Gender and Ethnicity-Based Differential Item Functioning on the Myers-Briggs Type Indicator

Melissa B. Gratias

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

Master of Science
in
Psychology

Robert J. Harvey, Chair
Roseanne J. Foti
Neil M.A. Hauenstein

May 7, 1997
Blacksburg, Virginia

Keywords: Differential Item Functioning, Item Bias, Item Response Theory

Gender and Ethnicity-Based Differential Item Functioning on the Myers-Briggs Type Indicator

Melissa B. Gratias

(ABSTRACT)

Item Response Theory (IRT) methodologies were employed in order to examine the Myers-Briggs Type Indicator (MBTI) for differential item functioning (DIF) on the basis of crossed gender and ethnicity variables. White males were the reference group, and the focal groups were: black females, black males, and white females. The MBTI was predicted to show DIF in all comparisons. In particular, DIF on the Thinking-Feeling scale was hypothesized especially in the comparisons between white males and black females and between white males and white females. A sample of 10,775 managers who took the MBTI at assessment centers provided the data for the present experiment. The Mantel-Haenszel procedure and an IRT-based area technique were the methods of DIF-detection.

Results showed several biased items on all scales for all comparisons. Ethnicity-based bias was seen in the white male vs. black female and white male vs. black male comparisons. Gender-based bias was seen particularly in the white male vs. white female comparisons. Consequently, the Thinking-Feeling showed the least DIF of all scales across comparisons, and only one of the items differentially scored by gender was found to be biased. Findings indicate that the gender-based differential scoring system is not defensible in managerial samples, and there is a need for further research into the study of differential item functioning with regards to ethnicity.

Acknowledgements

First and foremost, I would like to thank R.J. Harvey who was a wonderful and patient committee chair. I learned more about research in the months we worked together than in any class I could have taken in graduate school. I thank him for being not only an advisor but a teacher in this venture. I thank him for never complaining about the hours we spent in front of his computer, and for showing “concern and respect for me as an individual.” This document would be vacuous if it were not for him. I would also like to express my appreciation to Roseanne Foti and Neil Hauenstein for the time they took to understand what I was trying to accomplish here, and for being an invisible presence every step of the way because I was always trying to anticipate what questions they would ask about everything.

I thank my supportive and loving husband for taking care of the house and of me while this process was underway and for being unselfish and undemanding. I thank my family for believing in me always, and I thank God for the strength He gave me to complete what you now hold in your hands.

Table of Contents

List of Tables

List of Figures

Gender and Ethnicity-Based Differential Item Functioning
on the Myers-Briggs Type Indicator

As society continues to expand into the new information age, the ways in which this information is obtained become increasingly important. Observed mean differences between subgroups on psychological tests which result in consistently lower (or higher) scores for some groups have been a subject of debate in both the empirical and popular literature. Further, movements to keep speech and literature "politically correct" have encouraged gender and ethnicity-neutral terminology that makes the wording of test items even more crucial to overall test performance. The effects that this changing society have had on the field of psychological testing are staggering. Subgroup norming, separate cutoff scores for nonminority and minority applicants, test score banding, and abandonment of certain psychological instruments have been solutions explored in order to equate the observed mean differences between some subgroups (Cascio, 1991).

In the past decade, Item Response Theory (IRT) has emerged as not only a new way to score and interpret psychological tests, but also as an effective method of detecting item bias in these instruments. In the context of Item Response Theory, item bias is referred to as differential item functioning (DIF). When two subgroups (e.g., women and men) are matched with respect to the construct a particular test is measuring (e.g., ability), a bias-free item will show no differences in the item endorsement rates for the subgroups. However, an item exhibiting DIF will perform differently for the subgroups. In other words, when two test-takers who have the same ability but are members of different subgroups (e.g., gender or ethnicity) do <u>not</u> have the same probability of answering an item correctly, the particular item is showing DIF (Hambleton, Swaminathan, & Rogers, 1991).

An important distinction that should be made in the debate of bias in testing is the difference between differential item functioning and impact. Impact refers to actual performance discrepancies that result in mean differences between subgroups where the between-group ability distributions are consistently different (Dorans & Holland, 1992). In addition, similar items will have similar impact. For example, Asian-Americans will typically score higher than Caucasians, and men will score higher than women on a typical Scholastic Aptitude Test (SAT) Mathematics item, and this impact will probably carry over to other math items as well (Dorans & Holland, 1992). By contrast, DIF occurs when subgroups who have been matched on ability still consistently perform differently on an item. In short, impact refers to overall group-mean-based ability differences that affect item performance, whereas DIF is an item-level difference that occurs between examinees of comparable ability levels (Dorans & Holland, 1992). Discovering that there are mean performance differences between subgroups (i.e. impact) does not necessarily mean that the test itself is biased in a DIF sense (Thissen, Steinberg, & Gerrard, 1986).

Another distinction that needs to be made is that between item and test bias. IRT methods are designed to detect item bias. In fact, in a single test, some items may be biased toward a majority group, and other items may be biased toward the minority group. Hence, when item bias is aggregated (into test bias), there may be no effects on subgroup performance because the individual item biases may have canceled each other out (Rudner, Getson, & Knight. 1980). However, detection of item bias (DIF) is important for test construction and revision, and Item Response Theory allows researchers to explore these issues at an item level as well as at a total score level.

Item Response Theory Basics

Assumptions

There are two primary postulates of IRT: (a) Examinee performance on a test item is a function of latent traits, or abilities; and (b) the graphical relation between examinees' latent traits and their probabilities of answering an item correctly is in the form of a monotonically increasing function called an item characteristic curve (ICC). In other words, item performance depends on latent traits (e.g., ability), and as the level of the latent trait increases, the probability of a correct response either increases or stays the same (Hambleton et al., 1991). In IRT models, the underlying latent trait is referred to as theta (θ), which is conceptually similar to a "true score" in Classical Test Theory. The graph of an item characteristic curve has, on its x-axis, θ (expressed typically as a Z-score ranging from -3 to +3), and on its y-axis, the probability of a correct response (PCR).

There are also several assumptions about the data to which IRT models are applied. The first assumption is that of unidimensionality, that one ability (latent trait) is measured by a test. In order for this assumption to be adequately met in an IRT model, a set of test data must consist of a "dominant" factor from which overall test performance results (Hambleton et al., 1991). Local independence, while related to unidimensionality, is the assumption that when ability (the latent trait) is held constant, there should be no relation between examinees' responses to different items (Hambleton et al., 1991). In other words, the underlying latent trait the test purports to measure should be the only factor that has an overall influence on responses to test items, and when that latent trait is statistically controlled, there should be nothing consistently affecting item performance, and thus, the items should be uncorrelated (independent).

When there is an adequate fit between an IRT model and a set of test data, there are several desirable results such as test-free measurement. Test-free measurement implies that the estimates of examinee ability are assumed to be the same even if a different set of items is used (barring measurement errors), and item parameter estimates will be identical for different groups of examinees (except for sampling errors; Hambleton et al., 1991). This property of invariance of item and ability parameters is one of the advantages of IRT models.

IRT Models

The one-parameter logistic model, also known as the Rasch model, explains the relationship (i.e. the ICC) between levels of the latent trait (θ) and probability of a correct response on the item in terms of the difficulty of the item. In IRT terminology, item difficulty is referred to as the b parameter. An item's b parameter (difficulty) is the point on the ability scale corresponding to the location on the S-function item characteristic curve where the probability of a correct response is 0.5 (Hambleton et al., 1991). In other words, b is determined by first locating the point on the ICC that corresponds to a 50% chance of getting the item right (0.5 PCR on the y-axis), and then determining the value of θ (on the x-axis) that corresponds to that point on the ICC. Items that are difficult will have higher b values and will be located at the right or higher end of the θ scale, which indicates that a greater level of ability is required in order to answer them correctly. Conversely, easy items will have lower b parameter values, will stay to the left (lower) end of the θ scale, and will require less ability to answer them correctly (Hambleton et al., 1991). In short, for the one-parameter model, an item is defined by its left-right orientation on the ability (θ) scale (i.e. its difficulty).

The two-parameter logistic model makes use of the b parameter just as in the one-parameter model, but adds an additional element which indicates how well an item separates examinees into different θ levels. The a parameter used in the two-parameter model is called the item discrimination parameter and is equal to the slope of the ICC when it is at its steepest (Hambleton et al., 1991). Not all items are going to be equally discriminating with regards to the latent construct. An item with a very steep slope (large a parameter) is going to separate examinees into ability levels much more effectively than an item with a flatter slope because as you progress from one level of θ to the next, the corresponding PCR will change much more dramatically for the item with the steep slope than for the item with a lesser slope.

The three-parameter logistic model builds upon the two-parameter model by adding a pseudo-chance-level parameter (c). The c parameter is the (possibly nonzero) value of the lower asymptote of the item characteristic curve and is indicative of the probability that an examinee with a very low θ score (e.g., low ability) would answer an item correctly. Put simply, the c parameter is designed to account for performance at the low end of the ability continuum (Hambleton et al., 1991), in particular, modeling the fact that even some very low scoring individuals will get an item correct.

Advantages of IRT over Classical Test Theory

Classical Test Theory (CTT) models state that an examinee's observed score consists of his/her true score plus error. IRT has a similar interest in determining an examinee's true score (latent trait score). However, CTT approaches are limited in that examinee ability is defined in terms of a particular test, and the difficulty of that test is determined by the ability of the examinees who take it. This circularity of item and examinee characteristics in CTT branches into the estimation of reliability and validity as well because the test and item characteristics change as the examinee pool changes. Item Response Theory models, contrary to CTT models, are falsifiable in that they may or may not be appropriate for a particular data set (Hambleton et al., 1991). IRT models do not suffer from the limitations of CTT described above because item and ability parameters are invariant under a linear transformation (i.e., it is possible to change the means and variance estimates for different subgroups so that they lie on the same metric). Estimates of item parameters obtained from different examinee groups will be the same, and estimates of examinee ability do not depend on the pool of items administered (except for sampling or measurement errors; Hambleton et al., 1991).

