Differential Functioning by High and Low Impression Management Groups on a Big Five Applicant Screening Tool

by

Brennan Daniel Cox

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 14, 2010

Keywords: faking, impression management, personality, selection, measurement equivalence, differential functioning

Approved by

Adrian Thomas, Chair, Associate Professor of Psychology
Daniel Svyantek, Associate Professor of Psychology
Jacqueline Mitchelson, Assistant Professor of Psychology
William Buskist, Distinguished Professor in the Teaching of Psychology

Abstract

The degree to which applicant personality test faking constitutes a real world threat is a topic of considerable debate among industrial and organizational psychologists. Researchers have investigated the faking problem using a variety of methodologies, but have found inconclusive results. One method for studying faking involves the use of impression management scales, which are designed to detect individuals' use of intentional response distortion. However, most scales designed to detect applicant faking are too lengthy, too general, or otherwise impractical for use in applied settings.

The current applied research involved the development, implementation, and validation of an eight-item impression management scale for use with the Fitability 5a, a Big Five personality test used for screening job applicants. Applicants' (n = 21,017) scores on the new scale were found to have satisfactory reliability and correlated as one might expect with the five personality scales. Applicants considered to be "fakers" produced meaningful score differences on the agreeableness, conscientiousness, neuroticism, and openness scales, but not the extraversion scale.

Additional tests for measurement equivalence were performed using the item response theory-based differential functioning of items and tests framework developed by Raju, van der Linden, & Fleer (1995). Most personality items (35 of 55) demonstrated differential item functioning (DIF). Only items on the extraversion scale did not exhibit significant DIF. Significant differential test functioning (DTF) was found for each of the

scales that contained DIF items. Correction for DTF by eliminating items with significant DIF was impossible, as DIF was uniform across all items in that the high impression management group demonstrated a higher probability of responding positively to the items (or negatively, for neuroticism) than the low impression management group. These findings suggest that applicant faking is a real world threat to the Fitability 5a, because impression management strongly affected the construct validity of personality measure.

Acknowledgments

I would like to thank Adrian Thomas for his guidance in developing this study and preparing the manuscript. Without his patience, none of this project would have been possible. I would also like to thank Jackie Mitchelson for her statistical assistance, as well as Dan Svyantek, Bill Buskist, and Alan Walker for their reviews and feedback. They helped make this project a rewarding learning opportunity. To my wife Asha, my daughter Annabella, my parents, family, and friends: Thank you for your love and support. You gave me air as I suffocated myself in this project, and I could not have done it without you.

Table of Contents

Abstract ii

Acknowledgements iv

List of Figures viii

List of Tables ix

Chapter 1. Introduction 1

Chapter 2. Literature review 7

A brief history of personality testing for selection 7

The Big Five 7

Personality testing for personnel selection 9

The faking problem 12

Can personality tests be faked? 14

Do applicants fake? 15

Is faking a problem in the real world? 21

Summary of the faking problem 24

Measurement equivalence and the DFIT framework 25

Measurement equivalence 25

The DFIT procedure 26

DFIT and the faking problem 29

The current research 33

Chapter 3. Study 1: Development of the Fitability 5a impression management scale 36
Method.....38
Item development.....38
Participants and procedure.....39
Sample A participants and procedure.....40
Sample A measures.....41
Sample A analyses and results.....42
Sample B participants and procedure.....43
Sample B measures.....43
Sample B analyses and results.....43
Discussion.....45
Chapter 4. Study 2: Is Faking a Problem for the Fitability 5a?.....48
Method.....49
Participants.....49
Measures.....50
Procedure and analyses.....50
Results.....53
DFIT on the Fitability 5a.....54
Agreeableness.....55
Conscientiousness.....55
Extraversion.....56
Neuroticism.....56
Openness.....57

DFIT on the impression management scale 57
Discussion 58
Chapter 5. General discussion 60
Contribution 1: The Fitability 5a impression management scale 60
Contribution 2: Faking matters in the real world 63
Limitations and future directions 68
Implications 74
References 77
Appendices 88
A. Figures 88
B. Tables 93
C. Measures 109

List of Figures

1. Item response function (item characteristic curve).....90

2. Category response function for a five category item.....91

3. Scree plot and Eigenvalues for impression management scale development.....92

List of Tables

1. Validity of selection tests commonly used for predicting overall job performance 94
2. DFIT research on personality test faking and measurement equivalence.....95
3. Items, factor loadings, and internal consistency estimates for the four-factor solution.....96
4. Correlations among the four factors and the BIDR scales.....97
5. Impression management item and scale means and standard deviations 98
6. Scale means and standard deviations for the total sample and high/low impression management subgroups.....99
7. Correlations among the variables for the total sample.....100
8. Correlations among the variables for the high impression management group.....101
9. Correlations among the variables for the low impression management group.....102
10. Agreeableness item means and NCDIF, CDIF, and DTF values for fakers versus non-fakers 103
11. Conscientiousness item means and NCDIF, CDIF, and DTF values for fakers versus non-fakers 104
12. Extraversion item means and NCDIF values for fakers versus non-fakers 105
13. Neuroticism item means and NCDIF, CDIF, and DTF values for fakers versus non-fakers 106
14. Openness item means and NCDIF, CDIF, and DTF values for fakers versus

non-fakers 107

15. Impression management item means and standard deviations by gender and race ...108

Chapter 1

Introduction

Personality tests are becoming increasingly more popular for use in employment selection (Rothstein & Goffin, 2006; Viswesvaran, Deller, & Ones, 2007). Explanations for this growth are due largely to widespread acceptance of the five-factor model for organizing and describing personality constructs as well as evidence that personality measures predict job-relevant criteria without demonstrating adverse impact (Barrick & Mount, 2005). Despite these encouraging developments, most personality assessments used in selection are self-report and may therefore be susceptible to impression management (i.e., faking good). The extent to which applicant faking actually constitutes a real world problem is currently a topic of considerable debate in industrial and organizational (I/O) psychology (e.g., Morgeson et al., 2007; Ones, Dilchert, Viswesvaran, & Judge, 2007; Tett & Christiansen, 2007).

In 2004, a panel of current and former editors of *Personnel Psychology* and the *Journal of Applied Psychology* gathered at the annual Society for I/O Psychology conference to discuss the issue of personality test faking. In an effort to reach closure, these experts concluded that (a) applicants can and do fake on personality tests, (b) faking can affect criterion-related validity, (c) efforts to correct for faking do not resolve this problem completely, and (d) for some jobs, presenting a false, but favorable impression may actually be a desirable applicant characteristic (Morgeson et al., 2007). Tett and

Christiansen (2007) added to these conclusions that (a) applicants tend to fake to different degrees, (b) individual differences in faking can upset the rank-order of applicants, and (c) faking attenuates, but may not necessarily destroy personality test validity. Ones et al. (2007) responded to these declarations with expressed concern that some of the panel's conclusions may be unwarranted because of the different methodologies used in the studies reviewed by the panel. To date, most researchers have examined the issue of personality test faking using traditional methodologies, such as comparing mean scores of laboratory participants directed to respond honestly and then fake good, or by conducting meta-analyses comparing applicants to non-applicants (Ones et al., 2007). The results of studies are largely mixed, with some studies concluding that faking is a problem and others not. The debate over faking provides a need for additional research examining the faking problem using alternative methodologies than those typically used in faking studies, as well as research investigating the faking phenomenon as it occurs in real world settings and with real world job applicants.

Most research investigating the consequences of faking examines whether faking disrupts the criterion-related validity of personality tests. One lesser-used approach that may inform the debate involves testing to determine if faking disrupts the measurement properties of personality tests, such as the construct validity of these tests. It would be considerably problematic, for instance, if personality tests (or test items) were found to function differently for individuals that fake versus those that do not fake (or fake to a lesser degree). This form of measurement bias refers to the concept of measurement equivalence and holds profound implications for organizations that use the results of personality tests for making employment-based decisions.

Measurement equivalence occurs when "the relations between observed scores and latent constructs are identical across relevant groups" (Drasgow & Kanfer, 1985, p. 662). In the presence of measurement equivalence, the items on a personality test should be equally accurate for individuals who fake versus those who do not fake. Thus, two sets of applicants who have the same standing on a latent personality construct (e.g., conscientiousness) should score identically on a test of this construct, regardless of their respective differences in faking. It should not matter if one set of applicants fakes good and the other does not – the relationship between their standing on the construct and their observed scores should remain identical.

Alternatively, in the absence of measurement equivalence, a test could potentially favor members of one group over the other: Two sets of applicants with identical levels of conscientiousness would respond differently to the items on a conscientiousness measure. Applicants engaging in impression management, for instance, might respond more favorably to desirable items than applicants not engaging in impression management, despite having equal standing on the latent construct. The key issue with measurement equivalence is not whether faking results in inflated scores on personality tests (though this outcome is likely to occur), but whether faking affects how respondents interpret the test or test items. Such a scenario would compromise the validity of the personality measure and make the interpretation and use of scores for making selection decisions impossible. Demonstrating measurement equivalence between fakers and non-fakers is therefore an issue of high practical importance for organizations that use personality tests for selection.

Researchers can test for measurement equivalence at the item-level or across an entire scale or test by conducting analyses based on differential item functioning (DIF) or differential test functioning (DTF), respectively. Although numerous procedures are available for assessing DIF and DTF, the differential functioning of items and tests (DFIT) procedure developed by Raju, van der Linder, and Fleer (1995) has proven particularly useful for organizational researchers. The DFIT procedure not only identifies items with significant DIF, but it also determines the effects of eliminating such items on the overall functioning of the test. Thus, psychometricians can use the DFIT procedure to evaluate and potentially correct for differential responding by members of different groups, thereby increasing measurement equivalence. In addition, because the DFIT procedure works with dichotomous as well as polytomous models, it applies to most measures used in employment contexts, including personality tests.

Previous researchers have used the DFIT procedure to examine the measurement equivalence of personality tests used for selection across groups of fakers and non-fakers. Flanagan and Raju (1997) applied this technique on the extraversion scale of the 16-PF and Henry and Raju (2006) used the DFIT procedure to examine measurement equivalence on an empirically derived conscientiousness scale of the California Psychological Inventory (CPI). In both studies, the researchers evaluated item-level and scale-level scores on the personality measures for differential functioning by comparing high and low/average scorers on the impression management scales included with these measures (used to represent fakers and non-fakers, respectively). With the exception of a few minor differences, they found that the measures functioned in the same manner for each group. Thus, both studies concluded that faking, as measured with impression

management scales, might not be a significant problem for personality tests used for selection.

The DFIT studies by Flanagan and Raju (1997) and Henry and Raju (2006) were limited, however, in three key ways. First, both studies examined only one personality dimension; as such, the conclusions reached apply only to the scales used in these studies. It is possible that applicant faking could affect the validity of alternative personality tests that contain additional or different scales. Second, neither the 16-PF nor the CPI measure the five-factor model of personality, the most commonly accepted model of personality. Henry and Raju even had to derive their conscientiousness scale empirically from items intended for other CPI scales in order to assess this Big Five construct. It is possible that a personality test designed to assess the Big Five factors directly could produce different results. Finally, as a third limitation, the response scale for the16-PF uses a three level forced-choice format and the response scale for the CPI uses a dichotomous True/False format. Therefore, individuals who take these tests are restricted to a narrow set of response options, that could produce a restriction of range in their overall scores. It is possible that a personality test that uses a response scale with more than three options could influence the degree to which applicants fake. One goal of the current research was to assess these possibilities by applying the DFIT framework to investigate the impact of real world applicant faking on all five scales of a true Big Five measure of personality, one that uses a polytomous, five-category response format.

In this dissertation, the measure of interest was the Fitability 5a. The Fitability 5a is a Big Five personality test "specifically designed for job applicant populations" (Lucius, 2003, p. 40). Thousands of job candidates complete this measure each month for

employment screening purposes, including applicants to Fortune 500 companies. Observations since the induction of the Fitability 5a, however, indicate that applicants and non-applicants score differently on this measure. One potential explanation for these score differences is that some applicants might be faking on the Fitability 5a in order to increase their desirability to the hiring organization. Although many personality tests (e.g., the 16-PF and CPI) have custom scales that detect applicant faking, no such measurement device existed for the Fitability 5a. Therefore, in Study 1, a scale was developed for assessing impression management on the Fitability 5a. Next, in order to investigate the influence of faking on the Fitability 5a, in Study 2 DIF and DTF analyses were performed across high and low impression management groups to evaluate whether faking affects the measurement equivalence of the Fitability 5a's scales.

Chapter 2

Literature Review

A Brief History of Personality Testing for Selection

Organizational researchers have investigated the use of personality tests in employment contexts for over 100 years. In reviewing this body of work, Barrick, Mount, and Judge (2001) outlined two distinct phases. The first phase, which lasted from the early 1900s to the mid-1980s, was largely pessimistic and is often summarized using Guion and Gottier's (1965) cautionary conclusion: "It is difficult… to advocate with a clear conscience, the use of personality measures in most situations as a basis for making employment decisions about people" (p. 160). This sentiment was justified for several reasons, including the fact that the researchers of this period lacked a proper system for managing the vast complexity of personality traits used to describe people (Barrick et al., 2001). The past 30 years, however, has seen a surge of support for the use of personality measures in the workplace. The second phase of Barrick et al.'s (2001) history, which continues today, developed from years of converging evidence in support of the unifying five-factor model for classifying personality traits.

The Big Five. Early efforts to produce a definitive taxonomy for organizing personality attributes began with Galton's (1884) lexical hypothesis, which proposed a complete catalog for all personality traits could come from sampling the vocabulary (i.e., lexicon) people use to describe each other. In applying Galton's theory, Allport and

Odbert (1936) performed an exhaustive dictionary search that produced a list of nearly 18,000 personality-type adjectives, which they synthesized into a more manageable list of 4,500 distinct personality traits. Cattell (1957) later condensed this list to 171 terms and eventually, via factor analysis, derived 16 comprehensive factors that he considered fundamental to describing normal personality (Cattell, Eber, & Tatsuoka, 1970). Several efforts to replicate Cattell's work, however, were largely unsuccessful, as many researchers were unable to derive more than five general factors of personality (e.g., Borgatta, 1964; Fiske, 1949; Norman, 1963; Smith, 1967; Tupes & Christal, 1961). At the time, the leading personality theorists overlooked these findings in support of more established theories. Indeed, it was not until Goldberg's 1981 lexical analysis that the five-factor model, or the "Big Five," gained popular applied acceptance (Digman, 1990).

The Big Five personality factors include agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience. Each of these major traits subsumes a larger number of more specific traits, which often appear as contrasting adjectives to characterize high and low scorers on Big Five measures (McCrae & Costa, 1987). Agreeableness represents several opposing trait comparisons, including the degree to which a person is good-natured versus irritable, courteous versus rude, lenient versus critical, flexible versus stubborn, and sympathetic versus callous. Traits associated with conscientiousness include dependable, hardworking, and organized, versus careless, unreliable, and lazy. Extraversion concerns one's interpersonal style, with high scorers (i.e., extraverts) tending to be sociable, energetic, and assertive, while low scorers (i.e., introverts) tend to be more reserved, lonesome, and quiet. The neuroticism factor (or emotional stability, conversely) represents a person's tendencies toward negative

emotions, like whether one is generally calm, relaxed, and hardy, instead of worrying, nervous, and vulnerable. Lastly, the openness factor characterizes individuals as intellectually curious, creative, and daring, versus conventional, cautious, and straightforward.

The introduction of the Big Five marked the beginning of what Barrick et al. (2001) called the renaissance of personality testing. As these authors discussed, the majority of studies conducted since the mid-1980s have used some variant of the five-factor model to conceptualize personality. This model generalizes across cultures and rating formats (e.g., self, peer, observer), and evidence suggests that these traits are heritable and stable over time (Costa & McCrae, 1992). Although the Big Five should not be considered the end-all model of personality (McAdams, 1992), there is a consensus among trait theorists that these five factors can be used to describe all human personality traits effectively (Cervone & Pervin, 2007). The introduction of the five-factor model provided hope in light of Guion and Gottier's (1965) early warnings—once researchers and practitioners had an agreed upon system for categorizing traits, they seemed less hesitant in using personality tests for employment purposes.

Personality testing for personnel selection. Perhaps the most compelling argument in support of the use of personality tests for employment decision-making came in the early 1990s, when separate meta-analyses conducted by Barrick and Mount (1991) and Tett, Jackson, and Rothstein (1991) concluded that organizations can use Big Five measures to predict job-relevant performance criteria. In their study, Barrick and Mount examined the Big Five in relation to job proficiency, training proficiency, and personnel data (e.g., salary level, turnover, status change, tenure) for various occupational groups.

Of all of the Big Five dimensions, conscientiousness demonstrated the most consistent relationships with *all* criterion types and for *all* occupational groups. For this reason, Barrick and Mount recommended that conscientiousness be considered the primary personality variable of interest for organizational researchers and personality test practitioners (Mount & Barrick, 1995).

In addition to their evidence promoting the use of conscientiousness for workplace research, Barrick and Mount (1991) also determined that each of the remaining Big Five factors were valid predictors for *at least one* criterion variable and for *at least one* occupational group. Extraversion, for example, predicted all three performance criteria for individuals employed in both sales and management positions. Thus, for these two occupational groups, being outgoing, sociable, and assertive (i.e., extraverted), as opposed to inactive, quiet, and reserved (i.e., introverted), was associated with better performance on the job, in training, and across personnel data (e.g., higher pay, less turnover). In addition, scores on the openness factor predicted training proficiency for all job categories, including sales, management, police, skilled/semi-skilled, and professional occupations. This finding suggested that individuals who are intellectually curious and willing to change (i.e., open to experience) tend to perform better in training than individuals who are suspicious or narrow-minded, regardless of occupation.

