# The Development of a Computer Adaptive Test
# Of the Five Factor Model of Personality:
# Applications and Extensions

Reagan Brown

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Psychology

Robert J. Harvey (chair)
Neil Hauenstein
Sigrid Gustafson
Roseanne Foti
Jeffrey Facteau

June 20, 1997
Blacksburg, Virginia

Keywords: Computer Adaptive Testing, CAT, Personality Testing, Faking Detection, Intentional Response Distortion

# The Development of a Computer Adaptive Test
# Of the Five Factor Model of Personality:
# Applications and Extensions

Reagan Brown

## (ABSTRACT)

Although not universally accepted, much of the field has converged upon the Five Factor Model (FFM) of personality as constituting a comprehensive taxonomy of normal personality functioning.  A weakness common to all personality inventories is excessive length, which can result in examinee fatigue, and ultimately, poor data quality.  Computer adaptive testing offers a solution to the test fatigue problem by administering only the items that are informative for that examinee on a given scale.  A new test based upon the FFM of normal personality administered in a computer adaptive fashion was constructed.  Reliability and validity evidence were obtained, with favorable results.  New approaches to the detection of intentional response distortion were explored with mixed results including some promising findings in need of cross-validation.  Response latencies were able to discriminate between honest and faking subjects, but the findings were unable to clarify the issue of whether faking is an easier or more laborious cognitive process than honest answering.  New directions in computer adaptive personality testing research are proposed.

## Table of Contents

Personality measurement has been both a problem and a source of innovation for the field of psychology. It has been problematic in that many different techniques and measurement instruments, often redundant, have been developed. All of these techniques have limitations of some type. Innovation results in new attempts, techniques, and hope for future improvements in the validity of personality measurement. One of these innovations is the emergence of the Five Factor Model (FFM) of personality. Although not new, the widespread acknowledgment of the existence of the FFM of personality has served to organize many of the previous efforts to measure normal personality functioning. Its help has been in the form of a reduction in the chaos and confusion associated with the multitude of measurement instruments.

## The Birth of the Five Factor Model

The genesis of the FFM of personality began with Allport and Odbert (1936) who culled all of the personality-relevant terms from an unabridged dictionary containing over 500,000 entries. This was an attempt to be comprehensive in developing a starting point for the future construction of a personality taxonomy. The guiding principle behind this taxonomy is the lexical hypothesis, which holds that all of the “socially relevant and salient personality characteristics have become encoded in the natural language” (John, 1990, p.67). Given that the ultimate goal of the FFM is to concisely describe *normal* personality functioning, the lexical hypothesis seems to be an acceptable guiding principle. In essence, the lexical hypothesis is the fundamental assumption behind the FFM. Its invalidity jeopardizes any possible validity that the FFM might have as a comprehensive and efficient taxonomy of normal personality functioning.

Allport and Odbert’s (1936) work culminated in a list of 18,000 terms. They further categorized this list into groups, a quarter of which were traits. Traits were defined as stable, internal, causal tendencies. Cattell (1943) further reduced this list of traits through rational and empirical methods to a mere 35 variables. Much of Cattell’s attempt at empirical reduction was handicapped by the computational limitations of the time. When making the reduction from 171 clusters to a smaller set, Cattell attempted to identify the structure through factor analysis, a process that yielded a correlation matrix consisting of 14,535 coefficients, much too daunting for a true empirical analysis via the paper-and-pencil calculation methods of the time.

Cattell’s (1943) own factor analyses of his 35 variables revealed 12 factors, a number that Cattell would later change after subsequent analyses. His varying factor solutions eventually led to the development of the Sixteen Personality Factor Questionnaire (16PF; Cattell, Cattell, & Cattell, 1993). Reanalysis of Cattell’s original 35 variables as well as his 16PF have resulted in a five factor solution. Fiske (1949) was the first to find a five factor solution using Cattell’s trait descriptors. Tupes and Christal (1961, reprinted in 1992) replicated Fiske’s results across eight samples of data, finding only five replicable factors. Using either the same trait list, or a list derived from Cattell’s list, Norman (1963b), Borgatta (1964), Digman and Takemoto-Chock (1981), and Krug and Johns (1986) all found five factor solutions in their analyses. These results gave rise to McCrae and John’s (1992) comment that the “fiveness” of the FFM “… is an empirical fact, like the fact that there are seven continents on earth or eight American presidents from Virginia” (p. 194).

## Criticisms of the Five Factor Model

Comments that any research finding is an “empirical fact” invite and encourage criticism from every researcher with an unconvinced mind. Block (1995) published his manifesto of problems with the FFM citing a litany of theoretical and technical inadequacies with its development. The essence of Block’s (and others) criticism of the FFM is its comprehensiveness. Specifically, researchers reliant upon earlier trait-reducing efforts have perpetuated a mistake resulting in too many (Eysenck’s position, 1992) or too few factors (Cattell, Block, and others). That is, if 100 studies all discovered five factors in Cattell’s list of traits, then the veracity of all 100 studies could be undermined by an error in Cattell’s original work. Moreover, other instruments that were devised and drew their structure from Cattell’s list (or treat Cattell’s instrument as a criterion) perpetuate errors present in the original work. These instruments, although appearing to offer independent support for the FFM, would only serve to confirm the same problems, limitations, and errors inherent in the source.

Fortunately, truly independent research is available and confirms the FFM. In the most ambitious study, Norman (1967) eschewed not only Cattell’s work, but also that of Allport and Odbert (1936) by starting over with a new dictionary search of personality traits. Various rational reduction strategies resulted in a list of some 1,600 terms. Norman did little empirical work with this list, opting instead to further assign the adjectives to poles of the FFM.  Goldberg (1990) resumed Norman’s work by first expanding the list to include 1,710 terms, then used abbreviated lists of the common terms, which he further classified into 131 tight synonym clusters.  Factor analyses of these terms from multiple samples revealed five replicable factors matching the previous five factors.

More independent evidence comes from Conley (1985) and Field and Millsap (1989, cited in John, 1990) who analyzed trait lists from the 1930’s.  Conley’s analysis focused on a list predating Allport and Odbert’s (1936) massive search whereas Field and Millsap’s analysis used a list developed in a guidance study just shortly after Allport and Odbert.  Except for the omission of a Conscientiousness factor in the guidance list, both analyses resulted in the traditional five factors:  Agreeableness, Conscientiousness, Emotional Stability, Openness to Experience, and Surgency.  Peabody (1987) set out to achieve adequate and uniform representation in a classification project of Norman’s list.  Thus, each of the remaining terms avoided overrepresentation of a factor.  Factor analysis of these 57 trait descriptors also resulted in a five factor solution. Additionally, Peabody and Goldberg (1989) compared the 57 traits with Cattell’s 35 traits.  Results showed that Cattell’s list adequately mapped onto Peabody’s list with only slight underrepresentation (in Cattell’s list) occurring at some of the lower-level factor components.

Independent confirmation of the FFM also occurs outside of the adjective-checklist mode of measurement.  In an unpublished study, Chaplin and John (1989, cited in John, 1990) instructed 300 college students to describe themselves in positive and negative terms in a free-response format. The 10 most frequently generated items (e.g., friendly, caring, happy, selfish) easily mapped onto the FFM. Theoretically derived personality questionnaires also show concordance with the FFM. The Myers-Briggs Type Indicator (MBTI, Myers & McCauley, 1985), an instrument modeled after Jungian type theory, includes four scales which clearly correspond to four of the five FFM dimensions (McCrae & Costa, 1989). Moreover, unscored items on the MBTI load onto the missing Neuroticism factor (Harvey, Murry, & Markham, 1995). Jackson's Personality Research Form (PRF, 1984), an instrument designed to measure Murray's (1938) needs in an objective format, also demonstrates a five factor solution upon analysis (Paunonen, Jackson, Trzebinski, & Fosterling, 1992; Stumpf, 1993).

Finally, support for the FFM independent of Cattell's (1943) work is provided by studies involving non-English languages. Researchers in both The Netherlands (John, Angeleitner, & Ostendorf, 1988) and Germany (Ostendorf, 1990, cited in Goldberg & Saucier, 1995) have applied the lexical techniques of dictionary-searching to both languages, resulting in trait questionnaires. These are not merely translations from English instruments into other languages, but rather an analogue of the same process used to generate the English questionnaires. Resultant factor analyses revealed only five replicable factors for instruments in both languages. Thus, it appears that the generalizability of the FFM is not an artifact due to researcher error, but rather a reflection of personality structure.

Based on the accumulated evidence, one must ask: Is the FFM comprehensive in its organization and description of personality? The ultimate answer to this question is purpose-specific. If one is interested in describing the broad aspects of normal personality structure, then the FFM is clearly adequate in terms of content validity. If however, one desires to tap a specific trait (e.g., locus of control) not covered by the FFM, then an instrument based upon the FFM is not appropriate. “Small and specific factors may be important in some contexts, but they are not useful in organizing a broad taxonomy” (Costa & McCrae, 1995, p. 217). In addition to addressing the adequacy issue, this statement also raises the *importance* issue. Given that the FFM summarizes the broadest dimensions, are these dimensions also the most important dimensions of normal personality functioning? Conversely, are they so broad as to be trivial? The answer to this question can only be determined by the relations of the FFM to external variables.

### The Five Factor Model and Job Performance

Personality inventories based on the FFM have a spotty history as a predictor of job performance. Enough evidence of their predictive ability has accrued to generate two meta-analyses. Barrick and Mount (1991) found Conscientiousness to be the only dimension of note across a number of jobs (mean corrected validity = .22). Extraversion had the second highest average validity (corrected validity = .13) across all occupations. The other three dimensions had corrected validities of less than .10. Using stricter criteria for inclusion, Tett, Jackson, and Rothstein (1991) calculated corrected validities (corrected for criteria unreliability only) of -.19 for Neuroticism, .13 for Extraversion, .24 for Openness to Experience, .28 for Agreeableness, and .15 for Conscientiousness. Optimistically speaking, the FFM has room for improvement in the prediction of job performance.

A fundamental weakness of meta-analysis is that it focuses on main-effects only. Individual-level (i.e., occurring at the person-level, not the group-level) moderation is impossible to assess with meta-analytic procedures unless interaction terms are reported. This is an unfortunate deficiency present in both meta-analyses of the FFM personality tests. FFM-based inventories offer a great deal of promise when moderated analyses are investigated. Barrick and Mount (1993) found that managerial autonomy moderated the relationship between both Extraversion and Conscientiousness with job performance (multiple R’s .31 and .36, respectively). In an SEM analysis, Barrick, Mount, and Strauss (1993) further supported the role of Conscientiousness as a predictor of job performance by demonstrating that it was a valid predictor even after controlling for cognitive ability, Extraversion, and goal-setting behaviors. In short, the FFM does have a place in the prediction of job performance. Future investigations, exploring more moderated models as well as non-linear effects, are likely to expand and improve the role of FFM-based personality inventories in personnel selection.

### Future Directions for Personality Testing: The Present Study

Many of the popular FFM-based personality inventories share a common weakness, excessive length. Consider the following instruments and their number of items: NEO-PI-R (Costa & McCrae, 1992), 240 items; 16PF (Cattell, et al., 1993), 185 items; Hogan Personality Inventory (HPI, Hogan & Hogan, 1986), 206 items; and MBTI, 95-290 items (the most common forms have 136-166 items). Although most personality test items are short, answering questions with any degree of thought and reflection taxes the examinee.  This examinee fatigue that develops during the course of testing results, almost inevitably, in less thoughtful (careless?) response behavior towards the latter portion of the test.  This is not a trivial issue.  The MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) employs a validity scale specific to the latter portions of the test for one purpose, the detection of random response behavior due to test fatigue. Although fatigue is a problem with all testing, it is exacerbated in personality testing in which the ego is not challenged, achievement motivation is not activated, and logical connections between a given test item and real-world consequences may be difficult to draw.  In short, because there is minimal motivation and accountability, cognitive loafing is likely to occur.

Computer adaptive testing offers a solution to the length/test-fatigue problem by tailoring the test to the examinee’s trait level.  Rather than administering all of the test’s items to the examinee, computer adaptive testing gives only the items that offer information for that particular examinee. Item selection is based on an examinee’s estimated ability level for a given scale.  Ability estimates are updated after each item, reflecting the pattern of responses to all of the items administered to that point.  Item administration ceases when a specified criterion is achieved.  This criterion usually reflects a minimal standard error of measurement for that scale, but can conceivably be anything the researcher desires (e.g., specific number of items, time, minimal ability estimate shift).  By this process, computer adaptive testing does not force examinees to answer items that are meaningless to them (e.g., excessively hard items for one low in ability, items measuring facets of highly extraverted behavior for an extremely introverted subject).

Historically, computer adaptive testing has been limited to ability testing; only recently has it been applied to personality tests. Waller and Reise (1989) demonstrated that the use of computer adaptive testing reduced test length by over 50% for 90% of subjects on a measure of absorption (personality characteristics related to hypnotic susceptibility). Thus, it appears that a substantial reduction in the number of items needed for personality measurement could be achieved by use of computer adaptive testing.

Computer programs are currently available to administer tests in an adaptive format. The only requirements are that the item parameters (based on Item Response Theory: difficulty, discrimination, and pseudo-guessing) are known at the time of computer adaptive administration. Therefore, any test that will be administered in a computer adaptive fashion must first be administered to large samples in its full version, typically in a paper-and-pencil format. In contrast I have developed a new measure of personality based on the FFM and have administered it in a computer adaptive format. The instrument consists of forced-choice adjective pairs. Unlike typical forced-choice items, which pair only diametrically opposed traits such as “silent vs. talkative”, this test pairs trait descriptors that are much closer on the dimensional scale (e.g., silent vs. reserved). This allows for finer distinctions, greater measurement precision, and better measurement at the extreme positions of the scale.

Validation of the instrument has proceeded through the multitrait-multimethod technique proposed by Campbell and Fiske

(1959). Convergent validity has been determined by correlations with same-scale scores on another personality inventory, the NEO-PI-R, and same-scale acquaintance ratings. Discriminant validity has been examined through an analysis of different-scale correlations across the two external data sources. Acquaintance (or peer) ratings are a commonly employed method for validation of normal personality dimensions (Cheek, 1982; McCrae & Costa, 1987). The use of acquaintance ratings in addition to the NEO-PI-R is crucial to establishing the validity of the instrument because acquaintance ratings offer evidence independent of the correlation-inflationary effects of common response sets (i.e., common-method variance). That is, while every human making a judgment is prone to various response sets, it is extremely unlikely that two different raters, across a number of people, will possess the same response set. This use of separate raters serves to provide an unbiased estimate of the convergent validity of the instrument.

The use of acquaintance ratings with the FFM of personality has some interesting history. Passini and Norman (1966) found that a five factor solution emerges not only when people rate themselves, but also when they rate strangers. Interestingly, this result is cited by both supporters and non-supporters of the FFM of personality. Supporters feel that it points to the universality of the FFM whereas detractors claim that the results reflect the existence of commonly-shared, superficial stereotypes of personality functioning and nothing more. Subsequent studies (Funder & Colvin, 1988) demonstrated that although the factor patterns were the same, stranger ratings were much less accurate than acquaintance ratings. Acquaintances ratings correlated with each other (inter-rater agreement) and with self-ratings at .27 on average whereas stranger ratings correlated with self-ratings at .05 (inter-stranger agreement was .15). Thus, although it may be true that the FFM's factor pattern is at least partially reflective of shared personality stereotypes, the existence of a true five factor personality structure is supported by the fact that the accuracy of observer ratings increases with increased exposure and familiarity of the person being rated.

### Reliability and the Use of Computer Adaptive Testing

Computer adaptive testing has an interesting application to examinee motivation. As mentioned, any mechanism can be used to stop the administration of items. One potential mechanism is a minimal shift of ability estimates across a number of items. During careless responding behavior, the ability estimate will shift (or "bounce around") considerably. Computer adaptive administration with a "minimal shift" cut rule will likely administer all of the test's items when response behavior is careless. Forewarning examinees of this phenomenon offers examinees an interesting motivational tool for thoughtful response behavior. Specifically, if examinees are told "The more thoughtful you are, the shorter the test will be," they now have the incentive to expend cognitive resources throughout the test. This is not an empty promise, but rather reflects psychometric principles at work when a "minimal shift" cut rule is employed.

Consider the average motivation level of the examinee to respond in a thoughtful manner under the two conditions. First, for the full-test condition, longer tests will have lower "thoughtfulness" on average. Conversely, a computer adaptive administration with the aforementioned warning/instructions is likely to have high and uniform thoughtful response behavior. One has a case in which better quality data are associated with a shorter test. In classical test theory terms, this translates to reduced random error variance, or higher reliability. Thus by use of a computer adaptive administration mode, one can potentially *improve* a test's reliability rather than reduce it.

That it is possible to obtain higher reliabilities with shorter tests is counter to one of the central tenants of classical test theory which states that given equal item quality, longer tests are more reliable (Croker & Algina, 1986). This principle is quantified in the Spearman-Brown prophecy formula which demonstrates the exact pattern by which increased length results in higher reliability. Why are longer tests more reliable? The answer lies in the same process that governs any random event; given additional opportunities, any deviant pattern has greater opportunity to be negated, or offset. In questionnaire terms, if I am unsure about my exact stance on a given item (e.g., I am undecided between a "2" or "3" on a 7-point Likert scale), I must force an answer (okay, I'll go with "2"). Given another opportunity in the same situation, there is a chance I might offset the previous chance decision by choosing the other available alternative (this time I'll choose "3"). After enough items, all of the random error associated with these chance splits (random error has many other causes) would cancel out at the scale score level.

### The Faking Issue

Intentional response distortion, or faking, is a weakness common to all self-report instruments (e.g., interviews, integrity tests, bio-data, personality tests). Even ability tests are susceptible to faking, although only in a downward direction, which would offer little assistance in job selection. With the classic emphasis in selection placed on avoiding false positives, most faking research has focused on people faking good, or describing themselves in a more positive fashion than they actually are. The motivational underpinnings of faking behavior are obvious. An applicant faking good is more likely to be selected than one answering honestly but in a less positive direction.