However, differential item functioning, in hypothesizing that items will perform differently depending on subgroup membership, is essentially predicting that this assumption of invariance of item parameters for different examinees will be violated. If DIF is not found, then the researcher can conclude that item parameters are equal across subgroups, and there is no item-level bias. If DIF is found on a test or scale, then the item parameters are not equal across subgroups and the researcher may decide to exclude the item or use differential scoring by subgroup such as is done on the T-F scale of the MBTI in terms of gender.

In estimating the reliability of a test, CTT not only assumes that reliability is constant across the entire range of observed scores, but that reliability is computed through the use of parallel tests. Parallel tests are very difficult to construct, and reliability coefficients computed in a CTT framework are often replete with unknown biases (Hambleton et al., 1991). In addition, CTT assumes that there is a single standard error of measurement across all possible test scores. IRT does not assume that tests are equally precise across all levels of the latent trait. Standard error functions can be computed across all levels of θ to show how much precision can be expected in estimating test scores at each θ level. In addition, a test information function (TIF) can be computed that shows how much measurement precision is provided by a test across all θ levels rather than a single, global reliability coefficient (Harvey & Thomas, 1996).

Lastly, Classical Test Theory is limited in that it can only provide test level information. There is no consideration of how examinees perform on individual items (other than via statistics such as the item p value). It is sometimes essential to be able to design tests with items targeted toward specific ability levels. IRT models allow a test developer to design items that, for example, discriminate well among high ability examinees (Hambleton et al., 1991). In short, IRT models, because they provide item level information, are far superior to CTT models for many testing applications, especially those that seek to examine the performance of individual test items.

Differential Item Functioning

Definitions

Test items are designed to provide information about the examinee. Difficult items are designed to be more demanding, and easy items are less so. However, sometimes test items carry with them demands other than those intended by the test developer (Scheuneman & Gerritz, 1990). When personal attributes, such as gender or ethnicity systematically affect examinee performance on an item, the result can be differential item functioning. The following is an actual item from the Scholastic Aptitude Test from December 1977 that exhibited gender-based DIF. It is an item from the verbal analogy section of the test,

DECOY : DUCK :: (A) net : butterfly (B) web : spider (C) lure : fish
(D) lasso : rope (E) detour : shortcut.

Knowledge of hunting and fishing jargon is essential in order to answer this item correctly, which is why DIF analyses showed this item was substantially biased against females (Dorans & Holland, 1992). The preceding item is an example of DIF that is indicative of bias because the question was designed to assess verbal reasoning, but gender is a large factor in answering it correctly (e.g., Carlton & Harris, 1992).

The courts have viewed DIF as a mean difference in majority and minority performance on an item (resulting in adverse impact; Hambleton et al., 1991). Psychometricians define DIF more precisely as a situation where individuals who have the same ability, but are members of different subgroups, do not have the same probability of a correct response to an item (Hambleton et al., 1991). Operationally, when the item characteristic curves for two or more subgroups are different, the item is showing DIF (Hambleton et al., 1991). There are several methods for assessing DIF. The two that will be utilized in the present study are the Mantel-Haenszel procedure and an IRT area-based method.

The Mantel-Haenszel Procedure

The Mantel-Haenszel (M-H) procedure was originally used to match subjects retrospectively on cancer risk factors in order to study current cancer rates (Mantel & Haenszel, 1959). The procedure has since been adapted to study differential item functioning and is now the primary DIF detection device used at the Educational Testing Service (ETS; Dorans & Holland, 1992). The M-H method works by first dividing subgroups into the reference group (e.g., males) and the focal group (e.g., females). The focal group is of primary interest in the analysis and is compared to the reference group after being matched on θ (Uttaro & Millsap, 1994). The total test score usually serves as the θ estimate, and the performance (i.e. item endorsement rates) of the reference and focal groups is compared at unit intervals of θ weighted by the number of examinees at each level (Scheuneman & Gerritz, 1990). From this comparison, an odds-ratio estimator can be calculated, and a X^2 test of significance can be carried out to assess the presence of DIF.

To assess the degree of DIF present, the odds-ratio estimator can be transformed onto the ETS "delta metric" (Δ; Dorans & Holland, 1992). The Δ statistic represents the difference in item difficulty for the reference and focal groups after the total score has been taken into account (Scheuneman & Gerritz, 1990). The advantage of using the Δ statistic to classify degree of DIF present is that the ETS has defined the values of it into a classification scheme delineated by Dorans and Holland (1992). A Δ value of 0.0 indicates no DIF, a positive value indicates DIF favoring the focal group (e.g., females), and a negative Δ value reflects DIF that favors the reference group (e.g., males). More specifically, there are three possible degrees of DIF: (a) negligible DIF, where X^2 is nonsignificant or the absolute value of Δ is less than 1.0; (b) intermediate DIF, where X^2 is significant and Δ is between 1.0 and 1.49 in absolute value; and (c) large DIF, where X^2 is significant and the absolute value of Δ is 1.5 or larger (Dorans & Holland, 1992).

The Mantel-Haenszel technique is ideal because it does not rely solely on the X^2 statistic, which can be overly sensitive when large samples are used, which is customary in DIF analyses. The Δ statistic not only complements the X^2 statistic, but also allows assessments of the degree of DIF to be made. The one limitation of the M-H procedure is that it may lack power to detect DIF that is not uniform across the range of θ scores (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994). Uniform DIF, which the M-H technique detects with good accuracy (Swaminathan & Rogers, 1990), occurs when two ICC's differ, but are more or less parallel (Hambleton et al., 1991). Uniform DIF is likely to occur when two ICC's have different b (difficulty) parameters and similar a (discrimination or slope) parameters (Swaminathan & Rogers, 1990). Nonuniform DIF, where the M-H technique is less powerful, occurs when there is an interaction between θ level and subgroup membership (Swaminathan & Rogers, 1990), and the result is that the ICC's for the two subgroups cross at some θ value (Hambleton et al., 1991). Before the crossover point, the item is favoring one subgroup, and after the ICC's cross, the item starts to favor the other group, so the biases could cancel themselves out, and the item shows no net DIF for the M-H technique (Harvey & Greenberg, 1996). Nonuniform DIF is likely to occur when the two ICC's have similar b parameters and different maximum slopes (a parameters; Swaminathan & Rogers, 1990). In order to assess nonuniform as well as uniform DIF, a second DIF statistic, based on the unsigned area between the item characteristic curves (ICCs) will be computed.

Area Between Item Characteristic Curves

Whereas the Mantel-Haenszel procedure falls short in its inability to detect nonuniform DIF, IRT area-based statistics are powerful in detecting nonuniform DIF. Area-based statistics rest on the premise that when an item is unbiased, the ICCs for two subgroups are identical (Rudner et al., 1980), and the area between the curves is zero. However, when an item is biased, the ICCs are not the same, the area between the curves is not zero, and DIF is present (Hambleton et al., 1991).

The most important aspect of area-based statistics is also the most difficult to attain. In order to accurately calculate the area between two item characteristic curves, both curves must be on the same metric, otherwise, observed large areas may be due to scaling differences rather than actual DIF. This problem is referred to as the "linking" problem (Harvey & Greenberg, 1996), and it arises whenever item parameters are estimated using data from two different subgroups (samples) of examinees (e.g., males and females; Stocking & Lord, 1983). Item bias studies using IRT area-based methods will always require subgroup parameters to be linked, and several approaches to doing this have been proposed.

In iterative linking techniques, parameter estimates are linked (placed on the same metric), and item bias statistics are computed initially using all items on the test. Several items may be found to be biased, so in the subsequent iteration, parameter estimates are relinked using only the items found to be unbiased. Item bias statistics are then recomputed for all items (even those found to be biased) using the new parameter estimates. This process continues until the same set of items is flagged as biased on two successive iterations (Drasgow, 1987). Iterative linking techniques have an advantage over non-iterative ones because biased items are not used when parameters are relinked (Drasgow, 1987). Stocking and Lord (1983) have developed equations for linking parameters from subgroups, and Drasgow (1987) derived analogous procedures which agree “surprisingly well” (p. 27) with the more complex techniques of Stocking and Lord.

The numerical procedure for computing the area between two item characteristic curves can be applied after the parameters have been linked. First, the ability (θ) range must be divided into k intervals of width Δθ (e.g., z = 0.01). Then, rectangles centered around the midpoint of each interval are constructed. Third, the values of the ICCs (i.e. the corresponding probabilities of a correct response) at the midpoint of each interval are determined. Next, the absolute value of the difference between the two PCRs is calculated for each interval. Finally, the difference between the PCRs is multiplied (weighted) by the interval width and summed across all levels of θ (Hambleton et al., 1991). Once the area statistics have been computed for each item, a cutoff level can be established, and biased items can be flagged. Visual inspection of the graphs of the ICCs will reveal whether the DIF is uniform (i.e. parallel ICCs) or nonuniform (i.e. ICCs that cross).

The Myers-Briggs Type Indicator

The Myers-Briggs Type Indicator (MBTI) is currently the most widely used personality instrument in non-clinical populations (Myers, 1993). Over three million MBTI’s are administered annually in the United States (Myers, 1993). Its potential uses to industrial and organizational psychologists include, but are not limited to: self-development, career development and exploration, organizational development (OD), team building, problem solving, and management and leadership training (Myers, 1993).

Psychological Type

The MBTI is based on Carl Jung’s theory of psychological type. The central premise of Jung’s theory is that an active mind is always involved in one of two mental activities: perceiving, (i.e. taking in information); or judging, which involves organizing the information and drawing conclusions. Jung further divided these two mental activities. He saw two opposite ways to perceive. One way is by sensing, where one becomes aware of things through the five senses, and the other way is the process of intuition, where perception occurs indirectly through the unconscious. There were also two opposite ways to judge according to Jung. Thinking involves logical processes aimed at an impersonal finding, while feeling is aimed toward appreciation and giving things a subjective value. Jung also observed differences in where people preferred to focus their attention. A focus on the external world of people, things, and experiences is called extraversion, and a focus on the inner world of reflections is introversion (Myers, 1993; Myers & Myers, 1980).

These Jungian processes were incorporated into the MBTI, which consists of four preference scales, each with two opposite poles: Extraversion-Introversion (E-I), where people prefer to focus attention; Thinking-Feeling (T-F), the way people take in information; Judging-Perceiving (J-P), how people prefer to make decisions; and Sensing-iNtuition (S-N), how people choose to orient themselves to the outside world (Myers, 1993). Scores on the four scales of the MBTI are dichotomized at their scale midpoints, and an examinee's categorical "type" is determined by the side of the cutoff where their score on that dimension falls. The personality profile the MBTI gives consists of the four letters that depict the types that corresponded to examinees' scores (e.g., an "ENFP" denotes an Extraverted, iNtuitive, Feeling, Perceiving type).

