake, a research design often employed in faking research, particularly with validity scale research (e.g., Graham, Watts, & Timbrook, 1991). Zickar and Drasgow (1996) compared the responses from test takers instructed to answer honestly versus fake using an appropriateness index called LRx, for which they failed to find any support. They did not, however, examine the Z3 or the F2 indices. Their study, however, raises an interesting issue regarding the manner in
which faking was induced. Zickar and Drasgow used faking conditions that they referred to as adlib faking. The participants in the adlib faking condition were instructed to fake in a way to make themselves appear in the best light possible (i.e., fake maximally). The problem with this methodology is that faking maximally is not likely the behavior of a typical job candidate. Rather, it is more likely applicants fake just good enough to obtain the job while still declaring (in their responses) a realistic set of behaviors. We refer to this style of faking as realistic faking. Ultimately, the distinction between the styles of faking is an empirical issue. Hauenstein (1998) reported that maximal faking has very different effects on the item parameters than does realistic faking. In short, faking research, whether it involves validity scale research or appropriateness measurement, should induce faking in a realistic manner.

The present study examines participants who were instructed to fake realistically in addition to those instructed to fake maximally and compares the responses of both to a group of subjects instructed to answer honestly. We restrict our hypotheses to the FFM scales of agreeableness and conscientiousness because those scales were found to have the best predictive validity in each of the meta-analyses (Barrick & Mount, 1991; Tett et al., 1991). Moreover, agreeableness and conscientiousness possess the highest face validity, and their items have clear socially desirable responses for job applicants.

Hypothesis 1a: \( Z_3 \) scores for the subjects instructed to fake realistically will be lower than the \( Z_3 \) scores for the subjects instructed to answer honestly.

Hypothesis 1b: \( Z_3 \) scores for the subjects instructed to fake maximally will be lower than the \( Z_3 \) scores for the subjects instructed to answer honestly.
Hypothesis 2a: F2 scores for the subjects instructed to fake realistically will be higher than the F2 scores for the subjects instructed to answer honestly.

Hypothesis 2b: F2 scores for the subjects instructed to fake maximally will be higher than the F2 scores for the subjects instructed to answer honestly.

Finally, we compare appropriateness indices from the realistic and maximal faking groups against each other. Because both Drasgow, Levine, and McLaughlin (1991) and Zickar and Drasgow (1996) found that fakers are easier to detect when faking is more prevalent or extreme, we expect that test takers faking maximally will have appropriateness indices scores more indicative of faking (lower scores for the Z3 and higher scores for the F2) than test takers faking realistically.

Hypothesis 3a: Z3 scores for the subjects instructed to fake realistically will be lower than the Z3 scores for the subjects instructed to fake maximally.

Hypothesis 3b: F2 scores for the subjects instructed to fake realistically will be higher than the F2 scores for the subjects instructed to fake maximally.
Method

Participants

A total of 190 undergraduate students from a large southeastern university participated in the study. Participants received extra credit in exchange for their participation. Participants were randomly assigned to one of three groups: honest responding, realistic faking, or maximal faking. All data were collected anonymously.

Instruments

The personality questionnaire was developed by Brown (1997) and consisted of 127 items designed to measure the five-factor model of personality. The number of items per scale ranged from 16 for openness to experience to 33 for conscientiousness, with 26 items for agreeableness. All items were of a dichotomous forced choice format (e.g., “talkative versus shy”). In the event that the participants felt that neither of the options adequately described themselves, they were instructed to pick the option that comes closer to their behavior. Test retest reliability estimates for the scales range from .69 for agreeableness to .90 for extraversion, with .84 for conscientiousness (Brown). Convergent validity coefficients between the five scales of the questionnaire and the same five scales from the NEO-FFI (Costa & McRae, 1992) ranged from .35 for openness to experience (agreeableness was second lowest at .54) to .80 for conscientiousness (Brown).

Procedure

The honest group was instructed to answer honestly. Additionally, they were reminded that they had not given any identifying information and, thus, there was no
reason for not being completely candid. The maximal faking group was instructed as follows.

"This is a measure of normal personality. Please do not answer honestly, but rather distort your responses to present the best possible image of yourself." Instructions for the realistic faking group were as follows.