The faking problem is exacerbated by two experimental findings. First, item subtlety is negatively correlated with criterion-related validities (Holden & Jackson, 1979). That is, items high in face validity exhibit higher correlations with criteria. However, high face validity also allows for easier fakability (Bornstein, Rossner, Hill, & Stepanian, 1994). Thus, the test developer is placed in a quandary. The potential to reap higher criterion-related validities is jeopardized by the greater ease and likelihood of intentional response distortion.

### Empirical Investigations of Faking

The first issue addressed in the realm of intentional response distortion is whether people are capable of faking. Empirical investigations uniformly demonstrate that regardless of the specific test used, subjects are able to distort their responses in a desired direction (Dunnette, McCartney, Carlson, & Kirchner, 1962; Thornton & Gierasch, 1980; Kelly & Greene, 1989; Krahe, 1989, LoBello & Sims, 1993; Thumin & Barclay, 1993; Ni, 1995).

Given that respondents are capable of faking in a desired direction, what are the effects of this response distortion on the criterion-related validities associated with use of the test? Empirical examinations of this topic have yielded three categories of results. First, some investigations have failed to detect differences in validities under faking and normal conditions (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). That is, faking failed to moderate criterion-related validities. In a slightly different style of faking study, Christiansen, Goffin, Johnston, and Rothstein (1994) compared the predictive validities of the 16PF uncorrected for faking with the validity of the same scores after a routine, empirically derived faking correction was implemented. Correcting for faking had little effect on the criterion-related validities of the scale scores.

Other investigators (Dunnette, et al., 1962) concluded that faking has a negative effect on predictive validity. Ni (1995), in a study of college GPA prediction, found that faking good decreased the criterion-related validity of some scales of the California Psychological Inventory (CPI, Gough, 1987). This confirms the common-sense notion that any construct-irrelevant variance (i.e., faking) will only serve to attenuate the systematic relationship between predictor and criterion constructs. Others have approached the negative aspects of faking from an integrity test perspective. Specifically, the validity scales associated with the personality tests are correlated with negative job behaviors (absenteeism, disciplinary actions, etc.). Costello, Schneider, and Schoenfeld (1993) found positive correlations between faking-good scale scores on the MMPI and days missed due to disciplinary actions despite range restriction in the criterion variable.

The contrasting view is offered by those who feel that faking can actually raise criterion-related validity coefficients. The logic operating behind this perspective is that any work-related behavior (test-taking, job performance, etc.) involves adopting a role (Kroger, 1974). Those people who are able to adapt to the role requirements in one situation are better capable of assuming a new role in another situation. Thus, those who are able to diagnose the correct role needed for good test performance (i.e., fake) may also identify and adopt the correct role and role-related behaviors needed for good job performance (i.e., be a high performer on the job). Although no investigation has found evidence that tests taken under faking conditions exhibited higher criterion-related validities than tests taken under honest conditions, two studies have offered indirect support for this beneficial role-taking aspect of faking behavior. Ruch and Ruch (1967) used MMPI scales to predict sales personnel performance. Significant validity coefficients between the MMPI scales and performance were actually reduced when corrections for faking were made to the scores (a customary practice with MMPI scores). Similar to the integrity test approach of faking research, Elliott (1981) found positive correlations between 16PF lie scale scores and stability for a variety of careers. Finally, Ni (1995), who found detrimental effects of faking on the validities of some CPI scales, found that faking good increased the criterion-related validity of other scales of the CPI. In light of the many conflicting research findings, no clear conclusions can currently be drawn about the effects of intentional response distortion on criterion-related validity.

### Faking and Construct Validity

Any time an instrument measures a construct other than the intended construct its construct validity has been, at least partially, contaminated. As such, measurements and predictions made using the instrument are subject to any number of biases (systematic errors) in any direction (capable of raising or lowering a score). From this perspective, faking poses a serious threat to the construct validity of personality tests. This of course is the theoretical explanation for the attenuation of validity coefficients under faking conditions.

To date, the only empirical investigations of the effects of faking on the construct validity of the Big Five personality model were offered by Schmit and Ryan (1992) and Cellar, Miller, Doverspike, and Klawsky (1996). Both investigations found that applicants exhibited a different factor structure from that found in volunteer populations. Specifically, applicants demonstrated, in addition to the customary five factors, a sixth factor reflecting an ideal candidate (high in Conscientiousness and Agreeableness; low in Neuroticism). Schmit and Ryan concluded that the factor structure did not generalize across populations. A more realistic interpretation of the results is that the sixth factor represents an effort on the part of motivated test-takers to fake good. In essence, it is faking behavior (not the use of a non-volunteer population) that affects the construct validity of the FFM of personality functioning.

### Faking Detection

Given that faking contaminates the construct validity of personality tests and the effects of faking in criterion-related validity studies are unclear, determining when applicants are faking is of paramount importance if personality tests are to be used for selection purposes.  Hough et al. (1990) reviewed the literature dealing with methods of faking detection.  Their conclusion was that the most effective strategy of reducing faking was through a warning to examinees that faking will be detected. Kluger and Colella (1993) extended much of the warning research, finding that warning reduced faking on obvious items only.  That is, a warning is effective at reducing faking on the items that traditionally have the highest validities.  Although warning reduces faking, a motivated examinee is still capable of faking.

No one strategy is capable of detecting faking, including the ubiquitous validity scale.  Although many personality tests employ validity scales, none are as detailed and researched as the MMPI’s scales (Rogers, 1984).  The MMPI employs three scales which can be used in isolation or in conjunction with each other to detect response distortion.  Additionally, because the MMPI’s items are subtle (another method used to combat intentional distortion), the examinee’s chore when trying to distort his responses is more difficult.  In effect, the MMPI enjoys two methods for reducing and detecting faking.  In spite of these safeguards, Kroger and Turnbull (1975) found that respondents were capable of faking a desired profile without detection.  Validity scale research extends beyond the MMPI.  Using the Eysenck Personality Inventory (Eysenck & Eysenck, 1968), both Dunnette, Koun, and Barber (1981) and Velicer and Weiner (1975) found that subjects could fake without validity scales detection.

The final method employed to combat faking is that of matching item responses on social desirability (Norman, 1963a). The logic in this case is because both alternatives are equally desirable (or undesirable) there is no reason to distort one’s response. This technique is also susceptible to intentional distortion on the part of the test-taker (Waters, 1965). Equal desirability of item alternatives does not preclude a motivated applicant from distorting her responses. The aforementioned methods have one characteristic in common. They focus on *what* the test-taker chooses for an answer. Given that they are all deficient in their ability to prevent and identify faking, perhaps we should adjust our focus from *what* the respondent answers to *how* the respondent answers the test items.

## Effects of Faking From an IRT Perspective

Faking research typically instructs subjects to distort their responses in some way. Faking instructions can be classified into two categories: distort your responses to present a profile to appear as angelic as possible, or pretend as if you are applying for a job and answer accordingly. Experiments using the two types of faking instructions have yielded disparate results for the various faking instructions. Hauenstein (1997), in an experiment using the two types of faking conditions as well as an honest condition, found that subjects produced markedly different profiles. The honest and “fake as good as possible” (or maximal faking) conditions yielded profiles at opposite ends of the continuum, whereas the realistic faking condition (i.e., the faking mindset of a job applicant) produced results that lie somewhere in between the two. It appears that when the instructions relate to realistic faking conditions, respondents only fake on items when they feel that they can evade detection, or when their distorted response seems believable.

This occasional faking that occurs in realistic conditions has an interesting consequence when applied to computer adaptive testing. A pattern of responses consisting of distorted answers in conjunction with honest answers is inconsistent or aberrant. An aberrant response profile results in fluctuating ability estimates, forcing a computer adaptive program to administer more items in order to obtain a stable ability estimate (Hambleton, Swaminathan, & Rogers, 1991). As a result, the number of items that an examinee must answer can function as an index of faking. However, an aberrant response pattern can also arise from simple carelessness on the part of the examinee. Thus, although the number of items administered scale is indicative of a lack of careful, honest responding, one cannot conclude from high numbers of items that intentional response distortion did occur.

## Use of Latencies as an Indicator of Deception

Response latency, or the time it takes one to respond to a question, offers another method for faking detection. Research using latencies have been plagued by mixed results. Latencies have been successfully used for the detection of deception in interview based studies. Both Harrison, Hwalek, Raney, and Fritz (1978) as well as Kraut (1978) found that longer latencies (more time between the end of the question and start of the answer) were associated with deception on the part of the respondent. However, the cognitive processes associated with answering a verbal question and answering a personality test item are not isomorphic. Perhaps the fundamental difference between the two is the depth of detail that the answer entails. Personality test questions require only a choice among alternatives, not a (presumably) logical explanation.

The use of latencies as an indicator of intentional response distortion on personality inventories has been investigated in a number of studies with mixed results. The investigations typically fall into two categories, within-subjects designs and between-subjects designs. In the within-subjects design (Kluger, Reilly, & Russell, 1991), examinees take the test twice under honest and faking conditions. Unfortunately, differential familiarity with the items results in massive order effects, such that latencies on the second condition are shorter than those on the first, regardless of manipulation. As a result, Kluger et al. concluded that response distortion had no effect on response time latency.

Other investigators have employed a between-subjects design in which subjects, once randomly assigned to groups, are told to answer the items in one of two response sets, honest or faked. Hsu, Santelli, and Hsu (1989) found shorter latencies associated with faking good on the part of respondents. Holden and colleagues has run a number of studies (between-subjects design) investigating latency times and faking behaviors (Holden, Kroner, Fekken, & Popham, 1992; Holden, 1995) using a design in which respondents’ scores are fully standardized within the person. Although this standardization completely removes any *absolute* levels of latency associated with faking behaviors, it does reveal some interesting *relative* latency information. In both of his studies, he found that subjects faking good took a significantly longer amount of time when answering “true” to items worded in a negative direction. In essence, when admitting faults (Holden’s studies employed realistic faking instructions), subjects attempting to present overly positive self-images deliberated over the decision slightly longer than for other types of items. This illustrates a key weakness of faking detection via validity scales such as social desirability.  Subjects faking good know that they should still endorse negative items occasionally; however, the process of complying with (or “beating”) the validity scales (in terms of latencies) actually yields the most diagnostic information relevant to faking detection.

## Cognitive Models of Response Distortion

At the heart of the response latency issue are the cognitive processes associated with honest versus distorted answers.  Based upon his results, Holden et al. (1992) proposed a cognitive model to explain the process of response distortion.  Their model states that items congruent with a person’s response set are likely to be answered faster than normal response times (shorter latencies). Items incongruent with a response set (e.g., a person faking good answering negatively worded items) are likely to be answered slower than normal (longer latencies).  Two problems in Holden’s research undermine Holden et al.’s theory.  First, positively worded items should be answered faster by a person faking good. This is not what Holden et al. found.  In all of their examinations, positively worded items were answered by the faking group at roughly the same speed as the honest group, not faster.

The second issue relates to the apparent contradiction between Holden’s findings and those of other researchers (e.g., Hsu et al., 1989) who found shorter latencies under all faking conditions.  The resolution lies in Holden’s experimental design. Specifically, Holden standardized all of the response latencies within each subject to control for differences in reading ability. This standardization removed all absolute levels of latency differences between the conditions.  It could be the case that all groups had shorter (or longer) latencies for all items than the honest group while still maintaining the pattern of results found by Holden.  In short, Holden’s results illuminate the relative effects of faking but in no way address the nature of the overall relationship between latencies and response distortion.

Rogers, Kuiper, and Kirker (1977) proposed a cognitive model for test-taking behavior that dictates faking behavior will be associated with shorter response latencies.  During encoding and recall of information, subjects can use a variety of techniques. Rogers et al. studied two techniques that are particularly relevant for examinees, self-referent and semantic.  Self-referent encoding “… Can be seen as a process involving the schema of self. This process involves the interaction between previous experience with personal data and new stimulus input” (p. 679).  The honest answering of personality test items is postulated to invoke self-referent processing.  Ideally, subjects scan their history to assess their fit with the item’s available alternatives.  Taking Hogan’s (1991) view that personality test items represent a self-presentation process rather than a self-report process, a subject would place herself in the context of a future event and ask, “Would I do that?”  In contrast with both is semantic processing, which is a judgment of the item’s definition, or meaning.  During semantic processing, no relationship between the item’s content and subject’s personality perceptions is assessed.  The item’s is assessed on a desirable/undesirable continuum, and an answer is selected.  Clearly semantic processing represents a streamlined approach to test-taking behavior.  Consistent with this interpretation, Rogers et al. found shorter latencies associated with semantic processing versus self-referent processing.

Applying Rogers et al.'s model to the response distortion issue, self-referent processing is analogous to honest answering whereas semantic processing is hypothesized to be the process operating during faked answers. Although Hsu et al.'s (1989) results are consistent with this model, it is not universally supported. It is clear that test-taking behavior is too complex to allow for a simple relationship between latencies and intentional response distortion. Recall that in Hauenstein (1997), realistic faking profiles differed from both maximal faking and honest profiles. Each one of these instructional sets may invoke the use of a different cognitive model. Specifically, maximal faking instructions most likely invoke semantic processing whereas honest instructions most likely invoke self-referent processing. But consider the case of realistic faking conditions, which are clearly of the greatest concern to consumers of personality tests. Does intentional response distortion streamline the cognitive processes involved with realistic faking of personality tests? The results found by Hauenstein as well as those of Holden suggest that at least in some cases faking results in more judicious thought than honest answering, implying that faking results in longer latencies. It is important to note that these longer latencies may only exist for a subset of items. That is, realistic faking instructions may be invoking longer latencies for some items *and* shorter latencies for others. If this is the process, then the absolute value of response latency will provide the best index of intentional response distortion.

Hsu et al. (1989) explored the possibility that latencies differed based on item subtlety by comparing response latencies for subtle and obvious items on the MMPI, obtaining shorter latencies for both types of items when faked. Although item subtlety does not appear to be related, Hsu et al.'s results do not rule out the existence of another process forcing a subset of items to have longer latencies while the rest of the items display shorter latencies under realistic faking conditions. The strength of the absolute value approach to faking detection is that it does not assume all people fake through the same process. That is, some subjects might follow the shorter latency approach found by Hsu et al. whereas a small subset of subjects might be more thoughtful (longer latencies) when intentionally distorting their responses. By using absolute values of latencies, all that matters is that the realistic faking group is answering in a different manner than an honest group, not simply faster or slower.

### Faking Detection and the Present Study

Given that aberrant response patterns and absolute values of response latencies should both be associated with realistic faking efforts on the part of the examinee, the use of both indices should result in even better identification of faking. That is, an interaction of the two measures should result in even more accurate identification of faking than their simple main effects. Recall that aberrant response patterns (resulting in more items being administered) can result from intentional distortion as well as from carelessness on the part of the examinee. Latencies however, should only be high during intentional faking. Multiplying the two factors together will yield high scores only when both variables are at high levels. That is, the interaction term will be maximized only when an aberrant pattern is present and latencies are high, the two characteristics hypothesized to covary with intentional response distortion only. Thus, the interaction of the two variables should offer the best available index of faking.

Hypothesis 1: Same-scale correlations between self-ratings on the computer adaptive FFM model of personality inventory and self-ratings on the NEO-PI-R will converge.

Hypothesis 2: Same-scale correlations between self-ratings on the computer adaptive FFM model of personality inventory and peer ratings will converge.

Hypothesis 3: Subjects instructed to fake realistically will be administered more items than subjects instructed to answer honestly and subjects instructed to maximally fake.

Hypothesis 4: Subjects instructed to fake realistically and subjects instructed to maximally fake will display larger absolute values of latencies than subjects instructed to answer honestly.

Hypothesis 5: The product of the absolute values of latencies with the number of items administered will be higher for subjects instructed to fake realistically than for subjects told to maximally fake and subjects told to answer honestly.

## Method

### Study 1: Test Development and Calibration

The FFM instrument constructed for this study was developed through the results of previous examinations of the lexical hypothesis (Tupes & Christal, 1992; Peabody, 1987; Peabody & Goldberg, 1989; Goldberg, 1990). Marker adjectives for each pole of each scale of the FFM were used as a starting point for a synonym and antonym search. A total of 40 items were written for each scale in the forced-choice format (e.g., extroverted vs. introverted). Twenty items per scale were constructed with words that were polar opposites of each other. The remaining 20 items contained word pairs that were similar in meaning, requiring the subject to make fine distinctions between the choices. The 200 item paper-and-pencil instrument contained the 40 items per scale in a random order, with approximately 50% of the items reverse-coded.

### Subjects

Subjects for the calibration sample consisted of friends and volunteers. All responses were anonymous and many were returned by mail to the experimenter. Less than one-third of the sample consisted of students. The rest of the sample were employed persons spanning a wide range of occupations. No subject received monetary or scholastic credit for their participation.

### Apparatus

Analyses were executed on IBM compatible personal computers. Bilog version 3.07 (Mislevy & Bock, 1990) and SAS version 6.10 (SAS Institute Inc., 1988) were used for data analysis.

### Analysis

Once a minimum of 300 completed forms were received, the initial calibration sample was analyzed for use in the computer adaptive validation study. A total of 302 forms were collected. Two subjects’ responses were deleted due to coding errors (data coded outside of valid range). Additional data were deleted due to excessive missing items. Roth (1994) indicates that missing data for any given subject may bias results when more than five percent of the items are blank. Levels less than five percent are not likely to have any biasing effects, regardless of how the missing data are handled (e.g., mean insertion, imputing missing data, etc.). For the PTI, this cutoff translates to 10 items.

Deletion of forms with more than 10 missing items removed 7 subjects from the pool. The remaining 293 subjects' responses were again checked for missing data. Many missing items resulted from light markings on the opscan forms and were entered after a check of the original opscans. Upon completion, 21 items across the 293 subjects were truly left blank. Those items were randomly inserted via a coin flip.

Once the items were coded properly, internal consistency analyses were conducted for each scale to check for unidimensionality and coding errors. Three-parameter IRT analyses were then conducted for each scale. Items offering less than .1 information units on a given scale were rejected. Based on a content examination of these items, many of them were retried on other scales, with little success. Additionally, six items were used on two scales, reflecting the non-orthogonal nature (in a definitional sense) of the FFM of personality. This process resulted in a 133 item scale. Finally, the test's instructions were prepared as a 14 item instructional "subtest."