Gender Issues and the Myers-Briggs Type Indicator

The Thinking-Feeling scale of the Myers-Briggs Type Indicator is the only one of the four scales that maintains a differential scoring system for males and females (Harris & Carskadon, 1988). It has been noted, however, that there are considerable male-female mean differences on the T-F scale despite the differential keying (Harvey & Greenberg, 1996). Differential item functioning means that a man and a woman will not have the same probability of endorsing an item on the T-F scale of the MBTI when they have the same true score on the T-F dimension. Therefore, the act of scoring this item differently for the man and the woman leads to the logical conclusion that the two examinees are not expected to have the same probability of endorsing the item. In utilizing this differential scoring system for some items on the T-F scale, the test publishers are essentially predicting that DIF will be present on this subscale. The rationale for the differential scoring system is that social desirability issues and cultural pressures influence women, even if they have a clear preference for Thinking in their behavior and attitudes, to endorse the Feeling responses disproportionately more often than their male counterparts (Myers & McCaulley, 1985).

As of 1991, two-thirds of men preferred Thinking, and more than six in ten women preferred the Feeling pole of the T-F scale (Hammer & Mitchell, 1996). Also in the Hammer and Mitchell (1996) study, slightly more women preferred Sensing (71%) and Judging (61%) as compared to men (64% and 55%, respectively). The differential male-female scoring weights on the T-F scale changed from Form F to Form G of the MBTI to purportedly keep up with changes in the population, but the new scoring weights have not been well received by some (e.g., Harris & Carskadon, 1988).

McCarley & Carskadon (1986) performed a preliminary study on the perceived accuracy of specific elements of type descriptions among males and females. They not only found few sex differences among the items, but also pointed out that men and women do not differ in how accurately they perceive their type descriptions to be, and they questioned the differential scoring of the T-F scale (McCarley & Carskadon, 1986). Harris & Carskadon (1988) found that the new scoring weights for the T-F scale of Form G were less accurate for males and less likely to predict males' self-reports of type than the old scoring weights used on Form F. This confusion about the relation between gender and the T-F scale has branched into the examination of test-retest reliabilities of the MBTI as well. In 1979, Carskadon found that the test-retest reliabilities on the T-F scale with a seven-week interval were better for females (.87) than males (.48). In 1982, the pattern reversed when Carskadon found that the coefficient was higher for males (.91) than for females (.56) with a five-week interval. The latter study pointed out the need for more research into the interactions of sex and type.

Ethnicity and Age Issues and the Myers-Briggs Type Indicator

The question of relations between type and ethnicity or age has not been extensively addressed in the literature. Hammer & Mitchell (1996) studied the distribution of MBTI types in the population and found that, compared to the total sample of Whites, African Americans, and Hispanics, there were slightly more Introverts, and slightly less Sensors, Thinkers, and Judgers among Whites. The African American sample consisted of more Sensing (81%) and Thinking (70%) types than the total sample, which contained 70% and 53% of these types respectively. The Hispanic sample, which was half male and half female, contained 54% Feeling types and 54% Extraverts (Hammer & Mitchell, 1996).

Hammer and Mitchell (1996) collected data from persons aged 18 to over 80. They divided the sample into children under age 18 and adults age 18 and over. In contrast to the adult sample, there were more Extraverts (58% vs. 46%), more Perceivers (58% vs. 42%), and a closer balance between Sensors (56% vs. 68%) and Intuitives (44% vs. 32%) among the children under 18 (Hammer & Mitchell, 1996). In a study more focused on making age comparisons, Cummings (1995) divided his 15 to 60+ age sample into eight groups in order to make type comparisons. Results from a sample of over 86,000 subjects revealed curvilinear relationships over age groups for Sensing and Thinking types with the greatest numbers of Sensors in the young and old extremes, and the most Thinkers in the middle age ranges (Cummings, 1995). In addition, women tended to become significantly more Extraverted from the youngest to oldest age groups (Cummings, 1995). Although these findings may cast some doubt on the type theory assumption of type invariance throughout life, they may be due to differential willingness to self-report, or even generation effects (Cummings, 1995). Nevertheless, none of the gender, ethnicity, or age differences reported above will affect the present study because Item Response Theory approaches to test scoring involve first matching subjects on the variable of interest before conducting any analyses that compare their performances, such as assessing differential item functioning. Problems that plague studies of mean differences do not affect item-level analyses where subjects can be matched as to their levels of the latent trait.

Type theory (e.g., Myers & Myers, 1980) suggests that one's type is an innate predisposition that can be fostered or hindered by early childhood experiences. Type development begins early in life in a very simple form, and through processes of differentiation and adaption to the environment, one's preferences eventually take hold (Myers & Myers, 1980). If type development is hindered for some reason, true preferences will still emerge as one passes through adulthood (Myers & Myers, 1980). In short, although type is inborn, it may not come to full and complete expression until adulthood, and thus, type measurement in early adulthood may not be as accurate as in later adulthood when preferences have had ample time to crystallize. Although the present study cannot support or refute type theory, its premises may help explain any age-based DIF that is found.

<u>The Myers-Briggs Type Indicator and Item Response Theory</u>

<u>Connecting the terminology.</u> Item Response Theory can be easily applied to the MBTI by translating some of the traditional ability-oriented terminology to that which can apply to personality tests as well (Harvey & Thomas, 1996). The latent construct (θ) will refer to the set of four bipolar constructs the MBTI measures, keeping in mind that the traditional prediction-ratio method of scoring the MBTI is a direct analog of the IRT θ scores (Harvey & Thomas, 1996). Although personality instruments have no "correct" responses, IRT methods only require that a test have a dichotomous scoring system. This system can easily be applied to the MBTI by arbitrarily selecting one of the two poles from each scale to be the *keyed* pole (e.g. I, N, F, and P). In this context the PCR becomes the likelihood that an examinee will respond to an item *in the keyed direction*. When considering the x-axis of the ICC, positive θ values will be associated with I, N, F, or P (the keyed direction), and negative θ values will represent E, S, T, or J (the non-keyed pole). A value of θ = 0.0 will serve as the cutoff point between the two poles. In short, anywhere the word "correct" or "right" appears in the context of IRT, substituting "a response in the keyed direction" makes the discussion applicable to the MBTI (Harvey & Thomas, 1996).

Harvey and Thomas (1996) examined the relationship between MBTI preferences and item endorsement rates and found that not only were these relationships nonlinear, but the items differed in the location on the θ scale where they were maximally informative and the amount of discrimination they provided. The three-parameter model was judged to be the most appropriate representation of the dynamic relationship between type and item responses (Harvey & Thomas, 1996). In the context of the three-parameter model, item difficulty, <u>b</u>, will determine an examinee's tendency to endorse an item in the keyed direction. MBTI items with high b parameters are the ones that the I's, N's, F's, or P's (the keyed pole) will tend to endorse, and items with low b parameters will be endorsed in the keyed direction even by those whose preferences are E, S, T, and J (the non-keyed pole). The a parameter will signify that some items are stronger indicators of an examinee's type than others. Ideally, the b parameter will be located on or near the cutoff point between the dichotomous poles (i.e. at $\theta = 0.0$), and the items would demonstrate large a parameters. The c parameter (pseudo-chance-level) will represent the probability that an examinee located at the non-keyed end of the pole (E, S, T, or J) will endorse the item in the keyed direction (I, N, F, or P) at nontrivial rates (Harvey & Thomas, 1996).

IRT-based research on the MBTI. Seminal work in the area of IRT-based scoring methods and the MBTI was performed by Harvey and Murry (1994). Results from this study painted an ambivalent picture of the MBTI. Although the IRT-based scoring system produced similar results as the traditional MBTI prediction-ratio system, a larger than desirable number of MBTI items had low a (discrimination) parameters, and the standard error was too large for a dichotomous scoring scale to be valid (Harvey & Murry, 1994). However, IRT scoring also revealed that the point at which maximum information could be obtained was located near the scale cutoff, and the bimodal distribution sought by MBTI researchers was obtained, showing that most examinees's scores rested at either end of the poles with few people at the cutoff (Harvey & Murry, 1994). Harvey, Murry, and Markham (1994) found that as the MBTI gets translated into short forms, IRT analyses reveal that there is appreciably less information provided by the test, particularly at the critical type cutoff values. Thomas and Harvey (1995) found that writing new items and using IRT scoring systems substantially improves the scales of the MBTI.

To date, Greenberg (1993) and Harvey and Greenberg (1996) are the only studies assessing DIF in the MBTI. Greenberg (1993) used a pseudo-IRT approach to assessing gender-based DIF on the T-F scale. Results revealed only two items exhibiting gender-based DIF, one on the E-I scale, and another on the J-P scale (Greenberg, 1993). The Harvey and Greenberg (1996) study also assessed gender-based DIF in the MBTI with the T-F scale being the hypothesized DIF target. Mantel-Haenszel and area-based DIF statistics found sizable DIF on all four MBTI scales with the T-F scale showing the least DIF, complementing the Greenberg (1993) study (Harvey & Greenberg, 1996).

The Present Study

The Greenberg (1993) data was a subset of the Harvey and Greenberg (1996) data, and subjects in both studies were college students participating in an experiment for course credit. The present study sought to expand the procedures used in Harvey and

Greenberg (1996) using a large sample of managers who took the MBTI as a part of training and development programs at an assessment center. There were profound age differences between the sample used in this study and that of the Harvey and Greenberg (1996) study, and because type development theory (Myers & Myers, 1980) infers that types should be much clearer in older ages, the sample of younger college students may not have been optimal for assessing psychological type. In addition, measures obtained under more realistic conditions such as in an assessment situation where motivation is presumably higher may be more reliable and valid than measures obtained under pure research conditions (Cascio, 1991). In the literature currently available, the MBTI has not been examined for DIF that is not gender-based. Therefore, the aim of the current study was to examine the MBTI for gender, ethnicity, and age-based differential item functioning among managers. In the end, however, only the gender and ethnicity variables turned out to be appropriate for the analyses.