In the same year as Barrick and Mount's (1991) influential meta-analysis, Tett et al. also concluded that the Big Five factors contribute to the prediction of job performance. These researchers extended Barrick and Mount's findings by calculating the average validity of the Big Five factors taken together as well as by examining the

potential moderating role of whether organizations used a job analysis to determine their choice of personality test. Their analyses produced a corrected mean scale validity of .38 for studies that relied on a job analysis to select the appropriate personality test to use, compared to .29 for studies that did not rely on job analysis. Thus, Tett et al. concluded that the criterion-related validity of personality tests used in selection is higher when there is a conceptual link between the test and the position under study, as determined through a job analysis. Taken together, the meta-analyses by Barrick and Mount and Tett et al. provided sufficient empirical support to counter Guion and Gottier's (1965) early warnings. As Tett et al. stated in their concluding remarks: "Personality measures have a place in personnel selection" (p. 732).

With renewed confidence in the practical validity of personality tests, organizational researchers have conducted thousands of studies to determine the overall value of these measures in job-related contexts. In a prototypic example, Schmidt and Hunter (1998) compared the predictive validity of conscientiousness scales versus 18 other selection procedures. Their results (see Table 1) indicated that conscientiousness measures produced an average validity coefficient of .31 for predicting overall job performance. Although this estimate was lower than the .51 validity coefficient for tests of general mental ability (*g*), Schmidt and Hunter noted that tests of *g* do not explain the total variability in job performance *and* they have a history of producing adverse impact, or unintentional discrimination against members of protected groups (e.g., racial minorities). For this reason, organizations are encouraged to supplement their selection battery with measures that are not *g*-loaded, such as personality tests (Gatewood & Feild, 2001).

Organizations rely on personality tests for employment selection because personality tests reliably predict job outcomes (e.g., Barrick & Mount, 1991; Tett et al., 1991; Hurtz & Donovan, 2000) without demonstrating adverse impact (Hogan & Holland, 2003; Hogan, 2005). Because personality test scores do not correlate with scores on *g*-based tests (Ackerman & Heggestad, 1997), and even explain additional variance in employee performance beyond the variance accounted for by *g* (Schmidt & Hunter, 1998), they can add value to most selection test batteries currently in use. That is, organizations can use personality measures in conjunction with other measures to increase the prediction of job performance as well as to help reduce adverse impact resulting from their selection batteries (Hough, Oswald, & Ployhart, 2001).

In sum, personality tests have a place in selection research as well as practice. Measures of the Big Five are particularly useful for four main reasons. First, they assess a wide range of personality traits (Cervone & Pervin, 2007). Second, they correlate with a wide range of job-related performance variables (Ones et al. 2007). Third, they apply to a wide range of occupations (Barrick & Mount, 1991). Fourth, Big Five personality test scales are unlikely to produce adverse impact in employment selection (Foldes, Duehr, & Ones, 2008). Therefore, although the history of workplace personality testing has had its share of criticism (see Barrick et al., 2001), it seems worthwhile to pursue the practice of personality testing for employment selection.

The Faking Problem

Despite major advances in work-related personality test theory, research, and application over the last 100 years, one considerable threat to the validity of these instruments remains. Most personality tests used by organizations are self-report, and

therefore may be susceptible to biased responding. Unlike *g*-loaded or job knowledge tests, which consist of items with one correct answer choice and which require a particular ability level to answer correctly on a consistent basis, most personality tests contain self-report items with response choices that vary in desirability depending on the testing context. Given the high-stakes nature of most selection contexts, some applicants may be inclined to distort their responses to personality tests in order to appear more attractive to the hiring organization. This possibility represents *the faking problem*, the next hurdle for personality researchers to overcome.

Applicant faking constitutes "a conscious effort to manipulate responses to personality items to make a positive impression" (Zickar & Robie, 1999, p. 551). The assumption behind the faking problem is that some applicants may be able to manipulate their responses to personality tests in such a way as to appear more attractive to the hiring organization and thereby increase their chances of gaining employment over more honest and potentially more qualified applicants. Any faulty hiring decisions that result from applicant faking have the potential to affect the organization negatively, as it is unlikely that the selected "fakers" will be able to uphold their false impression forever. Given the potential implications of the faking problem, it is no surprise that faking research has received a surge of popularity in the past several years (Griffith & McDaniel, 2006). Most faking research attempts to answer three general questions: Can personality tests be faked? Do applicants fake? Is faking a problem in the real world? Researchers have investigated these questions using a variety of methodologies and have produced mixed and therefore inconclusive results (e.g., Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Hogan, Barrett, & Hogan, 2007; McFarland & Ryan, 2000; Morgeson et al.,

2007; Ones et al., 2007; Tett et al., 2007). A review of this literature demonstrates a need for additional research investigating the faking problem.

Can personality tests be faked? Most personality tests used by organizations are self-report. Applicants, then, have the potential to inflate their responses to these measures in order to appear more desirable to the hiring organization. To assess the degree to which examinees can fake on a personality test, researchers typically conduct lab studies using some form of direction manipulation. For example, McFarland and Ryan (2000) instructed student participants to respond to the items on the NEO Five Factor Inventory (Costa & McCrae, 1989) under two experimental conditions. In one condition, they instructed students to answer the items "as honestly as possible"; in the other condition, they instructed students to answer "in such a way as to make [them] look as good an applicant as possible for a job [they] would want" (p. 815). They demonstrated that examinees are capable of producing statistically significant score changes on all five scales of the personality test. The effect sizes (Cohen's *d*) for these changes were 1.82 for conscientiousness, 1.66 for neuroticism, 1.06 for agreeableness, 0.98 for extraversion, and 0.19 for openness. Thus, with the exception of the openness to experience factor, student participants produced significant and meaningful changes in their personality tests scores when instructed to fake good.

Winkelspecht, Lewis, and Thomas (2006) conducted a similar lab-based directed faking study. These researchers found that when they instructed participants to fake good, participants not only produced higher scores, but they also rose to the top of the score distribution. Thus, they concluded that, in a top-down selection scenario, individuals who

fake good on personality tests are capable of improving their chances of obtaining employment over more qualified, honest respondents.

Lab studies using faking instructions are useful for demonstrating the extent to which examinees can fake on a personality test as well as potential consequences of faking on personality test validity. However, lab studies are limited in that students instructed to fake good may not represent true applicant behavior. In lab studies, there are no negative consequences for faking. Therefore, when participants in these studies receive instructions to *fake good* or *present the most favorable impression*, they are likely to maximize their faking efforts. It is unlikely that individuals in a true selection setting respond in this manner, given the high stakes nature of the selection context and the potential repercussions for being caught faking. Of course, to determine applicant behavior, there is no better source of data than actual job applicants.

Do applicants fake? To investigate whether actual job applicants fake, researchers tend to use one of three general methodologies: (a) between-subject designs comparing applicants to non-applicants, (b) within-subject designs comparing individuals in applicant and non-applicant conditions, and (c) research designs that use scales designed to detect applicant faking. Each of these methodologies contributes to the personality test faking literature in different ways, as discussed in the following examples.

One method for determining whether (and to what extent) actual job applicants fake on personality tests involves making score comparisons between samples of applicants and non-applicants (e.g., students, incumbents). The rationale behind this between-subject methodology is that if applicants score differently than non-applicants,

then these score differences may be attributable to applicant faking. Birkeland et al. (2006) conducted a meta-analysis of studies using this approach, and discovered that, across all job types, applicants scored significantly higher than non-applicants on measures of conscientiousness (d = .45), emotional stability (d = .44), openness (d = .13), and extraversion (d = .11). A comparison of these effect sizes to those obtained in McFarland and Ryan's (2000) direction manipulation study suggests that actual job applicants do not inflate their scores to the same magnitude as students instructed to fake good. Nevertheless, Birkeland et al.'s study provided evidence that actual job applicants tend to score differently than non-applicants, presumably because some members of the applicant group fake good. This study represents the general findings of between-subject studies of faking comparing applicants to non-applicants.

Between-subject studies comparing applicants and non-applicants are limited, however, in that score differences between the applicant and non-applicant groups could arguably be due to true group differences or other factors beyond faking effects. Therefore, some faking researchers have adopted within-subject methodologies to compare scores from the same individuals within applicant and non-applicant conditions.

As an example of the within-subject methodology, Ellingson, Sackett, and Connelly (2007) compared individuals' scores on the California Personality Index (CPI; Gough & Bradley, 1996) based on one of four naturally occurring test-retest conditions (i.e., selection-development, selection-selection, development-selection, or development-development). Individuals' scores from the selection context represented the faking condition while their scores from the development context represented the non-faking condition. Thus, any score differences between the selection and development conditions

were attributed to the effects of faking. After controlling for time between test administrations as well as feedback effects, Ellingson et al. estimated an average personality score change effect size of 0.08 for the two faking conditions. Therefore, they concluded that applicant faking was not a problem for the CPI.

In another within-subjects study, Griffith, Chmielowski, and Yoshita (2007) examined score changes on a conscientiousness scale for individuals who took the measure first as job applicants and later for research purposes. In the research condition, participants received instructions first to respond honestly and then to fake good. Consistent with previous direction manipulation studies, participants significantly inflated their scores on the conscientiousness scale when instructed to fake good, such that their scores on this measure were uncorrelated with as well as significantly more positive than scores from the respond honestly and applicant conditions. Mean scores from the applicant and respond honestly conditions, however, were significantly correlated with one another ($r = 0.50$, $p < .001$, $d = 0.61$), which suggested that most applicants were either not faking or were not faking to a great extent.

On the surface, the within-subject studies by Ellingson et al. (2007) and Griffith et al. (2007) appear to suggest that, although individuals are capable of faking on personality tests, they do not tend to do so to a troublesome degree in applied contexts. However, there are a few standout characteristics of these studies that might suggest otherwise.

Ellingson et al. (2007) compared individuals in selection and developmental conditions and concluded that faking is not a problem on the CPI, but their study is limited in two key ways. First, unlike most personality measures that use polytomous

response scales (e.g., Likert-type scales with 5-7 response options), the CPI features a dichotomous true-false response scale. As such, test-takers are limited in the degree to which they can distort their responses on the CPI without completely misrepresenting their true personalities. Whereas polytomous response scales allow users to inflate their responses to a small degree (e.g., from *moderately agree* to *strongly agree*), users of the CPI must transition from one extreme of the scale to the other (i.e., from *true* to *false*) in order to fake their responses. A second limitation to this study involves the data analysis. Specifically, these researchers averaged scores on all 20 of the CPI's scales together before correcting for test-retest time delay and feedback effects. After making score corrections and averaging the 20 CPI scale scores together, these researchers found minimal changes in scores. However, prior to correction, the test-retest effect sizes for the individual CPI scales were as high as 0.64, which suggests that at least some scales exhibited substantial score changes. Upon further consideration, Ellingson et al.'s conclusions are limited in that they may only apply to dichotomous personality tests, which constitute only a small percentage of personality tests used by organizations, and potentially unfounded because they computed score differences across the CPI's scales, rather than within the scales individually.

Additional characteristics warrant identification in the Griffith et al. (2007) study. These researchers focused their within-subjects study on a single measure of conscientiousness and determined that, although individuals' scores from applicant and respond honestly conditions correlated significantly with one another, they did not correlate significantly with scores from the faking condition. However, by comparing the rank order of the top 10 applicants across conditions, Griffith et al. demonstrated that

many of their participants would have stood less a chance of being hired had selection decisions been based on their respond honestly scores. Indeed, six of the top 10 scorers in the applicant condition exhibited score changes of more than half a standard deviation when asked to respond honestly, with an average effect size of 0.71. Thus, although applicant scores correlated significantly with responses from the honest condition, the two conditions resulted in vastly different rank orderings of participants' conscientiousness scores. If the organization in this study made its selection decisions using a top-down strategy, there is a good chance that a different group of individuals would have gained employment had the current group of applicants provided more honest responses than those they actually provided.

As a group, the within-subjects studies of applicant personality test faking provide mixed results. Although they are methodologically more powerful than between-subjects designs, because they eliminate any true group differences, it appears that differences in measurement and analytic procedures result in different findings. As an alternative research design, some researchers have chosen not to define faking using score changes between response conditions, but have instead operationalized faking using scales designed to assess socially desirable response patterns. The assumption behind this alternative methodology is that some measures of social desirability may serve to identify applicants that intentionally manipulate their responses to personality test items, or otherwise "fake good" on these tests.

Traditionally, researchers have conceptualized social desirability as a unidimensional construct that describes individuals' tendencies to present themselves favorably on self-report items (e.g., Crowne & Marlowe, 1960; Ones, Viswesvaran, &

Reiss, 1996). However, Paulhus (1984) recognized that there are actually two distinct dimensions of social desirability that differ based on intention. The first dimension is *self-deception enhancement*, which Paulhus described as an unintentional or natural tendency to consider oneself favorably, but falsely. With self-deception, individuals truly believe their positive self-enhancements, regardless of the accuracy of these beliefs. In contrast, the second form of social desirability, termed *impression management*, represents a deliberate misrepresentation of oneself. This dimension more closely relates to the concept of faking because it involves purposeful response distortion. Indeed, these concepts are so similar that researchers often use the terms *faking* and *impression management* interchangeably in the personality testing literature (e.g., Hogan et al., 2007; Mueller-Hanson, Heggestad, & Thornton, 2006).

To assess the dual conceptualization of social desirability, Paulhus (1984, 1991; 1998) constructed the Balanced Inventory of Desirable Responding (BIDR), a 40-item inventory containing two subscales, one for self-deception enhancement and one of impression management. According to Paulhus (1984), researchers can use the BIDR in conjunction with self-report personality tests to control for the effects of dishonest responding (i.e., impression management scores). Li and Bagger's (2006) meta-analysis of studies using the BIDR and Big Five personality measures found that applicant scores on the impression management scale correlated most highly with conscientiousness (ρ = .42, SD = .11) and agreeableness (ρ = .42, SD = .11), the two personality dimensions considered most related to job performance (see Barrick & Mount, 1991; Tett et al., 1991). These results suggested that applicants who score highly on the most desirable personality scales also tend to endorse impression management items, which could be

troublesome for organizations selecting on personality, as applicants' scores may not represent honest responding. However, when Li and Bagger partialed impression management from personality, the criterion-related validity of the personality assessments remained essentially unchanged. Other researchers using score-correction methodologies have found similar results (e.g., Barrick & Mount, 1996; Christiansen, Goffin, Johnston, & Rothstein, 1994; Griffith, Malm, English, Yoshita, & Gujar, 2006), and these findings have been used as evidence that faking is not a legitimate concern when it is operationalized using scores on impression management scales (e.g., Ones et al., 1996). Nevertheless, the BIDR and related scales remain in use today, as some researchers believe they assess a construct of considerable importance to employment contexts (e.g., Smith & Ellingson, 2002; Viswesvaran, Ones, & Hough, 2001).

Is faking a problem in the real world? Organizational researchers have used each of the methodologies described above not only to determine if and to what extent applicants fake, but also to determine if faking actually matters. Studies like Griffith et al.'s (2007) examine the effects of faking on the criterion-related validity of personality tests, such as whether faking affects the rank-order of applicants. Other studies investigate the effects of faking on the construct validity of personality tests. For example, using the direction manipulation methodology, Ellingson, Sackett, and Hough (1999) compared scores from a respond honestly and a fake good group on a multidimensional personality test used for selection by the U.S. Army. A confirmatory factor analysis (CFA) for the respond honestly condition confirmed the intended factor structure for the test. However, the CFA for the faking condition did not support the hypothesized factor structure of the instrument. A follow-up analysis revealed that the

data from the faking group actually supported a unidimensional model of personality. Therefore, Ellingson et al. (1999) concluded that personality test faking can affect construct validity by reducing (or eliminating) the factor structure of the test. Because changes in the factor structure would change the interpretation of the test results, the use of this measure for making employment decisions would be highly questionable.

In another directed faking study, Douglas, McDaniel, and Snell (1996) compared the performance appraisal ratings of a sample of honest respondents to a sample of respondents instructed to fake good and found significant differences in the criterion-related validity for each group (.31 and -.01, respectively). These researchers concluded that, when a large number of people fake, the predictive validity of personality tests lowers substantially.

Using a between-subjects design, Schmit and Ryan (1993) compared the structure fit of a Big Five personality test using data from applicants and non-applicants. Their CFA confirmed the five-factor structure for the non-applicant data, but not for the applicant data. A follow-up analysis revealed that the applicant group produced a sixth factor consisting of items intended for each of the five personality scales and to which agreement indicated an extremely positive self-bias (e.g., hard-worker, likable, committed). Schmit and Ryan labeled this factor the "ideal-employee factor" (p. 970) and concluded that faking has the potential to disrupt the construct validity of personality tests by introducing unintended scales to these measures.

Using the faking scale methodology, Rosse, Stecher, Miller, and Levin (1998) administered a Big Five measure and the BIDR to samples of applicants and incumbents. Applicants scored significantly higher than incumbents on agreeableness,

conscientiousness, emotional stability, and extraversion, as well as on the impression management scale of the BIDR, which suggested that some applicants intentionally distorted their responses. Similar to Griffith et al. (2007), Rosse et al. then rank ordered the applicants by their personality scores and found that, when using a top-down selection strategy, applicants with the highest impression management scores were overrepresented at the top of the distribution. Indeed, if the top 5% of applicants were hired, seven of eight would have impression management scores considered extreme (i.e., 3+ standard deviations above the mean). This ratio dropped to 9 of 16 when the top 10% were hired, though this proportion still equated to over 50% of selected applicants having impression management scores indicative of severe intentional distortion.