## Study 2: Validation and Reliability

### Validation Test Selection and Construction

Two tests were required for validation of the PTI, a self-completed instrument and a peer rating form. Because many published FFM personality tests exist for the purpose of self-ratings, a variety of features were available. The NEO-FFI (Costa & McCrae, 1992) is a well-researched short form of the NEO-PI-R that offers simple FFM scores. The test contains 60 items and requires only 10-15 minutes for a focused subject to complete. The NEO-FFI was chosen for two reasons. First, its items are in sentence format (e.g., "I enjoy trying new foods") rather than adjective-pairs. A second asset for the test is that its ratings are made through a five point Likert-type format (1 = strongly agree, 2 = agree, … 5 = strongly disagree) as opposed to the forced choice format of the PTI. Given that the NEO-FFI was designed to measure the same traits as the PTI, it offers an alternate form of the PTI, employing a slightly different method of measurement.

Campbell and Fiske (1959) posit that independence of methods (necessary to evidence construct validity) varies along a continuum from complete independence to total dependence. Accordingly, two sets of self-ratings are far too similar to be considered truly independent methods. However, self and peer ratings do constitute a different method of measurement. Thus, a peer rating form of the PTI was constructed. Items were chosen for the peer rating form through the IRT analyses conducted during the calibration analysis. The best six items per scale (in terms of item information) were identified for consideration in the peer rating form. Some substitutions to slightly less informative items were made to broaden scale content. Although the ratings were made on a five point unanchored Likert-type scale, the instructions helped define the scale points.

### Subjects

Volunteers and undergraduate students participated in the validation sample. Undergraduate subjects participated in exchange for class credit. Additionally, subjects in the peer rating sample were entered in a $50 lottery.

### Apparatus

Computer based data collection were executed using IBM-compatible Dell Pentium computers. Subjects responded to each test item with the standard mouse.  CAT (Harvey, 1996), a program designed for computer adaptive test administration, administered the test.  CAT selects items based on which items offer the most information for a given ability (theta) estimate.  Although a stopping rule can be invoked to cease item administration, no rule was actually used during the actual administration of items in order to explore a variety of *post hoc* stopping rules.  CAT recorded the number of items administered, which items were administered, the order of administration, the item-level responses, and the response latency for each item.

### Procedure

Before testing, the experimenter assured the subjects that although their names and phones numbers were needed for the study, all data would be kept confidential.  Subjects completed both the PTI in computer adaptive format and the NEO-FFI in paper and pencil format.  Although subjects were instructed that administration of the PTI would cease "once the computer has a stable estimate of your [the subject's] score," the test actually administered all 133 of the items.  Follow-up conversations with many of the subjects indicated that they believed the test had indeed stopped early.  This was most likely due to the fact that completion of the 133 items required roughly 10-15 minutes per subject, a rather short test time.  After completion of both tests, subjects were asked to provide the names and phone numbers of close friends.  The subjects were told these friends would be asked to rate the subjects.  Additionally, they were told that all responses would be kept confidential.  That is, their friends would not see their test scores and they would not see their friends' ratings of them.

The friends (or peers) were contacted by phone and asked to participate in the peer rating study. Their tasks were simply to ask a secretary in the Psychology department for a friend rating form, complete the form, and return it to the secretary.

Test-retest reliability data were collected by contacting the subjects a minimum of two weeks after their initial testing session for the purpose of taking the PTI a second time.

### Development of the Stopping Rule

From a practical perspective, the ideal stopping point for any computer adaptive test would be one in which the standard error of measurement for a given person’s score is minimized, while not administering items with little ability to reduce the standard error of measurement at that theta location. In essence, one wants to administer only the items that offer non-trivial amounts of information at a person’s theta location. Setting a general cutoff (i.e., “stop administering items when the standard error of measurement is below .7”) for all test takers is a poor approximation of this goal because any test has a varying “best possible” standard error at different points along the theta scale. Figure 1 illustrates this process in which the extreme locations have much higher standard errors of measurement than the middle locations of the theta scale. Using the aforementioned general cutoff would force test-takers at the extreme theta locations to take all of the test’s items, in spite of the fact that most of these items do not increase our confidence in the location of their scores (i.e., does not reduce the standard error of measurement).

The solution is to apply a stopping rule that is sensitive to the varying levels of measurement precision along the theta scale.

Figure 2 illustrates the process. Any subset of a test’s items will produce a standard error curve that is not as low, even at its best point, as the curve produced by the full set of items. What is important is that the “subset standard error curve” is almost as good as the “full standard error curve” at the theta point of interest. Additional items will offer information at other points along the theta scale, but will have little effect at the theta location of interest. Once the “subset curve” reaches a fixed percentage of the “full curve,” item administration stops.

For this study, I chose a 90% mark for the stopping point. That is, item administration stopped as soon as the standard error of a given person’s theta score reached 90% of the best possible standard error at that theta location. One subject whose theta score is in a middle location (e.g., Subject B in Figure 2) might have her test administration stop as soon as their standard error reaches .4 s.e. units. Another subject, however, whose theta score is at an extreme location (e.g., Subject A) may find his item administration cease with their standard error of measurement at 1.0 s.e. units (obviously, poorer precision) simply because that is *almost* the best that the test can do at that theta location.

## Analysis

The NEO-FFI and the peer rating form were scored via a standard Likert scoring format (i.e., position 1 = 1 point, position 5 = 5 points). An analysis of the NEO-FFI and peer rating data revealed that two subjects each left one item in the peer-rating form data blank. The mean value was inserted. The PTI was scored by maximum likelihood estimation of theta scores at a given stopping point. Thus, the ideal subject has three different scores (one per test) for each of the five scales. In reality peer rating data were collected for 46% of the subjects, leaving the other 54% with only PTI and NEO-FFI data. Analyses of the validation data were conducted using Pearson correlations in the multitrait-multimethod format (Campbell & Fiske, 1959).

## Study 3: Faking Study

### Second Item Calibration

Because the faking study began after completion of the validation study, I was able to use late arriving data from the original paper and pencil administration of the PTI as well as PTI computer adaptive data from the validation sample to increase the size of the calibration sample. The new data allowed for a second calibration of the item parameters based on a larger sample, which consisted of 166 new subjects (459 total). Of the 81 subjects whose data were from late arriving paper-and-pencil versions of the test, one was deleted due to non-valid responses. Additionally, 7 items of the 80 remaining subjects' responses were left unanswered and were entered randomly via a coin flip.

The total sample for the second calibration sample consisted of 459 subjects. Three-parameter IRT analyses were executed for each scale. No changes in item selection or scale composition were made, only the item parameters were updated. A new version of the PTI was entered into the computer with the new item parameter estimates in place of the old ones.

### Subjects

Undergraduate students served as subjects for the faking study. Subjects received extra credit in exchange for their participation. Subjects were randomly assigned to one of three groups: honest, realistic faking, or maximal faking.

## Procedure

Because no other data were needed from the subjects, all responses for the faking study were anonymous. Instructions for the subjects in the all three conditions were, in part, the “normal” instructions for the instrument. The key aspects of these instructions are as follows: “Because of the adaptive design of the test, your answers to questions will determine the future questions you receive. Thoughtful, open, and honest answers will result in your having to answer fewer items. In short, by answering each item carefully, you can assure yourself of taking the shortest possible test.”

In addition to the normal instructions, the maximal faking group was instructed: “This is a measure of normal personality. Please do not answer honestly, but rather distort your responses to present the best possible image of yourself.” Instructions for the realistic faking group were the normal instructions followed by:

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

Finally, the subjects in all three conditions were told to avoid disturbing their fellow test-takers. They were instructed to start at the same time, proceed quietly, raise their hands if they had any problems, and leave the room quietly when finished. These precautions were necessary due to the use of response latency as a dependent variable. Any disturbance due to distraction will introduce irrelevant variance into the latency measure.

After completion of the test, the subjects were debriefed in the hall outside of the room. Subjects in the realistic and maximal faking conditions were queried by the experimenter as to whether they had remembered to fake according to instructions throughout the entire test. Two subjects admitted they had answered in an honest, normal fashion (as opposed to faking) at some point during the test. Their data were identified and deleted.

### Analysis

Three primary dependent variables were used in the faking analysis: theta score at the point at which item administration stopped (based on a stopping-rule; which was imposed after the fact), number of items administered at that point, and the response latency for each item. Once a stopping-rule is established, it is a simple matter to determine the number of items at that point and the maximum likelihood estimate of theta at that point.

Computation of the response latencies is a relatively complicated matter. Two problems must be addressed. The first is that of differential item length and the second is individual differences in reading speed. The differential item length issue was addressed by standardizing response latencies for each item across subjects. This places the response latency for each item on a common scale. Standardization was accomplished by randomly selecting (without replacement) 25 subjects’ data from the honest sample. The latency means and standard deviations from the 25 subjects were computed for each item across subjects. It is important to note that a given subject provided a data point for a given item if and only if that item was a part of that subject’s estimated theta score. For example, a stopping rule may result in only the first 60% of a person’s responses being used to estimate their theta score. The remaining 40% of the items are treated as missing data (i.e., they were never administered). This is critical because many factors (relevance, test fatigue, etc.) can cause a person’s response and style of response to change based on the item’s position in the test. In short, items administered after a stopping point are irrelevant (e.g., an introverted person answering items designed for measuring extraversion) and would bias the latency data.

Consideration of this issue reduced the n-size available for the standardization of any given item from 25 (the total number of subjects in the standardization sample) to 10-15. Additionally, there were 1-2 items per scale that had a standardization n-size of less than 5 subjects. These items were deleted from latency analysis. Finally, an examination of the maximum latencies for each item revealed some large outliers. These were cases in which a subject clearly began daydreaming. To prevent this from being an influencing factor on the standardization and actual scoring of the latencies, all latencies larger than 15 seconds were set to missing. Thus, at the end of the standardization phase, each latency reported for each subject is in z-score form based on a standardization of 10-15 honest subjects whose data constitutes a separate sample from the honest dataset.

Individual differences in reading speed were controlled by the computerized administration of the test’s instructions. The instructions were administered 1-2 sentences at a time. Below the sentences, the subject simply clicked with the mouse on the “Click here to continue” button. Each instruction segment was literally entered as a test item with only one response. Although referred to as “instructional items,” they were simply the test’s instructions in sentence form, not sample test items with a choice among alternatives. In addition to being informative, the instructional items serve as a measure of reading speed and response time. A composite time from a number of items provides a reasonably stable estimate (for the record, coefficient alpha of instructional latencies = .86; 10 items). In total, the test includes 14 instructional items. For the purpose of obtaining a stable estimate, the first three items were dropped. It is during this point that the subjects may be acclimating themselves to the testing apparatus. Additionally, the last item was dropped because some subjects were observed stopping on the last item without proceeding. As always, latencies greater than 15 seconds were set to missing. Finally, the latencies of the first five items of the actual test were set to missing (once again, to avoid negative effects of the “familiarization” period).

In short, to cope with the problems inherent with using latencies as a dependent variable, every subject’s response was corrected twice. The first correction was the standardization due to differential item lengths and the second was the correction for differential reading speed. Note that this process differs from

Holden’s (Holden et al., 1992; Holden, 1995) double standardization procedure which removes all absolute latency levels. The procedure used in this experiment remained sensitive to cases in which subjects uniformly spent longer (or shorter) times on each item, a situation that Holden’s procedure was incapable of detecting.

Analysis of the latency data for each of the faking groups proceeded in one of two ways: unsigned (absolute values) or signed (negative values allowed to occur). Both methods proceeded in similar fashions. At the item level, the subject’s reading speed control (the average of instructional items) was subtracted from her standardized response time. The absolute value of this number offered an index, for each item, of how different the subject’s latency was from their normal reading speed (faster or slower). Smaller values indicated a subject who was answering at roughly the same rate as they normally read. Larger values signaled that a subject was answering in a drastically different speed (faster or slower) from his normal reading rate. These values were then averaged by scale, providing five latency scores per subject. Means were examined by group and tested via ANOVA’s for all three groups. Additionally, mean differences between the honest and the realistic faking groups (the groups highest in relevance) were examined via t-tests.

Signed (i.e., raw) latency analysis at the scale level is a similar process. The only difference from before is that the absolute value of the corrected item latencies is not computed. Positive (meaning slower answering than normal) and negative (faster answering) values are allowed to cancel each other when averaged across items. As before, the averages are then computed for each scale. The results for any two subjects (on a given scale) can be interpreted as "Subject A (on average) answered the items slightly faster than his normal rate, but Subject B answered the items much slower (on average) than her normal rate." As before, the data were analyzed via ANOVA's and t-tests.

Finally, a discriminant analysis was executed using two variables as predictors: overall averages of the five latency scale scores (computed from the previously mentioned method), and overall averages of the number of items presented. Composites of the five individual scales were chosen because the use of only two predictors (vs. 10) minimizes shrinkage effects. The discriminant analysis (and planned t-tests) were executed on the honest and realistic faking groups only. The maximal group was omitted (from these analyses) because it is a faking condition not present in real job application scenarios. That is, the honest and realistic faking groups are the only groups that occur outside of the laboratory, thus they are the groups of real importance.

## Results

### Study 1: Test Development and Calibration

Tables 1-5 display the items and internal consistency analysis results for each scale. Tables 6-10 list the items and item parameter estimates for each scale. Figures 3-7 display the test information functions for each scale. Figures 8-12 display the test standard error curves for each scale. All of these results were obtained from the final calibration sample of 459 subjects.

As can be seen in Tables 1-5, all of the scales display adequate internal consistency reliability. Problems surface for the scales when the item parameters in Tables 6-10 are examined.

Although each of the scales have a number of highly discriminating items (a’s > 1), Agreeableness and Emotional Stability items lose a great deal of information due to their high pseudo-guessing (c) parameters.  In spite of this information loss, Emotional Stability still displays adequate precision (Figure 5, Figure 10) because it has a sufficient number of good items covering a wide theta range.  Agreeableness (Figure 3, Figure 8) lacks the same quality of items and offers acceptable precision only in a limited theta range(-.7 to +.3).  Finally, Openness to Experience (Figure 6, Figure 11) displays the same limited precision problem as Agreeableness but for a different reason.  Openness is a scale with almost no information loss due to guessing (Table 9).  It simply has only 16 items, all providing their maximum information at or near the same theta location.  This results in a scale that provides unacceptable precision across most of the theta range.

### Study 2:  Validation and Reliability

Subjects in the validation sample were asked to list the names of multiple friends.  A total of 8 subjects were rated by more than one peer.  Agreement between these pairs of raters was examined.  All raters demonstrated acceptable agreement, allowing their ratings to be averaged into one composite rating for the 8 subjects.

Table 11 shows the multitrait-multimethod matrix for the validation effort.  The PTI scales in this matrix were scored with the 90% standard error stopping rule.  The percent of items administered using this rule were: 54% for Emotional Stability, 55% for Surgency, 58% for Openness to Experience, 55% for Agreeableness, and 64% for Conscientiousness (58% across all five scales).  The diagonal of the matrix contains reliability coefficients for the three tests. Reliabilities for the NEO-FFI are coefficient alphas reported in the manual (n = 1539, Costa and McCrae, 1992, p. 53). Reliabilities for the peer rating form are also internal consistency reliabilities and were computed simply from the sample of peer rating data (n = 47). Finally, reliabilities in the PTI diagonal are test-retest reliabilities (n = 64). The average time delay between the two administrations of the computer adaptive PTI was 22.6 days. Only one subject's inter-test interval was less than 19 days.

As Table 11 shows, convergent validity for all of the scales except Agreeableness are strong. Discriminant validity is acceptable for Emotional Stability, Conscientiousness, and Surgency. Given the previously mentioned limitations with Agreeableness and Openness to Experience, it is not surprising that they are the only scales with convergent and/or discriminant validity problems.

The test-retest reliabilities for all of the scales except Agreeableness are very strong (better than .75). Making the reliabilities more impressive is that scores from both administrations follow the 90% standard error stopping rule and thus are based on roughly 58% of the item pool.

A serious concern whenever one uses a shortened version of a test is that the scores from the short form might change drastically if all of the available items are administered. This concern is extremely relevant to the computer adaptive PTI because (on average) 58% of the total items were administered to the subjects (under the 90% s.e. stopping rule). As a check for problems caused by the shortened nature of the test, scores were computed for each subject using all 133 of the test's items and were correlated with the original 90% s.e. theta scores.  Table 12 displays the intercorrelation matrix using the 90% stopping rule and the full scale theta scores.  All of the correlations are greater than .90 and indicate little or no score change as more items are administered.

Study 3:  Faking Study

Table 13 lists the average scores in z-score units for each of the faking groups by PTI scale.  Table 14 lists the average number of items administered to each group by scale (Hypothesis 3), as well as an overall average of the five scales for each group.  ANOVA's (across all groups) and t-tests (between the honest and realistic groups) for each of the five scales and an overall average of the scales revealed significant differences for Conscientiousness ($F = 4.6$, $p < .05$; $t = 2.9$, $p < .05$) Emotional Stability ($F = 5.3$, $p < .05$; $t = 3.3$, $p < .05$), Openness to Experience ($F = 3.6$, $p < .05$; $t = 1.6$, $p = .11$), and the overall composite variable ($F = 6.9$, $p < .05$; $t = 3.5$, $p < .05$).

During the course of the previous analysis, the "minimal shift" stopping rule was operationalized and executed.  Based upon the same logic as the number of items administered analysis (greater theta score change during test administration for the realistic faking group due to inconsistent response styles characteristic of that faking strategy), a theta movement statistic was computed.  Specifically, theta movement is the change in theta location between any two points during item administration.  Greater change is indicative of aberrant response patterns.  For this analysis, the theta location at the 80% s.e. mark was subtracted from the theta location at the 90% s.e. mark. The absolute value of the difference was then computed offering a theta movement statistic. Small movements (regardless of direction) indicate little theta change, and thus a consistent response style. Larger movements should be associated with an inconsistent response style, as is assumed to be occurring during realistic faking behavior.

Table 15 displays the average theta movement for each scale by group. ANOVA’s indicated that mean differences across all groups for each scale did not approach significance (all p’s > .05).