Based on the literature cited thus far, the hypotheses for this study were as follows. To the extent that DIF was found in any comparison that had a male versus female component, it would be primarily located on the thinking-feeling (T-F) scale, particularly on the items that are differentially scored by gender. Due to the fact that the MBTI had never been previously assessed for ethnicity or age-based DIF, hypotheses for these two variables were largely exploratory. Differential item functioning for ethnicity was predicted, although where it would be located was unknown. Significant age-based DIF was also hypothesized because type development theory (e.g., Myers & Myers, 1980) predicts that younger and older examinees may be differentially inclined to endorse some items, or there might have been different, age-related tendencies to endorse an item or not.

If sizable levels of DIF are found, one of two situations can result. If the DIF consistently favors one group over another in any of the three categories (gender, ethnicity, and age), then this cumulative item bias will result in scale bias (Drasgow, 1987). In this circumstance, the flagged MBTI scale or scales will systematically and possibly inaccurately assign types based not only on the level of θ an examinee has, but also due to personal characteristics such as gender or ethnicity. The other possible result if sizable DIF is found in the MBTI is that the items may be biased in different directions, and when the effects are pooled across scales, the item biases may cancel each other out resulting in no cumulative measurement bias (Drasgow, 1987; Harvey & Greenberg, 1996).

Method

Participants and Instrument

10,775 managers who participated in leadership development workshops at assessment centers worldwide provided the data for the present experiment. In this sample there were 7,012 white males, 3,085 white females, 348 black males, and 330 black females. The original sample contained 13,766 observations. However, there was a nontrivial amount of missing data. For this reason, the sample used in the present study was restricted to data sets where there were five or less missing values on the Form F scored items. This process yielded the sample of 10,775 managers.

Form F of the MBTI was administered and 94 of the 166 items were scored (number 68 will be dropped as it permits more than one response alternative; the remaining items are not scored). In order to use IRT scoring, a value of 1 was keyed to indicate a response in the I, N, F, or P direction, and 0 represented a response toward the E, S, T, or J poles. Traditional MBTI preference scores were not computed because Harvey and Greenberg (1996) found that the interpretations of results were identical whether IRT score estimates or preference scores were used.

Analyses

Analyses were conducted similarly to the procedures used in Harvey and Greenberg (1996). In the IRT analyses, the three-parameter logistic model was fitted to the data, item parameters were estimated with the BILOG (version 3.07; Mislevy & Bock, 1990) program, and the Expected A Posteriori (EAP) scoring method was used within the BILOG program to estimate θ.

Following a recommendation of Dorans and Holland (1992), melting-pot DIF analyses were performed. The usual "marginal DIF analysis" involves assessing gender, then assessing ethnicity separately. Melting-pot DIF recognizes the potential interactions among the variables and allows tests of combinations of variables (e.g., black women vs. white men). DIF may be more prevalent in the smaller subgroups that tend to be ignored by marginal DIF analyses (Dorans & Holland, 1992). In addition, if there is, for example, a within-race interaction on the basis of gender (e.g., white males x white females), then aggregating across race is not warranted.

Breaking the sample down into sex x race x age comparisons allowed a determination of the appropriate focal and reference groups to be made. Having a large enough (e.g., over 300 subjects) sample size in each cell was the important issue. The range of the age variable was severely restricted, and thus it turned out to be untenable as the sample was primarily middle-aged, and a meaningful division into age groups could not be justified (see Table 1). With these considerations, subgroups were formed with the focal groups being black females, black males, and white females, and the reference group being white males.

To assess DIF, the Mantel-Haenszel procedure and an IRT-based area method were used in order to detect both uniform and nonuniform DIF. The M-H technique was implemented as described by Dorans and Holland (1992). First, subgroup members were matched based on their total scores for each scale. The matching variable was the 0.2 z unit divisions of the continuous θ scale, as was done in Harvey and Greenberg (1996). Once it was verified that there were no empty cells along the θ scale in any subgroups, a X^2 statistic was calculated for each item to see if there were any overall differences in item endorsement rates across subgroups. Then, the X^2's p value, as well as the M-H odds-ratio estimator (transformed to lie on the ETS delta [Δ] metric described by Dorans and Holland, 1992), were examined to assess the degree of DIF present. Positive values of Δ favored the focal group (i.e. they were more likely to endorse the item in the keyed direction), and negative Δ values favored the reference group. The degree of DIF was determined by the level of the ETS classification system at which the specific Δ value fell, as described previously.

In order to detect nonuniform DIF, a second statistic was calculated based on the unsigned area between the item characteristic curves obtained from the different subgroups (e.g., Hambleton et al., 1991). Once again, the θ metric was divided into intervals to estimate the area between the ICCs, but at a smaller size of z = 0.01. The parameters were linked using techniques employed in Drasgow (1987) based on the Stocking and Lord (1983) equations. The unweighted distributions of item difficulties were used to derive the linking coefficients that transformed the focal groups' a and b parameters onto the reference group's θ metric (Drasgow, 1987). An iterative technique was used until the list of biased items had stabilized. Because there are no specified significance tests or rules-of-thumb for the IRT area-based statistic as for the M-H, a rough cutoff value was calculated in the iterative technique to identify biased items. This cutoff value for flagging biased items was developed by randomly selecting subsets of approximately the same size of the BF and BM subgroups (n = 348) from the white male subgroup. Then, ICCs were computed for each of the four WM subsets, and areas between these ICCs were calculated. Because all four subsets came from the same subgroup (WM), any area differences would be due to sampling error. The cutoff values for each scale in the area iterations were determined by selecting the largest area found between two WM subsets for each scale

The second iteration began at this point. The items on each scale flagged as biased were removed when the item parameters were relinked so that the biased items would not affect the linking coefficients. After the new linking coefficients were derived, new area statistics were computed. Once the iterations were complete, the final areas were determined. The cutoff values used for the area statistics at this stage of analysis were based on the items flagged by the M-H procedure as well as on prior research (Harvey

& Greenberg, 1996; Uttaro & Millsap, 1994). Interpreting the area-based statistic in this manner was slightly more subjective than with the M-H procedure. Items were sorted in descending order in terms of the area between the ICCs. Next, items were identified that had large area differences but small M-H values, which is indicative of non-uniform DIF.

To assess the cumulative measurement bias in the scales themselves, the approach used in Drasgow (1987) was followed. Test-characteristic curves (TCCs), which represent the expected item endorsement rate as a function of the latent trait, were computed by aggregating the subgroups' ICC's. The TCCs revealed whether the direction of DIF for the scale items consistently favored one subgroup over another thus producing considerable scale-level bias. The other possibility was that the item-level biases canceled themselves out producing very little scale-level bias (e.g., Drasgow, 1987). Although the TCC item endorsement values do not translate in a one-to-one fashion with either traditional MBTI scoring or IRT scoring, the examination of the TCC gives a reasonable index of the amount of cumulative measurement bias in the four MBTI scales (Harvey & Greenberg, 1996).

Lastly, it was not necessary to conduct analyses of the dimensionality of the MBTI in order to justify the use of IRT methodologies. An assumption of IRT is that the MBTI's scales must be unidimensional. A previous study by Harvey, Murry, and Stamoulis (1995) factor analyzed the MBTI and found that the four-factor model of the MBTI is strongly supported as the most plausible representation of the instrument, and the dichotomous nature of the items has no effect on its dimensionality. Therefore, the assumption is met and IRT methodologies were applied to the MBTI.

Results

Several items on various scales were found to contain DIF in the comparisons made. There were four subgroups and three comparisons: white males (WM) versus black females (BF), white males versus black males (BM), and white males versus white females (WF). White males were the reference group and the other three were the focal groups. Results from the Mantel-Haenszel analyses can be seen in Tables 2 through 13.

Item parameters were calculated for the area statistics using BILOG (see Tables 14 - 17). Tables 18 - 33 give the classical item statistics for each subgroup on all four scales of the MBTI. In order to calculate the area between the reference and focal groups' ICCs, the a and b parameters of the focal groups were placed on the same metric as the reference group (see Tables 34 - 37). Equations found in Drasgow (1987) were employed, and two iterations were necessary. After the first iteration, cutoff values were determined for the BF and BM subgroups by using WM subsets as described previously. For the E-I scale, the largest area between any 2 WM subsets was .45, so this was used as the cutoff for the WM-BF and WM-BM comparisons on the E-I scale because for any area below this value, sampling error could not be ruled out as the cause. The cutoff value for S-N was determined to be .35, for T-F, it was .42, and for J-P, the cutoff value for the WM-BF and WM-BM comparisons was .47. For the WM-WF comparisons on all four scales, sample sizes were larger and therefore more stable, so “natural breaks” in the distribution of area statistics were found, and the items above the breaks on each scale were flagged as biased. However, the “cutoff” for the WF comparisons were roughly the same as for the BF and BM subgroups. Because the same items were found to be biased in the second iteration as in the first, no further iterations were necessary, and the area statistics computed in the second iteration were treated as the final DIF results for the IRT area-based method. The final areas were sorted in descending order and compared to the M-H Δ’s (see Tables 38 - 49).

<u>White Males versus Black Females</u>

<u>Mantel-Haenszel analyses.</u> Tables 2 - 5 show the group means, Δ, and X^2 statistic obtained in the WM-BF comparison. According to the ETS criteria, “large” DIF was found on the E-I, S-N, T-F, and J-P scales in 4 (18.2% of the scale items), 3 (11.5%), 2 (8.7%), and 3 (13%) of each scales’ items, respectively. Combining the items that exhibited either “intermediate” or “large” DIF shows that there were 11 (50%), 7 (26.9%), 4 (17.4%), and 6 (26%) biased items on the four MBTI scales in the measurement of white males versus black females. Across all scales, a total of 12 items showed “large” DIF and a total of 16 items showed “intermediate” DIF in the WM-BF comparison.

<u>IRT Area-Based Method.</u> The parameters for item 147 (on the T-F scale) could not be calibrated for the WM-BF comparison because of its low item-total correlation (see Table 27), so no area-based results are available for this item. For the E-I scale, the top 6 items with the highest areas were also flagged as biased by the Mantel-Haenszel (see Table 38), and the range of these six was 0.387 to 0.838 . On the S-N scale, the highest nine areas were above 0.40, with the first three also flagged by the M-H (see Table 39). On the T-F scale the three highest areas are all above 0.50 and the top two were also found biased by the M-H (see Table 40). Lastly, the six highest area items on the J-P scale ranged from 0.254 to 0.652, and all of these items were also flagged by the M-H (see Table 41).