Most recently, Peterson, Griffith, O'Connell, and Isaacson (2008) utilized a within-subjects design to investigate whether applicant faking predicted individuals' engagement in counter-productive work behaviors (CWBs) once they were hired. A preliminary analysis found a non-significant correlation between individuals' conscientiousness scores as applicants and their CWB scores as incumbents. However, after removing data from the participants who exhibited statistically significant conscientiousness score changes between test administrations, Peterson et al. obtained a statistically significant improvement in the criterion-related validity of their measure, which suggested that faking affected the criterion-related validity of the conscientiousness scale. Interestingly, applicants' scores on the social desirability scale were unrelated to their conscientiousness scale score changes, suggesting that social desirability scales may not be useful indicators of actual applicant faking. However, because these researchers used a unidimensional social desirability scale (i.e., the

Marlowe-Crowne short form; Crowne & Marlowe, 1960), they were unable to discuss these findings in terms of impression management alone (i.e., in the absence of self-deception enhancement). Nevertheless, applicants' social desirability scores correlated significantly with their CWB scores as incumbents, thereby supporting the argument that some social desirability response patterns may legitimately predict meaningful work outcomes (e.g., Smith & Ellingson, 2002; Viswesvaran et al., 2001).

Summary of the faking problem. Previous research on personality test faking has assumed a number of forms. Some studies investigated whether tests *can be* faked while others asked if they *actually are* faked. To address these concerns, researchers have sampled from non-applicants, true applicants, or some combination of the two. The results of these studies vary by sample, context, and test, but the general conclusions are that personality tests are susceptible to faking and that faking occurs in actual selection settings (e.g., Birkeland et al., 2006; Griffith et al., 2007; Rosse et al., 1998; Viswesvaran & Ones, 1999). Although researchers cannot determine the true prevalence of applicant faking (estimates range between 15% and 63% of applicants fake; e.g., Dunnette, McCartney, Carlson, & Kirchner, 1962; Donovan, Dwight, & Hurtz, 2003; Griffith et al., 2007), the knowledge that even some applicants fake presents the more critical question: Is faking a problem?

Research on the consequences of applicant faking has produced mixed results. Some studies suggest that faking may disrupt the construct and criterion-related validity of personality tests (e.g., Schmit & Ryan, 1993; Ellingson et al., 1999; Douglas et al., 1996; Rosse et al., 1998). However, methodological weaknesses and confusion regarding the interpretation of many faking studies may limit their applied value (Ones et al., 2007).

Additional research utilizing applied samples and comparatively more advanced analytic techniques than the majority of the existing faking research may offer new perspectives on the issue of applicant faking and its associated effects on personality measures.

Measurement Equivalence and the DFIT Framework

Measurement equivalence. A principle assumption behind psychological measurement is that a well-developed scale (or item) may be used to make inferences about some unobservable characteristic(s) of the test taker. These unobservable characteristics, or *latent constructs*, include abilities (e.g., intelligence), attitudes, and traits (e.g., personality characteristics). With valid psychological measures, individuals' raw scores provide observable indicators of their standing on the latent construct as measured by the scale or item (Lord & Novick, 1968). It follows that, with valid psychological measures, any two people with identical standing on a latent construct should produce the same expected scale or item-level scores (Drasgow & Kanfer, 1985). This testing property represents the concept of measurement equivalence.

Measurement equivalence occurs when "the relations between observed scores and latent constructs are identical across relevant groups" (Drasgow & Kanfer, 1985, p. 662). Measurement equivalence is necessary in order to make score comparisons across different groups of test takers. In the context of employment selection, a test must demonstrate measurement equivalence if the organization intends to use scores on the test to make selection decisions. For tests that lack measurement equivalence (i.e., result in differential functioning), the comparison of scores for members of different groups becomes difficult, if not impossible. Researchers can assess measurement equivalence (or differential functioning) at the item-level as well as the scale-level. Among the statistical

techniques available to assess measurement equivalence, the differential functioning of items and tests (DFIT) framework proposed by Raju et al. (1995) has proven particularly useful for organizational researchers.

The DFIT procedure. The DFIT framework is based in item response theory (IRT), which assumes a nonlinear relationship between individuals' latent trait/ability levels, termed *theta levels,* and their observed scores on a test item or scale (Lord & Novick, 1968). Different IRT models exist for different types of items. For dichotomously scored items (i.e., items with two response alternatives; e.g., correct/incorrect, true/false), IRT models estimate the probability that an individual will respond to the item successfully based on the individual's theta level and the characteristics of the test item, such as the item difficulty and item discrimination parameters (Hambleton & Swaminathan, 1985). This relationship is represented graphically as the item characteristic curve, or item response function (IRF; see Figure 1). Most personality tests, however, do not feature dichotomous scoring formats, opting instead for multiple response categories. Such tests require polytomous IRT models.

Because polytomous items feature multiple response categories, there are no true correct or incorrect responses to these times. Rather than estimate the probability of answering the item successfully, as found in dichotomous IRT models, polytomous IRT models estimate the probability that an individual will respond to each response category given the individual's theta level (Oshima, Kushubar, Scott, & Raju, 2009). Thus, for every response category, polytomous IRT models estimate a separate IRF, which are termed category response functions (CRFs; see Figure 2). As depicted in Figure 2, at any given theta level, the sum of the probabilities of the CRFs should equal one. For a more

detailed description of dichotomous and polytomous IRT models, see Oshima et al. (2009).

Organizational researchers have been relatively slow to incorporate IRT models into their research partly because IRT models require complex statistical equations (Ellis & Mead, 2002). However, computer programs are becoming increasingly available for estimating and equating IRT item parameters, thereby simplifying the analysis of measurement equivalence for work-related tests. The DFIT computer program based on Raju et al.'s (1995) framework, for instance, recently published its eighth edition (Oshima et al., 2009). Raju et al.'s DFIT procedure is one of the many IRT-based methods for assessing measurement equivalence, with the major difference being the DFIT procedure assess the differential functioning of both items (DIF) and tests (DTF), and applies to dichotomous as well as polytomous models. These qualities make the DFIT framework appropriate for examining the measurement equivalence of most personality measures used by organizations for personnel decision making.

The DFIT procedure provides an estimate of measurement equivalence (or differential functioning, conversely) by comparing the item parameters of two subgroups of respondents: the focal group and reference group. Users of the DFIT methodology define group membership, and can create groups based on any relevant characteristics of the examinees (e.g., race, gender, intelligence). After placing the item parameters on the same scale as the focal group, the DFIT program compares the IRFs or CRFs for each group, depending on whether the item is dichotomous or polytomous. Measurement equivalence exists when the response functions are identical. If the response functions

differ, then there is evidence of differential functioning or measurement inequivalence (Oshima et al., 2009).

The DFIT program produces three indices of interest. The first index is the noncompensatory DIF (NCDIF) index, which is an item-level estimate of differential functioning that focuses on each item individually. In calculating NCDIF, the DFIT program assumes all other items are free from DIF. Users of this program tend to rely on the NCDIF index when making decisions regarding item-level data. The second index is the compensatory DIF (CDIF) index, which is an item-level estimate of differential functioning that does not assume all other items are free from DIF. Unlike with NCDIF, the CDIF index accounts for correlated differential functioning among test items; therefore, items that exhibit DIF in opposing directions (i.e., one item favors the focal group and another favors the reference group) can cancel each other out using the CDIF indices (Henry & Raju, 2006). The third index calculated by the DFIT program is the DTF index, which is a scale-level estimate of differential functioning. Users of the DFIT program that are concerned with the scale-level performance of a measure tend to use the DTF index in conjunction with the CDIF indices to delete items individually until the scale no long exhibits significant DTF. This strategy results in the deletion of fewer items than if one were to remove items based on the NCDIF values; however, it also permits the resulting modified scale to contain items with significant CDIF (i.e., assuming these items exhibit DIF in opposing directions).

As a prototypic example of the application the DFIT methodology to personality testing, Mitchelson, Wicher, LeBreton, and Craig (2009) recently evaluated the Abridged Big Five Circumplex of personality traits for differential functioning by gender and

ethnicity. Of the personality measure's 45 scales, 17 displayed NCDIF for gender and 28 displayed NCDIF based on ethnicity (33 of 45 scales exhibited DIF, altogether). Findings of DIF were not uniform in that some items favored women while others favored men, and some items favored Caucasians while others favored African Americans. Therefore, in many cases the CDIF indices cancelled each other out, resulting in no evidence of DTF for gender and evidence of only two scales exhibiting DTF by ethnicity. Depending on one's level of focus (i.e., item-level or scale-level), evidence of differential functioning in this study may or may not be viewed as problematic. As most organizations consider scale-level scores in making personnel decisions, evidence of DTF on only 2 of 45 may not pose a serious threat to test validity. However, if one's theoretical orientation considers each individual item as a test in and of itself, then the finding that 33 of 45 scales exhibited DIF may be an issue of sizeable concern. Either way, this study demonstrated the practical use of the NCDIF, CDIF, and DTF indices and provided encouragement for organizational researchers interested in using the DFIT methodology for assessing measurement equivalence on personality items and scales.

DFIT and the faking problem. Most organizational studies that employ the DFIT methodology compare members of protected groups (e.g., men versus women, Caucasians versus racial minorities) as a means of determining compliance with equal employment laws. Relatively few studies have applied the DFIT framework to examine measurement equivalence among other relevant groups. Nevertheless, the DFIT methodology appears in a limited selection of research investigating applicant faking by comparing groups of fakers versus non-fakers on a variety of personality tests (e.g., Flanagan & Raju, 1997; Henry & Raju, 2006; Robie, Zickar, & Schmit; 2001; Stark,

Chernyshenko, Chan, Lee, & Drasgow, 2001; Zickar & Robie, 1999). Table 2 provides a summary of the methodologies and results of these studies.

As Table 2 indicates, DFIT researchers investigating the measurement equivalence of personality tests across groups of fakers and non-fakers have employed a variety of methodologies, often with conflicting results. For instance, these researchers have grouped fakers and non-fakers using (a) fake good and respond honest instructions, (b) high and low/average impression management scores, (c) applicants and non-applicants, or (d) a combination of the latter two grouping methods. The personality measures examined thus far include the 16-PF, the military's ABLE test, the Personal Preference Inventory, and an empirically derived conscientiousness scale from the CPI. Overall, the results of these studies suggest that faking can disrupt the measurement equivalence of some personality measures, but that DIF and DTF are more likely to occur when researchers instruct examinees to fake good or when they compare applicants to non-applicants. DFIT studies comparing high to low/average impression management groups indicate that personality measures tend to function equivalently for members of these groups (Flanagan & Raju, 1997; Henry & Raju, 2006; Stark et al., 2001). A closer examination of the limitations of these studies, however, provides a need for additional DFIT research on the faking problem before one can draw any conclusions from the extant research employing this methodology.

Flanagan and Raju (1997) and Stark et al. (2001) suggested that faking is not a problem for the 16-PF when grouping respondents using scores on the 16-PF's impression management scale. However, when comparing groups of applicants and non-applicants, Stark et al. found evidence of differential functioning. One potential limitation

to these studies that might explain these divergent results is that the 16-PF uses a three-category response scale. Thus, there is little room for individuals to engage in impression management without completely distorting their responses. It may therefore come as little surprise that analyses comparing groups based on the 16-PF's impression management scale did not produce evidence of DIF or DTF, as respondents are restricted in their ability to fake good on this scale. An impression management scale with more than three response options, such as a five-point Likert-type scale, may well produce different results, because it would permit individuals to inflate their responses without misrepresenting their beliefs entirely.

This same issue is also a notable limitation to Henry and Raju's (2006) study, as the CPI uses a dichotomous, True/False response scale. Again, respondents are restricted in the degree to which they can fake on the CPI, which might explain why these researchers did not find faking to affect the measurement equivalence of their empirically derived conscientiousness scale. Henry and Raju's choice of personality measurement, too, may limit the results of their study. Although conscientiousness is an important job-related personality variable, the CPI does not directly assess this construct. Researchers may obtain different results using a more established measure, one designed with the explicit purpose of assessing conscientious.

Using the military's ABLE personality test, Zickar and Robie (1999) demonstrated that when instructed to fake good individuals are capable of disrupting the measurement equivalence of a personality test. These researchers found DTF on two of the three scales they examined when comparing groups instructed to fake good and respond honestly. However, direction manipulation studies are historically limited in that

respondents who are instructed to fake good under controlled conditions tend to fake more than applicants in a true selection context. Thus, although this study demonstrated that DTF can result from faking, it may not represent true applicant behavior. Researchers may obtain different results using samples of actual job applicants rather than individuals instructed to behave like applicants.

Finally, Robie et al. (2001) concluded that the PPI functions equivalently for applicants and incumbents: They found little evidence of DIF or DTF in their study. However, their sample size was below the recommended sample size for performing IRT-based analyses, which may have biased the results. Robie et al.'s focal group consisted of 999 applicants and their reference group consisted of 796 incumbents. Ideally, IRT-based analyses require 200 respondents per response option (Drasgow, 1989). Because the PPI uses a five-point rating scale, each group would have needed 1,000 participants in order to meet these recommendations. Researchers may obtain different results using larger samples than used by Robie and colleagues.

Research using the DFIT procedure to test for measurement equivalence on personality measures across groups of fakers and non-fakers has produced mixed results, with some studies finding considerable evidence of DIF and DTF (e.g., Stark et al., 2001; Zickar & Robie, 1999) and others not (e.g., Flanagan & Raju, 1997; Henry & Raju's, 2006; Robie et al., 2001). One potential reason for these differences derives from different conceptualizations of faking. Some studies conceptualized faking using impression management scores, while others compared applicants to non-applicants or honest to fake-good groups. Another reason for a lack of converging evidence concerns measurement issues. For instance, some studies used personality scales with relatively

few response options, thereby minimizing the degree to which respondents could fake good. Also, studies such as Robie et al.'s did not achieve the requisite sample size for the analyses they performed. As Table 2 indicates, relatively few faking-related DFIT studies have focused on the same model of personality, let alone the same measure. The studies described above followed some combination of Cattell's framework (Cattell et al., 1970), the five-factor model (Goldberg, 1981), Hogan's socioanalytic theory (1982), folk concepts (Gough & Bradley, 1996), and models of managerial job performance (Davis, Skube, Hellervik, Gebelien, & Sheard, 1996). Although the Big Five is the most commonly accepted model, none of the existing DFIT research examining fakers and non-fakers has done so with a true Big Five measure.

The Current Research

This dissertation developed from both a research-oriented as well an applied need. From a research perspective, this dissertation sought to investigate the issue of applicant personality test faking further through application of the IRT-based DFIT methodology. This methodology is statistically more rigorous than the majority of faking research, as it analyzes differences in response functioning between fakers and non-fakers as opposed to differences in total scores. One reason why this methodology does not appear more often in faking studies is that IRT analyses require considerably large sample sizes. The current research adds value to the faking literature by providing data from over 20,000 real world job applicants who completed the measures in a real world selection context.

From an applied perspective, this dissertation developed from the need to determine whether applicants fake on the Fitability 5a personality inventory, and, if so, to determine if applicant faking was a problem for this measure. Thus, in Study 1 an

impression management scale was developed for exclusive use with the Fitability 5a to assess applicant faking on this test. The Fitability 5a impression management scale is unique to this Big Five measure in that its items were written to appear similar to the Fitability 5a's items and feature the same response format. These properties make it less likely that applicants will identify the impression management items for what they truly are, and will instead perceive them to assess the traditional Big Five traits instead. In Study 2, the new impression management scale was implemented in an actual selection context in which applicants completed the Fitability 5a as part of the job candidate screening process. After dividing the applicant sample into high and low scorers on the impression management scale, measurement equivalence was examined across these groups on all 55 items and five scales of the Fitability 5a to determine if faking produced DIF or DTF on this measure. This latter effort also served to validate the impression management scale.

This research differs from similar DFIT research investigating the effects of applicant faking on the measurement equivalence of personality-based selection measures in a number of ways. First, the sample not only exceeded the recommended sample size for IRT-based research, but also consisted of real-life job applicants completing the measures in a true selection context. Second, this study conceptualized faking using high and low scorers on an impression management scale designed exclusively to detect applicant faking using a five-point Likert-type scale. Third, this study assessed the measurement equivalence of all five factors of a true Big Five personality test; that is, none of the scales were derived empirically or post hoc, as in the case of Henry and Raju (2006). For these reasons, this dissertation serves as a logical next step to investigating

the faking problem using scores on a custom-built impression management scale and the IRT-based DFIT methodology, thereby addressing both research-oriented and applied needs.

Chapter 3

Study 1: Development of the Fitability 5a Impression Management Scale

Organizations have at their disposal a variety of measures for assessing the Big Five factors of personality: agreeableness, conscientiousness, extraversion, neuroticism, and openness to experience. Although they may be psychometrically sound, many of these measures consist of hundreds of items, which could make them impractical for use in some employment contexts. As a personality assessment alternative, the 50-item Fitability 5 was developed with the specific goal of assessing the Big Five factors in a quick, yet accurate manner. Lucius (2003) demonstrated that this instrument's five scales converged onto the Big Five taxonomy, with scale score correlations ranging from .69 to .88 with established five-factor model personality measures (Barrick & Mount, 1994; John & Donahue, 1994). The Fitability 5 scales also demonstrated sufficient evidence of internal consistency (coefficient alphas ranging from .74 to .85) and criterion-related validity, with employees' scores correlating significantly with such job-related variables as worker well-being, self-esteem, satisfaction, motivation, and productivity. However, a comparison of applicants' scores with those who completed the Fitability 5 for developmental purposes revealed considerable score differences on 14 of the measure's 50 items. Under suspicion that members of the applicant sample may have intentionally distorted their responses to create a more favorable impression, the Fitability 5 underwent revision to become more resistant to faking. The revised measure, the Fitability 5a, was

considered “new and improved” and “specifically designed for job applicant populations” (Lucius, p. 40).