The final set of analyses focused on the response latencies. To reiterate, the unsigned (absolute value) latency analysis corrected the item latencies for reading speed (the instructional pages) at the item level, retaining the absolute value of the difference. Group averages were then calculated for each scale. Table 16 lists the mean latencies for each scale by faking group (Hypothesis 4). ANOVA’s (for all groups) and t-tests (for the honest and realistic groups) between the group means did not approach significance for each of the five scales (all p’s > .05).

The group means for each scale of the raw (or signed) latency analysis are listed in Table 17. ANOVA’s (for all groups) and t-tests (between the honest and realistic groups) approached (or achieved) significance for Agreeableness (F = 1.98, p = .14; t = -1.88, p = .06), Emotional Stability (F = 2.53, p = .08; t = -2.32, p < .05), as well as an average of all five scales (t = -1.63, p = .11).

A series of discriminant analyses was executed using only the honest and realistic faking groups for the previously discussed reasons. The first discriminant analysis predicted group membership from the (across scale) averages of the number of items administered and raw (i.e., signed) latencies. That is, each subject provided two predictors, each of which was an average of the five scale scores. The design resulted in a successful classification rate of 63.8% which is significantly greater than chance (F = 6.94, n = 106, p < .05).

An interaction term was created from the product of each predictor (Hypothesis 5). Analysis of variance of this new variable revealed no significant differences among group means (all p’s > .05). Nevertheless, a discriminant analysis with the interaction term added was executed. The results showed a slight *decrease* in classification rate (63.73%) when the interaction term was added. Thus, the interaction between the number of items and latency did not offer increased predictability and will not be discussed further.

In an effort to improve the discriminant analysis classification rates, the two predictors (each an average of the five scales) were altered in order to remove their non-contributing components. That is, the “overall average number of items administered” was previously the average of all five scales. However, the scale means by group indicate (see Table 14) that the Agreeableness mean difference is in the opposite (non-significantly) direction from the other four scales. Thus, the new “overall average number of items administered” is based on the average of the other four scales (Conscientiousness, Emotional Stability, Openness to Experience, Surgency). In a similar manner, the overall latency average was reduced to only the scales that approached significant latency group differences in the previously executed group latency t-tests. In short, the latency average became an average of Agreeableness and Emotional Stability latencies (i.e., an average of two). The discriminant analysis was executed with the two revised predictors demonstrating improved prediction in the form of a higher classification rate (67.4%). Table 18 lists the results from the discriminant analysis of the revised predictors. This new result is obviously significantly better than chance (F = 10.46, n = 106, p < .05), but not significantly better than the results of the original discriminant analysis.

Predicted values (i.e., y-hat’s) from this analysis were examined. A quick examination revealed that 9 of the 10 lowest predicted scores were from honest subjects, and 8 of the 10 highest scores were from realistic faking subjects. Figure 13 displays a chart of all of the results showing the group separation at the extremes. This is an important finding for the test’s future use in an applied setting. If a stringent cut-off score were to be set for “faked” vs. “not faked”, incidence of false positives or negatives would be near zero. Finally, it is recognized that this final discriminant analysis represents *post hoc* predictor revision and must be cross-validated in order to be considered superior to the original version.

## Discussion

### General Test Performance

Like many tests, the PTI performed adequately in the middle ranges of the theta scale only. In graphical terms, the curves in Figures 3-12 were simply too peaked. It was the unfortunate failure of many of the extreme items (i.e., designed to measure the extreme locations of the theta scale) that doomed the PTI to this position. Generally, the extreme items could be classified into two groups, neither of which performed well. The first consists of two words that are too similar in meaning to meaningfully distinguish from one another. The alternative format used two words whose definitions were distinct, but invariably one of the words would be too extreme to be a desirable selection. Not all of the items of the extreme variety performed as poorly as the above descriptions, just 60-70% of them.

Beyond the failure of a class of unused items, lie problems in two of the five scales, Openness to Experience and Agreeableness. The multitrait-multimethod matrix displays fair discriminant and convergent validity for all of the PTI's scales except these two. Openness is a scale that correlates with almost everything. Problems occurred with the Openness scale during item analysis when *a priori* facets of Openness failed to correlate with each other. Most surprisingly, curiosity items (an Openness staple) failed to load onto the general Openness construct. This left the scale with only 16 useable items, all of which measured Openness well in the .0 to +.5 theta range, but poorly everywhere else. This scale needs refinement. Other instruments (e.g., MBTI, NEO-PI-R) have little trouble using a broader Openness construct, suggesting that problems encountered with the PTI's development of an Openness scale are not permanent barriers.

Solving the Agreeableness scale's problems is commensurate with better measurement at the extreme theta locations. That is, the Agreeableness scale simply suffers from a social desirability problem. Socially desirable answering of Agreeableness items can be witnessed by the fact that a person with an average score (i.e., theta near zero) has answered “yes” to 79% of the items. This clearly restricts the range of the scale, which can only hinder test performance. The addition of more and better items at the upper end of the scale (items that people with typical levels of Agreeableness will not endorse) can solve this problem and improve the scale’s measurement precision.

Isolation of the above problems with the PTI illuminates the test’s strengths. Table 19 displays a revised version of the multitrait-multimethod matrix in which the problem scales (Openness and Agreeableness have been removed). Their removal improves the test’s discriminant validity to acceptable levels. Obviously, the scales are not orthogonal, but that simply reflects the FFM’s obliqueness. When considering only these scales, it is clear that the trait variance far exceeds the method variance.

Beyond an examination of the test’s performance at the scale level is the application of the FFM of performance in the form of a computer adaptive test. The concept of stopping item administration when a person’s standard error of measurement reaches a fixed percentage of the best possible standard error is novel. Based on the validity correlations and correlations with full-scale test scores, it can only be termed an unqualified success. Sparing the test-taker 42% of the test’s items with a maximum standard error loss of 10% is a clear improvement in personality testing. Given that the lowest correlation between the 90% s.e. scores and full-scale scores is .94, very little can be gained by the administration of the remaining 42% of the items. It is also important to note that the generally high test-retest reliability coefficients are also computed using the 90% s.e. scores. In sum, the standard error percentage stopping rule appears to provide the ultimate balance of test administration efficiency with measurement precision.

### Faking – Number of Items Administered

The faking manipulations (honest, realistic, or maximal) had their predicted effect. The maximal faking group generally had the best scores on scales possessing a clear best point (i.e., not Openness). At the other end of most of the scales were the honest subjects. In the middle of the two points were the realistic fakers who were forced to walk a fine line between self-presentation and self-preservation. The results are consistent with other faking studies (Ni, 1995; Hauenstein, 1997), and indicate successful experimental manipulations.

The analysis of the number of items administered to each subject produced a few significant results, but unfortunately none were in the hypothesized direction. Thus, tracking the number of items administered did help discriminate among faking groups, but it helped for reasons not postulated. Given that four of the five scales behaved in this manner, it is unlikely that these results were a product of sampling error. These results suggest that the causal processes are the opposite of the hypothesized causal factors. That is, test-takers are *more* consistent when faking realistically or maximally, a condition thought only to occur for maximal faking. Given that there is more inconsistency (as demonstrated by more items administered) under honest answering, a redefinition of faking behavior must be examined. If the hallmark of faking (realistic or otherwise) is consistency, then the test taker must be attending to that one aspect of her pattern of answers. This strongly suggests that a characteristic of honest personality description is acknowledgement of the inherent inconsistencies in real behavior. Realistic (or maximal) faking washes out those “wrinkles”.

Analysis of the theta movement variable was without success. It is difficult to reconcile the fact that the number of items administered analysis resulted in significant differences whereas the theta movement analysis (based upon the same principles) failed to evidence even the slightest mean differences. The most likely explanation is that the window of opportunity needed to identify theta movement is so large that it encompasses most of a scale’s items. That is, movement might have to be tracked from an early point in the test to one of the final items in the scale. At that point, it is unlikely that the theta movement would offer any predictability not already captured by the number of items administered.

### Faking – Latencies

Consideration of many previous research efforts led to the concept of using the absolute value of the item latencies as a measure of faking. Given that the absolute values produced (absolutely) dismal results, it appears in retrospect that the role of response latency had been unnecessarily complicated. Differences between the faking group are being obscured, not revealed, by the use of the absolute value. Recall that their use was necessitated by the notion that some items might have longer latencies whereas other items might be associated with shorter latency within a faking subject’s response pattern. Clearly this is an overcomplicated view of response latency.

Analysis of the simpler signed latencies did reveal consistent means patterns across scales. The honest group had the shortest latencies, whereas the realistic faking group had the longest latencies, with the maximal fakers somewhere in between the two. Comparisons between the honest and realistic fakers revealed only modest effect sizes (the reason why significant differences were rare), but as with the number of items administered analysis, the pattern of differences was consistent across all five scales. This pattern of latencies supports the model of test-taking that dictates faking adds a cognitive step to the process. This was expected for the realistic faking group, but it is surprising that the maximal faking group also followed this pattern given that their job was the simplest of all.

These results (the signed latencies and number of items administered) do a credible job of discriminating between honest and realistic faking subjects. This of course has natural applicability whenever personnel selection is based on personality tests. The only concern is that these results may be inflated by sampling error. Consideration of the fact that the better predictor in the discriminant analysis (number of items administered) featured results in the opposite direction from the hypothesized direction and that the other predictor (latency) was modified before it succeeded in boosting correct classifications to higher levels, is enough to give one cause for concern that cross-validation might reduce the discriminant ability of the two predictors to trivial levels. However, it is important to note that these results were obtained with more than a 50-1 subjects to predictors ratio, and the better predictor achieved significance before it was tweaked to improve results (recall that Agreeableness was dropped from its composite). Nevertheless, these results must be cross-validated.

### A Potentially Serious Confound

It was expected that the three faking groups would not differ on reading speed (instructions) given that they were randomly assigned to groups.  However, a check of the instructional means (Table 20) revealed significant differences between the honest and faking groups (F = 4.44, p < .05).  The honest group spent almost one half of a second longer on each instructional item than did the other two groups, who had almost the same means.  If this difference reflects nothing other than sampling error, then all is well with the previous results and conclusions.

If, however, the longer latencies are the result of an experimental confound, then previous latency conclusions would be incorrect.  The confound in question could be related to the experimental manipulation.  Although all three groups received the normal instructions in the form of computer items (these were used to control for reading speed), the realistic and maximal faking groups received their manipulation instructions verbally before testing began.  To these subjects (and not the honest subjects), the normal computer administered instructions may have appeared less important, and were read more quickly as a result.  These groups had no reason to pay less attention to the instructional items because at no other point were they given the specific task-relevant information contained therein.  Nevertheless, those groups did read at a faster rate.  The confound could arise from a situation in which subjects from all three groups were actually (on average) equal in reading speed, but the experimental manipulations caused the faking subjects to display shorter instructional times.  Thus, when each test item is corrected for reading speed, the correction biases results, with each member of the honest group having their scores corrected .5 seconds more than the faking groups.

Table 21 displays the means when a control for reading speed is not used (granting that the group reading speeds were actually equal negates the need for a correction as long as group-level effects are examined). Oddly, the three scales that were previously not significant in the signed latency analysis now achieve (or approach) significance (Agreeableness F = 2.4, p = .09; Conscientiousness F = 7.73, p < .05; Surgency F = 3.2, P < .05) while previously significant scales no longer approach significance (p’s > .05). In essence, two scales were traded for three. An average of these three scales was run in a discriminant analysis with the previously used average of the number of items presented, resulting in a correct classification rate of 68.9%.

It appears that even if the manipulation instructions unintentionally affected the subjects’ behavior, the ability of response latencies to aid in the identification of realistic faking is not adversely affected. The interpretation of the cognitive processes of faking do change, however. Faking (both realistic and maximal) now results in shorter latencies, on average, supporting a different cognitive model of faking. Specifically, this model states response distortion streamlines the decision making process, making test-taking less effortful than honest answering.

## Future Directions

Ambiguities and mistakes arising from this study can be eliminated with one carefully executed follow-up study. This study would feature two critical aspects. First, the standardization sample (which put all of the items on a common latency metric) would be increased from an n-size of 10-15 subjects per item to 50 subjects per item. Larger sample sizes are obviously more likely to correctly approximate population means and standard deviations due to the reduced effect of outliers. Use of a larger standardization sample would serve to reduce more of the random error in the latencies, thereby increasing effect sizes of the already significant scales.

The second change would be an alteration in the manipulation instructions in order to eliminate the possibility of the previously discussed confound. The honest group would receive instructions similar to the realistic group before starting the test. If any biasing outcome were to arise from these instructions, its effects would be spread equally over both groups. Additionally, these instructions would serve to maximize realism by instructing the honest subjects to pretend as if: they are applying for a job, they are taking this test as part of the application process, the test has mechanisms designed to detect faking, and anyone identified as faking will not be considered for the job. To increase examinee motivation, a financial incentive would be offered for subjects with the lowest faking score. Similarly, the realistic faking group would receive the same instructions as before, but also with a financial incentive (highest score on the test without being identified as faking).

Other future directions involve the previously discussed improvement of measurement precision at the extreme theta locations. One of the biggest potential benefits from this improvement is in the number of items administered variable. Increased measurement sensitivity dictates that inconsistent answers will result in greater theta change, causing the honest group to complete more items before the stopping point. In short, this will lead to larger effect sizes for the number of items variable and hence, better discrimination between the honest and realistic faking groups.

One final problem that must be addressed in future studies is faking identification via use of latency in low verbal ability samples. Will low ability subjects be disproportionately falsely identified as faking? There is no reason why they should be falsely identified at a higher rate, but the implications of a false assumption in this context could prove to be extremely detrimental in a disparate impact analysis.

# References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. Psychological Monographs, 47, (No. 211).

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. (1993). Autonomy as a moderator of the relationships between the big five personality dimensions and job performance. Journal of Applied Psychology, 78, 111-118.

Barrick, M. R., Mount, M. K., & Strauss, J. P. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. Journal of Applied Psychology, 78, 715-722.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. Psychological Bulletin, 117 (2), 187-215.

Borgatta, E. F. (1964). The structure of personality characteristics. Behavioral Science, 9, 8-17.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L., (1994). Face validity and fakability of objective and projective measures of dependency. Journal of Personality Assessment, 63, 363-386.

Butcher, J. N., Dahlstrom, V. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). Minnesota Multiphasic Personality Inventory-2 manual for administration and scoring. Minneapolis, MN: University of Minnesota Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. Journal of Abnormal and Social Psychology, 38, 476-506.

Cattell, R. B., Cattell, A. L., & Cattell, H. E. (1993). The 16 personality factor questionnaire. (5th ed.). Psychological Corporation: San Antonio, TX.

Cellar, D. F., Miller, M. L., Doverspike, D. D., & Klawsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measure of personality based on the five factor model. Journal of Applied Psychology, 81, 694-704.

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. Journal of Personality and Social Psychology, 43, 1254-1269.

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.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlational analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Conley, J. J. (1985). Longitudinal stability of personality traits: A multitrait-multimethod-multioccasion analysis. Journal of Personality and Social Psychology, 49, 1266-1282.

Costa, P. T., & McCrae, R. R., (1992). Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory

(NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R., (1995). Solid grounds in the wetlands of personality: A reply to Block. Psychological Bulletin, 117 (2), 216-220.

Costello, R. M., Schneider, S. L., & Schoenfeld, L. S. (1993). Applicants’ fraud in law enforcement. Psychological Reports, 73, 179-183.

Croker, L. & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth: Harcourt Brace Jovanovich.

Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: Re-analysis and comparison of six major studies. Multivariate Behavioral Research, 16, 149-170.

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, 13-14.

Dunnette, S., Koun, S. & Barber, P. J. (1981). Social desirability in the Eysenck Personality Inventory. British Journal of Psychology, 72, 19-26.

Elliott, A. G. (1981). Some implications of lie scale scores in real-life selection. Journal of Occupational Psychology, 54, 9-16.

Eysenck, H. J. (1992). Four ways five factors are not basic. Personality and Individual Differences, 13, 666-672.

Eysenck, H. J., Eysenck, S. B. (1968). Manual: Eysenck Personality Inventory. San Diego, CA: Educational and Industrial Testing Service.

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

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*, 149-158.

Goldberg, L., R. (1990). An alternative “description of personality”: The big five factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.

Goldberg, L. R., & Saucier, G. (1995). So what do you propose we use instead? A reply to Block. *Psychological Bulletin, 117* (2), 221-225.

Gough, H. G. (1987). *The California Psychological Inventory administrator’s guide*. Palo Alto, CA: Consulting Psychologists Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harrison, A. A., Hwalek, M., Raney, D. F., & Fritz, J. G. (1978). Cues to deception in an interview situation. *Social Psychology, 41*, 156-161.

Harvey, R. J. (1996). *CAT: A computer adaptive testing program*. Blacksburg, VA: Personnel Systems & Technologies Corporation.

Harvey, R. J., Murry, W. D., & Markham, S. E. (1995). *A “Big Five” scoring system for the Myers-Briggs Type Indicator*. Paper presented at the 10th annual conference of the Society for Industrial and Organizational Psychology, Orlando.

Hauenstein, N. M. (1997). A faking study: A study of faking. Manuscript in preparation.

Hogan, R. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 873-919). Palo Alto, CA: Consulting Psychologists Press.

Hogan, R., & Hogan, J. (1986). The Hogan Personality Inventory manual. Minneapolis: National Computer Systems.

Holden,, R. R. (1995). Response latency detection of fakers on personnel tests. Canadian Journal of Behavioural Science, 27, 343-355.

Holden, R. R., & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. Journal of Consulting and Clinical Psychology, 47, 459-468.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. Journal of Personality and Social Psychology, 63, 272-279.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effects of response distortion on those validities [Monograph]. Journal of Applied Psychology, 75, 581-595.

Hsu, L. M., Santelli, J., & Hsu, J. R. (1989). Faking detection validity and incremental validity of response latencies to MMPI subtle and obvious items. Journal of Personality Assessment, 53, 278-295.

Jackson, D. N. (1984). Personality Research Form manual (3rd ed.). Port Huron, MI: Research Psychologists Press.

John, O. P. (1990). The “Big Five” factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. Pervin (Ed.), Handbook of personality: Theory and research (pp. 66-100). New York: Guilford Press.