<u>White Males versus Black Males</u>

<u>Mantel-Haenszel analyses.</u> Inspecting Tables 6 - 9 reveals several biased items in the WM-BM comparison. In terms of the E-I, S-N, and T-F scales, there were 3 (13.6%), 2

(7.7%), and 1 (4.3%) items showing "large DIF. No items exhibited "large" DIF on the J-P scale for this comparison. However, all four scales contained items with "intermediate" DIF, and when combined with the "large" DIF items, there were 8 (36.4%), 6 (23.1%), 3 (13%), and 3 (13%) biased items on each scale, respectively. Comparing the response patterns of white males and black males across scales, a total of 6 items contained "large" amounts of DIF, and 14 had "intermediate" DIF.

IRT Area-Based Method. As with item 147 for the WM-BF comparison, item 122 on the T-F scale could not be calibrated for the WM-BM comparison because of a low correlation with the total test score (θ), and no area information will be available for this item as well. The first six items that had the highest area on the E-I scale for this comparison agreed with the Mantel-Haenszel results, and they ranged from 0.345 to 0.446 (see Table 42). On the S-N scale, the top five areas also corresponded with M-H results but ranged from 0.345 to 0.611 (see Table 43). The highest area for this comparison on the T-F scale was 0.680 ("large" M-H Δ as well), and the next highest was 0.362 (see Table 44). The J-P scale appeared like the T-F scale in that the highest area statistic (0.528; "intermediate" M-H Δ) was also considerably larger than the second highest (0.346; see Table 45).

White Males versus White Females

Mantel-Haenszel analyses. Tables 10 - 13 show the results from the M-H analyses and the items found to be biased for each scale of the MBTI. "Large" DIF was seen in 3 (13.6%) E-I items, 2 (7.7%) S-N items, and 3 (13%) J-P items. There were no T-F items exhibiting "large" DIF in this comparison. "Large" and "intermediate" DIF together flags 8 (36.4%) E-I items, 6 (23.1%) S-N items, 1 (4.3%) T-F item, and 8 (34.8%) J-P items. Across scales for the white male versus white female comparison, there are 8 items overall with "large" DIF and 15 items exhibiting "intermediate" DIF.

IRT Area-Based Method. The best agreement between M-H and area in any comparison was on the E-I scale for WM-WF. The eight largest areas were the only items flagged by the Mantel-Haenszel analyses. The top three ranged from an area of 0.458 to 0.661 (all three had "large" M-H Δ as well), and the fourth through the eighth (0.284) were labeled with "intermediate" M-H Δ values on the E-I scale (see Table 46). However, the M-H Δ values were more scattered throughout the hierarchy of areas for the S-N scale. The two M-H Δ values considered "large" by the ETS had area values of 0.471 and 0.273. The top four areas on the S-N scale, though, all were above 0.320 (see Table 47). There was only one item on the T-F scale flagged by the M-H procedure, and it was the second highest area (0.284). The highest WM-WF area for the T-F scale was a low 0.29 (see Table 48). The Mantel-Haenszel procedure flagged eight items in the J-P scale as biased for this comparison, and all of these were scattered throughout the twelve largest area statistics. The top three items were flagged by both procedures and ranged from 0.485 to 0.691 (see Table 49).

Uniform and Non-Uniform DIF

The M-H statistic is primarily sensitive to DIF that is uniform across all levels of θ (e.g., parallel ICCs) and may not be sensitive to non-uniform DIF (Hambleton & Rogers, 1989; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994). It was for this reason that the IRT area statistic was calculated because it is particularly sensitive to non-uniform DIF (e.g., ICCs that cross at some point). When M-H Δ values are small but area statistics are large, non-uniform DIF may be occurring. In these situations, it is necessary to examine the item characteristic curves for the subgroup comparison in question to see if they cross at some point.

By examining the subgroup ICCs for items that have not been flagged by the M-H procedure but whose areas are still within the range of the areas of the items that were flagged by it, it is possible to find non-uniform DIF. For the E-I scale, most of the items containing "large" or "intermediate" DIF by M-H standards had areas above 0.30. However, not all items with areas above 0.30 were flagged by the M-H analyses. Item 77 on the E-I scale is a prime example of uniform DIF (see Figure 1). It had an area of 0.661 for the WM-WF comparison, and an area of 0.838 for the WM-BF comparison. It also had positive, "large" Δ values. Both black and white females endorse the item in the Introverted direction (preferring "theater" over "party") almost twice as often as similar white males, even when the women score strongly on Extraversion (i.e., negative θ scores). However, E-I item 160 is indicative of non-uniform DIF for the WM-BF comparison (see Figure 2). The area statistic is 0.316, but the Δ value is a low -0.12 (and has a nonsignificant X^2). Examining the WM and BF ICCs for this item, it is evident that the ICCs cross near the cutoff score (θ = 0.0) that is used to assign MBTI types. The result is that Extraverted black women tend to endorse this item in the 'I' direction ("People close to me know how I feel only when I tell them.") more often than Extraverted white men, but Introverted white men endorse this item in the 'I' direction more often than Introverted black women. The differences that occur on the Extraverted side of the cutoff cancel out the differences on the Introverted side, resulting in no "net" DIF according to the M-H statistic.

On the S-N scale, item 165 is a good example of uniform DIF (see Figure 3). For the WM-BF comparison, the area is 0.671 and the Δ value is 1.466, and for the WM-BM comparison, the area is 0.503 and the Δ statistic is 1.686. Looking at Figure 3, the BF and BM ICCs are both above the WM ICC across the entire range of θ which indicates that black men and women are more likely to say that they prefer to be "original" (the

‘N’ response) rather than “conventional” (the ‘S’ response) no matter what their type more often than white men. Item 119 on the S-N scale is indicative of non-uniform DIF (see Figure 4). In the WM-BF comparison, the area statistic is 0.502 and the Δ value is 0.728 (“negligible” DIF according to ETS). The WM and BF ICCs cross at θ = -0.8, and before that point, white men (who have relatively strong ‘S’ preferences) respond more often in the ‘N’ (preferring “figurative” over “literal”) direction on this item, and after that point, black women (who are weaker “S’s” or any level of “N’s”) respond more often in the ‘N’ direction than white males.

Item 81 on the T-F scale is an example of uniform DIF for the WM-BF and WM-BM comparisons (see Figure 5). The areas between the subgroup ICCs are 1.225 and 0.68, and the Δ values are 3.279 and 2.197, respectively. The two black subgroups respond consistently more often in the ‘F’ direction (preferring “blessings” over “benefits”) across the θ scale compared to white men. Item 26 on the T-F scale is an example of non-uniform DIF for the WM-BM comparison (see Figure 6). The area is 0.34 and the Δ statistic is a small 0.022 (and it has a nonsignificant X^2). The ICCs cross at θ = 0.3. Before that point, black males are responding in the ‘F’ direction more often (“Values sentiment more than logic.”), and after that point, white males are responding in the ‘F’ direction more often.

Lastly the J-P scale, also shows some items with uniform or non-uniform DIF. Uniform DIF is exemplified by item 74 for the WM-WF comparison (see Figure 7). Across the scale of θ, white women are more likely to endorse the ‘P’ response (preferring “spontaneous” over “systematic”) than white males matched on θ. Non-uniform DIF on the J-P scale is apparent in item 153 for the WM-BM comparison (see Figure 8). The area statistic is 0.242 and the Δ value is 0.613 (and it has a nonsignificant X^2). Figure 8 reveals that the ICCs for these two subgroups cross at θ = -0.3 with white males who are primarily Judgers favored to the left of that point (giving the ‘P’ response of “often forget things until much later”), and with black males giving the ‘P’ response more often to the right of the crossover point.

Scale-Level Bias

Across comparisons, 8 (36.4%) E-I items, 5 (19.2%) S-N items, 2 (8.7%) T-F items, and 5 (21.7%) J-P items were flagged as exhibiting “large” DIF using the Mantel-Haenszel procedure and the ETS classifications. In addition, 15 (68.2%), 12 (46.2%), 7 (30.4), and 12 (52.2%) items on each scale, respectively, exhibited either “intermediate” and/or “large” DIF in one or more comparisons. Tables 50 - 53 illustrate the items that were biased for each scale across the three subgroup comparisons.

It is obvious that many of the individual MBTI items on various scales contain DIF, but the major concern of test developers is the cumulative effects of item-level biases. When administering a paper and pencil test, how much bias is there in the total score? The answer to this question lies in test characteristic curves. Figures 9 - 12 present the TCCs for the four MBTI scales. An inspection of these figures reveals that there is some degree of scale-level bias in each scale. However, most is negligible and located at the extreme ends of the θ scale. For example, the TCC for the E-I scale (see Figure 9) indicates that white males and black females who are extreme Introverts ($\theta = -2.0$) will, on average, differ by about a question and a half. On the S-N scale (see Figure 10), these same two groups will again differ by 1 1/2 questions (i.e., black females endorse 1 1/2 more 'I' responses than the white males at this level). On the T-F scale (see Figure 11), it is evident that white men and white women differ very little in their item endorsement rates, and black women at both extremes of the θ scale only differ from white men by about one and one-half items, which seriously questions the practice of differential scoring by gender. The J-P scale (see Figure 12) shows perhaps the lowest level of scale-level bias of all four scales. The only differences evident are in the $-1 > \theta > 0.5$ range where black women and men endorse about one less item in the 'P' direction on average than white males.

Test Information

To assess the impact of removing the biased items from each MBTI scale on the information the scale can provide, test information functions (TIFs) were calculated for each scale both with and without the items deemed biased by the M-H and area analyses. Figures 13 - 20 show the TIFs and standard errors across all levels of θ both with and without biased items for each MBTI scale. Items were removed from the second calculations of the TIFs and standard error functions by determining which items both the Mantel-Haenszel and IRT area-based methods flagged as containing the most DIF in the comparisons made. On the E-I scale, items 15, 47, 58, 77, and 87 were removed. The two TIFs for this scale (see Figure 13) differ by about one unit of information at the highest peak (around $\theta = 0.2$). The standard error functions for the E-I scale (see Figure 14) are virtually equal around the scale cutoff point ($\theta = 0.0$) but differ as they approach the extreme ends of the θ scale.