Fitability Systems has hosted the Fitability 5a for online applicant screening purposes for more than 5 years, and currently administers the test to over 10,000 applicants a month, including Fortune 500 firm applicants. Although the test developers designed this measure to be more resistant to faking than the original Fitability 5, observations since the revision have revealed that applicants respond differently than incumbents and volunteers on the Fitability 5a as well. Rather than revise the items a second time, Fitability Systems contacted members of the industrial and organizational program at Auburn University, including the current author, to determine an alternative method for resolving their potential faking problem. The agreed upon solution provided the rationale for Study 1 of this dissertation: The development of the Fitability 5a impression management scale.

Test administrators often rely on impression management scales to detect examinees’ engagement in intentional response distortion (as opposed to self-deception enhancement; Paulhus, 1984). Impression management measures such as Paulhus’ (1984, 1998) BIDR have been used extensively in academic and applied research. However, some practitioners may be hesitant to use the BIDR for personnel selection due to its impractical length (40 items), because it was designed for general purposes (i.e., not designed for employment contexts, per se), or because it utilizes a unique response scale (1 = *Not true* to 7 = *Very true*). Because existing off-the-shelf impression management scales did not fit needs of Fitability Systems, the development of a new impression management scale, one that could detect applicants’ use of impression management in as

few items as possible, was deemed necessary. Further, because this scale would be included as the sixth scale of the Fitability 5a, its items would need to appear similar to items on the Fitability 5a in terms of length, reading level, and response format. As part of the validation process, scores on this scale would also need to exhibit a significant and positive relationship with scores of a measure designed to assess the same construct. Therefore, Paulhus' (1998) BIDR was also used in Study 1 for construct validation.

Method

Item development. To assure the Fitability 5a impression management scale would appropriately assess the construct of interest, relevant theory guided the scale development process (DeVellis, 2003). Thus, it was necessary to distinguish between the two forms of social desirability: impression management and self-deception enhancement (Paulhus, 1998). Because impression management represents an individual's attempt to misrepresent his or her true self deliberately, items were constructed specifically to assess the intentional form of social desirability.

Sixty impression management-type items were developed for the initial item pool. Although most items were original, existing impression management items (e.g., the International Personality Item Pool, 2001) and scales (e.g., the impression management scale of Paulhus' BIDR, 1998) served as guides during the item-writing process. Care was taken to avoid items that were lengthy, difficult to read, double-barreled, double negative, or ambiguous. Of the original 60 items, half were reverse coded. Four experts with backgrounds in individual differences and personality testing reviewed the initial item pool. The following set of instructions was provided for completing this task:

As part of the scale development process, subject matter experts are required to review the initial pool of items for the scale. Please examine the following items for grammar, clarity, and conciseness. Also, please evaluate whether each item fits the format of the Fitability test items as well as determine if each item is relevant for assessing impression management response patterns. Please make note of any items that require modification or elimination.

Based on the experts' feedback, 25 items were modified and five items were eliminated altogether. Reasons for item modification and elimination included the following:

- Too wordy
- Improper grammar
- Too negative sounding
- Too personal sounding
- Stated as an absolute (e.g., "I always," "I never")
- Unclear wording
- Did not appear to address impression management

This process resulted in a final item pool consisting of 55 items for administration to the developmental sample.

Participants and procedure. Participants were 300 undergraduate students (78.3% female, 86.7% Caucasian) who took part in the study for extra credit in psychology courses at Auburn University. Participants were divided into two distinct samples (Samples A and B) that completed the measures under different instructional sets. Participants completed these measures by following a Web link to SurveyMonkey, an online survey program used to host the measures. All data were collected anonymously.

Participants received extra credit if they accessed the SurveyMonkey program, even in the absence of completing the measures. The following sections describe the methodologies used for Sample A followed by Sample B.

Sample A participants and procedure. Sample A consisted of 54 (77.8% female, 89.1% Caucasian) participants. These participants completed the measures based on the direction manipulation paradigm used by researchers at Auburn University (e.g., Winkelspecht, et al., 2006; Teague & Thomas, 2008) in their investigations of personality test faking and applicant response patterns. The specific instructions used for Sample A appeared to participants as follows:

> ***Respond Honestly instructions****: The test you are about to take will be used as an aid for making a hiring decision for the position of SALESMAN/SALESWOMAN 1. Please carefully adhere to the following request: ANSWER ALL QUESTIONS HONESTLY*
>
> ***Most Favorable Impression instructions****: The test you are about to take will be used as an aid for making a hiring decision for the position of SALESMAN/SALESWOMAN 1. Please carefully adhere to the following request: ANSWER ALL QUESTIONS IN A WAY THAT YOU BELIEVE WILL PRESENT THE MOST FAVORABLE IMPRESSION OF YOU AS A SALESMAN/SALESWOMAN.*

Data from the first instruction condition (Respond Honestly) represented the respondents' true scores. Data from the second instruction condition (Most Favorable Impression) represented the respondents' scores as job applicants maximizing their use of impression

management. The results from Sample A assisted in identifying the items that were most susceptible to score changes based on the two instruction conditions.

Sample A measures. Participants completed three different measures: the Fitability 5a, the impression management item pool, and a modified version of the BIDR. Items for each of these measures appear in Appendix A.

First, participants completed the Fitability 5a, a 55-item Big Five personality test containing scales ranging from 10 to 12 items. Previous internal consistency estimates the measure's five scales ranged from .69 to .75 (Lucius, 2003). For the current study, the items on the Fitability 5a used a five-point Likert-type response scale ranging from 1 = *Strongly disagree* to 5 = *Strongly agree*.

Second, participants completed the 55 items constructed for the Fitability 5a impression management scale (described above). Participants responded to these items using the same five-point Likert-type response scale as the Fitability 5a (1 = *Strongly disagree* to 5 = *Strongly agree*).

Third, participants completed a modified version of Paulhus' (1998) BIDR, one of the most widely used, reliable, and valid social desirability scales in existence (Li & Bagger, 2006). The unmodified version of the BIDR consists of 40 items divided equally between two 20-item subscales: self-deception enhancement (e.g., *I never regret my decisions*) and impression management (e.g., *I never swear*). These subscales differ in that self-deception enhancement measures unintentional self-promotion whereas impression management measures purposeful response distortion. Respondents indicated their level of agreement with each item using a seven-point Likert-type scale (1 = *Not true* to 7 = *Very true*). The modification made for the current study consisted of

eliminating two items for being inappropriate to an employment context. These items were *I have sometimes doubted my ability as a lover* and *I never read sexy books or magazines*.

Sample A analyses and results. Separate analyses were conducted for Samples A and B. Sample A data were analyzed first to identify which items were most susceptible to score changes based on the two instruction conditions. Because Sample A participants responded to the impression management items honestly and by presenting the most favorable impression, analyses using these data would inform the scale development process by determining which items respondents were best able to fake good. Sample A analyses would also benefit the scale development process by identifying which items were least susceptible to score changes between the two instructional conditions. Items that produced little to no score changes between the instruction conditions would be eligible for elimination from the initial item pool, as these items would be considered least susceptible to faking.

An examination of Sample A's descriptive statistics for the Respond Honestly condition revealed no departures from normality. However, the data from the Most Favorable Impression condition displayed several departures from normality, as the item scores tended to be biased toward the high end of the response scale (*Strongly agree*). These results suggested that respondents were endorsing the items intended for the new scale more often in the Most Favorable Impression condition than in the Respond Honestly condition.

Next, a paired-samples t-test and effect size (Cohen's *d*) analysis were conducted to determine whether Sample A respondents were able to change their scores

significantly and meaningfully based on the different instruction conditions. Forty-eight items exhibited medium to large effect size differences between the two instructional conditions. These items were considered optimal for inclusion in the final scale as these items produced the most meaningful score differences between the Respond Honestly and Most Favorable Impression conditions. The seven items that produced small effect size differences were removed from all additional analyses, with the rationale that these items would be unlikely to identify impression management behavior. Following the Sample A analyses, the impression management item pool consisted of 48 items.

Sample B participants and procedure. Sample B consisted of 247 (78.1% female, 83.8% Caucasian) undergraduate students. These participants completed the measures using the instructions typically provided with the Fitability 5a. Specifically:

> ***Instructions:*** *Below are a series of statements that broadly describe an individual's personality. Indicate whether you agree or disagree with each statement as it applies to you by "clicking" on the appropriate response. There are no right or wrong answers, nor is there an "ideal" response for each question. Attempting to misrepresent your true personality may actually work against you. The best approach is to simply respond truthfully. Do not think too much about your answer - go with your first impression.*

Sample B measures. Sample B participants completed the same measures as the Sample A participants: the Fitability 5a, the impression management item pool, and a modified version of the BIDR.

Sample B analyses and results. An examination of the descriptive statistics from Sample B indicated that all items were normally distributed. However, 10 items exhibited

biased responses patterns (e.g., respondents did not use the entire scale; responses favored either *Strongly disagree* or *Strongly agree*), and were therefore removed from the item pool. Following this elimination, 38 items remained.

The remaining 38 items were analyzed using exploratory factor analysis (EFA). According to DeVellis (2003), EFA is the "best means of determining which group of items, if any, constitute a unidimensional set" (p. 94). Before conducting the EFA, the data from Sample B were evaluated for appropriateness. The ratio of observations to variables was moderate, but acceptable (6.5:1). This assumption was confirmed using Bartlett's test of sphericity (X^2 = 2,657.04, *df* = 703, p < .001), which determined that the correlations, when taken collectively, were significant. Further, the overall Kaiser-Meyer-Olkin measure of sampling adequacy (MSA) provided that the items were correlated and that the pattern of variable correlations was suitable for EFA (*MSA* = .80). An examination of the partial correlation matrix found that no one variable had a partial correlation with any other variable greater than ±.5. Taken together, these checks established that all 38 items met the requirements for EFA (Hair, Black, Babin, Anderson, & Tatham, 2006).

The next stage in conducting the EFA was to determine the number of factors to extract. Although the items for the new scale were intended to assess a single construct (i.e., impression management), it was considered possible that different subsets of items might yield different conceptualizations of the latent construct. The rationale for this possibility came from Schmit and Ryan (1993), who suggested that faking-based scales may consist of items that are directed toward different target objects. Using principal components analysis and the scree plot method (see Figure 3), it was decided to retain

four factors. Next, an EFA was conducted with a forced four-factor solution using direct oblimin rotation (maximum likelihood extraction), as both theory and the previous analyses suggested that the resulting solution would produce correlated factors. The resulting pattern matrix produced four factors, with 10 items on the first factor, 4 on the second, 6 on the third, and 11 on the fourth. The four-factor solution accounted for 30.4% of the total variance. This initial four-factor solution was examined for potential improvements based on alpha-if-item-deleted estimates and item similarity. Based on these criteria, two items were eliminated from Factor 1 and two items were eliminated from Factor 4. This revision yielded four separate factors for consideration as the new impression management scale. Conceptual interpretation of these four factors suggested that they differed based on the target of the respondents' use of impression management, which included social conventions, reputation, responsibility, and emotions, respectively (see Table 3 for factors and items).

To assess the construct validity of the four potential scales, a correlational analysis was performed to determine whether scores on these scales correlated significantly with scores on the BIDR (Table 4). Based on these results, it was concluded that the first impression management scale, which targeted social conventions, was the most similar to the BIDR's impression management scale as well as most dissimilar to the BIDR's self-deception enhancement scale. Therefore, the eight items from the social conventions scale were adopted for use as the Fitability 5a impression management scale.

Discussion

Applicants and non-applicants scored differently on the Fitability 5a personality test, and one potential explanation for these score differences is that applicants are faking

good by inflating their responses on this test to appear more favorable to the hiring organization. The intentional use of false, but favorable response patterns is the defining characteristic of the construct known as impression management. Therefore, in Study 1, the principle goal was to develop a scale for detecting applicants' use of impression management on the Fitability 5a.

By following the scale development procedure offered by DeVellis (2003), an eight-item impression management scale was constructed for detecting applicant faking on the Fitability 5a. The items for this scale included:

1. I never listen in on other people's private conversations.
2. I always tell the truth.
3. I have lied to get myself out of trouble. (r)
4. I rarely gossip.
5. I sometimes talk bad about my friends behind their back. (r)
6. I find it easy to resist temptations.
7. I sometimes break the rules to get ahead. (r)
8. I always know why I do the things I do.

This scale is appropriate for use with the Fitability 5a because it employs the same response scale and sentence structure, thereby allowing its items to appear similar to the items on the Fitability 5a. Because it consists of only eight items, the new impression management scale fits with the Fitability 5a's goal of offering a brief assessment of job candidates. In addition, because scores on this scale correlated significantly with scores from an existing impression management scale, there is preliminary evidence of construct validity. However, being that the new impression management scale was designed to

assess applicants' response patterns, it required application in a real world selection context to demonstrate whether it truly operates as intended. Therefore, Study 2 provided an applied context for administering the impression management scale to a sample of real world job applicants. According to theory, applicants who engage in faking should score differently on the Fitability 5a's scales than applicants who do not fake. To assess the extent to which faking affected applicants' personality measurement, the IRT-based DFIT procedure was used to test the measurement equivalence of the Fitability 5a across high and low impression management groups.

Chapter 4

Study 2: Is Faking a Problem for the Fitability 5a?

Organizational researchers have long been concerned with the issue of applicant personality test faking. In their investigation of the faking problem, researchers have employed a variety of methodologies, including direction manipulations, comparisons of applicants to non-applicants, and incorporating impression management scales into the selection battery. Much of this research relied on traditional analytic techniques based on classical test theory, such as making mean score comparisons between groups of fakers and non-fakers. Methodologies based on IRT, however, offer a more sophisticated and considerably superior technique for comparing different groups of respondents. Some organizational researchers have already incorporated IRT-based methods, such as the DFIT framework (Raju et al., 1995), into the investigation of the faking problem. These studies produced mixed results, with some studies indicating faking is a problem and others not. Study 2 attempted to build on this research by examining the affects of applicant faking on the Fitability 5a using the DFIT procedure, a large applied sample of job applicants, and an impression management scale designed exclusively for selection contexts.

Study 1 entailed the development of the Fitability 5a impression management scale and provided preliminary evidence of construct validity by demonstrating that students' scores on this scale correlated significantly with their scores on the impression

management scale of the BIDR. To determine if applicant faking is a problem for the Fitability 5a, however, actual job applicants that score highly on the impression management scale would need to respond differently to the Fitability 5a compared to applicants that do not score highly on the impression management scale. Raju et al.'s (1995) DFIT procedure provides an appropriate test for differential functioning across groups of fakers and non-fakers. Therefore, the purpose of Study 2 was to test the Fitability 5a for differential item functioning (DIF) and differential test functioning (DTF) across groups of high and low scorers on the Fitability 5a impression management scale using a sample of actual job applicants. If members of each group respond differently to the items or scales of the Fitability 5a, there would be reason to believe that faking disrupts the construct validity of this Big Five personality measure. Such an outcome would render the use of the Fitability 5a for personnel selection questionable, as users would not be able to compare scores on this test between those who fake and those that do not fake.

Method

Participants. Participants were 21,017 applicants to a large automotive parts and service company. All participants applied to work in locations within the United States between March and May 2009. Participants completed the measures through an online applicant screening program hosted by Fitability Systems. As part of the application process, participants voluntarily provided demographic information. Of the 20,910 participants who reported their gender, 18,222 (87.1%) were men and 2,688 (12.9%) were women. Of the 16,630 participants who reported their race, 10,761 (64.7%) were Caucasian (non-Hispanic), 3,099 (18.6%) were African American, and 1,819 (10.9%)

were Hispanic. The positions applied for included technician/specialist (8,353; 39.7%), management (7,051; 33.5%), and customer service (5,613; 26.7%).

Measures. Participants completed the Fitability 5a personality test and the Fitability 5a impression management scale (described in Study 1) as part of a larger selection battery containing additional selection tests not considered for the current study. The organization that provided access to the applicant participants based their selection decisions in part on applicants' scores on the Fitability 5a. The organization did not use applicants' scores on the impression management scale for making employment decisions; these data were collected for research purposes only.

Procedure and analyses. Descriptive statistics and correlations among the variables for the total sample and the high and low impression management subgroups were obtained to provide a general overview of the data prior to conducting analyses for measurement equivalence.

Because the DFIT methodology is based on IRT, the first step in testing for measurement equivalence was to ensure the data meet the assumptions of IRT. IRT assumes that measurement scales are unidimensional (Hambleton & Swaminathan, 1985). Therefore, a principal-axis factor analysis was conducted on each of the Fitability 5a scales as well as the impression management scale in order to test for the IRT assumption of unidimensionality of scales. To satisfy this assumption, the first factor of each scale would need to account for at least 20% of the variance (Reckase, 1979).

Next, in order examine measurement equivalence on the personality measure between fakers and non-fakers, the total sample of respondents was divided into two subgroups based on their impression management scores. Previous researchers have used

a variety of cutoff points for identifying high and low impression management groups. High score cutoffs have included the median (Stark et al., 2001), top quartile (Henry & Raju, 2006), and the top 15th percentile of scores (Flanagan & Raju, 1997). Low score cutoffs have included score ranges (e.g., scores between the 15th and 85th percentile; Flanagan & Raju) as well as scores below the 50th percentile (Henry & Raju). For the current study, the frequency distribution of impression management scores was examined to determine the most representative high and low scoring groups. Each subgroup of respondents consisted of 5,000 participants, because this value was the maximum sample size permitted by the statistical software used to analyze the data. The high scoring group (i.e., faking group; focal group) had an average impression management score of 4.91 (*SD* = 0.09) and the low scoring group (i.e., non-faking group; reference group) had an average impression management score of 3.58 (*SD* = 0.24). Demographic characteristics for these groups were consistent with the total sample of participants.