John, O. P., Angeleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical view of trait taxonomic research. European Journal of Personality, 2, 171-203.

Kelly, D. B., & Greene, R. L. (1989). Detection of faking good on the MMPI in a psychiatric inpatient population. Psychological Reports, 65, 747-750.

Kluger, A. N., & Colella, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. Personnel Psychology, 46, 763-780.

Kluger, A. N., Reilly, R. R., & Russell, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? Journal of Applied Psychology, 76, 889-896.

Krahe, B. (1989). Faking personality profiles on a standard personality inventory. Personality and Individual Differences, 10, 437-443.

Kraut, R. E. (1978). Verbal and nonverbal cues in the perception of lying. Journal of Personality and Social Psychology, 36, 380-391.

Kroger, R. O. (1974). Faking in interest measurement: A social-psychological perspective. Measurement and Evaluation in Guidance, 7, 130-134.

Kroger, R. O., & Turnbull, W. (1975). Invalidity of validity scales: The case of the MMPI. Journal of Consulting and Clinical Psychology, 43, 48-55.

Krug, S. E., & Johns, E. F. (1986). A large scale cross-validation of second-order personality structure defined by the 16PF. Psychological Reports, 59, 683-693.

LoBello, S. G., & Sims, B. N. (1993). Fakability of a commercially produced pre-employment integrity test. Journal of Business and Psychology, 8, 265-273.

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.

McCrae, R. R., & Costa, P. T. (1989). Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. Journal of Personality, 57, 17-40.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. Journal of Personality, 60, 175-215.

Mislevy, R. J. & Bock, R. D. (1993). Bilog 3: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Murray, H. A. (1938). Explorations in personality. New York: Oxford University Press.

Myers, I. B. & McCauley, M. H. (1985). Manual: a guide to the development and use of the Myers-Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press.

Ni, Y. (1995). The effect of intentional response distortion on the accuracy of predictive inferences of personality inventories. Unpublished Master’s Thesis.

Norman, W. T. (1963). Personality measurement, faking, and detection: An assessment method for use in personnel selection. Journal of Applied Psychology, 47, 225-241.

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.

Norman, W. T. (1967). On estimating psychological relationships: Social desirability and self-report. Psychological Bulletin, 67, 273-293

Passini, F. T. & Norman, W. T. (1966). A universal conception of personality structure? Journal of Personality and Social Psychology, 4, 44-49.

Paunonen, S. V., Jackson, D. N., Trzebinski, J., & Fosterling, F. (1992). Personality structure across cultures: A multimethod evaluation. Journal of Personality and Social Psychology, 62, 447-456.

Peabody, D. (1987). Selecting representative trait adjectives. Journal of Personality and Social Psychology, 52, 59-71.

Peabody, D. & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. Journal of Personality and Social Psychology, 57, 552-567.

Rogers, R. (1984). Toward an empirical model of malingering and deception. Behavioral Science and the Law, 2, 93-112.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. Journal of Personality and Social Psychology, 35, 677-688.

Roth, P. L. (1994). Missing data: a conceptual review for applied psychologists. Personnel Psychology, 47, 537-560.

Ruch, F. L., & Ruch, W. W. (1967). The K factor as a (validity) suppresser variable in predicting success in selling. Journal of Applied Psychology, 51, 201-204.

SAS Institute Inc. (1988). SAS/STAT User's Guide, Release 6.03 Edition. Cary, NC: SAS Institute Inc.

Schmit, M. J., & Ryan, A. M. (1992). The big five in personnel selection: Factor structure in applicant and nonapplicant populations. Journal of Applied Psychology, 78, 966-974.

Stumpf, H. (1993). The factor structure of the Personality Research Form: A cross-national evaluation. Journal of Personality, 61, 27-48.

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

Thornton, G. C., & Gierasch, P. F. (1980). Fakability of an empirically derived selection instrument. Journal of Personality Assessment, 44, 48-51.

Thumin, F. J., & Barclay, A. G. (1993). Faking behavior and gender differences on a new personality research instrument. Consulting Psychology Journal, 45, 11-21.

Tupes, E. C., & Christal R. E. (1992). Recurrent personality factors based on trait ratings. Journal of Personality, 60, 225-251. (Original work published in 1961).

Velicer, W. F., & Weiner, B. J. (1975). Effects of sophistication and faking sets on the Eysenck Personality Inventory. Psychological Reports, 37, 71-73.

Waller, N. G., & Reise S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. Journal of Personality and Social Psychology, 57, 1051-1058.

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

Table 1

Internal Consistency Analysis for Agreeableness

```
Cronbach Coefficient Alpha
for STANDARDIZED variables:  0.822119
n = 459

Item        Correlation
Number      with Total    Alpha if Dropped
------------------------------------
1           0.239001         0.820813
2           0.209421         0.821987
3           0.369670         0.815551
4           0.215094         0.821763
5           0.260809         0.819943
6           0.058656         0.827877
7           0.490690         0.810568
8           0.397411         0.814419
9           0.362113         0.815859
10          0.252040         0.820294
11          0.372422         0.815439
12          0.322519         0.817464
13          0.364197         0.815774
14          0.432993         0.812957
15          0.457927         0.811928
16          0.451092         0.812211
17          0.377805         0.815220
18          0.245796         0.820542
19          0.355809         0.816115
20          0.425054         0.813284
21          0.518240         0.809419
22          0.331567         0.817098
23          0.379547         0.815149
24          0.472043         0.811343
25          0.365973         0.815702
26          0.423331         0.813355
```

Table 2

Internal Consistency Analysis for Conscientiousness

```
Cronbach Coefficient Alpha
for STANDARDIZED variables:  0.867638
n = 459

Item     Correlation
Number   with Total   Alpha if Dropped
-------------------------------------
1        0.492025         0.861387
2        0.455884         0.862243
3        0.311683         0.865616
4        0.179749         0.868644
5        0.545710         0.860108
6        0.490204         0.861431
7        0.196150         0.868271
8        0.268867         0.866605
9        0.523663         0.860635
10       0.421733         0.863048
11       0.433898         0.862762
12       0.289872         0.866121
13       0.264202         0.866712
14       0.351221         0.864698
15       0.460494         0.862134
16       0.430357         0.862845
17       0.487426         0.861497
18       0.221773         0.867686
19       0.432907         0.862785
20       0.356898         0.864566
21       0.286621         0.866196
22       0.299472         0.865899
23       0.441777         0.862576
24       0.383019         0.863956
25       0.508710         0.860991
26       0.450617         0.862368
27       0.315567         0.865526
28       0.463420         0.862065
29       0.335500         0.865064
30       0.238419         0.867304
31       0.204082         0.868090
32       0.569768         0.859532
33       0.402980         0.863488
```

Table 3

Internal Consistency Analysis for Emotional Stability

```
Cronbach Coefficient Alpha
for STANDARDIZED variables:  0.844751
n = 459

Item     Correlation
Number   with Total    Alpha if Dropped
-------------------------------------
1        0.350265         0.840525
2        0.555131         0.834155
3        0.363396         0.840123
4        0.381328         0.839572
5        0.441731         0.837706
6        0.320337         0.841439
7        0.454672         0.837304
8        0.425192         0.838218
9        0.389392         0.839324
10       0.298205         0.842112
11       0.357744         0.840296
12       0.144694         0.846717
13       0.438658         0.837801
14       0.587567         0.833128
15       0.295627         0.842190
16       0.277701         0.842733
17       0.215734         0.844600
18       0.356790         0.840325
19       0.525893         0.835076
20       0.453897         0.837328
21       0.340621         0.840820
22       0.293502         0.842255
23       0.323568         0.841340
24       0.359731         0.840235
25       0.511134         0.835540
26       0.324237         0.841320
27       0.135524         0.846989
28       0.087858         0.848395
29       0.385648         0.839439
30       0.410772         0.838664
```

Table 4

Internal Consistency Analysis for Openness to Experience

```
Cronbach Coefficient Alpha
for STANDARDIZED variables:  0.807481
n = 459

Item     Correlation
Number   with Total    Alpha if Dropped
-------------------------------------
1        0.586920         0.784739
2        0.605259         0.783433
3        0.455303         0.793940
4        0.565836         0.786232
5        0.465298         0.793251
6        0.496118         0.791117
7        0.240963         0.808312
8        0.534058         0.788469
9        0.319681         0.803120
10       0.212878         0.810140
11       0.497836         0.790998
12       0.265470         0.806706
13       0.244688         0.808068
14       0.344386         0.801470
15       0.295033         0.804756
16       0.415471         0.796667
```

Table 5

Internal Consistency Analysis for Surgency

```
Cronbach Coefficient Alpha
for STANDARDIZED variables:  0.857634
n = 459

Item     Correlation
Number   with Total   Alpha if Dropped
-------------------------------------
1        0.528473         0.849007
2        0.412136         0.852448
3        0.398368         0.852851
4        0.500353         0.849844
5        0.491916         0.850095
6        0.581166         0.847427
7        0.291415         0.855955
8        0.279609         0.856294
9        0.375403         0.853522
10       0.518777         0.849296
11       0.357072         0.854056
12       0.445665         0.851463
13       0.567304         0.847844
14       0.121819         0.860770
15       0.416891         0.852309
16       0.289705         0.856004
17       0.491334         0.850112
18       0.410585         0.852494
19       0.298982         0.855737
20       0.543619         0.848554
21       0.553596         0.848255
22       0.168309         0.859463
23       0.626289         0.846065
24       0.207615         0.858350
25       0.140648         0.860242
26       0.187252         0.858927
27       0.361611         0.853924
28       0.360350         0.853960
```

Table 6

Item Parameter Estimates for Agreeableness

N = 459

| Item | Discrimination (a) | Difficulty (b) | Pseudo-Guessing(c) |
|---|---|---|---|
| 1 | 0.97521 | 0.63430 | 0.45913 |
| 2 | 0.46418 | -0.59568 | 0.36746 |
| 3 | 1.03827 | -1.58992 | 0.33538 |
| 4 | 1.06833 | 0.55343 | 0.49439 |
| 5 | 0.63680 | -0.04151 | 0.37440 |
| 6 | 0.64452 | 2.24014 | 0.50000 |
| 7 | 1.34181 | -1.04464 | 0.27219 |
| 8 | 0.78972 | -0.83158 | 0.25944 |
| 9 | 1.07974 | -0.12188 | 0.36393 |
| 10 | 1.05075 | 0.45176 | 0.45620 |
| 11 | 0.81593 | -1.00748 | 0.29062 |
| 12 | 0.92365 | -0.15876 | 0.37323 |
| 13 | 0.90232 | -0.25467 | 0.36801 |
| 14 | 1.25291 | -0.20884 | 0.30993 |
| 15 | 1.28185 | -0.85333 | 0.30803 |
| 16 | 1.49050 | -1.19118 | 0.31771 |
| 17 | 0.90993 | -1.12734 | 0.31558 |
| 18 | 0.49603 | -1.44411 | 0.31749 |
| 19 | 1.20279 | 0.10168 | 0.37285 |
| 20 | 1.15417 | -0.17339 | 0.31955 |
| 21 | 1.65675 | -1.01588 | 0.29026 |
| 22 | 0.76921 | 0.00751 | 0.30188 |
| 23 | 1.30000 | -1.68442 | 0.30066 |
| 24 | 1.32979 | -0.14281 | 0.27446 |
| 25 | 0.96427 | -1.10040 | 0.35302 |
| 26 | 1.08555 | -0.83678 | 0.29981 |

Table 7

Item Parameter Estimates for Conscientiousness

N = 459

| Item | Discrimination (a) | Difficulty (b) | Pseudo-Guessing(c) |
|---|---|---|---|
| 1 | 1.34332 | -1.22818 | 0.14562 |
| 2 | 0.98634 | -1.29882 | 0.09716 |
| 3 | 0.49701 | -0.40184 | 0.07793 |
| 4 | 0.46865 | 1.87179 | 0.12856 |
| 5 | 1.19507 | -0.59570 | 0.09474 |
| 6 | 0.95442 | -0.80981 | 0.08715 |
| 7 | 0.35923 | 0.90149 | 0.10115 |
| 8 | 0.44788 | -0.47741 | 0.10320 |
| 9 | 1.11036 | -0.91274 | 0.07986 |
| 10 | 0.75105 | -0.52327 | 0.09017 |
| 11 | 0.81739 | -0.09839 | 0.09422 |
| 12 | 0.50379 | -0.84312 | 0.10647 |
| 13 | 0.54143 | -2.48514 | 0.10124 |
| 14 | 0.62411 | -1.54170 | 0.08862 |
| 15 | 0.84787 | -0.53077 | 0.09207 |
| 16 | 0.85061 | -0.71189 | 0.12987 |
| 17 | 0.97660 | -1.22065 | 0.07226 |
| 18 | 0.41653 | 1.08534 | 0.10820 |
| 19 | 0.80946 | -1.24552 | 0.08900 |
| 20 | 0.58303 | 0.15756 | 0.07433 |
| 21 | 0.45582 | 0.05448 | 0.09746 |
| 22 | 0.55261 | -1.94687 | 0.09214 |
| 23 | 0.82531 | -0.93302 | 0.08903 |
| 24 | 0.68983 | -1.40532 | 0.09924 |
| 25 | 1.12133 | -1.04841 | 0.08978 |
| 26 | 0.78995 | -0.40866 | 0.06831 |
| 27 | 0.53940 | -0.47713 | 0.09702 |
| 28 | 0.95103 | 0.00784 | 0.09733 |
| 29 | 0.56790 | -0.50932 | 0.08655 |
| 30 | 0.38542 | -0.35350 | 0.09635 |
| 31 | 0.35818 | 0.30757 | 0.10649 |
| 32 | 1.19639 | -0.82351 | 0.06491 |
| 33 | 0.73024 | -0.75740 | 0.10059 |

Table 8

Item Parameter Estimates for Emotional Stability

N = 459

| Item | Discrimination (a) | Difficulty (b) | Pseudo-Guessing(c) |
|---|---|---|---|
| 1 | 0.72530 | -0.57827 | 0.22158 |
| 2 | 1.73291 | -0.45393 | 0.21156 |
| 3 | 0.81327 | -1.49531 | 0.27514 |
| 4 | 0.77408 | -0.68061 | 0.21574 |
| 5 | 0.94235 | -0.65739 | 0.21503 |
| 6 | 0.71629 | -1.55734 | 0.25338 |
| 7 | 1.06348 | -1.27431 | 0.22561 |
| 8 | 1.01375 | -0.54383 | 0.32154 |
| 9 | 0.99407 | -0.65697 | 0.33699 |
| 10 | 0.98216 | 0.71929 | 0.35121 |
| 11 | 0.81085 | 0.02989 | 0.30057 |
| 12 | 1.62766 | 1.52983 | 0.37626 |
| 13 | 0.78609 | -0.53002 | 0.22516 |
| 14 | 1.69144 | -0.38979 | 0.18258 |
| 15 | 0.66689 | 0.35431 | 0.25883 |
| 16 | 1.18680 | 0.88659 | 0.32643 |
| 17 | 1.52702 | 1.29628 | 0.33370 |
| 18 | 1.02213 | -1.28821 | 0.31731 |
| 19 | 1.37136 | -0.34584 | 0.26208 |
| 20 | 1.16661 | -0.85487 | 0.26178 |
| 21 | 0.84932 | -1.12336 | 0.30384 |
| 22 | 0.76423 | -1.64711 | 0.26790 |
| 23 | 0.81265 | -1.62099 | 0.26184 |
| 24 | 0.84154 | -1.05303 | 0.28707 |
| 25 | 1.78148 | -0.05578 | 0.25019 |
| 26 | 0.80685 | -0.22402 | 0.33062 |
| 27 | 0.79068 | 1.98736 | 0.32769 |
| 28 | 1.09544 | 1.79264 | 0.42170 |
| 29 | 0.94781 | -1.32331 | 0.27642 |
| 30 | 0.81188 | -0.13008 | 0.19422 |

Table 9

Item Parameter Estimates for Openness to Experience

N = 459

| Item | Discrimination (a) | Difficulty (b) | Pseudo-Guessing(c) |
|---|---|---|---|
| 1 | 1.34197 | 0.15237 | .000722 |
| 2 | 1.65069 | 0.25057 | .000382 |
| 3 | 0.90239 | 0.75972 | .000438 |
| 4 | 1.47363 | 0.39485 | .000336 |
| 5 | 0.80167 | 0.56451 | .000842 |
| 6 | 1.01450 | 0.24397 | .000396 |
| 7 | 0.35776 | -0.51599 | .001068 |
| 8 | 1.22111 | 0.71618 | .000224 |
| 9 | 0.54258 | -1.12899 | .000985 |
| 10 | 0.39239 | -1.08803 | .001127 |
| 11 | 0.93038 | 0.45220 | .000429 |
| 12 | 0.38659 | -0.63023 | .001221 |
| 13 | 0.50095 | 1.88106 | .001533 |
| 14 | 0.51562 | 0.56552 | .000751 |
| 15 | 0.46709 | 0.75282 | .002129 |
| 16 | 0.83590 | 0.95400 | .000556 |

Table 10

Item Parameter Estimates for Surgency

N = 459

| Item | Discrimination (a) | Difficulty (b) | Pseudo-Guessing(c) |
|---|---|---|---|
| 1 | 1.11900 | -0.75635 | 0.02687 |
| 2 | 0.72951 | -1.01903 | 0.04228 |
| 3 | 0.69084 | 0.58643 | 0.02390 |
| 4 | 0.94851 | -0.82318 | 0.03903 |
| 5 | 0.87150 | -0.29976 | 0.03485 |
| 6 | 1.24334 | -0.34953 | 0.02354 |
| 7 | 0.51860 | 1.75500 | 0.01882 |
| 8 | 0.40762 | 1.05924 | 0.04564 |
| 9 | 0.66700 | -1.31964 | 0.03406 |
| 10 | 1.00932 | -0.60768 | 0.02993 |
| 11 | 0.54526 | 0.56687 | 0.03167 |
| 12 | 0.82305 | 0.30434 | 0.05241 |
| 13 | 1.19459 | -0.28458 | 0.02455 |
| 14 | 0.26965 | 1.57539 | 0.06583 |
| 15 | 0.66541 | -0.07954 | 0.03414 |
| 16 | 0.42692 | 0.59142 | 0.04514 |
| 17 | 0.88047 | -0.53035 | 0.02358 |
| 18 | 0.93324 | -1.43648 | 0.04741 |
| 19 | 0.45052 | -0.53664 | 0.03825 |
| 20 | 1.04154 | -0.45047 | 0.01936 |
| 21 | 1.08790 | -0.06750 | 0.02144 |
| 22 | 0.30493 | -1.03706 | 0.04794 |
| 23 | 1.68505 | -0.31336 | 0.03816 |
| 24 | 0.47232 | -2.55399 | 0.04460 |
| 25 | 0.79948 | 2.22971 | 0.14679 |
| 26 | 0.37197 | 1.30300 | 0.06865 |
| 27 | 0.57839 | -0.29340 | 0.03453 |
| 28 | 0.59567 | -0.89002 | 0.04832 |