Test information and standard error functions were computed for the S-N scale both with all items and with items 98, 107, 112, and 165 removed. The two TIFs (see Figure 15) for this scale differ by about one unit of information at the peak of the curves (around $\theta = 0.0$, the scale cutoff). The two standard error functions (see Figure 16) differ very little across all but the the extreme positive pole of the θ scale. Examining Figure 17 it is evident that the T-F scale is the most platykurtic of all four scales. Both

TIFs peak at θ = 0.8, which is not as desirable as a peak close to the cutoff point of θ = 0.0. However, the removal of items 81, 89, and 122 did not affect the T-F TCCs too much with a difference of only one-half of a unit of information at the peaks. In addition, the standard error functions (see Figure 18) for the T-F scale with and without the biased items differ only trivially across all levels of θ. Lastly, item numbers removed from the J-P scale for these analyses were: 60, 74, 97, and 113. The two TIFs (see Figure 19) peaked at θ = 0.6 and differed there by one and one-half units of information. The standard error functions (see Figure 20) were relatively uniform across the levels of θ.

Discussion

Past research (e.g., Harvey & Greenberg, 1996) using samples of college students found item-level bias on the MBTI, and to some extent, scale-level bias as well. The present study replicated these results in many ways but varied in terms of the use of a large sample of managers in a motivational context and the melting-pot analyses that resulted from crossing gender and race. This methodology was able to delineate the interactions between gender and ethnicity/race that are ignored in marginal DIF analyses where variables are analyzed separately. Hypotheses which predicted gender and ethnicity-based DIF were supported, but hypotheses about more DIF being located on the T-F scale were not.

The Thinking-Feeling scale is the only scale on the MBTI that is still differentially scored by gender. However, examination of the TCCs for this scale (see Figure 11) reveals that white males and females are virtually indistinguishable, and when white males and black females differ, it is not by much and is located in the extremes of the θ scale where fewer people tend to score. In addition, the T-F scale had the <u>fewest</u> biased items as indicated by the Mantel-Haenszel procedure and some of the smallest IRT area statistics of any scale (particularly in the WM-WF comparison; see Table 48). Only one of the seven T-F items flagged by the M-H procedure was one of the items that was differentially scored by gender (see Table 52), and the Δ-value there was <u>negative</u>! This means that white women disproportionately more often endorsed the Thinking (not Feeling) response than their white male counterparts, which is contrary to the reasoning behind the differential scoring system that women will disproportionately endorse Feeling items (e.g., Myers & McCaulley, 1985). This evidence, along with the findings of Harvey and Greenberg (1996) and Greenberg (1993), suggests that the differential scoring system on the T-F scale should be seriously reconsidered if not abandoned. Significant DIF on this scale has not been consistently found in college student samples, and now it is also lacking in larger managerial samples. Future research must determine how generalizable these findings are to other samples.

Another area where this study contributes to the literature is in the evaluation of differential item functioning with regards to ethnicity. There are several caveats to be made on this issue before discussing the results, however. First, the sample sizes of black males and females were minimally acceptable for IRT analyses (n = 348 and 330, respectively). Also, there were few members of these subgroups at the extremes of the θ scale, so the results may be unstable there. In addition, the sample used in the present study was not generalizable to the population of blacks in general. Black managers (and white managers for that matter) may not behave in the same manner as a representative sample of black (white) people. Given the above stipulations, one may conclude from the results presented in this study that there may be a problem with ethnicity-based differential item functioning on the MBTI. Many items on several scales were found to be biased in the WM-BF and in the WM-BM comparisons. More research needs to be done in this area particularly. This study has exposed the first layer of the ethnicity question, but other studies with larger and more representative samples of African-Americans need to continue from this point.

The implications of the fact that the African-American subgroups used in this study are not generalizable depend on the research question being addressed. If a large, representative sample of blacks is obtained in future research, there are a multitude of ways it can be divided in order to explore DIF (e.g., black professionals, unemployed, male, female, etc.). Perhaps as the sample was subdivided again and again, most if not all items could show bias in one or more comparisons. Then, the researcher would be in a quandry on what to do with all the items that showed bias. The answer may lie in not only the research questions but also the particular uses of the test. A possible remedy may be to restrict DIF analyses to subgroups for which the MBTI may realistically be used. However, any questions of external validity are empirical ones, and the goal of the present study was to open the door to DIF research in terms of ethnicity. Future studies may assess the generalizability of ethnicity-based DIF.

To the extent that DIF was found on scales other than T-F, the item-level bias did not result in significant scale-level bias. Although this may not pose problems for the paper-and-pencil administrations of the MBTI, it has great potential to alter scores of computer-adaptive test (CAT) administrations of the instrument. CAT is advantageous in that applicant scores can be estimated with fewer items because the test itself adapts (gives different items) to different examinees based on their responses to earlier test items (Harvey & Greenberg, 1996). However, if an examinee, through a specific pattern of responding, happens to be given a series of biased items, his or her score can be seriously altered at the scale level.

The question of why some of the items on the MBTI were biased and others were not could be the focus of another entire study. However, the answers probably lie in the semantics of the items and the meaning that certain subgroups may attach to them disproportionately more often than others. For example, black and white women (whatever their actual E-I preference) both endorsed the 'I' response of "theater" over "party" more often than did white men. Perhaps women generally picture a romantic trip to the theater and dressing nicely while the white men imagine a boring evening in an uncomfortable tuxedo (whether they are Extraverts or Introverts). Although this is merely speculation, it points to the kind of reasoning that may be required to answer the question of why some items are biased.

The fact that item-level bias was found at all points out that, for the sample used in the present study, the IRT assumption of invariance of item parameters was not met for some of the items; many seemed to be fine. Item parameters are not supposed to be group-dependent (Hambleton et al., 1991). However, DIF analyses by their very natures seek out group differences in item response patterns. Because there were group differences in the probabilities of endorsing items in the keyed directions in the present study, it is evident that this assumption does not hold for all data sets.

Speculating on what can be done about the biased items themselves leads to at least two possible conclusions: rewriting new items to replace the biased ones, or simply eliminating the biased items from the instrument. Although Thomas and Harvey (1985) did not assess the MBTI for DIF, they found that writing new items that paralleled the content of existing low-performing items resulted in a significant increase in the information provided by the test. What is more relevant to the research questions in the present context is: are the new "high performance items" also less biased? If the answer is "yes," then rewriting some of the items on the MBTI represents a viable option for reducing item-level (and to some extent, scale level) bias.

The other option for dealing with the biased items would be to eliminate them from the instrument. Some problems with this approach are that having fewer items can lower reliability, raise standard errors, and alter what the test purports to measure. It was for the purpose of examining this option that the TIFs were calculated. However, not all biased items were removed, just the ones flagged as "most" biased by the two procedures. The fact that the TIFs for the four scales did not change tremendously from the full-scale versions to the reduced versions suggests that the items that are most biased may not be the best items on the scale. This suggests that maybe some items need to be replaced for other reasons than the fact that they exhibit DIF.

The results of the present study and methodology used may have implications for other psychological instruments and other contexts. In Connecticut v. Teal (1982), the

Supreme Court ruled that a "bottom line" lack of adverse impact in a multiple hurdle selection system is not a sufficient defense against disparate treatment in individual selection components. Putting this in IRT terminology, one may conclude that, even if test characteristic curves reveal no cumulative bias, item bias is still a serious consideration. Therefore, test developers should attempt to minimize bias at the item-level so the test users will not be vulnerable to "bottom line" lawsuits. However, this applies to other psychological instruments more than it does to the MBTI as its developers warned against using it in selection contexts (Myers & McCaulley, 1985).

Although the dissenting Justices in Connecticut v. Teal (1982) declared that "Employers need not develop tests that accurately reflect the skills of every individual candidate" (p. 2539), it should be at least a goal of behavioral scientists to develop tests that are as fair as possible. Differential item functioning means that some people, because they are members of a certain subgroup with traditions and cultures that are distinct from other subgroups, will have a disadvantage on some test items before they read the questions. This is not to say that test items will always favor the majority groups over the minorities, nor will all members of a subgroup respond to a question in the same way, but, as the Justices giving the majority opinion in Connecticut v. Teal (1982) stated, "Congress never intended to give an employer license to discriminate on the basis of race or sex merely because he favorably treats other members of the employees' group" (p. 2528). It would seem, then, that whether bias occurs at the item, scale, or total test level, behavioral scientists have an ethical and perhaps legal responsibility to find and reduce it if at all possible.

References

Carlton, S.T., & Harris, A.M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. Princeton, NJ: Educational Testing Service.

Carskadon, T.G. (1979). Test-retest reliabilities of continuous scores on Form G of the Myers-Briggs Type Indicator. Research in Psychological Type, 2, 37.

Carskadon, T.G. (1982). Sex differences in test-retest reliabilities of continuous scores on Form G of the Myers-Briggs Type Indicator. Research in Psychological Type,5, 46-47.

Cascio, W.F. (1991). Applied psychology in personnel management. (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Connecticut v. Teal (1982). 457 U.S. 440.

Cummings, W.H. III (1995). Age group differences and estimated frequencies of the Myers-Briggs Type Indicator preferences. Measurement and Evaluation in Counseling and Development, 28, 69-77.

Dorans, N.J., & Holland, P.W. (1992). DIF detection and description: Mantel-Haenszel and standardization. (RR-92-10). Princeton, NJ: Educational Testing Service.

Drasgow, F. (1987). Study of the measurement bias of two standardized tests. Journal of Applied Psychology, 72, 19-29.

Greenberg, S.E. (1993). Differential item functioning on the Myers-Briggs Type Indicator. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.

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

Hammer, A.L., & Mitchell, W.D. (1996). The distribution of MBTI types in the US by gender and ethnic group. Journal of Psychological Type, 37, 2-15.

Harris, A.H., Carskadon, T.G. (1988). Comparative validity of the old and new scoring weights on the MBTI thinking-feeling scale. Journal of Psychological Type, 15, 54-62.