As the Fitability 5a uses a polytomous response scale with five ordinal response categories, Samejima's (1969) graded response model was used for performing the DFIT analyses. The computer program Multilog 7.03 (Thissen, Chen, & Bock, 2003) was used to estimate the parameters for each response option for each item and estimated the latent trait scores (θ) for each respondent on each item. The Equate 2.1 computer program (Baker, 1997) was used to transform the parameter estimates for the faking and non-faking groups as needed to place them onto a common metric.

Following this recalibration of the item parameters, DFIT analyses were conducted using the DFITps6 computer program (Raju, 2000). This program provides indices of compensatory and non-compensatory differential item functioning (CDIF and

NCDIF, respectively), as well as an estimate of differential test functioning (DTF). The NCDIF values were examined for each personality test item separately to ascertain the presence of DIF using a critical value of .096 for determining statistical significance at the .01 alpha level (as recommended by Raju et al., 1995). The critical values for determining statistically significant DTF differed by the number of items in each scale, and were calculated by multiplying the number of items in each scale by .096. For the 10-item openness scale, the critical value for statistically significant DTF was set to .960. For the 11-item agreeableness, conscientiousness, and neuroticism scales, the critical value for statistically significant DTF was set to 1.056. For the 12-item openness scale, the critical value for statistically significant DFT was set to 1.152. In the presence of DTF, the DFITps6 program removed items with statistically significant CDIF one at a time until there was no longer DTF on the scale.

After testing for measurement equivalence on the Fitability 5a's scales across faking and non-faking groups, the same set of procedures were completed using the total sample of respondents to assess for measurement equivalence on the impression management scale based on participants' gender and race. These analyses were performed to determine if the items on this scale functioned differently for male versus female applicants as well as for Caucasian applicants versus applicants that identified themselves as African American or Hispanic. Because there were fewer than 5,000 women participants in the total sample, all female participants were considered in the gender analyses. However, because over 18,000 participants identified themselves as men (and due to the sample size constraints of the DFITps6 program), two random samples of 5,000 male respondents were used when conducting the gender-based analyses to ensure

the estimates were stable and not subsample dependent. There were also fewer than 5,000 African American and Hispanic participants in the total sample; therefore, all participants from these groups were considered. Because over 10,000 participants identified themselves as Caucasian (and due to the sample size constraints of the DFITpsa6 program), two random samples of 5,000 Caucasian respondents were used when conducting the race-based analyses to establish the estimates were stable and not subsample dependent.

Results

Impression management item and scale means and standard deviations for the total sample and subsamples groups by gender and race appear in Table 5 for norming purposes. Scale means and standard deviations for the five Fitability 5a scales and the impression management scale using the total sample as well as the high and low impression management subsamples appear in Table 6. For each scale, applicants from the high impression management subgroup scored higher, on average, than applicants from the low impression management subgroup, with the exception of the neuroticism scale, which demonstrated the opposite trend. Estimates of the effect sizes (Cohen's *d*) of these score differences indicated that these score differences were practically meaningful (i.e., exhibited large effect sizes greater than .80) with the exception of the extraversion scale, which exhibited a small effect (d = .19).

Correlations among the five Fitability 5a scales and the impression management scale appear in Tables 7, 9, and 9 for the total sample of applicants and the high and low impression management subgroups, respectively. In all cases, applicant's impression management scores were positively and significantly correlated with their scores on the

agreeableness, conscientiousness, extraversion, and openness scales, and negatively and significantly correlated with their scores on the neuroticism scale. However, given the large sample sizes used in the current analyses, even small, non-meaningful relationships were likely to be found statistically significant. The correlations between applicants' impression management scores and their extraversion scores, for instance, were .09, .05, and .04 for the total sample, high impression management group, and low impression management group, respectively. Though considerably weak, each of these values were statistically significant at (at least) the .05 level.

To examine the Fitability 5a for measurement equivalence using the DFIT methodology, tests were performed to satisfy the IRT assumption of unidimensionality scales. Results of principle-axis factor analysis indicated that the first component of each of the five personality scales and the impression management scale exceeded Reckase's (1979) cutoff of at least 20% of the variance, which satisfied the unidimensionality assumption and thereby supported the use of the remaining IRT-based analyses.

Results of the DFIT analyses for assessing measurement equivalence between fakers and non-fakers on the Fitability 5a personality test appear in the following sections alphabetically by. Next, results of the analyses for assessing measurement equivalence on the Fitability 5 impression management scale appear in the following order: Males versus Females, Caucasians versus African Americans, and then Caucasians versus Hispanics.

DFIT on the Fitability 5a. The following sets of results are based on analyses that tested for measurement equivalence on each of the Fitability 5a's personality scales when comparing subgroups of respondents (n = 5,000 each) identified as fakers (high impression management; focal group) and non-fakers (low impression management;

reference group). Items were equated by putting them on the scale of the focal group, which in this case was the faking group. Items were considered to have statistically significant DIF at an alpha level of .01 if the NCDIF values exceeded the .096 cutoff score recommended by Raju et al. (1995) for items with five response options. The cutoff scores for determining DTF are provided separately for each scale analysis below, as these cutoff values vary by the number of items in the scale. For scales with significant DTF, items with the most significant CDIF were removed one at a time until there was no longer DTF for the scale. (Note: One can remove items based on the NCDIF or CDIF values, as both produce a final scale free of DTF. However, there latter method results in fewer items deleted and is therefore preferable if one wishes to retain as many items as possible on the final DTF-free scale).

Agreeableness. Agreeableness item means by subgroup, as well as the NCDIF, CDIF, and DTF values appear in Table 10. Applicants identified as fakers scored higher on average on all items from this scale. When controlling for ability level across subgroups of fakers and non-fakers, 8 out of 11 agreeableness items showed statistically significant NCDIF. The agreeableness scale also exhibited significant DTF, as the uncorrected DTF index (19.51) exceeded the critical value for an 11-item scale (1.056). Items with statistically significant CDIF were deleted from this scale one at a time until the scale no longer exhibited DTF. The final DTF-free agreeableness scale consisted of three items (6, 10, and 22).

Conscientiousness. Conscientiousness item means by subgroup, as well as the NCDIF, CDIF, and DTF values appear in Table 11. Applicants identified as fakers scored higher on average on all items from this scale. When controlling for ability level across

subgroups of fakers and non-fakers, 10 out of 11 conscientiousness items showed statistically significant NCDIF. The conscientiousness scale also exhibited significant DTF, as the uncorrected DTF index (34.17) exceeded the critical value for an 11-item scale (1.056). Items with statistically significant CDIF were deleted from this scale one at a time until the scale no longer exhibited DTF. The final DTF-free conscientiousness scale consisted of two items (4 and 50).

Extraversion. Extraversion item means by subgroup and NCDIF values appear in Table 12. Applicants identified as fakers scored higher on average on all items from this scale. When controlling for ability level across subgroups of fakers and non-fakers, none of the extraversion items showed statistically significant NCDIF. There was no evidence of statistically significant DTF for the extraversion scale as the uncorrected DTF index (0.68) was well below the critical value for a 12-item scale (1.152). Therefore, no items were deleted from this scale, and no CDIF or DTF data were necessary for Table 12.

Neuroticism. Neuroticism item means by subgroup, as well as the NCDIF, CDIF, and DTF values appear in Table 13. Applicants identified as fakers scored lower on average on all items from this scale. When controlling for ability level across subgroups of fakers and non-fakers, 8 out of 11 neuroticism items showed statistically significant NCDIF. The neuroticism scale also exhibited significant DTF, as the uncorrected DTF index (44.20) exceeded the critical value for an 11-item scale (1.056). Items with statistically significant CDIF were deleted from this scale one at a time until the scale no longer exhibited DTF. The final DTF-free neuroticism scale consisted of three items (3, 15, and 48).

Openness. Openness item means by subgroup, as well as the NCDIF, CDIF, and DTF values appear in Table 14. Applicants identified as fakers scored higher on average on all items from this scale. When controlling for ability level across subgroups of fakers and non-fakers, 9 out of 10 openness items showed statistically significant NCDIF. The openness scale also exhibited significant DTF, as the uncorrected DTF index (20.20) exceeded the critical value for a 10-item scale (0.96). Items with statistically significant CDIF were deleted from this scale one at a time until the scale no longer exhibited DTF. The final DTF-free openness scale consisted of two items (20 and 44).

DFIT on the impression management scale. The remaining results are based on the analyses that tested for measurement equivalence on the impression management scale based on gender (males versus females) and race (Caucasians versus African Americans and Caucasians versus Hispanics). For the gender-based analyses, two random subsamples of 5,000 male participants were tested, as over 18,000 applicants identified themselves as male and the computer programs used to analyze the data set sample size constraints to 5,000 subjects per comparison group. Similarly, for the race-based analyses, two random subsamples of 5,000 Caucasian participants were tested, as over 10,000 applicants identified themselves as Caucasian. Analyses based on these subsamples were use to evaluate whether the estimates were stable and not subsample dependent.

For the gender comparisons, items were equated by putting them on the scale of the male respondents. For the race comparisons, items were equated by putting them on the scale of the Caucasian respondents. Items were considered to have statistically significant DIF at an alpha level of .01 if the NCDIF values exceeded the .096 (Raju et

al., 1995). The cutoff score for determining statistically significant DTF was set to .768 (based on .096 x 8 items).

Impression management item means and standard deviations for each group of respondents (i.e., two male subsamples, all females, two Caucasian subsamples, all African Americans, and all Hispanics) appear in Table 15. Based on the effect size estimates (Cohen's *d*), any differences in the means scores based on gender or race were negligible. Table 15 also contains the NCDIF values for the gender and race-based item-level comparisons when controlling for ability level in impression management. All NCDIF values were below the critical value of .096 and therefore non-significant. Thus, there was no evidence of DIF on the impression management scale based on gender or race. There was also no evidence of DTF on this scale based on gender or race, as all DTF values were below the critical value of .768 (therefore, the CDIF and DTF values do not appear in Table 15). Based on these results, the Fitability 5a impression management total scale and individual items function equivalently for male and female applicants as well as for Caucasian, African American, and Hispanic applicants.

Discussion

The DFIT procedure allows researchers to compare how individuals from different groups respond to items and scales. In the current study, the response functions on the Fitability 5a personality test were compared across subsamples of job applicants grouped according to their scores on the Fitability 5a impression management scale. This comparison was used to determine if fakers (i.e., high impression management scorers) responded differently to the Fitability 5a compared to non-fakers (i.e., low impression management scorers). Evidence of group differences would indicate that applicant faking

has the potential to disrupt the measurement equivalence of a Big Five personality test used for selection.

Study 2 revealed that applicants from the high and low impression management groups responded differently to the Fitability 5a. Of the 55 items on the Fitability 5a, 35 demonstrated significant DIF. Only the extraversion scale did not contain items with significant DIF. In all cases, DIF uniformly favored the high IM group. Of the Fitability 5's five scales, four demonstrated significant DTF. Only the extraversion scale did not demonstrate significant DTF. To produce a measure free of DTF required the elimination of the majority of items and would have resulted in a 3-item agreeableness scale, a 2-item conscientiousness scale, a 3-item neuroticism scale, and a 2-item openness scale. The extraversion scale would retain all 12 of its items, as this scale was free of DIF and DTF. From an applied perspective, these results provide strong support that applicant faking is a problem for this particular Big Five selection test.

Chapter 5

General Discussion

This dissertation makes two significant contributions to the investigation of applicant personality test faking. The first contribution met an applied need by developing an impression management scale for use with the Fitability 5a, a Big Five personality test used for employment selection. The second contribution provided empirical evidence to address the question: Does faking matter in the real world? The sections that follow provide a discussion of these contributions, followed by the limitations and implications of this research, including recommendations for future studies on applicant personality test faking.

Contribution 1: The Fitability 5a Impression Management Scale

Organizational researchers have long been concerned with the issue of applicant faking on personality tests used for selection. In investigating the prevalence and severity of the faking problem, researchers have relied on a variety of methodologies, including the use of impression management scales for identifying applicants that engage in intentional response distortion. Although there are numerous impression management scales available for use in organizational research, many of these measures are too lengthy for applied purposes or intended for general populations as opposed to job applicants. In addition, off-the-shelf scales tend to use their own unique response formats that might affect applicants' response patterns. If an impression management scale's

response format does not match the response format of the other tests in the assessment battery, it may cue respondents that the impression management items are assessing a different construct and may therefore evoke different response patterns. For these reasons, some personality tests contain custom impression management scales designed to detect applicant faking exclusively on their measure.

The first contribution of this dissertation met an applied need by developing a custom impression management scale for the Fitability 5a. The Fitability 5a is a Big Five personality test designed for screening job candidates quickly and accurately. The newly developed Fitability 5a impression management scale is custom in that its items adhere to the format of the Fitability 5a items in terms of reading level and response scale (i.e., 1 = *Strongly disagree* to 5 = *Strongly agree*). In addition, the impression management scale contains only eight items, so it is consistent with the Fitability 5a's goal of providing a timely assessment.

In developing the Fitability 5a impression management scale, items were constructed to resemble items on other impression management scales, such as Paulhus' (1998) BIDR, which ensured that the initial item pool addressed intentional response distortion rather than self-deception enhancement. To maximize the likelihood of detecting applicant faking, the item pool was administered to student participants using respond honest and fake good instructions. This direction manipulation identified the items that demonstrated the largest effect sizes attributable to faking. After eliminating items that were less likely to detect faking, a factor analysis was performed using data from a normal instructions sample to arrive at the final impression management scale. Students' scores on this scale correlated positively and significantly with Paulhus'

impression management scale, which suggested that the new measure assesses the construct of impression management.

The development of the Fitability 5a impression management scale satisfied the applied needs of Fitability Systems by providing a short and reasonably construct valid measure for assessing applicant faking. A key assumption behind impression management measurement is that examinees are unaware that they are responding to impression management items: Otherwise, test administrators and researchers could not use scores on these scales for identifying faking-related response patterns. Because the items on new impression management scale resemble the Fitability 5a' s items in terms of length, readability, and response format, it is believed that they could be added seamlessly to the personality measure without cueing examinees that the impression management items are assessing a sixth construct intended for the detection of faking.

Measurement equivalence was examined on the impression management scale across gender and race-based groups. These analyses indicated that males and females as well as Caucasians, African Americans, and Hispanics tended to respond similarly to the items on this scale. This finding provided further evidence of construct validity for the scale, as members of different groups should interpret a construct valid measure in the same way. Findings of measurement equivalence across gender and race also supports the use of Fitability 5a impression management scale with applicant populations, as items on this scale do not appear to favor members of groups protected by equal employment laws (e.g., women, racial minorities).

Organizational researchers disagree on the extent to which applicant faking should be considered a real world threat. The newly developed Fitability 5a impression

management scale has the potential to add to the investigation of applicant personality test faking by provided a measure designed exclusively for use with job applicants as opposed to general populations. Although there are a variety of off-the-shelf impression management scales for detecting intentional response distortion, most of these measures are designed for general purposes, contain a large set of items, utilize unique response scales, or otherwise do not incorporate well into personnel selection test batteries. Applied research investigating the faking problem using off-the-shelf impression management scales may be limited in that applicants may respond to these scales differently compared to scales designed exclusively for organizational contexts. As the Fitability 5a impression management scale was essentially designed as the sixth scale of this Big Five personality measure, it has the potential to evoke applicant response patterns that inform the faking debate in ways that off-the-shelf impression management scale do not.

Contribution 2: Faking Matters in the Real World

The second major contribution of this dissertation sought to answer the question: Does faking matter in the real world? To this end, the new Fitability 5a impression management scale was implemented in a real world selection setting consisting of actual job applicants. This applied investigation served to validate the new scale further by determining (a) whether high and low scorers on the impression management scale scored differently on the Fitability 5a's five scales and (b) whether faking affected the measurement equivalence of the personality measure.

A comparison of the mean score differences between high and low impression management applicants on the Fitability 5a's five scales revealed considerable score

differences on all but the extraversion scale in terms of effect size. Those who scored higher on the impression management scale scored higher on the agreeableness, conscientiousness, and openness scales and lower on the neuroticism scale. Based on these score comparisons, one may conclude that the applicants who engaged in impression management scored in the more desirable direction compared to applicants who did not engage in impression management. These results served to validate the new scale, as producing favorable responses is a hallmark characteristic of impression management behavior.

The finding that some scales exhibited mean score differences and others did not is not unusual in the personality test faking literature. Although some research suggests that all Big Five scales are equally susceptible to applicant faking (e.g., Viswesvaran & Ones, 1999), other studies suggest that this outcome is not the case (e.g., Birkeland et al. 2006). The results of the current research support the latter findings in that each of the Fitability 5a scales were differentially susceptible to applicant faking.

In their meta-analytic study, Viswesvaran and Ones (1999) suggested that, of all of the Big Five factors, agreeableness may be the least susceptible to faking. However, in the current study, the only scale that did not exhibit a meaningful effect for faking in was the extraversion scale. One potential explanation for why the high and low impression management applicants did not score differently on the extraversion scale is that the most favorable response to the items on this scale did not always fall in the same direction (i.e., toward extraversion or introversion). On the other four personality scales, members of the high impression management group consistently scored higher (or lower, for neuroticism) on all items compared to the low impression management group. This was not the case

for extraversion. An examination of the item-level mean scores for this scale (Table 12) reveals that the high impression management group produced higher scores on seven extraversion items and lower scores on five of these items. Essentially, applicants who engaged in impression management interpreted approximately half the items as more favorable when responding toward the high end of the response scale (i.e., toward extraversion) and approximately half the items as more favorable when responding toward the low end of the response scale (i.e., toward introversion). Thus, at the item level, applicants that faked tended to score differently on extraversion compared to applicants who did not fake. At the scale level, however, the net effect of these differences is that the item-level scores cancelled each other out, resulting in near equivalent scale-level scores for the high and low impression management groups.