Table 11

Multitrait-Multimethod Matrix for Personality Tests

| | NEO-N | NEO-E | NEO-O | NEO-A | NEO-C | peerE | peerS | peerO | peerA | peerC | PTI-E | PTI-S | PTI-O | PTI-A | PTI-C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NEO-N | [86] | **-35** | 00 | **-34** | **-46** | **-49** | -22 | 06 | -05 | -27 | **-69** | **-21** | 10 | **-22** | **-39** |
| NEO-E | | [77] | -11 | 18 | **27** | 09 | **32** | 02 | 22 | 13 | **31** | **67** | 07 | **26** | **29** |
| NEO-O | | | [73] | 03 | -02 | 02 | 01 | 05 | 06 | -07 | -07 | -07 | **35** | 07 | -08 |
| NEO-A | | | | [68] | **41** | 04 | -16 | -05 | 26 | **30** | **22** | -03 | -15 | **54** | **23** |
| NEO-C | | | | | [81] | 18 | 04 | **-28** | 21 | **61** | **40** | 00 | **-22** | 17 | **80** |
| peerE | | | | | | [85] | **33** | -17 | **47** | **36** | **34** | -06 | -06 | -06 | 10 |
| peerS | | | | | | | [83] | **37** | 21 | -24 | 13 | **44** | **35** | 04 | -02 |
| peerO | | | | | | | | [92] | -05 | **-63** | -01 | 17 | **50** | 21 | **-35** |
| peerA | | | | | | | | | [88] | 27 | 08 | 06 | **-34** | 22 | 15 |
| peerC | | | | | | | | | | [90] | 22 | -23 | **-52** | -07 | **47** |
| PTI-E | | | | | | | | | | | [76] | **33** | -15 | **36** | **48** |
| PTI-S | | | | | | | | | | | | [90] | **22** | **25** | 09 |
| PTI-O | | | | | | | | | | | | | [80] | 09 | **-28** |
| PTI-A | | | | | | | | | | | | | | [69] | 07 |
| PTI-C | | | | | | | | | | | | | | | [84] |

Note. All values are rounded to two decimal places and multiplied by 100.

Correlations significant beyond .05 are **boldfaced**.

Reliability coefficients are in brackets on the diagonal. Sample sizes for reliability coefficients: NEO-FFI (coefficient alpha) = 1,539 (Costa and McCrae, 1992, p. 53), Peer Rating Form (coefficient alpha) = 47, PTI (test-retest) = 64.

Sample size for correlations among tests: NEO-FFI and PTI = 100, NEO-FFI and Peer Rating Form = 46, PTI and Peer Rating Form = 47.

NEO-FFI Neuroticism is the (reverse coded) equivalent of PTI-Emotional Stability.

NEO-FFI Extraversion is the equivalent of PTI-Surgency.

Percent of Items Administered for PTI scales (using the 90% s.e. stopping rule): Emotional Stability = 54%, Surgency = 55%, Openness to Experience = 58%, Agreeableness = 55%, Conscientiousness = 64%.

Table 12

Intercorrelation Matrix for CAT PTI using 90% Stopping Rule and Full Scale Theta Scores

| | | All Items Administered (100%) | | | | |
|---|---|---|---|---|---|---|
| | | Ems | Sur | Opn | Agr | Con |
| Emotional Stability (90% s.e.) | Ems | **98** | **37** | -13 | **33** | **48** |
| Surgency (90% s.e.) | Sur | **36** | **97** | **25** | **28** | 07 |
| Openness to Experience (90% s.e.) | Opn | -15 | 19 | **96** | 14 | **-30** |
| Agreeableness (90% s.e.) | Agr | **41** | **25** | 12 | **94** | 08 |
| Conscientiousness (90% s.e.) | Con | **49** | 11 | **-28** | 10 | **99** |

Note. All values are rounded to two decimal places and multiplied by 100.

N = 101 for all correlations.

Correlations significant beyond .05 p-value are **boldfaced**.

Table 13

Faking Group Mean Scores on PTI Scales

Honest Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 42 | 0.1842857 | 1.2702758 |
| Conscientiousness | 42 | -0.2545238 | 0.8341966 |
| Emotional Stability | 42 | 0.0914286 | 1.2733163 |
| Openness to Experience | 42 | 0.0669048 | 1.4801326 |
| Surgency | 42 | 0.3309524 | 1.0561061 |

Realistic Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 64 | 0.3420313 | 0.8972925 |
| Conscientiousness | 64 | 1.5506250 | 1.3088877 |
| Emotional Stability | 64 | 1.2596875 | 1.2364478 |
| Openness to Experience | 64 | -0.0723437 | 1.1074506 |
| Surgency | 64 | 0.2200000 | 0.9732159 |

Maximal Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 49 | 0.7593878 | 0.9313418 |
| Conscientiousness | 49 | 1.1351020 | 1.2118445 |
| Emotional Stability | 49 | 1.3516327 | 0.9736087 |
| Openness to Experience | 49 | 0.4281633 | 0.9088561 |
| Surgency | 49 | 0.6259184 | 0.7707408 |

Table 14

Percent of Items Administered for Each PTI Scale by Group

Honest Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 42 | 0.5595238 | 0.0998637 |
| Conscientiousness | 42 | 0.6226551 | 0.0419050 |
| Emotional Stability | 42 | 0.5095238 | 0.1134064 |
| Openness to Experience | 42 | 0.5997024 | 0.0956958 |
| Surgency | 42 | 0.5510204 | 0.0688303 |
| All Scales | 42 | 0.5684851 | 0.0408476 |

Realistic Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 64 | 0.5751202 | 0.0797496 |
| Conscientiousness | 64 | 0.5894886 | 0.0643232 |
| Emotional Stability | 64 | 0.4239583 | 0.1378173 |
| Openness to Experience | 64 | 0.5703125 | 0.0908104 |
| Surgency | 64 | 0.5390625 | 0.0557362 |
| All Scales | 64 | 0.5395884 | 0.0428163 |

Maximal Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 49 | 0.5502355 | 0.0733144 |
| Conscientiousness | 49 | 0.6134818 | 0.0632415 |
| Emotional Stability | 49 | 0.4455782 | 0.1466752 |
| Openness to Experience | 49 | 0.5522959 | 0.0629501 |
| Surgency | 49 | 0.5604956 | 0.0526731 |
| All Scales | 49 | 0.5444174 | 0.0365050 |

Table 15

Theta Movement by Faking Group

Honest Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 42 | 0.2590476 | 0.4411947 |
| Conscientiousness | 42 | 0.2345238 | 0.4784487 |
| Emotional Stability | 42 | 0.1652381 | 0.1591170 |
| Openness to Experience | 42 | 0.2269048 | 0.2612959 |
| Surgency | 42 | 0.3152381 | 0.3924174 |

Realistic Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 64 | 0.2685938 | 0.5473163 |
| Conscientiousness | 64 | 0.3081250 | 0.4018682 |
| Emotional Stability | 64 | 0.2153125 | 0.3749157 |
| Openness to Experience | 64 | 0.2282813 | 0.4003076 |
| Surgency | 64 | 0.2670313 | 0.2990816 |

Maximal Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 49 | 0.3904082 | 0.6786689 |
| Conscientiousness | 49 | 0.2589796 | 0.3263066 |
| Emotional Stability | 49 | 0.2028571 | 0.3266560 |
| Openness to Experience | 49 | 0.2330612 | 0.2224317 |
| Surgency | 49 | 0.2291837 | 0.2106818 |

Table 16

Average Absolute Value of Latencies by Faking Group

Honest Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 42 | 1.1842629 | 0.5012028 |
| Conscientiousness | 42 | 1.1070687 | 0.4339774 |
| Emotional Stability | 42 | 1.1570622 | 0.5203134 |
| Openness to Experience | 42 | 1.1898331 | 0.4594270 |
| Surgency | 42 | 1.2004602 | 0.3992052 |

Realistic Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 64 | 1.1980390 | 0.5112191 |
| Conscientiousness | 64 | 1.1267183 | 0.4587203 |
| Emotional Stability | 64 | 1.2311288 | 0.6284950 |
| Openness to Experience | 64 | 1.2548521 | 0.7237492 |
| Surgency | 64 | 1.1323463 | 0.6262202 |

Maximal Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 49 | 1.0354801 | 0.5088065 |
| Conscientiousness | 49 | 1.0119220 | 0.5274624 |
| Emotional Stability | 49 | 1.1030307 | 0.6526759 |
| Openness to Experience | 49 | 1.1471420 | 0.5245631 |
| Surgency | 49 | 1.1042680 | 0.6671852 |

Table 17

Average Signed Latencies by Faking Group for Each Scale

Honest Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 42 | 0.1577924 | 0.9684018 |
| Conscientiousness | 42 | 0.1688837 | 0.7660247 |
| Emotional Stability | 42 | 0.0994696 | 0.8912421 |
| Openness to Experience | 42 | 0.1273068 | 0.8808867 |
| Surgency | 42 | 0.1260000 | 0.8682275 |
| All Scales | 42 | 0.1358905 | 0.7500915 |

Realistic Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 64 | 0.5104616 | 0.9033835 |
| Conscientiousness | 64 | 0.2740418 | 0.8271187 |
| Emotional Stability | 64 | 0.5281597 | 0.9839320 |
| Openness to Experience | 64 | 0.3235104 | 1.0957270 |
| Surgency | 64 | 0.3455239 | 1.0081048 |
| All Scales | 64 | 0.3963395 | 0.8786928 |

Maximal Faking Group

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 49 | 0.3055673 | 0.8704854 |
| Conscientiousness | 49 | 0.1466689 | 0.8604097 |
| Emotional Stability | 49 | 0.4020087 | 1.0040905 |
| Openness to Experience | 49 | 0.3610544 | 0.9192463 |
| Surgency | 49 | 0.2267323 | 1.0223987 |
| All Scales | 49 | 0.2884063 | 0.8597106 |

Table 18

Discriminant Analysis Results

| | Number of Observations and Percent Classified into GROUP: | | |
|---|---|---|---|
| From GROUP | 1 | 2 | Total |
| 1 | 31 | 11 | 42 |
| | 73.81 | 26.19 | 100.00 |
| 2 | 25 | 39 | 64 |
| | 39.06 | 60.94 | 100.00 |
| Total | 56 | 50 | 106 |
| Percent | 52.83 | 47.17 | 100.00 |
| Priors | 0.5000 | 0.5000 | |

Percent Correctly Classified: 67.37% F = 10.4601 df = 2,103 p < .05

Table 19

Revised Multitrait-Multimethod Matrix

|  | NEO-N | NEO-E | NEO-C | peerE | peerS | peerC | PTI-E | PTI-S | PTI-C |
|---|---|---|---|---|---|---|---|---|---|
| NEO-N | [86] | **-35** | **-46** | **-49** | -22 | -27 | **-69** | **-21** | **-39** |
| NEO-E |  | [77] | **27** | 09 | **32** | 13 | **31** | **67** | **29** |
| NEO-C |  |  | [81] | 18 | 04 | **61** | **40** | 00 | **80** |
| peerE |  |  |  | [85] | **33** | **36** | **34** | -06 | 10 |
| peerS |  |  |  |  | [83] | -24 | 13 | **44** | -02 |
| peerC |  |  |  |  |  | [90] | 22 | -23 | **47** |
| PTI-E |  |  |  |  |  |  | [76] | **33** | **48** |
| PTI-S |  |  |  |  |  |  |  | [90] | 09 |
| PTI-C |  |  |  |  |  |  |  |  | [84] |

Note. All values are rounded to two decimal places and multiplied by 100.

Correlations significant beyond .05 are **boldfaced**.

Sample size for correlations among tests: NEO-FFI and PTI = 100, NEO-FFI and Peer Rating Form = 46, PTI and Peer Rating Form = 47.

NEO-FFI Neuroticism is the (reverse coded) equivalent of PTI-Emotional Stability.

NEO-FFI Extraversion is the equivalent of PTI-Surgency.

Percent of Items Administered for PTI scales (using the 90% s.e. stopping rule): Emotional Stability = 54%, Surgency = 55%, Openness to Experience = 58%, Agreeableness = 55%, Conscientiousness = 64%.

Table 20

Mean Latencies for Instructional Items

```
Honest

  N          Mean       Std Dev       Minimum       Maximum
-----------------------------------------------------------
 42     0.3814447     0.8344893    -1.1588635     2.5451763
-----------------------------------------------------------


Realistic Faking

  N          Mean       Std Dev       Minimum       Maximum
-----------------------------------------------------------
 64    -0.0556009     0.8043849    -2.7049752     1.4653119
-----------------------------------------------------------


Maximal Faking

  N          Mean       Std Dev       Minimum       Maximum
-----------------------------------------------------------
 49    -0.0734163     0.8481013    -2.4200000     2.5207218
-----------------------------------------------------------
```

Table 21

Uncorrected Latencies for Each Scale by Group

Honest

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 42 | 0.5392370 | 0.8240360 |
| Conscientiousness | 42 | 0.5503284 | 0.6859905 |
| Emotional Stability | 42 | 0.4809142 | 0.7619007 |
| Openness to Experience | 42 | 0.5087515 | 0.8781732 |
| Surgency | 42 | 0.5074447 | 0.7643657 |

Realistic Faking

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 64 | 0.4548608 | 0.6500862 |
| Conscientiousness | 64 | 0.2184410 | 0.5680849 |
| Emotional Stability | 64 | 0.4725589 | 0.7607034 |
| Openness to Experience | 64 | 0.2679096 | 0.8703482 |
| Surgency | 64 | 0.2899231 | 0.6357794 |

Maximal Faking

| Scale | N | Mean | Std Dev |
|---|---|---|---|
| Agreeableness | 49 | 0.2321511 | 0.6601069 |
| Conscientiousness | 49 | 0.0732526 | 0.5213448 |
| Emotional Stability | 49 | 0.3285925 | 0.6660256 |
| Openness to Experience | 49 | 0.2876382 | 0.9635389 |
| Surgency | 49 | 0.1533160 | 0.6150349 |