Harvey, R.J., & Greenberg, S.E. (1996). Gender-based differential item functioning in the Myers-Briggs Type Indicator: Implications for employee selection and big-five inventories. Unpublished manuscript. Virginia Polytechnic Institute and State University.

Harvey, R.J., & Murry, W.D. (1994). Scoring the Myers-Briggs Type Indicator: Empirical comparison of preference score versus latent-trait methods. Journal of Personality Assessment, 62, 116-129.

Harvey, R.J., Murry, W.D., & Markham, S.E. (1994). Evaluation of three short-form versions of the Myers-Briggs Type Indicator. Journal of Personality Assessment, 63, 181-184.

Harvey, R.J., Murry, W.D., & Stamoulis, D.T. (1995). Unresolved issues in the dimensionality of the Myers-Briggs Type Indicator. Educational and Psychological Measurement, 55, 535-544.

Harvey, R.J., & Thomas, L.A. (1996). Using item response theory to score the Myers-Briggs Type Indicator: Rationale and research findings. Journal of Psychological Type, 37, 16-60.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

McCarley, N.G., & Carskadon, T.G. (1986). The perceived accuracy of elements of the 16 type descriptions of Myers and Keirsey among men and women: Which elements are most accurate, should the type descriptions be different for men and women, and do the type descriptions stereotype sensing types? Journal of Psychological Type, 11, 2-29.

Mislevy, R.J., & Bock, R.D. (1990). BILOG 3: Item analysis and test scoring with binary logistic methods [Computer software]. Mooresville, IN: Scientific Software, Inc.

Myers, I.B. (1993). Introduction to type: A guide to understanding your results on the Myers-Briggs Type Indicator (Rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

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

Myers, I.B., & Myers, P.B. (1980). Gifts differing. Palo Alto, CA: Consulting Psychologists Press.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). Biased item detection techniques. Journal of Educational Statistics, 2, 213-233.

Scheuneman, J.D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. Journal of Educational Measurement, 27, 109-131.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thomas, L.A., & Harvey, R.J. (1995, May). Improving the measurement precision of the Myers-Briggs Type Indicator. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Orlando.

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. Applied Psychological Measurement, 18, 15-25.

Table 1
Frequencies of the Age Variable

AGE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
15 - 21	104	0.8	125	0.9
22 - 23	58	0.2	183	1.3
24	39	0.3	222	1.6
25	64	0.5	286	2.1
26	101	0.7	387	2.8
27	128	0.9	515	3.8
28	148	1.1	663	4.9
29	191	1.4	854	6.3
30	254	1.9	1108	8.1
31	311	2.3	1419	10.4
32	393	2.9	1812	13.3
33	505	3.7	2317	17.0
34	530	3.9	2847	20.9
35	573	4.2	3420	25.1
36	502	3.7	3922	28.8
37	587	4.3	4509	33.2
38	623	4.6	5132	37.7
39	627	4.6	5759	42.3
40	634	4.7	6393	47.0
41	584	4.3	6977	51.3
42	553	4.1	7530	55.4
43	618	4.5	8148	59.9
44	594	4.4	8742	64.3
45	546	4.0	9288	68.3
46	532	3.9	9820	72.2
47	581	4.3	10401	76.5
48	577	4.2	10978	80.7
49	465	3.4	11443	84.1
50	386	2.8	11829	87.0
51	354	2.6	12183	89.6
52	299	2.2	12482	91.8
53	267	2.0	12749	93.7
54	187	1.4	12936	95.1
55	158	1.2	13094	96.3
56	122	0.9	13216	97.2
57	100	0.7	13316	97.9
58	86	0.6	13402	98.5
59	68	0.5	13470	99.0
60	30	0.2	13500	99.3
61	29	0.2	13529	99.5
62	16	0.1	13545	99.6
63	9	0.1	13554	99.7
64 - 65	25	0.1	13579	99.8
66 - 68	13	0.0	13592	99.9
69 - 78	8	0.0	13600	100.0

Table 2
Mantel-Haenszel Analyses: E-I White Males vs. Black Females

	Group Means				
Item	WM	BF	Δ	X^2	p-value
6	0.5684226	0.4316109	-1.81452	28.8977	0.00000
15	0.5763173	0.5136778	-0.64686	4.5733	0.03247
19	0.6550443	0.6261398	-0.33730	0.9889	0.32002
25	0.2358153	0.2256098	0.13104	0.0556	0.81352
33	0.4204432	0.3333333	-1.08476	10.3781	0.00128
41	0.6087845	0.5950156	-0.10118	0.0757	0.78319
47	0.2524760	0.4067278	1.90980	45.0901	0.00000
50	0.4360074	0.3666667	-1.54429	8.5548	0.00345
58	0.5579960	0.6595745	1.25536	15.9064	0.00007
66	0.3153283	0.3730887	0.86568	7.9595	0.00478
77	0.4333857	0.6181818	2.03972	51.0117	0.00000
87	0.5412084	0.5181818	-0.17931	0.2399	0.62429
92	0.3104039	0.4072948	1.44500	22.5839	0.00000
95	0.2914229	0.3708207	1.01635	11.9570	0.00054
106	0.4150136	0.4181818	0.17675	0.2990	0.58450
116	0.0982296	0.0848485	-0.06970	0.0004	0.98469
126	0.4770197	0.3951368	-1.07900	8.5348	0.00348
129	0.8361995	0.7477204	-1.30089	16.1321	0.00006
134	0.6056358	0.5440729	-0.72732	4.8570	0.02753
138	0.5627681	0.6443769	1.15989	13.4280	0.00025
148	0.4287567	0.4785276	0.84355	6.7741	0.00925
160	0.3355235	0.3151515	-0.12801	0.1351	0.71317

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 3
Mantel-Haenszel Analyses: S-N White Male vs. Black Female

	Group Means				
Item	WM	BF	Δ	X^2	p-value
2	0.3379937	0.3312883	0.29369	0.6038	0.43713
11	0.6098332	0.6177370	0.04850	0.0097	0.92163
17	0.5029484	0.5424242	0.55932	3.4178	0.06450
37	0.7022912	0.7222222	0.17637	20.6991	0.00001
53	0.6092860	0.5312500	-0.82612	8.2075	0.00417
64	0.8103028	0.7507692	-1.06332	10.5105	0.00119
70	0.6983208	0.6865204	-0.26190	0.5672	0.45136
73	0.8049650	0.8090909	-0.42780	0.8994	0.34295
76	0.4099343	0.3018293	-1.25847	12.9480	0.00032
78	0.4762925	0.5501520	0.94515	9.5498	0.00200
88	0.7554984	0.7469512	-0.57766	2.2543	0.13324
90	0.6128249	0.6534954	0.48763	2.1024	0.14707
98	0.6119275	0.6018237	-0.21566	0.4103	0.52184
102	0.6270896	0.6181818	-0.27313	0.5201	0.47078
104	0.3750536	0.3951368	0.93326	6.5475	0.01050
107	0.7498573	0.8636364	1.93943	19.7616	0.00001
112	0.2727403	0.1380368	-2.07533	24.2055	0.00000
115	0.1790661	0.0820669	-1.86906	13.2135	0.00028
117	0.5241439	0.6189024	1.20006	15.3804	0.00009
119	0.4679817	0.5015291	0.72773	4.7512	0.02928
121	0.7003290	0.6424242	-0.81897	7.4673	0.00628
128	0.4886185	0.4727273	-0.05532	0.0094	0.92282
140	0.8479038	0.8419453	-0.41169	0.9307	0.33467
145	0.5693431	0.5153374	-0.64150	4.5401	0.03311
149	0.5032923	0.4176829	-0.82999	8.2142	0.00416
165	0.5330756	0.6545455	1.46561	22.9337	0.00000

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 4
Mantel-Haenszel Analyses: T-F White Male vs. Black Female

Item	Group Means WM	BF	Δ	X^2	p-value
4	0.6295389	0.6513761	-0.80231	6.102	0.01350
26	0.1805516	0.3436533	1.11757	10.841	0.00099
29	0.4328487	0.4984802	-0.51334	2.446	0.11781
72	0.6662382	0.7545455	-0.54453	1.851	0.17364
79	0.1444682	0.2583587	0.57297	1.802	0.17943
81	0.3876268	0.7446809	3.27872	105.532	0.00000
84	0.4218795	0.4863222	-0.17026	0.285	0.59319
86	0.2400057	0.3799392	0.36530	0.856	0.35493
89	0.5699928	0.7416413	0.96145	6.786	0.00919
91	0.7571000	0.8000000	-0.17175	0.175	0.67541
93	0.4979971	0.4909091	-0.89393	9.638	0.00191
100	0.3698630	0.4559271	-0.18766	0.313	0.57583
103	0.3471865	0.4636364	-0.02271	0.000	0.99667
105	0.2386461	0.3586626	0.41044	1.527	0.21656
108	0.9074286	0.8658537	-1.74689	17.815	0.00002
111	0.4869267	0.5484848	-0.97095	8.782	0.00304
114	0.2866333	0.4030303	-0.36517	0.780	0.37715
120	0.7273247	0.8170732	0.26571	0.376	0.53985
122	0.2502142	0.1424242	-1.94439	25.432	0.00000
133	0.8219708	0.8393939	-0.64195	2.608	0.10635
147	0.0412386	0.0484848	-0.49157	0.434	0.50992
154	0.3001004	0.3588957	-0.56011	3.041	0.08120
158	0.1647817	0.1926606	-0.49208	1.669	0.19638