This is a test of normal personality functioning. Imagine that you are applying for a job. As part of the application process, you will be completing the following test, a measure of normal personality functioning. Please respond so as to maximize your chances of being hired. Therefore, do not answer the questions truthfully, but answer so that you will be hired. In short, fake this test so that you will get the job. This instrument has several features designed to detect faking. Do your best to avoid detection, while also doing your best to get the job.

Finally, participants in all three conditions were told to avoid disturbing their fellow test takers. After completion of the test, the subjects were debriefed in a separate room.

Subjects in the realistic and maximal faking conditions were questioned by the experimenter whether they had remembered to fake according to instruction throughout the entire test. Two subjects in the faking groups admitted they had answered in an honest, normal fashion (as opposed to faking) at some point during the test. They assisted the researchers in identifying their data, which were immediately deleted.

**Analyses**

$Z_3$ and $F_2$ indices were calculated for each participant for both the agreeableness and conscientiousness scales (see Appendix for complete formulas). For each coefficient type mean differences between groups was examined using one-way ANOVAs.
Results

Descriptive statistics for participants’ theta scores are displayed in Table 1. As can be seen, participants faked to increase their score when asked to do so. Surprisingly, participants completing the conscientiousness scale in the maximal faking condition faked less than did participants in the realistic faking condition.

Table 1

Descriptive Statistics for Theta Scores by Faking Condition

<table>
<thead>
<tr>
<th>Faking Condition</th>
<th>n</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honest</td>
<td>67</td>
<td>.191</td>
<td>1.14</td>
</tr>
<tr>
<td>Realistic Faking</td>
<td>69</td>
<td>.327</td>
<td>0.82</td>
</tr>
<tr>
<td>Maximal Faking</td>
<td>54</td>
<td>.638</td>
<td>0.72</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honest</td>
<td>67</td>
<td>-.101</td>
<td>.90</td>
</tr>
<tr>
<td>Realistic Faking</td>
<td>69</td>
<td>1.47</td>
<td>1.20</td>
</tr>
<tr>
<td>Maximal Faking</td>
<td>54</td>
<td>1.03</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Means and standard deviations of the $Z_3$ and the $F_2$ scores for each faking condition are shown in Tables 2 and 3. One-way ANOVAs were conducted to examine differences among means. The results located in Table 4 indicate that the data do not support our hypotheses for either the agreeableness scale, $F(2,187) = 1.25, p > .05$, or the conscientiousness scale, $F(2,187) = .12, p > .05$, thus no post hoc tests were conducted.
In short, the $Z_3$ index did not successfully discriminate among participants answering honestly, faking realistically, or faking maximally.

Table 2

*Descriptive Statistics for $Z_3$ Scores by Faking Condition*

<table>
<thead>
<tr>
<th>Faking Condition</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honest</td>
<td>.130</td>
<td>.58</td>
</tr>
<tr>
<td>Realistic Faking</td>
<td>.163</td>
<td>.53</td>
</tr>
<tr>
<td>Maximal Faking</td>
<td>.012</td>
<td>.51</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honest</td>
<td>-.054</td>
<td>.74</td>
</tr>
<tr>
<td>Realistic Faking</td>
<td>-.101</td>
<td>.80</td>
</tr>
<tr>
<td>Maximal Faking</td>
<td>-.042</td>
<td>.57</td>
</tr>
</tbody>
</table>

The results located in Table 5 indicate that data for the $F_2$ index do not support our hypotheses for either the agreeableness scale, $F(2,187) = 2.93, p > .05$, or the conscientiousness scale, $F(2,187) = 2.35, p > .05$, thus no post hoc tests were conducted. These results suggest that the $F_2$ index did not successfully discriminate among participants answering honestly, faking realistically, or faking maximally.
Table 3

*Descriptive Statistics for F2 Scores by Faking Condition*

<table>
<thead>
<tr>
<th>Faking Condition</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honest</td>
<td>1.00</td>
<td>.18</td>
</tr>
<tr>
<td>Realistic Faking</td>
<td>.960</td>
<td>.18</td>
</tr>
<tr>
<td>Maximal Faking</td>
<td>1.04</td>
<td>.17</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honest</td>
<td>1.01</td>
<td>.10</td>
</tr>
<tr>
<td>Realistic Faking</td>
<td>.933</td>
<td>.29</td>
</tr>
<tr>
<td>Maximal Faking</td>
<td>.963</td>
<td>.13</td>
</tr>
</tbody>
</table>