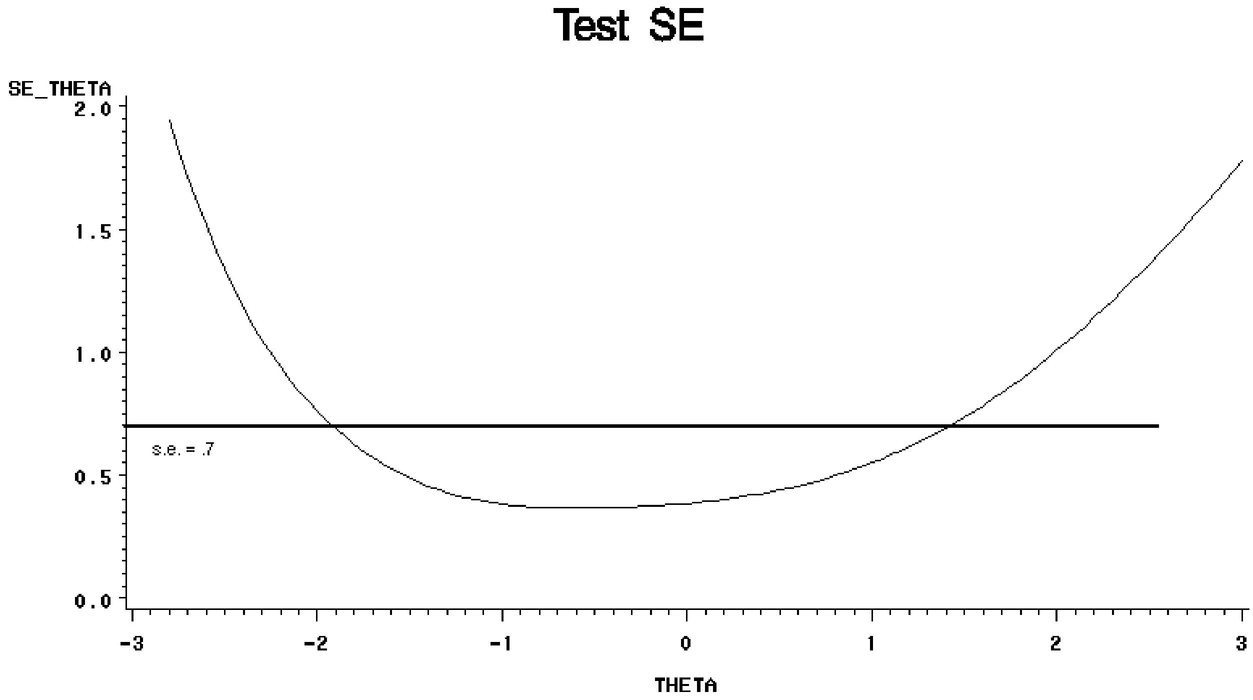

Figure 1

Typical Standard Error Curve with Simple Cut-Off Mark

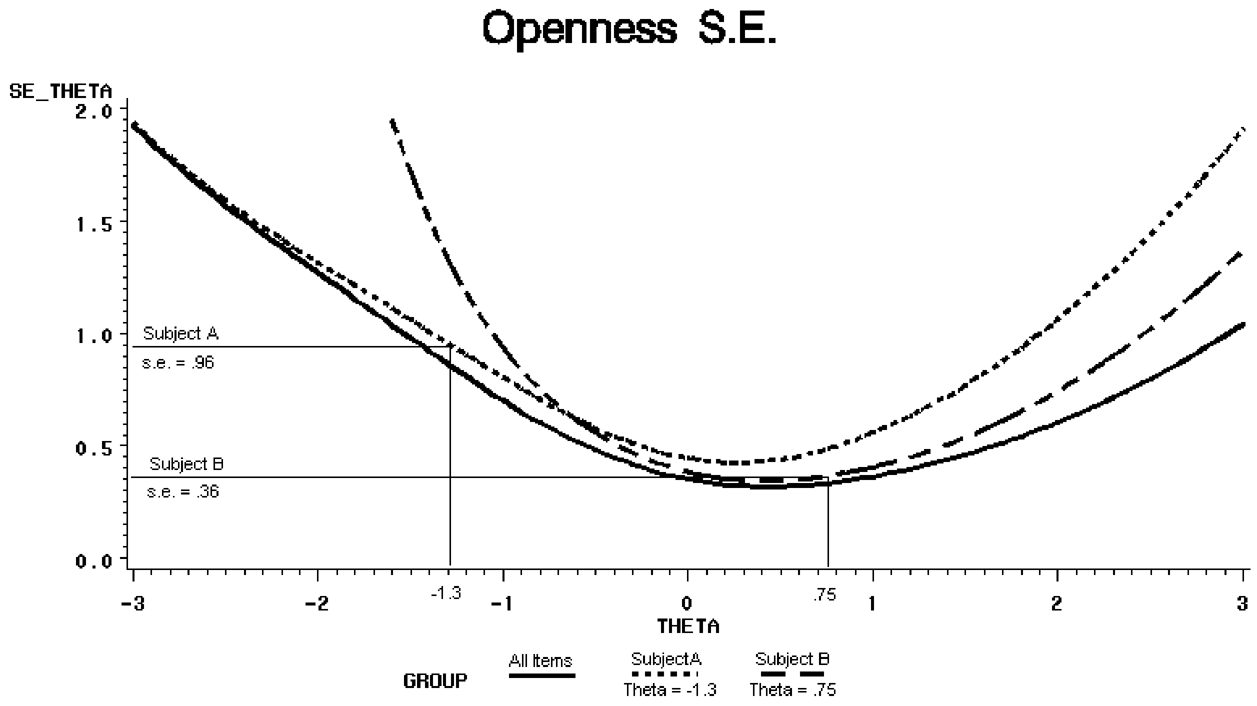

Figure 2

<u>Standard Error Curve for Entire Test and Curves for Two Subjects</u>

<u>Completing a Subset of Test Items in a Computer Adaptive Format</u>

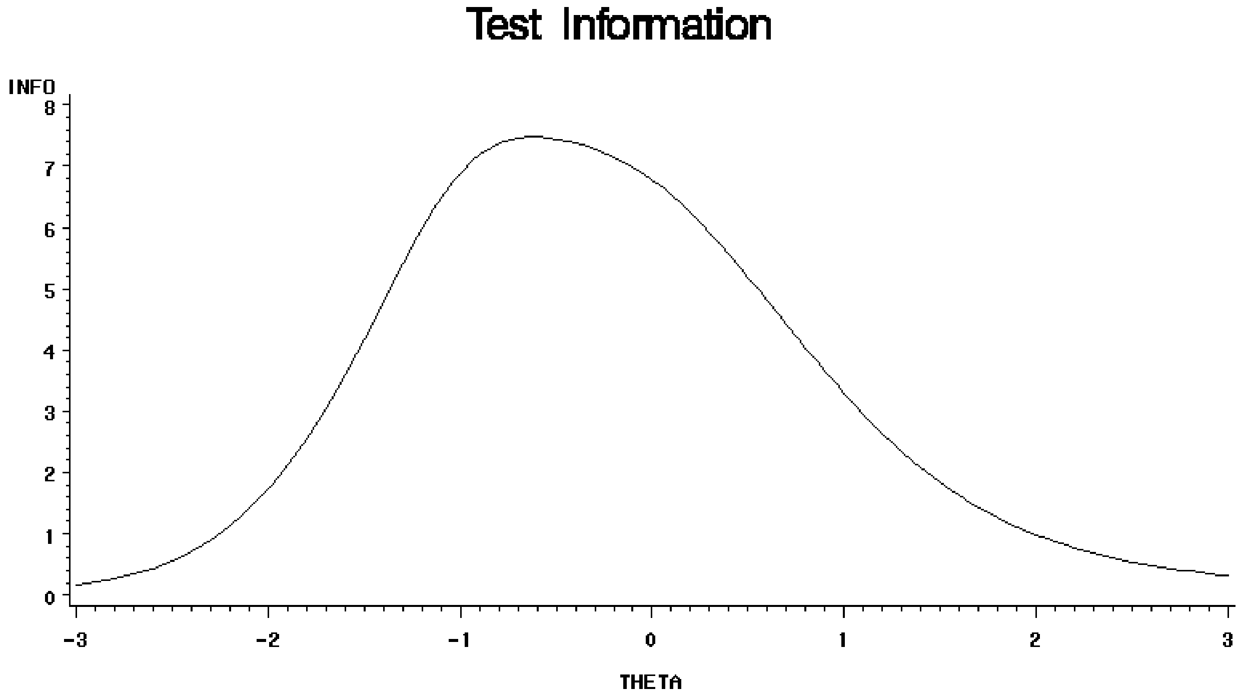

Figure 3

Test Information Function for Agreeableness

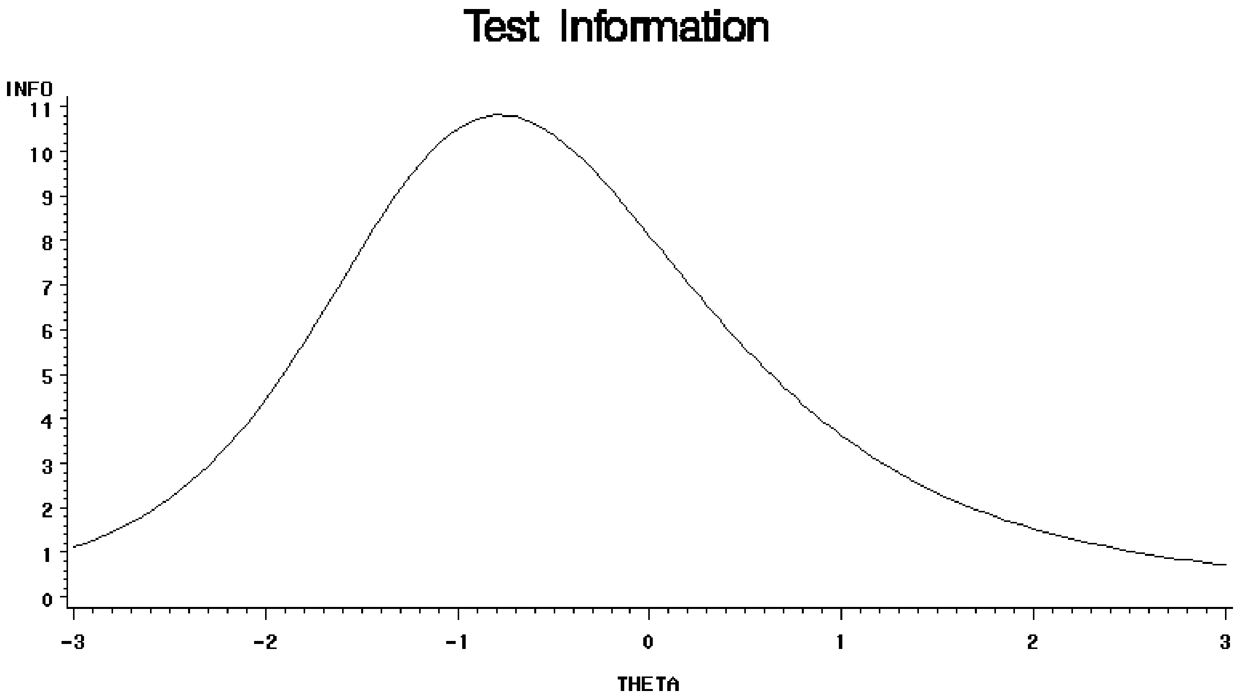

Figure 4

Test Information Function for Conscientiousness

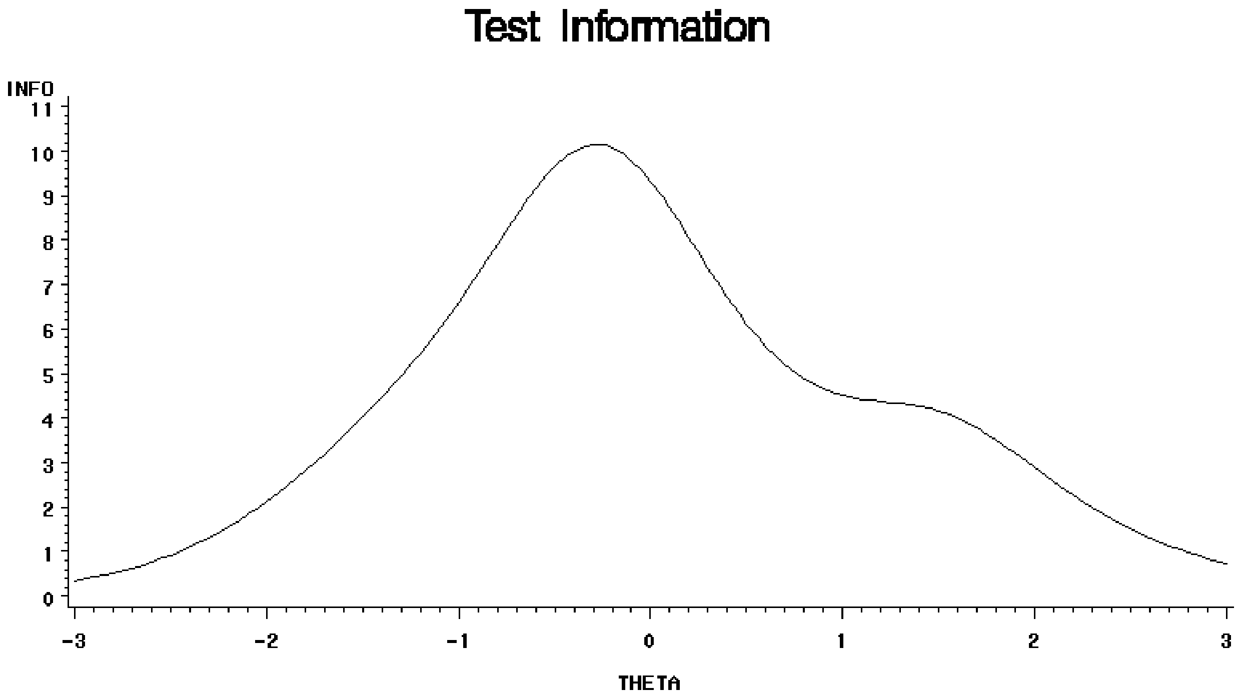

Figure 5

Test Information Function for Emotional Stability

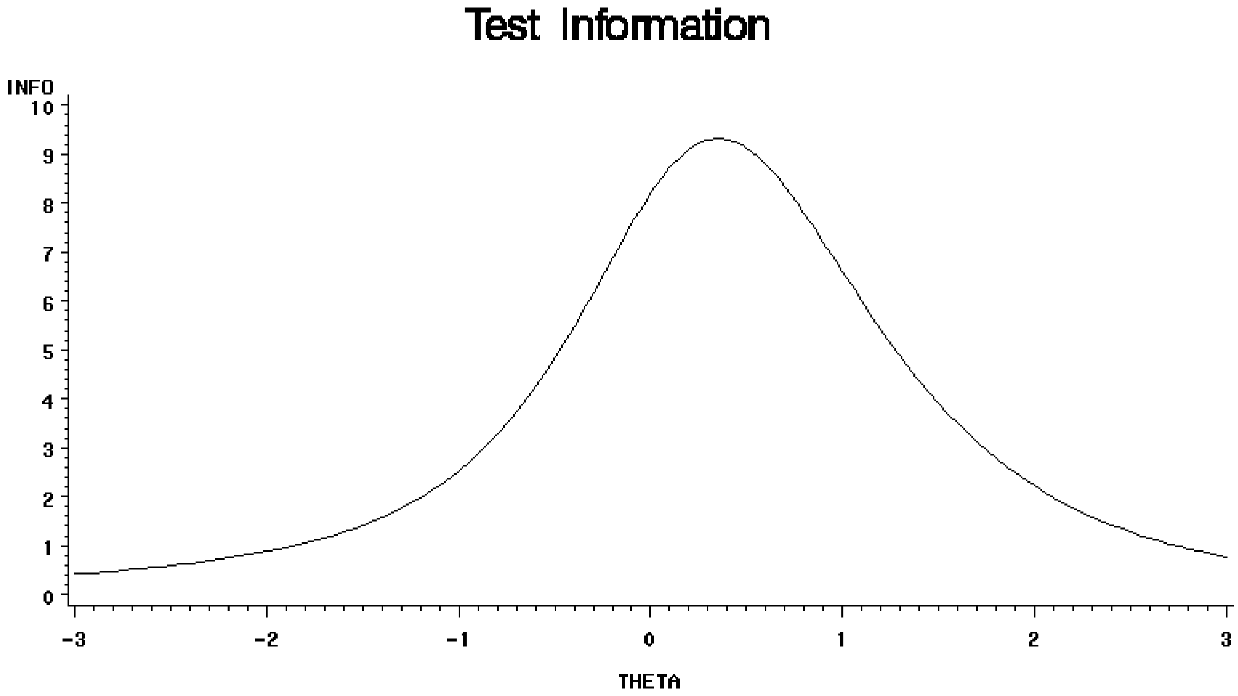

Figure 6

Test Information Function for Openness to Experience

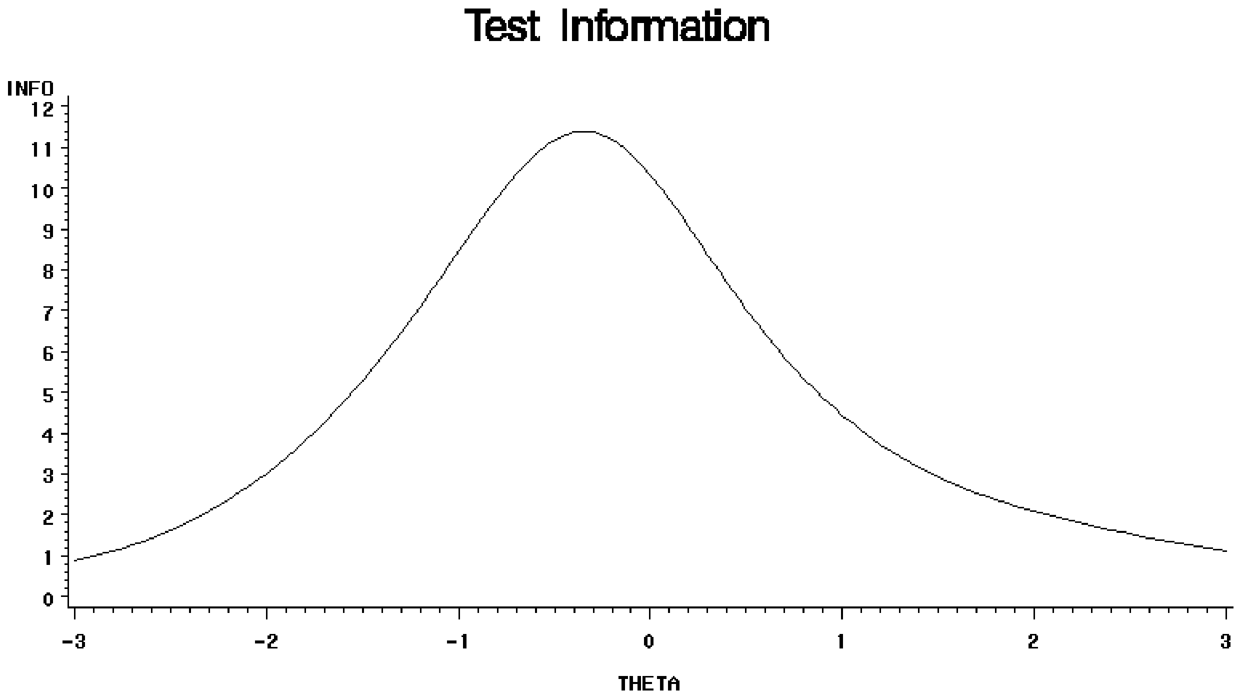

Figure 7

Test Information Function for Surgency

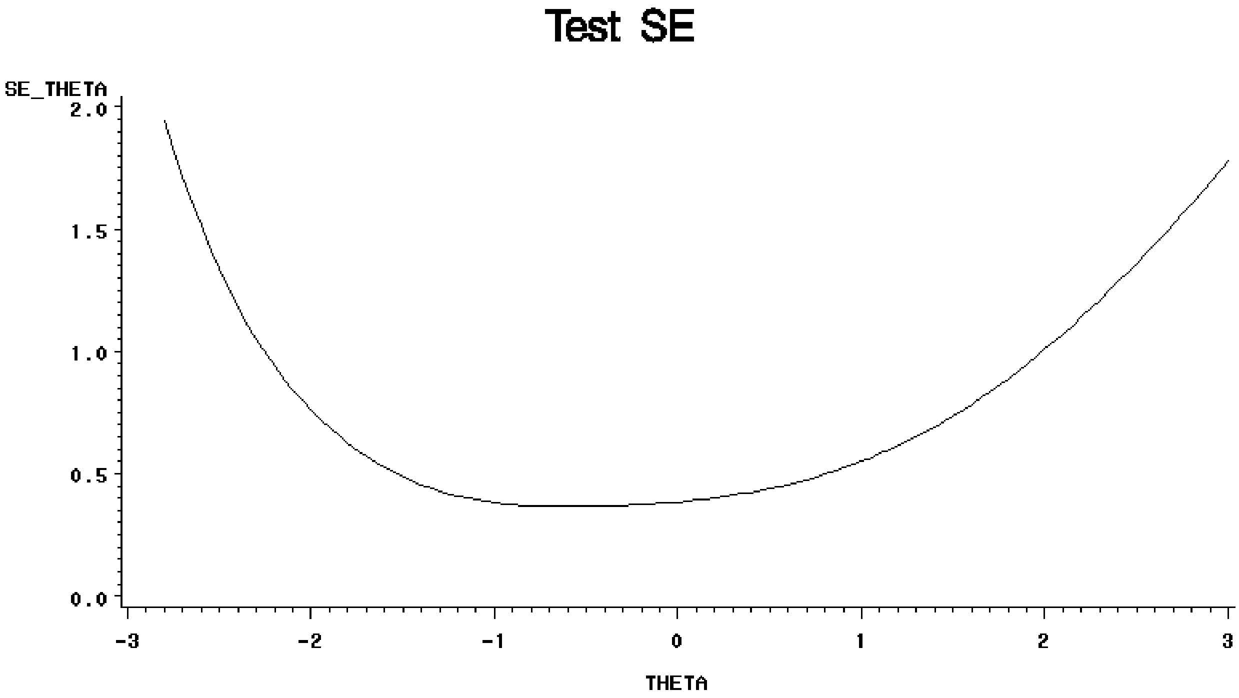

Figure 8

Test Standard Error Function for Agreeableness

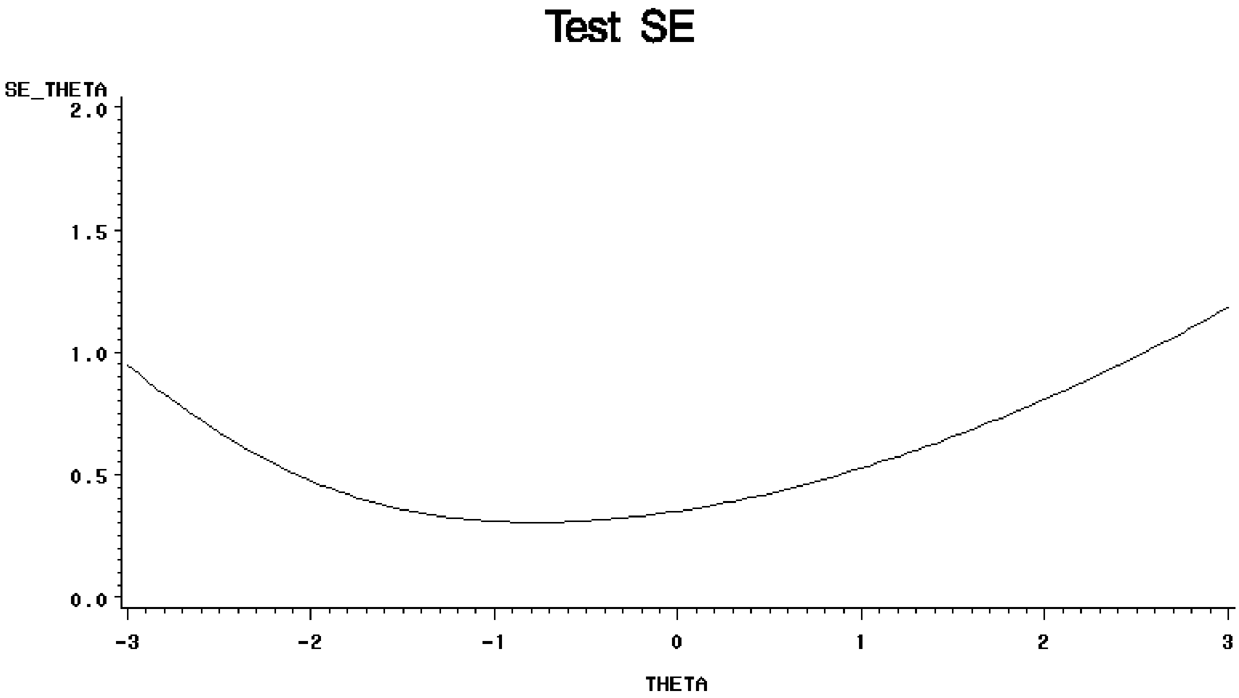

Figure 9

Test Standard Error Function for Conscientiousness

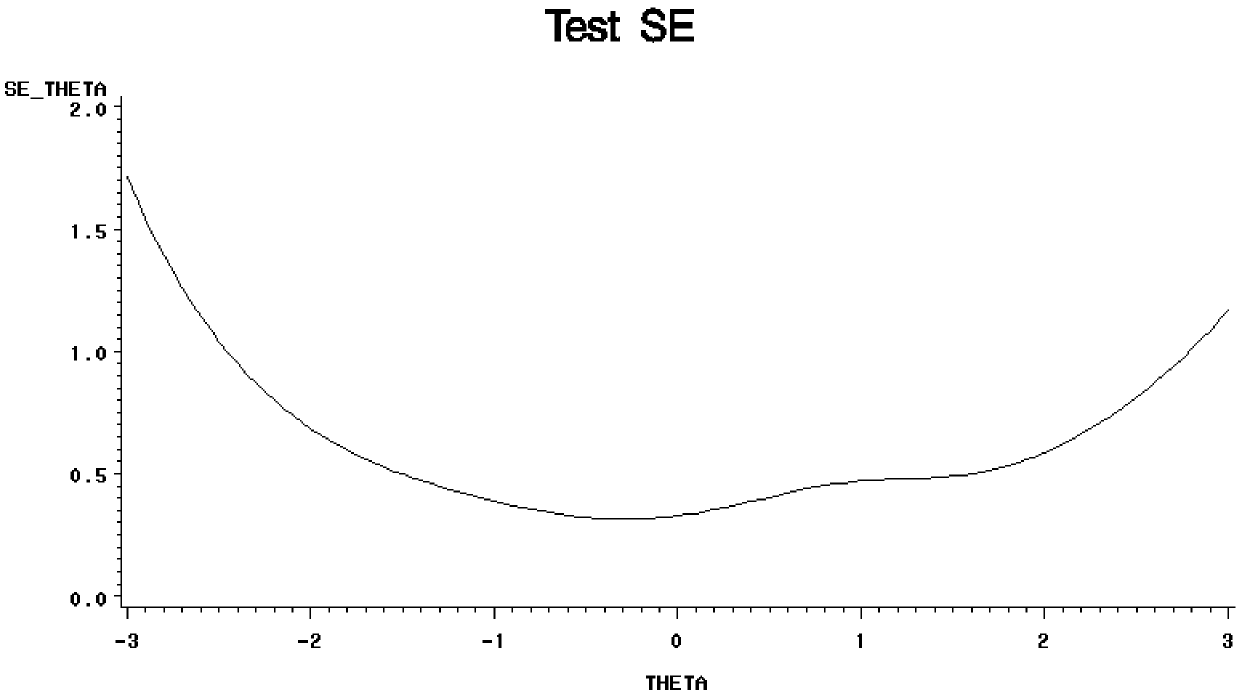

Figure 10

Test Standard Error Function for Emotional Stability

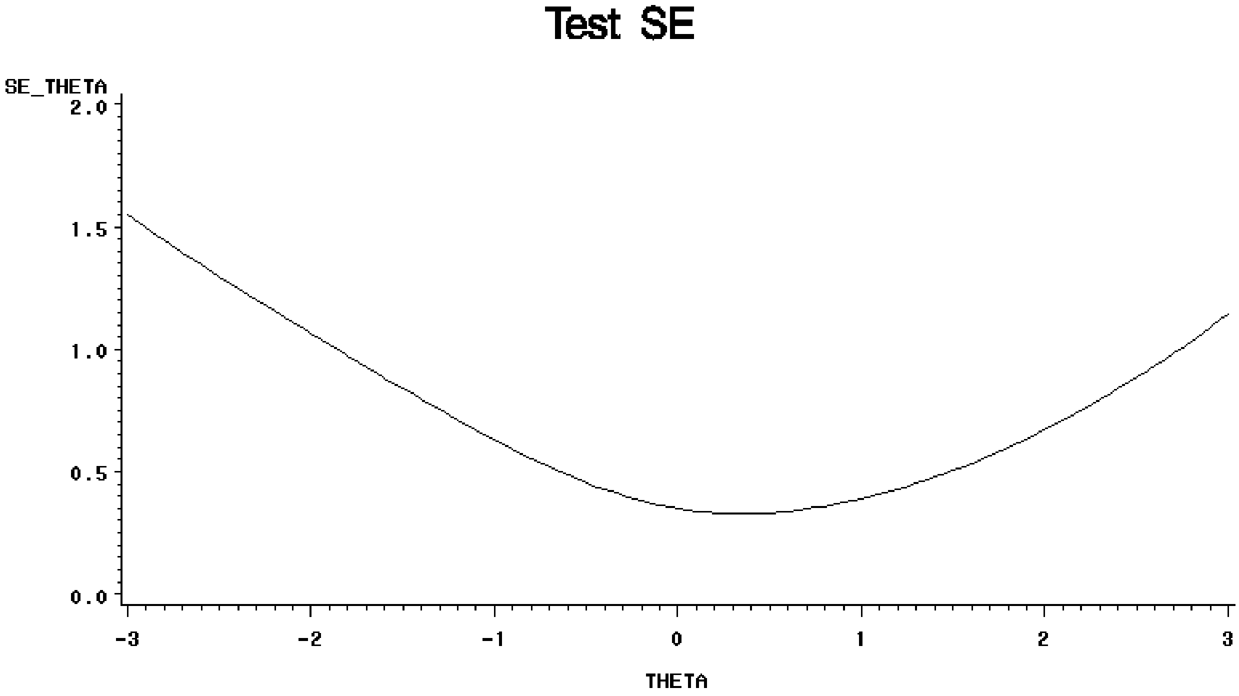

Figure 11

Test Standard Error Function for Openness to Experience

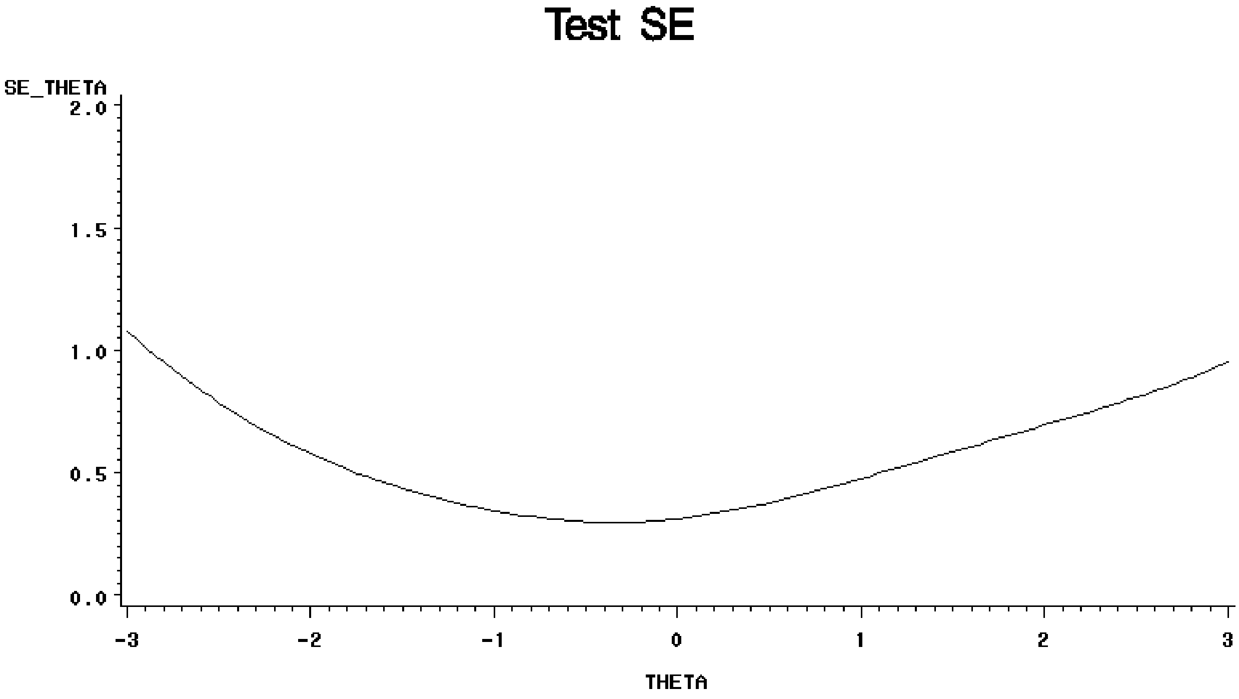

Figure 12

Test Standard Error Function for Surgency

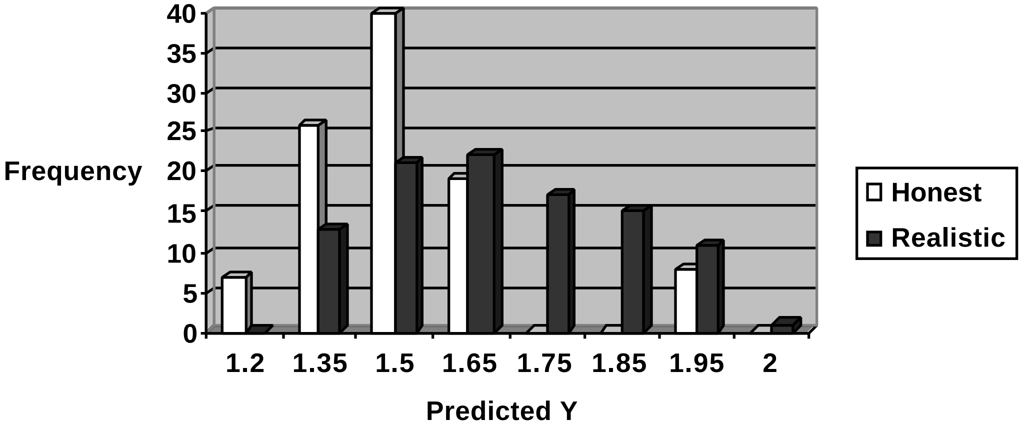

Figure 13

<u>Frequency of Predicted Faking Scores by Group</u>

## REAGAN D. BROWN

Western Kentucky University
1 Big Red Way
Department of Psychology
Bowling Green, KY 42101

## EDUCATION

**Doctor of Philosophy in Psychology**, June 1997
Concentration: Industrial/Organizational Psychology
Virginia Polytechnic Institute and State University (Virginia Tech)
**Dissertation:** The Development of a Computer Adaptive Test Of the Five Factor Model of Personality: Applications and Extensions

**Master of Science Degree in Psychology**, November 1995
Concentration: Industrial/Organizational Psychology
Virginia Polytechnic Institute and State University (Virginia Tech)
**Thesis:** An Examination of the Structure and Predictability of Myers-Briggs Type Indicator Preferences Using a Job Component Validity Strategy Based on the Common-Metric Questionnaire

**Bachelor of Arts Degree in Psychology**; Minor: Math, May 1993
The University of Texas at Austin

## HONORS/AFFILIATIONS

Student Affiliate of the American Psychological Association
Student Affiliate of the Society for Industrial/Organizational Psychology
Dean’s List, Spring 1992, University of Texas
President’s Scholarship Award, University of Texas
Robert Dedmon Scholarship Recipient, 1990-1993
National Merit Scholarship Finalist, 1989

## RESEARCH INTERESTS

· Selection: Validation and Fairness
· Personality Testing
· Validity Generalization
· Psychometrics
· Psychopathy
· Job Analysis
· Synthetic Validity
· Item Response Theory

## PUBLICATIONS/PRESENTATIONS

Brown, R. & Harvey, R.J. (April, 1996). Job-Component Validation using the MBTI and the Common-Metric Questionnaire (CMQ). Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology. San Diego.