The finding that fakers and non-fakers score approximately the same at the scale level, but differently at the item level on the extraversion scale supports the use of item-level analyses in personality test faking research. Organizations tend to base selection decisions on scale-level scores, which may explain why the majority of personality test faking research is conducted using scale-level scores. However, this practice has the potential to mask true differences among applicants that occur at the item level. Each personality test item is, in and of itself, a test of a latent personality construct. Therefore, item-level analyses are important to the investigation of the faking problem. The results of the current study support the need for additional research examining item-level score differences between fakers and non-fakers. Of course, item-level analyses require large samples sizes that are often difficult for organizational researchers to achieve, particularly large applied samples. As the current research contained data from over 20,000 actual job

applicants, it offers a relatively rare, but informative perspective on how applicant faking manifests in applied contexts.

Beyond mean score comparisons, the present research also provided more sophisticated item-level and scale-level analyses utilizing the DFIT methodology. At the item-level, significant differential item functioning (DIF) occurred on the majority of items for the agreeableness, conscientiousness, neuroticism, and openness scales. Only the extraversion scale exhibited measurement equivalence between the high and low impression management groups. This particular finding supports the results of Flanagan and Raju's (1997) study in which the extraversion scale of the 16-PF exhibited measurement equivalence between fakers and non-fakers. Flanagan and Raju's study was limited, however, in that it only examined this one scale of the 16-PF. In a follow-up study, Stark et al. (2001) tested for measurement equivalence on all 16 scales of the 16-PF. Unlike the current research, Stark et al. found little evidence of DIF on these scales when comparing applicants grouped by impression management scores.

One potential reason for why DIF was found in the current study but not in Stark et al.'s (2001) study is the choice of personality measurement. The Fitability 5a is a Big Five measure of personality that uses a five-category response scale. The 16-PF measures 16 different personality traits, only some of which map onto the five-factor model. In addition, the 16-PF uses a three-category response scale, which, as explained previously, may limit the degree to which applicants can fake good. That DIF occurred in the current Big Five study and not in previous research using the 16-PF emphasizes the need to evaluate applicant response patterns for all personality measures. The results of faking research investigating differential functioning by impression management groups do not

generalize across different personality measures. Because each personality measure contains different items, it is likely that different measures will be differentially susceptible to DIF.

At the scale level, results were similar to the mean score analyses in that differential test functioning (DTF) occurred for all of the Fitability 5a's scales with the exception of the extraversion scale. This finding suggests that some personality constructs may be more susceptible to DTF caused by impression management than other constructs, which is consistent with previous research investigating the measurement equivalence of personality scales across fakers and non-fakers. Zickar and Robie (1999), for instance, found significant DTF on two of three scales on the military's ABLE personality test. One explanation for these findings is that applicants may perceive certain scales to be more job-related, which may prime applicants likely to engage in impression management (Henry & Raju, 2006). Additional research may seek to investigate DIF and DTF by comparing individuals grouped according to whether they perceive each item or scale as being related to the position in question. There are also numerous individual difference variables that may explain further how and why different applicants, including fakers and non-fakers, respond differently to personality test items and scales. Teague and Thomas (2008), for instance, recently found that intelligence and mood state affect faking. Although the addition of moderating variables may complicate the examination of differential test functioning between fakers and non-fakers, it will likely produce more informative results than could be achieved in the current study.

In the presence of DTF, test administrators can eliminate items that exhibit significant DIF in order to achieve DTF-free scales. In the current study, the four scales

that exhibited DTF contained DIF items that uniformly favored the high impression management group. This finding limited the utility of the CDIF indices. Using the CDIF item deletion procedure to produce DTF-free scales resulted in a 3-item agreeableness scale, a 2-item conscientiousness scale, a 3-item neuroticism scale, and a 2-item openness scale. Thus, to achieve measurement equivalence on the Fitability 5a's scales between fakers and non-fakers, the vast majority of items on this test would need to be removed. As the Fitability 5a is already a brief personality assessment tool, the removal of multiple items would considerably reduce the validity of this test for employment contexts.

Overall, at the item level and the scale level, impression management severely impacted the measurement equivalence of the Fitability 5a, as the majority of test items exhibited DIF and the majority of scales exhibited DTF well beyond repair. These results contribute to the personality test faking literature by providing a relatively rare examination of DTIF on a Big Five personality test using high and low impression management groups from a large applied sample of job applicants. Although some researchers have concluded that faking is not a problem in real world settings (e.g., Hogan et al., 2003), the current study provides strong evidence to the contrary. Similar to Schmit and Ryan's (1993) factor analysis research, this dissertation suggests that impression management adversely affects the construct validity of personality measures to a severe degree.

Limitations and Future Directions

As with any study, there are limitations to the current research that deserve consideration. These limitations, in turn, give rise to directions for future research.

The first limitation concerns the sample used for developing the Fitability 5a impression management scale. The scale development sample consisted entirely of undergraduate student participants. It is possible that students respond differently than actual job applicants, even when instructed to respond as if they were in an employment context. Any differences between the developmental sample and the applied population for which the measure is intended have the potential to influence how respondents interact with the measure, and may therefore limit the degree to which the scale actually detects applicant faking. A more appropriate developmental sample would consist of actual job applicants as opposed to undergraduate student participants.

An additional limitation concerns the direction manipulation instructions provided to students. Specifically, students were instructed to respond in a manner that would present the most favorable impression as a salesman/saleswoman. Thus, the items that appear on the final impression management scale may be biased toward detecting faking for sales positions rather than other positions. Also, as noted by McFarland and Ryan (2006), not every student may want a sales job, which may affect their motivation to respond to items as if they were sales applicants. As an alternative, the instructions could have requested that students respond in a manner that would maximize their chances for obtaining their dream job (e.g., Mueller-Hanson, Heggestad, & Thorton, 2006). However, different students could interpret these instructions differently, which could further affect the validity of the measure when placed in an applied context.

The scale development phase of this research was also limited in that the factor analysis performed on the impression management items produced a four-factor solution, even though the impression management construct is theoretically unidimensional.

Researchers have yet to investigate if there are global and facet-level conceptualizations of impression management. However, research in the area of job satisfaction suggests that some psychological constructs lend themselves to overall and target-specific forms (e.g., Highhouse & Becker, 1993; Scarpello & Campbell, 1983). Schmit and Ryan's (1993) study suggested that faking-related scales may consist of items that assess a variety of constructs, which lends support to a possible multi-dimensional conceptualization of impression management. Future researchers may seek to explore this possibility. In the current research, the four potential impression management scales addressed four different target areas: social conventions, reputation, responsibility, and emotions. However, the impression management toward social conventions scale produced scores most similar to the BIDR's impression management scale and most dissimilar to the BIDR's self-deception enhancement scale. Therefore, this scale was adopted over the other potential choices as the final Fitability 5a impression management.

One could argue that the Fitability 5a impression management scale has not undergone sufficient validation tests to warrant its use in applied research. In part, fewer tests were performed on this scale than in common practice because the agreement established with Fitability Systems to develop the scale required immediate action. Although adequate tests for validation were performed on the impression management scale in the current research, further tests are recommended before this scale undergoes widespread use.

A lack of control over the applicant sample provided additional limitations for discussion. Although the applicants participating in this study all applied to the same organization, they did not all apply for the same position. Applicant participants applied

for work in three major job categories: technician/specialist, management, and customer service. However, the DFIT analyses were performed across all job types to maintain sufficient and equivalent sample sizes in each impression management group. It is possible that applicants engaging in impression management did so differently based on the job to which they were applying. An appropriate follow-up study would investigate this possibility by comparing impression management within job categories rather than across job categories.

Applicant participants also applied exclusively to locations within the U.S. Organizations, including the organization investigated in the current research, are becoming more global. Culture is a critical variable in organizational research (Rousseau & Fried, 2001) and has the potential to influence how individuals interpret and respond to test items (Mitchelson et al., 2009). Members of different cultures may be more or less inclined to engage in impression management, for instance, based on whether their cultures are individualistic versus collective or depending on their religious ideology. The closest the current study came to considering culture was the analyses testing for measurement equivalence on the impression management scale based on racial affiliation. Although there were no racial differences in response functioning on the impression management scale, it is possible that members of these groups interpreted the Fitability 5a items differently, which may have influenced the results. More advanced methodologies incorporating nested designs may offer an opportunity for future researchers to take a variety of demographic or group membership variables into account when assessing measurement equivalence for individuals in groups within groups.

In part, the limitations of previous DFIT research provided the rationale for the current study. For instance, no previous research had yet examined measurement equivalence on all five scales of a true Big Five measure of personality between high and low impression management groups. Other studies were limited in their use of two or three option response scales, which have the potential to restrict the degree to which respondents can fake good. In comparison to these earlier studies, most of which found little or no evidence of DIF or DTF, the current study investigated a Big Five measure that uses a five-point response scale and uncovered considerable evidence of DIF and DTF. Therefore, future research in this area may wish to replicate the current study using additional Big Five measures to determine whether the present results are more attributable to the five-factor model of personality or to the Fitability 5a test itself.

In addition, it may be of value to investigate whether the smaller response scales used by tests such as the CPI and 16-PF are responsible for measurement equivalence by testing the items on these measures with their traditional response format as well as with an expanded response scale consisting of additional options. An assumption made in the current study is that smaller response scales were partially responsible for findings of measurement equivalence between fakers and non-fakers in previous DFIT research. Thus, an appropriate test of this assumption might involve introducing a polytomous response scale to a traditionally dichotomous test to determine whether measurement equivalence is more a property of the test's items or the test's response scale.

One final limitation that permeates all phases of this research concerns the conceptualization of the faking variable. Researchers have conceptualized applicant faking as a trait variable, a situational variable, or both (e.g., Stark et al., 2001). Trait-

based faking studies view faking as an individual difference variable and tend to assess faking using self-report measures, such as impression management scales. The current study adopted this approach by defining and assessing faking as the tendency to present oneself favorably, but falsely. Situational studies, on the other hand, view faking as a behavior or response strategy that "may manifest itself differently in different situations" (Mueller-Hanson et al., 2006, p. 309). These studies tend to assess faking by comparing scores obtained under different response conditions, such as by comparing individuals' scores from applicant and non-applicant conditions.

Comparisons of results produced from these two conceptualizations of faking indicate that they produce similar, but not identical results (e.g., McFarland & Ryan, 2006; Stark et al., 2001), which has been noted as one reason the faking debate remains unresolved (Ones et al., 2007). The current and leading theoretical models of applicant faking tend to favor the faking-as-behavior approach (e.g., McFarland & Ryan; Mueller-Hanson et al., 2006). However, tests of situational models of faking do not lend themselves easily to true applicant samples, as they require data collection over multiple periods. Organizations may be reluctant to provide access to their current or potential employees for repeated testing, thereby presenting a considerable hurdle for researchers wishing to extend lab-based tests of situational models to applied settings. As trait-based studies of faking require only single administration of a faking-based measure to investigate this phenomenon, they are more practical for applied research. The current research developed out of an applied need to determine whether faking was a problem for the Fitability 5a, therefore the trait-based strategy was deemed most appropriate. Future researchers should seek to incorporate both trait-based and situational-based models of

faking into their investigations of the faking problem, or at least recognize the limitations and benefits of each approach.

Implications.

There are several implications of the current research. Presently, organizational researchers have mixed opinions as to whether applicant faking is a problem for applicant personality testing. Much of the research in this area has examined the effects of faking on the criterion-related validity of personality tests, such as whether applicant faking results in different rank ordering of applicants. An alternative consequence of applicant faking entails the degree to which the psychological meaning of personality test scores changes due to faking. For organizations to use personality tests for making employment decisions, their tests must demonstrate measurement equivalence. In the absence of measurement equivalence, test scores become impossible to interpret, as they do not carry the same psychological meaning for members of different groups. The finding of DIF and DTF on the Fitability 5a between high and low impression management groups, then, is an issue of considerable practical importance.

The Fitability 5a's items and scales for the latent constructs of agreeableness, conscientiousness, neuroticism, and openness functioned differently for applicants that engaged in impression management versus those that did not. Mean item and scale-level scores were not only more desirable for the high impression management group, but results of the DFIT analyses suggest that the items and scales measured different latent constructs for fakers versus non-fakers. In this sense, applicant faking destroyed the construct validity of the Fitability 5a, thereby rendering the use of this measure for making selection decisions impossible.

Questions abound as to the proper method for investigating the faking problem. Thus far, the different methodologies employed by faking researchers have produced mixed results, leading to increased confusion and debate. The current research investigated the faking problem using the IRT-based DFIT methodology and determined that applicants' engagement in impression management produced differential functioning on the items and scales of a Big Five selection test. These findings have considerable implications for the ongoing debate surrounding the faking problem and suggest that future research continue to utilize more advanced approaches to item and scale analyses, such as those offered by IRT, as well as applied samples of real world job applicants.

The degree to which the current findings generalize to other personality tests or applicant populations is unknown. The Fitability 5a assesses the most readily accepted model of personality, the five-factor model. However, the items on the Fitability 5a are sure to differ from the items of other Big Five measures; therefore, it is entirely possible that applicant faking is not a problem for other Big Five tests. Nevertheless, because the majority of the Fitability 5a's items and scales exhibited differential functioning between high and low impression management groups, it seems possible if not probable that at least some items and scales of other self-report Big Five personality tests with polytomous response formats will exhibit similar trends. Additional research aimed toward replicating the current study with other Big Five measures is necessary. Converging evidence of differential functioning between fakers and non-fakers on Big Five measures would call into question the use of these tests for employment decision-making. In light of the current findings, it may be time for organizational researchers and

practitioners to begin looking toward alternative methods of personality assessment that are less susceptible to applicant faking.

References

Ackerman, P. L., & Heggested, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121,* 219-245.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*, 211.

Baker, F. B. (1997). Equate 2.1: Computer program for equating two metrics in time response theory [Computer software]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M. R., & Mount, M. K. (1994). *Personal Characteristics Inventory technical manual*. Iowa City: University of Iowa, Department of Management and Organizations.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261-272.

Barrick, M. R., & Mount, M. K. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance, 18,* 359-372.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *Personality and Performance, 9*, 9-29.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317-335.

Borgatta, E. F. (1964). The structure of personality characteristics. *Behavioral Science, 12*, 8-17.

Cattell, R. B. (1957). Personality and motivation: Structure and measurement. *Journal of Personal Disorders, 19*, 53-57.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire* (16PF). Champaign, IL: Institute for Personality and Ability Testing.

Cervone, D., & Pervin, D. C. (2007). *Personality: Theory and research* (10th ed.). New York: John Wiley & Sons.

Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47*, 847-860.

Costa, P. T., Jr., & McCrae, R. R. (1989). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R.R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor (NEO-FFI) inventory professional manual*. Odessa, FL: PAR.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

Davis, B. L., Skube, C. J., Hellervik, L. W., Gebelein, S. H., & Sheard, J. L. (1996). *Successful manager's handbook*. Minneapolis, MN: Personnel Decisions International.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Donovan, J. J., Dwight, S. D., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*, 81-106.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). The validity of non-cognitive measures decays when applicants fake. *Academy of Management Proceedings,* 127-131.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistical model. *Applied Psychological Measurement, 13*, 77-90.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*, 662-680.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15*, 13-24.

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology, 92*, 386-395.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155-166.

Ellis, B. B., & Mead, A. D. (2002). Item analysis: Theory and practice using classical and modern test theory. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 324-343). Malden, MA: Blackwell.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44*, 329-344.

Flanagan, W.J., & Raju, N. S. (1997, April). *Measurement equivalence between high and average impression management groups: An IRT analysis of personality dimensions.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, St. Louis.

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. *Personnel Psychology, 61*, 579-616.

Galton, F. (1884). Measurement of character. *Fortnightly Review, 36*, 179-185.

Gatewood, R. D., & Feild, H. S. (2001). *Human resource selection* (5th ed.). Mason, Ohio: South-Western.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In Wheeler (Ed.), *Review of personality and social psychology, 1* (pp. 141-165). Beverly Hills, CA: Sage.

Gough, H. G., & Bradley, P. (1996). *The California Psychological Inventory manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Griffith, R., Chmielowski, T., Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36,* 341-355.

Griffith, R. L., Malm, T., English, A., Yoshita, Y., & Gujar, A. (2006). Applicant faking behavior: Teasing apart the influence of situational variance, cognitive biases, and individual differences. In R. L., Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 151-178). Greenwich, CT: Information Age.

Griffith, R. L., & McDaniel, M. (2006). The nature of deception and applicant faking behavior. In Griffin, R. L. and Peterson, M. H. (Eds.), *A closer examination of applicant faking behavior* (pp. 1-20). Information Age Publishing: Greenwich, CT.

Guion, R., M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18,* 135-164.

Hair, J. F., Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data Analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Henry, M. S., & Raju, N. S. (2006). The effects of traited and situational impression management on a personality test: An empirical analysis. *Psychology Science, 48,* 247-267.

Highhouse, S. & Becker, A. S. (1993). Facet measures and global job satisfaction. *Journal of Business and Psychology, 8*, 117-127.

Hogan, R. (1982). A socioanalytic theory of personality. In M. M. Page (Ed.), *1982 Nebraska symposium of motivation*, (pp. 55-89). Lincoln: University of Nebraska Press.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100-112.

Hogan, J. Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92,* 1270-1285.

Hogan, R. (2005). In defense of personality measurement. *Human Performance, 18*, 331-341.

Hough, L.M., Oswald, F.L., Ployhart, R.E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152-194.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The five revisited. *Journal of Applied Psychology, 85*, 869-879.

John, O. P., & Donahue, E. M. (1994). *The Big Five Inventory: Technical report of the 44-item version*. Berkeley: Institute of Personality and Social Research, University of California, 1994.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*, 131-141.