Table 4

*Analysis of Variance for Z2 Index*

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faking Condition</td>
<td>2</td>
<td>1.25</td>
<td>.29</td>
</tr>
<tr>
<td>Within Groups</td>
<td>187</td>
<td>(.293)</td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faking Condition</td>
<td>2</td>
<td>0.12</td>
<td>.89</td>
</tr>
<tr>
<td>Within Groups</td>
<td>187</td>
<td>(.513)</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Values in parentheses represent mean square errors.
Table 5

*Analysis of Variance for F2 Index*

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faking Condition</td>
<td>2</td>
<td>2.93</td>
<td>.056</td>
</tr>
<tr>
<td>Within Groups</td>
<td>187</td>
<td>(.032)</td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faking Condition</td>
<td>2</td>
<td>2.35</td>
<td>.099</td>
</tr>
<tr>
<td>Within Groups</td>
<td>187</td>
<td>(.039)</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Values in parentheses represent mean square errors.
Discussion

Regardless of the purpose of testing (i.e., personnel selection or clinical diagnostic) it is desirable to detect examinee faking. Appropriateness measurement is one method that offers hope for the detection of intentional response distortion. Unfortunately, the Z3 and F2 appropriateness indices failed to successfully detect examinee faking on the scales of conscientiousness and agreeableness, two constructs in the FFM. Moreover, considering that one of the faking conditions was a maximal faking manipulation, in which participants were instructed to fake without regard to potential detection, the utility of the Z3 and F2 indices appears bleak.

Theta scores varied by group, indicating that the participants did increase their scores by faking. Contrary to expectations, the maximal faking group did not have the highest theta scores for conscientiousness. Reasons for this outcome are unclear.

One explanation for the failure of the Z3 and F2 indices to detect faking is that test takers in the honest group were somewhat internally inconsistent in their responses to the test items. Given the differences in theta scores between the honest and faking groups (faking groups were 1.3 standard deviations higher for conscientiousness and 0.28 higher for agreeableness), internally inconsistent responding by honest test takers appears to be an unlikely explanation. Moreover, the group that would be expected to be the most internally consistent in their responses, the maximal faking group, had appropriateness indices scores that were the same as the honest group or indicative of greater amounts of internal inconsistency (e.g., the Z3 index for the agreeableness scale). Rather it appears to be more likely that the Z3 and F2 indices are simply unable to detect faking.
This study is important to the area of appropriateness measurement because of the use of real data. The majority of appropriateness measurement research has been conducted using simulation data generated by the researcher. Few studies have been conducted using real data (e.g., Zickar & Drasgow, 1996), but the current research differed from those studies in a number of ways. The major difference is the nature of the faking manipulation employed by the researchers. In the current study the participants were instructed to fake maximally as well as to fake realistically, whereas during the Zickar and Drasgow study participants were instructed to only fake maximally. Although, the type of faking manipulation did change the theta scores in this study, it did not have any effect on the extent of faking detection.
References


Appendix

Equations for $Z_3$ and $F_2$

$Z_3 = \frac{lo - M(\theta)}{(S(\theta))^{1/2}}$

where

$lo = \text{value of log-likelihood function at the examinee's estimated theta}$

$M(\theta) = \Sigma [P(\theta) \times \ln(P(\theta)) + (1 - P(\theta)) \times \ln(1 - P(\theta))]$

$S(\theta) = \Sigma [P(\theta) \times (1 - P(\theta)) \times \ln(P(\theta) / (1 - P(\theta)))]^2$

$P(\theta) = \text{probability of a correct response for three parameter model at the examinee's estimated theta}$

$F_2 = \frac{\Sigma [u - P(\theta)]^2}{\Sigma [P(\theta) \times (1 - P(\theta))]}

where

$P(\theta) = \text{probability of a correct response for three parameter model at the examinee's estimated theta}$

$u = \text{examinee's scored response (0 or 1) to item n}$