Brown, R. (April, 1997). An Investigation of Sex Bias in the Raven Advanced Progressive Matrices. Symposium presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology. St. Louis.

## RELATED EXPERIENCE

***Consulting***

T. H. Hill Associates: 11/95 to 1/97. Hired by a Houston (TX) based oil-field pipe inspection company to improve the current selection system through the use of an instrument designed to assess complex cognitive processing.

***Research***

Thesis Research: Robert J. Harvey, Ph.D. (chair), Virginia Tech

Designed and implemented a study to determine the feasibility of applying the job component validity technique to a personality inventory, the Myers-Briggs Type Indicator.

Other Research: Trained as a rater on the *Revised Psychopathy Checklist* interview, a semi-structured interview to assess psychopathy. Administered and scored interview with undergraduate subjects.

Preliminary Examination Research: Independently designed and executed an examination of sex bias linked to inadvertent measurement of spatial ability in the Raven Advanced Progressive Matrices.

Dissertation Research: Robert J. Harvey, Ph.D. (chair), Virginia Tech

Designed and validated a new computer adaptive measure of the Five Factor Model of Personality. Extensions included the development of a new test-stopping rule as well as new faking indices that focus on the examinee's pattern and style of responding rather than the simple endorsement of validity items.

***Teaching***

Graduate Instructor, Virginia Tech, Blacksburg, Virginia

· Social Psychology: 1/97-5/97. Co-taught an undergraduate section with a faculty member with the material divided according to our primary interests.

· Social Psychology: 8/96-5/97 (two sections). Taught undergraduate course covering the primary topics in Social Psychology.

***Teaching (continued)***

· Industrial/Organizational Psychology: 8/96-12/96. In conjunction with a faculty member, co-taught the senior-level undergraduate course in Industrial/Organizational Psychology. The teaching load was shared equally with the material divided according to our primary interests.

· Industrial/Organizational Psychology: 5/96-7/96. Taught senior-level undergraduate course covering the fundamental topics in Industrial/Organizational Psychology.

Graduate Teaching Assistant, Virginia Tech, Blacksburg, Virginia

· Laboratory for Motivation: 1/96-5/96. Instructed students in fundamentals of technical paper/lab report writing and basic data analysis on SAS system in context of experiments designed to demonstrate basic principles of motivation.

· Laboratory for Psychometrics: 8/95-12/95. Expanded upon material presented in the lecture portion of the course; instructed undergraduates on basic data entry, basic analytical procedures on SAS system, and application of psychometric analytical techniques presented in lecture.

· Laboratory Instructor for Introductory Psychology Course: 8/93-5/94. Engaged students in discussion of psychology-related readings, assigned written homework, administered quizzes, and reviewed tests.