Lucius, R. H. (April, 2003). *Technical report for the Fitability 5 personality profiles*. Atlanta, GA: Fitability Systems, LLC.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

McAdams, D. P. (1992). The five-factor model in personality: A critical appraisal. *Journal of Personality, 60*, 329-361.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.

Mitchelson, J. K., Wicher, E. W., LeBreton, J. M., & Craig, S. B. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits. *Education and Psychological Measurement, 69,* 613-635.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Mount, M. K., & Barrick, M. R. (1995). The big five personality dimensions: Implications for research and practice in human resource management. *Research in Personnel and Human Resource Management, 13*, 153-200.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychological Science, 48*, 209-225.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*, 574-583.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660-679.

Oshima, T. C., Kushubar, S., Scott, J. C., & Raju, N. S. (2009). *DFIT8 for Windows user's manual: Differential functioning of items and tests*. St. Paul, MN: Assessment Systems Corporation.

Paulhus, D. P. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609.

Paulhus, D. P. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (Eds.), *Measures of personality and social psychology attitudes* (pp. 17-59). New York: Academic Press.

Paulhus, D. L. (1998). *Manual for the Balanced Inventory of Desirable Responding (BIDR-7)*. Toronto, Ontario, Canada: Multi-Health Systems.

Peterson, M. H., Griffith, R. L., O'Connell, M. S., & Isaacson, J. A. (2008, April). Examining faking in real job applicants: A within-subjects investigation of score changes across applicant and research settings. In R. L. Griffith & M. H. Peterson (Chairs), *Examining faking using within-subjects designs and applicant data*. Symposium conducted at the 23rd Annual Conference for the Society for Industrial and Organizational Psychology: San Francisco, CA.

Raju, N. (2000). *Notes accompanying the differential functioning of items and tests (DFIT)* [Computer software].

Raju, N. S., van der Linden, & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187-207.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decision. *Human Performance, 14*, 187-207.

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

Rousseau, D. M., & Fried, Y. (2001). Location, location, location: Contextualizing organizational research. *Journal of Organizational Behavior, 22,* 1-13.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100-114.

Scarpello, V. & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology, 36,* 577-600.

Schmidt, F., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and non-applicant populations. *Journal of Applied Psychology, 78,* 966-974.

Smith, G. M. (1967). Usefulness of peer ratings of personality in educational research. *Educational and Psychological Measurement, 27,* 967-984.

Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87,* 211-219.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding. In B. Schneider, D. B., Smith, A. P. Brief, & J. P. Walsh (Eds.), *Personality and Organizations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Teague, S. & Thomas, A. (2008, April). *Intelligence and mood state influence faking behavior on personality tests*. Poster session presented at the annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt (2007). *Personnel Psychology, 60*, 967-993.

Tett, R. P., Jackson, D., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). *Multilog (version 7)* [Computer software]. Lincolnwood, IL: Scientific Software International.

Tupes, E. C., & Christal, R. E. (1961). Recurrent personality factors based on trait ratings. *USAF ASD Tech. Rep.* 61-97.

Viswesvaran, C., Deller, J, & Ones, D. S. (2007). Personality measures in personnel selection: Some new contributions. *International Journal of Selection and Assessment, 15*, 354-358.

Viswesvaran, C., & Ones, D. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.

Viswesvaran, C., Ones, D. S. & Hough, L.M. (2001) Do impression management scales in personality inventories predict managerial job performance ratings? *International Journal of Selection and Assessment, 9*, 277–289.

Winkelspecht, C., Lewis, P., & Thomas, A. (2006). Potential effects of faking on the NEO-PI-R: Willingness and ability to fake changes who gets hired in simulated selection decisions. *Journal of Business and Psychology, 21*, 243-259.

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

Appendix A

Figures

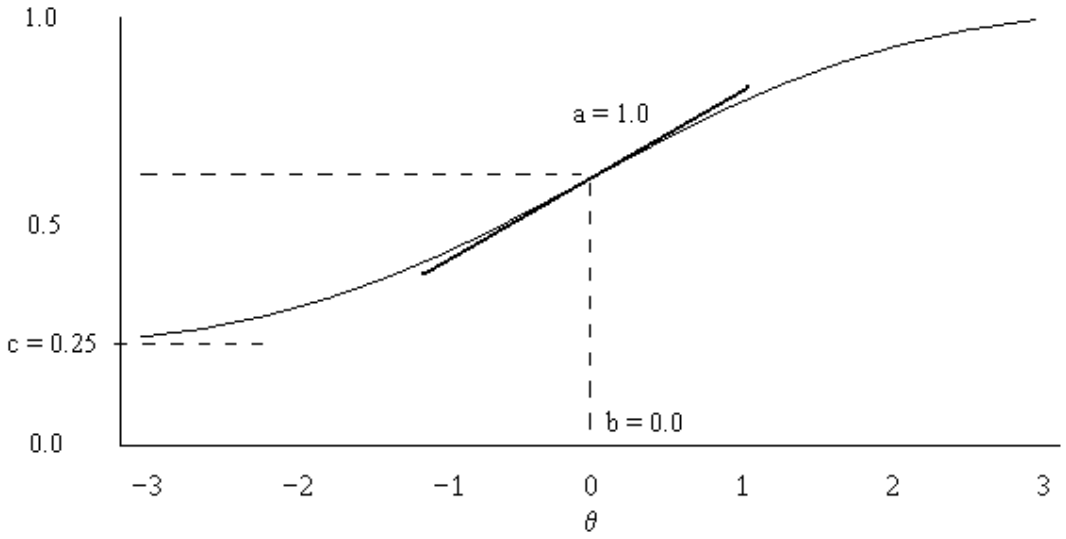
Figure Captions

Figure 1. Item Response Function (Item Characteristic Curve)

Figure 2. Category Response Functions for a Five Category Item

Figure 3. Scree plot and Eigenvalues for Impression Management Scale Development

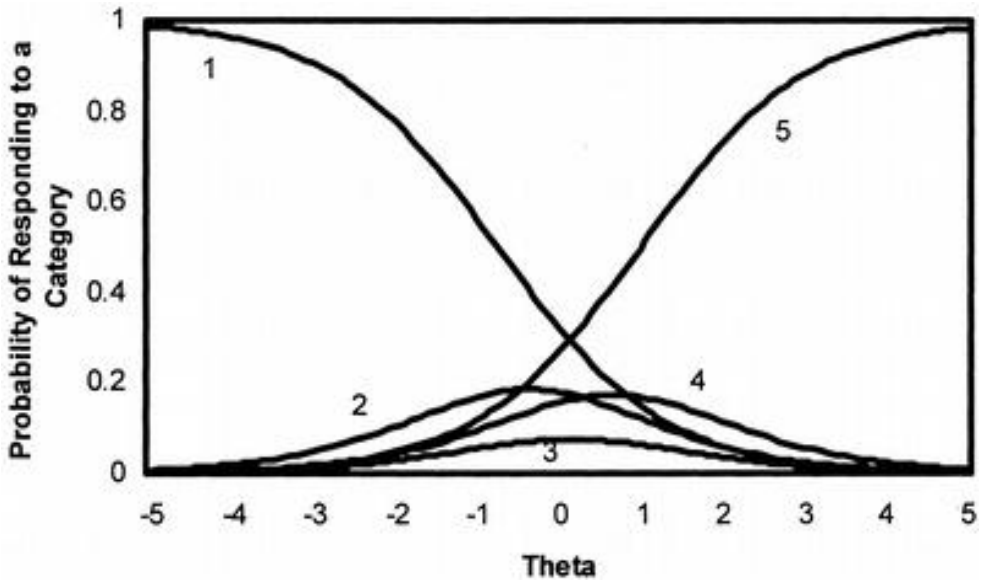
Figure 1



Parameters a, b, and c represent the item discrimination (slope), item difficulty, and lower asymptote, respectively.

Figure 2

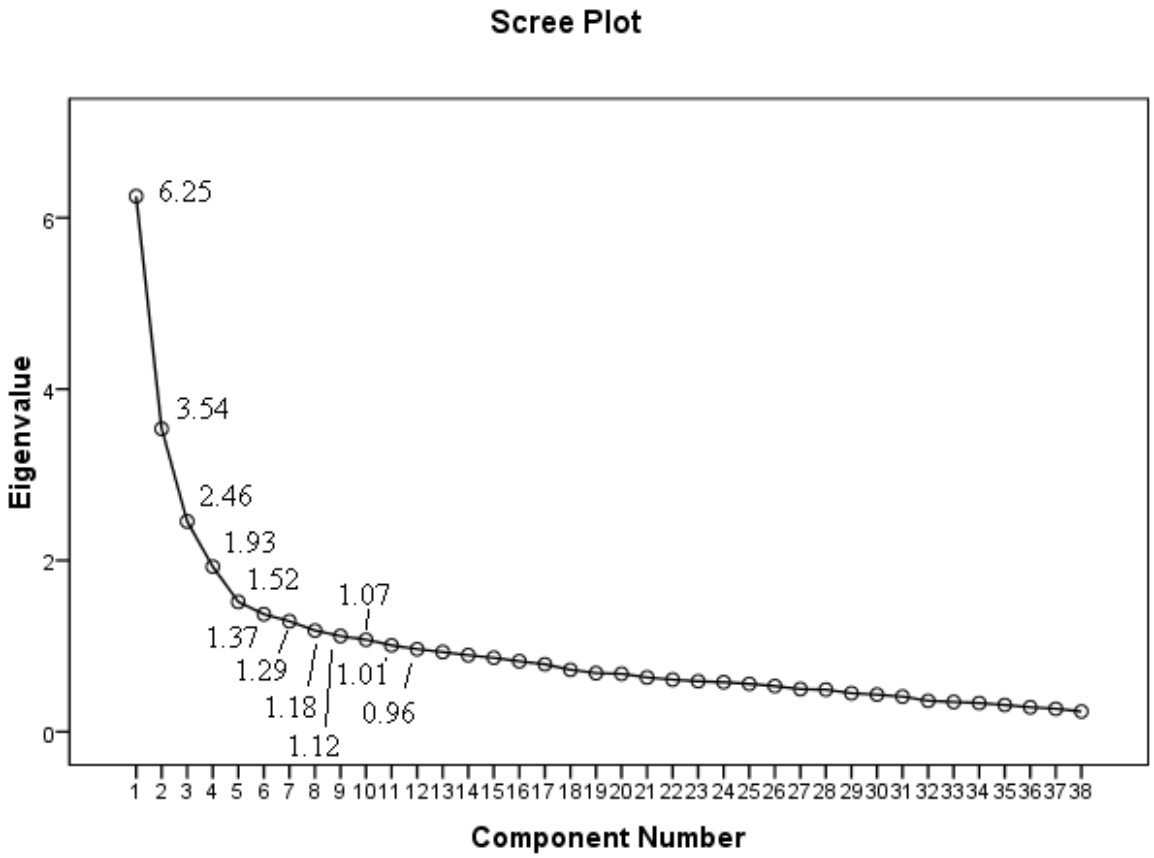
Category Response Functions for a Five Category Item



Response functions 1-5 represent the probability for selecting each of the five response choices.

Figure 3

Scree plot and Eigenvalues for Impression Management Scale Development



Appendix B

Tables

Table 1

Validity of Selection Tests Commonly Used for Predicting Overall Job Performance

Selection Test	Validity	Validity of Test Plus *g*	Gain in Validity
General mental ability (*g*)	.51	—	—
Work sample	.54	.63	.12
Structured interview	.51	.63	.12
Integrity	.41	.65	.14
Assessment centers	.37	.53	.02
Biographical data	.35	.52	.01
Conscientiousness	.31	.60	.09

Adapted from Frank L. Schmidt and John E. Hunter, "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings," *Psychological Bulletin* 124 (1998): 262-274.

Table 2

DFIT Research on Personality Test Faking and Measurement Equivalence

Study	Fakers vs. Non-Fakers	Personality Measure	Results
Flanagan & Raju (1997)	High vs. ave. IM	16-PF (extraversion)	ME
Zickar & Robie (1999)	Honest vs. fake instructions	ABLE	DIF & DTF
Robie et al. (2001)	Applicants vs. incumbents	PPI	ME
Stark et al. (2001)	a. Applicants vs. incumbents b. High vs. low IM	16-PF	DIF & DTF (a) ME (b)
Henry & Raju (2006)	a. Applicants vs. incumbents b. High vs. low IM	CPI (conscientiousness)	ME

IM = Impression management; ME = Measurement equivalence; DIF = Differential item functioning; DTF = Differential test functioning

Table 3

Items, Factor Loadings, and Internal Consistency Estimates for the Four-Factor Solution

Items	Factor Loadings			
Factor 1: Social Conventions (α = .73)				
I never listen in on other people's private conversations.	.52			
I always tell the truth.	.48			
I have lied to get myself out of trouble. (r)	.44			
I rarely gossip.	.44			
I sometimes talk bad about my friends behind their back. (r)	.36			
I find it easy to resist temptations.	.35			
I sometimes break the rules to get ahead. (r)	.34			
I always know why I do the things I do.	.33			
Factor 2: Reputation (α = .74)				
I never worry about what people think of me.		.72		
It does not upset me that some people do not like me.		.66		
It is easy to hurt my feelings. (r)		.48		
It really bothers me when people talk about me behind my back. (r)		.48		
Factor 3: Responsibility (α = .72)				
People often tell me I work too hard.			.43	
I am always responsible.			.59	
I keep all of my paperwork filed.			.57	
I can get a lot more tasks accomplished compared to others.			.65	
Too much planning makes life boring. (r)			.46	
I am very disciplined.			.64	
Factor 4: Emotions (α = .74)				
I often feel sorry for myself. (r)				.47
I sometimes try to get even rather than forgive and forget. (r)				.36
I tend to focus on the worst case scenario. (r)				.44
When someone criticizes my work, it feels like a direct attack on me as a person. (r)				.46
I have had emotional outbursts in public. (r)				.43
I get angry more than I should. (r)				.57
I usually get impatient if I have to wait. (r)				.34
I sometimes think that people are laughing at me. (r)				.60
I sometimes feel I am treated harshly without cause. (r)				.65

Table 4

Correlations among the Four Factors and the BIDR Scales

	IM (BIDR)	SDE (BIDR)	Scale 1	Scale 2	Scale 3	Scale 4
Impression Management (IM; BIDR)	1					
Self-Deception Enhancement (SDE; BIDR)	-.42**	1				
Scale 1: Social Conventions	.65**	-.53**	1			
Scale 2: Reputation	.03	-.42**	.39**	1		
Scale 3: Responsibility	.28**	-.21**	.22**	-.08	1	
Scale 4: Emotions	.43**	-.40**	.37**	.39**	-.01	1

** Significant at $p < .01$

Table 5.

Impression Management Item and Scale Means and Standard Deviations

	Total Sample		Male		Female		Caucasian		African American		Hispanic	
Item/Scale	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
I never listen in on other people's private conversations.	4.49	0.78	4.48	0.78	4.54	0.75	4.47	0.76	4.57	0.75	4.46	0.83
I always tell the truth.	4.02	1.37	4.04	1.16	3.95	1.24	4.08	1.08	3.94	1.31	3.89	1.33
I have lied to get myself out of trouble. (r)	4.17	0.89	4.18	0.88	4.14	0.94	4.10	0.89	4.28	0.90	4.30	0.85
I rarely gossip.	4.61	0.65	4.61	0.65	4.62	0.64	4.65	0.60	4.49	0.76	4.62	0.62
I sometimes talk bad about my friends behind their back. (r)	3.97	1.11	3.97	1.10	3.98	1.14	3.92	1.06	4.02	1.21	4.13	1.10
I find it easy to resist temptations.	4.66	0.68	4.66	0.68	4.69	0.66	4.64	0.68	4.72	0.65	4.74	0.63
I sometimes break the rules to get ahead. (r)	4.24	1.06	4.23	1.07	4.32	1.01	4.25	1.04	4.22	1.11	4.32	1.01
I always know why I do the things I do.	4.07	1.28	4.08	1.28	3.99	1.31	4.10	1.23	4.03	1.39	3.98	1.37
Total Scale	4.28	0.53	4.28	0.53	4.28	0.53	4.28	0.52	4.28	0.56	4.30	0.52

Table 6.

Scale Means and Standard Deviations for the Total Sample and High/Low Impression Management Subgroups

Sample/subgroup		Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness	Impression management
Total sample	Mean	4.08	4.18	3.48	2.34	4.00	4.28
	SD	0.46	0.43	0.51	0.50	0.48	0.53
High IM group	Mean	4.30	4.44	3.54	2.04	4.22	4.91
	SD	0.46	0.36	0.54	0.46	0.48	0.10
Low IM group	Mean	3.88	3.95	3.43	2.62	3.80	3.59
	SD	0.41	0.42	0.49	0.44	0.43	0.24
	Effect size	0.98	1.27	0.19	1.31	0.94	7.13

Total sample size = 21, 017. Impression management (IM) subgroups sample sizes = 5,000. Effect sizes estimated with Cohen's *d*.

Table 7. *Correlations among the Variables for the Total Sample*

	1	2	3	4	5	6
1. Agreeableness	1					
2. Conscientiousness	.41**	1				
3. Extraversion	.30**	.22**	1			
4. Neuroticism	-.26**	-.28**	-.04**	1		
5. Openness	.46**	.43**	.40**	-.25**	1	
6. Impression Management	.36**	.44**	.09**	-.46**	.34**	1

** Significant at $p < .01$

Table 8. *Correlations among the Variables for the High Impression Management Group*

	1	2	3	4	5	6
1. Agreeableness	1					
2. Conscientiousness	.31**	1				
3. Extraversion	.30**	.19**	1			
4. Neuroticism	-.15**	-.13**	.01	1		
5. Openness	.30**	.29**	.30**	-.11**	1	
6. Impression Management	.20**	.25**	.05**	-.26**	.29**	1

** Significant at $p < .01$

Table 9. *Correlations among the Variables for the Low Impression Management Group*

	1	2	3	4	5	6
1. Agreeableness	1					
2. Conscientiousness	.27**	1				
3. Extraversion	.31**	.16**	1			
4. Neuroticism	-.05**	-.09**	-.04**	1		
5. Openness	.39**	.26**	.36**	-.05**	1	
6. Impression Management	.18**	.28**	.04*	-.24**	.13**	1

* Significant at $p < .05$. *** Significant at $p < .01$.

Table 10.

Agreeableness Item Means and NCDIF, CDIF, and DTF Values for Fakers versus Non-fakers

Agreeableness Items	Faking Mean	Non-Faking Mean	NCDIF	CDIF	DTF
2. I am a charitable person	3.68	3.54	**0.247**	2.18 (2)	11.57
6. I am fanatical about finishing all tasks, no matter how trivial	4.77	4.52	0.047		
10. I am not particularly creative	3.03	2.74	0.047		
14. I am very careful with decisions, even ones others might think are…	4.60	4.17	**0.166**	1.79 (8)	0.47
22. I have a forgiving nature	4.55	3.93	0.068		
27. I like to have a plan and be organized before starting work	4.40	4.10	**0.211**	2.01 (5)	3.89
32. I often find myself taking charge of a situation or project	4.63	4.14	**0.208**	1.98 (6)	2.33
37. I sometimes talk too much	4.28	3.77	**0.274**	2.31 (1)	15.16
42. My mood is stable regardless of the situation	4.44	3.99	**0.238**	2.15 (4)	5.89
47. People often look to me to make important decisions	4.46	3.89	**0.200**	1.96 (7)	1.17
54. When someone asks for a favor it is hard for me to say no…	4.42	3.85	**0.239**	2.16 (3)	8.48

Item numbers correspond to their order on the Fitability 5a. NCDIF = noncompensatory differential item functioning. Bolded NCDIF values are statistically significant at the 0.01 alpha level. CDIF = compensatory differential item functioning. DTF = differential test functioning. Numbers in parentheses represent the order in which CDIF items were removed to achieve non-significant DTF. The critical value for DTF was 1.056 for this scale.

Table 11.

Conscientiousness Item Means and NCDIF, CDIF, and DTF Values for Fakers versus Non-fakers

Conscientiousness Items	Faking Mean	Non-Faking Mean	NCDIF	CDIF	DTF
4. I am careful in all of my decisions	4.83	4.42	**0.188**		
13. I am too busy to be reflective	4.51	4.06	**0.230**	2.79 (8)	1.25
18. I enjoy serious conversations about life and philosophy	4.75	4.21	**0.351**	3.41 (5)	7.49
25. I like to be the center of attention	4.85	4.38	**0.301**	3.07 (7)	2.69
30. I often analyze my thoughts and feelings	4.47	3.94	**0.464**	3.98 (1)	26.68
35. I often worry too much	3.70	3.49	**0.381**	3.54 (3)	15.25
40. I'd rather stay flexible than to always have everything planned out	4.80	4.13	**0.347**	3.43 (4)	11.02
45. Others see me as very social	4.59	3.97	**0.198**	2.57 (9)	0.53
50. Something has to be very important before I worry much about it	2.97	2.76	0.094		
52. Though I'm sometimes harsh, people appreciate that I "tell it like it is"	4.64	3.98	**0.344**	3.37 (6)	4.68
55. When traveling I tend to make plans well in advance	4.73	4.07	**0.434**	3.83 (2)	20.35

Item numbers correspond to their order on the Fitability 5a. NCDIF = noncompensatory differential item functioning. Bolded NCDIF values are statistically significant at the 0.01 alpha level. CDIF = compensatory differential item functioning. DTF = differential test functioning. Numbers in parentheses represent the order in which CDIF items were removed to achieve non-significant DTF. The critical value for DTF was 1.056 for this scale.

Table 12.

Extraversion Item Means and NCDIF Values for Fakers versus Non-fakers

Extraversion Items	Faking Mean	Non-Faking Mean	NCDIF
1. Good planning is more important than flexibility	4.13	3.91	0.008
5. I am curious about many different things	4.77	4.20	0.004
9. I am not moody	3.56	3.68	0.005
17. I don't like working with abstract concepts	4.70	4.20	0.003
21. I feel my best when I am around large groups of people	2.22	2.49	0.007
26. I like to clean my desk each day before leaving work	4.41	3.95	0.008
31. I often do favors for others	4.65	4.03	0.003
36. I seek thrills and excitement	3.54	3.15	0.007
41. It is ok to stop working on a job if you are getting nowhere with it	3.73	3.59	0.008
46. People know right away if I'm in a good or bad mood	1.93	2.44	0.002
51. Sometimes when I'm concerned or upset about something important…	2.69	2.79	0.002
53. When meeting someone new, I am usually the first to introduce myself	2.11	2.78	0.005

Item numbers correspond to their order on the Fitability 5a. NCDIF = noncompensatory differential item functioning.

Table 13.

Neuroticism Item Means and NCDIF, CDIF, and DTF Values for Fakers versus Non-fakers

Neuroticism Items	Faking Mean	Non-Faking Mean	NCDIF	CDIF	DTF
3. I am always willing to listen to my friends problems	1.68	2.30	0.078		
7. I am generally trusting	2.27	2.75	**0.902**	6.31 (2)	22.09
11. I am quick to forgive my friends	1.17	1.86	**0.196**	2.94 (8)	0.76
15. I can talk for long periods of time with friends, acquaintances…	3.11	2.87	0.081		
19. I enjoy telling jokes and stories at parties	1.29	2.16	**0.406**	4.23 (6)	3.19
23. I have an active imagination	1.97	2.65	**0.415**	4.27 (5)	5.84
28. I need to be around other people if I've been alone for several hours	1.31	2.23	**0.844**	6.01 (3)	14.48
33. I often get my own way	1.98	2.66	**1.007**	6.64 (1)	31.91
38. I take some time each week to organize my workspace	3.21	3.53	**0.227**	3.12 (7)	1.72
43. My mood often goes up and down	2.42	3.08	**0.555**	4.94 (4)	9.36
48. People say I worry about things that are not important	1.98	2.90	0.094		

Item numbers correspond to their order on the Fitability 5a. NCDIF = noncompensatory differential item functioning. Bolded NCDIF values are statistically significant at the 0.01 alpha level. CDIF = compensatory differential item functioning. DTF = differential test functioning. Numbers in parentheses represent the order in which CDIF items were removed to achieve non-significant DTF. The critical value for DTF was 1.056 for this scale.

Table 14.

Openness Item Means and NCDIF, CDIF, and DTF Values for Fakers versus Non-fakers

Openness Items	Faking Mean	Non-Faking Mean	NCDIF	CDIF	DTF
8. I am interested in other people's culture and perspectives	4.09	3.69	**0.219**	2.09 (5)	3.81
12. I am regarded as very, very nice, warm, pleasant and tender-hearted	4.74	4.21	**0.179**	1.88 (8)	0.40
16. I don't mind being criticized	4.18	3.82	**0.212**	2.07 (6)	2.22
20. I enjoy theoretical work	3.62	3.32	0.086		
24. I keep working on a task even when it appears that I'm not…	4.14	3.73	**0.322**	2.55 (1)	15.43
29. I never get upset when other people ridicule and tease me	4.44	3.82	**0.236**	2.18 (4)	5.84
34. I often think/rethink about how I should have said/done…	4.47	4.14	**0.214**	2.05 (7)	1.08
39. I will criticize someone in public if they deserve it	4.10	3.90	**0.278**	2.37 (2)	11.58
44. Other see me as kind and sympathetic	3.89	3.33	**0.144**		
49. People see me as creative and inventive	4.58	4.00	**0.255**	2.26 (3)	8.41

Item numbers correspond to their order on the Fitability 5a. NCDIF = noncompensatory differential item functioning. Bolded NCDIF values are statistically significant at the 0.01 alpha level. CDIF = compensatory differential item functioning. DTF = differential test functioning. Numbers in parentheses represent the order in which CDIF items were removed to achieve non-significant DTF. The critical value for DTF was 0.96 for this scale.

Table 15

Impression Management Item Means and Standard Deviations by Gender and Race

		Impression Management Items*							
Subgroup		56.	57.	58.	59.	60.	61.	62.	63.
A1. Males 1	Mean	4.48	4.02	4.19	4.62	3.97	4.65	4.25	4.12
	SD	0.79	1.17	0.87	0.62	1.10	0.69	1.05	1.25
A2. Males 2	Mean	4.48	4.04	4.19	4.62	3.97	4.66	4.23	4.11
	SD	0.78	1.16	0.88	0.64	1.11	0.67	1.07	1.26
B. Females	Mean	4.54	3.95	4.14	4.62	3.98	4.69	4.33	3.99
	SD	0.75	1.24	0.94	0.64	1.14	0.66	1.01	1.31
A1 vs. B	Effect size	0.08	0.06	0.06	0.00	0.01	0.06	0.08	0.16
	NCDIF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A2 vs. B	Effect size	0.08	0.07	0.05	0.00	0.01	0.05	0.10	0.17
	NCDIF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C1. Caucasians 1	Mean	4.47	4.09	4.09	4.64	3.90	4.63	4.25	4.09
	SD	0.77	1.06	0.90	0.60	1.07	0.69	1.04	1.23
C2. Caucasians 2	Mean	4.47	4.07	4.10	4.65	3.92	4.63	4.26	4.09
	SD	0.76	1.08	0.89	0.59	1.07	0.68	1.02	1.23
D. Afr Americans	Mean	4.57	3.94	4.28	4.49	4.02	4.72	4.22	4.03
	SD	0.75	1.31	0.90	0.76	1.21	0.65	1.11	1.39
C1 vs. D	Effect size	0.13	0.13	0.21	0.22	0.11	0.13	0.03	0.05
	NCDIF	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
C2 vs. D	Effect size	0.13	0.11	0.20	0.24	0.09	0.14	0.04	0.05
	NCDIF	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
E. Hispanics	Mean	4.46	3.89	4.30	4.62	4.13	4.74	4.32	3.98
	SD	0.83	1.33	0.85	0.62	1.10	0.63	1.01	1.37
C1 vs. E	Effect size	0.01	0.17	0.24	0.03	0.21	0.17	0.07	0.08
	NCDIF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C2 vs. E	Effect size	0.01	0.15	0.23	0.05	0.19	0.17	0.06	0.08
	NCDIF	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

*Impression management items appear in Appendix A. Item numbers correspond to their order on the Fitability 5a. NCDIF = noncompensatory differential item functioning

Appendix C

Measures

Modified 55 Item Pool for Inclusion in Scale Development

Respondents answer using the following 5-point scale

1	2	3	4	5
Strongly agree	Moderately agree	No opinion	Moderately disagree	Strongly agree

1. I am always good to others.
2. I often feel sorry for myself.
3. I always try to practice what I preach.
4. I am often the “peace maker” when arguments occur among my friends.
5. I sometimes try to get even rather than forgive and forget.
6. I usually say exactly what is on my mind.
7. I have little trouble making new friends.
8. People often tell me I work too hard.
9. I am the first to admit when I make a mistake.
10. I act differently around different people.
11. I am always responsible.
12. I tend to focus on the worst case scenario.
13. I am sometimes rude.
14. I sometimes tell lies.
15. I have never tried to cover up a mistake.
16. When someone criticizes my work, it feels like a direct attack on me as a person.
17. There have been occasions when I have taken advantage of someone to get ahead.
18. I have had emotional outbursts in public.
19. I get angry more than I should.
20. I keep all of my promises.
21. Being told not to do something makes me want to do it even more
22. I keep all of my paperwork filed.
23. I never worry about what people think of me.
24. It does not upset me that some people do not like me.
25. I have received too much change from a cashier without telling him or her.

26. I sometimes break the rules to get ahead.
27. I will do anything for others.
28. It is easy to hurt my feelings.
29. I can get a lot more tasks accomplished compared to others.
30. I sometimes take things that do not belong to me.
31. I find it easy to resist temptations.
32. Too much planning makes life boring.
33. I have used flattery to get ahead.
34. I always know why I do the things I do.
35. I am not concerned with making a good impression on people.
36. I usually get impatient if I have to wait.
37. I always tell the truth.
38. I never listen in on other people's private conversations.
39. I am a persistent and steady worker.
40. I sometimes talk bad about my friends behind their back.
41. I have lied to get myself out of trouble.
42. I have never revealed someone else's secret.
43. I am always willing to lend a hand.
44. It really bothers me when people talk about me behind my back.
45. I work hard on all jobs that I undertake.
46. It is important for me to do my best.
47. I am always nice to others.
48. If I have mistreated someone, I can hardly bear to face him or her again.
49. I get concerned when someone I am expecting does not show up on time.
50. I sometimes think that people are laughing at me.
51. I sometimes feel I am treated harshly without cause.
52. I rarely gossip.
53. I try to avoid using profanity.
54. I am very disciplined.
55. I have done things that I prefer to be kept secret.

Fitability 5a

Instructions: Below are a series of statements that broadly describe an individual's personality. Indicate whether you agree or disagree with each statement as it applies to you by selecting the appropriate response. There are no right or wrong answers, nor is there an "ideal" response for each question. Attempting to misrepresent your true personality may actually work against you. The best approach is to simply respond truthfully. Do not think too much about your answer – go with your first impression.

Items are rated on a five-point scale: 1 = *Strongly agree* to 5 = *Strongly disagree*

1 Good planning is more important than flexibility
2 I am a charitable person
3 I am always willing to listen to my friends problems
4 I am careful in all of my decisions
5 I am curious about many different things
6 I am fanatical about finishing all tasks, no matter how trivial
7 I am generally trusting
8 I am interested in other people's culture and perspectives
9 I am not moody
10 I am not particularly creative
11 I am quick to forgive my friends
12 I am regarded as very, very nice, warm, pleasant and tender-hearted
13 I am too busy to be reflective
14 I am very careful with decisions, even ones others might think are unimportant
15 I can talk for long periods of time with friends, acquaintances, coworkers… just about anyone
16 I don't mind being criticized
17 I don't like working with abstract concepts
18 I enjoy serious conversations about life and philosophy
19 I enjoy telling jokes and stories at parties
20 I enjoy theoretical work
21 I feel my best when I am around large groups of people
22 I have a forgiving nature
23 I have an active imagination
24 I keep working on a task even when it appears that I'm not making much progress
25 I like to be the center of attention
26 I like to clean my desk each day before leaving work

27 I like to have a plan and be organized before starting work
28 I need to be around other people if I've been alone for several hours
29 I never get upset when other people ridicule and tease me
30 I often analyze my thoughts and feelings
31 I often do favors for others
32 I often find myself taking charge of a situation or project
33 I often get my own way
34 I often think and rethink about how I should have said or done something better
35 I often worry too much
36 I seek thrills and excitement
37 I sometimes talk too much
38 I take some time each week to organize my workspace
39 I will criticize someone in public if they deserve it
40 I'd rather stay flexible than to always have everything planned out
41 It is ok to stop working on a job if you are getting nowhere with it
42 My mood is stable regardless of the situation
43 My mood often goes up and down
44 Other see me as kind and sympathetic
45 Others see me as very social
46 People know right away if I'm in a good or bad mood
47 People often look to me to make important decisions
48 People say I worry about things that are not important
49 People see me as creative and inventive
50 Something has to be very important before I worry much about it
51 Sometimes when I'm concerned or upset about something important, others don't seem to understand or care
52 Though I'm sometimes harsh, people appreciate that I "tell it like it is"
53 When meeting someone new, I am usually the first to introduce myself
54 When someone asks for a favor it is hard for me to say no – even if it is inconvenient
55 When traveling I tend to make plans well in advance

Balanced Inventory of Desirable Responding

Using the scale below as a guide, please respond to each statement to indicate how much you agree with it.

1 - - - - - 2 - - - - - 3 - - - - - 4 - - - - - 5 - - - - - 6 - - - - - 7
NOT TRUE SOMEWHAT TRUE VERY TRUE

1 My first impressions of people usually turn out to be right.
2 It would be hard for me to break any of my bad habits.
3 I don't care to know what other people really think of me.
4 I have not always been honest with myself.
5 I always know why I like things.
6 When my emotions are aroused, it biases my thinking.
7 Once I've made up my mind, other people can seldom change my opinion.
8 I am not a safe driver when I exceed the speed limit.
9 I am fully in control of my own fate.
10 It's hard for me to shut off a disturbing thought.
11 I never regret my decisions.
12 I sometimes lose out on things because I can't make up my mind soon enough.
13 The reason I vote is because my vote can make a difference.
14 My parents were not always fair when they punished me.
15 I am a completely rational person.
16 I rarely appreciate criticism.
17 I am very confident of my judgments.
18 It's all right with me if some people happen to dislike me.
19 I don't always know the reasons why I do the things I do.
20 I sometimes tell lies if I have to.
21 I never cover up my mistakes.
22 There have been occasions when I have taken advantage of someone.
23 I never swear.
24 I sometimes try to get even rather than forgive and forget.
25 I always obey laws, even if I'm unlikely to get caught.
26 I have said something bad about a friend behind his or her back.
27 When I hear people talking privately, I avoid listening.
28 I have received too much change from a sales person without telling him or her.
29 I always declare everything at customs.
30 When I was young I sometimes stole things.
31 I have never dropped litter on the street.

32 I sometimes drive faster than the speed limit.

33 I have done things that I don't tell other people about.

34 I never take things that don't belong to me.

35 I have taken sick-leave from work or school even though I wasn't really sick.

36 I have never damaged a library book or stole merchandise without reporting it.

37 I have some pretty awful habits.

38 I don't gossip about other people's business.