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 5

Mantel-Haenszel Analyses: J-P White Male vs. Black Female

Item	Group Means WM	BF	Δ	X^2	p-value
1	0.2926515	0.2530488	-0.04344	0.0009	0.97664
9	0.2905251	0.2848485	0.18926	0.2860	0.59277
13	0.4350527	0.4276923	0.39044	1.2453	0.26445
20	0.5695326	0.5613497	0.17794	0.3117	0.57666
27	0.4953485	0.3951368	-0.89910	6.0934	0.01357
35	0.3311911	0.3495441	0.57176	3.1042	0.07809
42	0.6582315	0.5775076	-0.68716	4.4328	0.03525
49	0.1637067	0.1155015	-0.69905	2.0118	0.15608
55	0.4180726	0.3799392	0.10855	0.0582	0.80931
60	0.3271711	0.2060606	-1.51771	15.8678	0.00007
74	0.4590351	0.5151515	1.29960	17.6984	0.00003
85	0.2724153	0.2401216	0.55046	1.2035	0.27262
94	0.1933229	0.1969697	0.51573	1.8650	0.17204
97	0.2822489	0.3757576	1.74341	29.6366	0.00000
99	0.7831360	0.7636364	-0.02405	0.0000	0.99696
109	0.5662479	0.5440729	0.15667	0.2029	0.65242
113	0.4448571	0.2666667	-1.80759	35.0649	0.00000
118	0.3557555	0.3799392	0.80054	6.2077	0.01272
124	0.4523980	0.3425926	-1.00354	10.8638	0.00098
132	0.3571633	0.2522796	-1.14600	9.2427	0.00236
142	0.3728717	0.3981763	0.73859	5.3575	0.02063
151	0.1986262	0.1398176	-0.80265	2.9835	0.08412
153	0.2630376	0.2324159	-0.16514	0.1844	0.66761

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 6

Mantel-Haenszel Analyses: E-I White Males vs. Black Males

	Group Means				
Item	WM	BM	Δ	X^2	p-value
6	0.5684226	0.4482759	-1.64302	27.0293	0.00000
15	0.5763173	0.6734104	1.17511	14.1892	0.00017
19	0.6550443	0.6213873	-0.50427	2.4113	0.12047
25	0.2358153	0.1700288	-1.08289	5.9745	0.01451
33	0.4204432	0.3936782	-0.24485	0.5193	0.47115
41	0.6087845	0.5953079	-0.18820	0.3483	0.55509
47	0.2524760	0.3045977	0.70748	5.5164	0.01884
50	0.4360074	0.3659942	-1.38652	9.1075	0.00255
58	0.5579960	0.6329480	0.79824	7.6523	0.00567
66	0.3153283	0.3786127	0.93525	9.1902	0.00243
77	0.4333857	0.4783862	0.47615	2.8632	0.09063
87	0.5412084	0.6570605	1.65718	23.3245	0.00000
92	0.3104039	0.3850575	1.17345	13.9763	0.00019
95	0.2914229	0.3017241	0.17368	0.2906	0.58983
106	0.4150136	0.4712644	0.73042	5.8896	0.01523
116	0.0982296	0.1066282	0.66729	1.8183	0.17751
126	0.4770197	0.3573487	-1.63102	24.7003	0.00000
129	0.8361995	0.7861272	-0.87370	6.9958	0.00817
134	0.6056358	0.5574713	-0.66184	4.5389	0.03313
138	0.5627681	0.6494253	1.07955	12.0863	0.00051
148	0.4287567	0.4568966	0.49899	2.4094	0.12061
160	0.3355235	0.3706897	0.49323	2.8058	0.09393

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 7

Mantel-Haenszel Analyses: S-N White Male vs. Black Male

Item	Group Means: WM	BM	Δ	X^2	p-value
2	0.3379937	0.3498542	0.87092	5.9120	0.01504
11	0.6098332	0.6408046	0.57195	3.3709	0.06636
17	0.5029484	0.4827586	0.07598	0.0383	0.84484
37	0.7022912	0.6472303	-0.52065	2.8328	0.09235
53	0.6092860	0.5539359	-0.42521	2.0404	0.15316
64	0.8103028	0.8074713	-0.03436	0.0007	0.97865
70	0.6983208	0.6889535	0.04171	0.0043	0.94766
73	0.8049650	0.7902299	-0.28347	0.4664	0.49464
76	0.4099343	0.3275862	-0.62466	2.8148	0.09340
78	0.4762925	0.5231214	0.86906	8.9912	0.00271
88	0.7554984	0.7550432	0.11357	0.0632	0.80150
90	0.6128249	0.6127168	0.28956	0.7918	0.37354
98	0.6119275	0.4827586	-1.38261	20.3712	0.00001
102	0.6270896	0.5344828	-1.02970	8.8418	0.00294
104	0.3750536	0.3275862	0.21407	0.2318	0.63018
107	0.7498573	0.8103448	1.22657	10.2967	0.00133
112	0.2727403	0.1220930	-2.27457	32.2067	0.00000
115	0.1790661	0.1642651	0.24926	0.2627	0.60827
117	0.5241439	0.5319767	0.38336	1.6953	0.19290
119	0.4679817	0.4438040	0.18777	0.3163	0.57382
121	0.7003290	0.7356322	0.52458	2.6903	0.10096
128	0.4886185	0.4710983	0.44986	1.5529	0.21271
140	0.8479038	0.8304598	-0.32302	0.6889	0.40653
145	0.5693431	0.4466859	-1.24669	15.3694	0.00009
149	0.5032923	0.3988439	-0.93593	9.3069	0.00228
165	0.5330756	0.6560694	1.68618	32.5722	0.00000

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 8

Mantel-Haenszel Analyses: T-F White Male vs. Black Male

	Group Means				
Item	WM	BM	Δ	X^2	p-value
4	0.6295389	0.6017442	-0.26301	0.6885	0.40669
26	0.1805516	0.1867816	0.02227	0.0007	0.97833
29	0.4328487	0.3602305	-0.88474	7.6445	0.00569
72	0.6662382	0.6350575	-0.18854	0.2236	0.63634
79	0.1444682	0.1522989	0.06594	0.0015	0.96918
81	0.3876268	0.5718391	2.19689	61.4143	0.00000
84	0.4218795	0.3988439	-0.18534	0.3352	0.56261
86	0.2400057	0.2241379	-0.25517	0.2948	0.58715
89	0.5699928	0.5086207	-0.66858	3.8417	0.04999
91	0.7571000	0.7672414	0.17864	0.2470	0.61920
93	0.4979971	0.4608696	-0.34622	1.5042	0.22003
100	0.3698630	0.3362069	-0.36217	1.1935	0.27462
103	0.3471865	0.3793103	0.60554	3.1658	0.07520
105	0.2386461	0.3054755	1.08717	10.7800	0.00103
108	0.9074286	0.8793103	-0.82015	3.4741	0.06234
111	0.4869267	0.4224138	-0.74639	4.4074	0.03578
114	0.2866333	0.2824207	0.05024	0.0020	0.96393
120	0.7273247	0.7385057	0.26930	0.5299	0.46664
122	0.2502142	0.2247839	-0.33339	1.0116	0.31452
133	0.8219708	0.7385057	-1.28944	16.4781	0.00005
147	0.0412386	0.0402299	-0.14566	0.0049	0.94403
154	0.3001004	0.3573487	0.89903	7.3124	0.00685
158	0.1647817	0.1936416	0.57420	2.2683	0.13205

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 9

Mantel-Haenszel Analyses: J-P White Male vs. Black Male

	Group Means				
Item	WM	BM	Δ	X^2	p-value
1	0.2926515	0.2427746	0.08651	0.0220	0.88196
9	0.2905251	0.2824207	0.35249	1.1032	0.29357
13	0.4350527	0.3901734	0.25795	0.5281	0.46742
20	0.5695326	0.5872093	0.80526	7.1992	0.00729
27	0.4953485	0.4466859	0.32922	0.7658	0.38151
35	0.3311911	0.3074713	0.23193	0.4818	0.48762
42	0.6582315	0.5936599	-0.01505	0.0004	0.98363
49	0.1637067	0.1695402	0.93354	5.2614	0.02180
55	0.4180726	0.3908046	0.61143	3.0850	0.07902
60	0.3271711	0.2890173	0.16379	0.1762	0.67462
74	0.4590351	0.3936782	0.05058	0.0082	0.92805
85	0.2724153	0.1982759	-0.47779	0.7014	0.40231
94	0.1933229	0.1235632	-0.82464	3.4782	0.06218
97	0.2822489	0.2327586	-0.10223	0.0461	0.83007
99	0.7831360	0.8092486	1.03474	7.8049	0.00521
109	0.5662479	0.4438040	-0.78154	6.7674	0.00928
113	0.4448571	0.2959770	-1.25104	18.1229	0.00002
118	0.3557555	0.2873563	-0.21875	0.3691	0.54351
124	0.4523980	0.3976945	-0.22855	0.5985	0.43916
132	0.3571633	0.2420749	-1.00833	6.9690	0.00829
142	0.3728717	0.3189655	-0.04247	0.0048	0.94503
151	0.1986262	0.1321839	-0.71555	2.4069	0.12080
153	0.2630376	0.2716763	0.61264	3.4453	0.06343

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 10
Mantel-Haenszel Analyses: E-I White Males vs. White Females

	Group Means				
Item	WM	WF	Δ	X^2	p-value
6	0.5684226	0.5293351	0.29394	5.198	0.02261
15	0.5763173	0.4044357	-1.54182	173.453	0.00000
19	0.6550443	0.6389703	0.45040	12.629	0.00038
25	0.2358153	0.2337240	0.98983	38.612	0.00000
33	0.4204432	0.3360495	-0.43194	10.591	0.00114
41	0.6087845	0.5475561	-0.27230	5.723	0.01674
47	0.2524760	0.3135011	1.09026	84.750	0.00000
50	0.4360074	0.3298634	-0.80130	17.410	0.00003
58	0.5579960	0.6612111	1.67889	198.948	0.00000
66	0.3153283	0.2106990	-1.02150	57.775	0.00000
77	0.4333857	0.5608964	1.76150	248.174	0.00000
87	0.5412084	0.4180221	-0.93128	57.928	0.00000
92	0.3104039	0.3510880	1.21346	92.807	0.00000
95	0.2914229	0.3652033	1.29464	124.070	0.00000
106	0.4150136	0.3663527	-0.06190	0.247	0.61940
116	0.0982296	0.0743024	-0.20885	0.879	0.34835
126	0.4770197	0.4107722	-0.02280	0.020	0.88765
129	0.8361995	0.7307441	-1.34677	114.681	0.00000
134	0.6056358	0.5598179	0.08422	0.442	0.50633
138	0.5627681	0.5079727	-0.02942	0.047	0.82860
148	0.4287567	0.3283142	-0.69059	28.394	0.00000
160	0.3355235	0.2411440	-0.82818	42.722	0.00000

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 11
Mantel-Haenszel Analyses: S-N White Male vs. White Female

	Group Means				
Item	WM	WF	Δ	X^2	p-value
2	0.3379937	0.3950049	0.29662	4.820	0.02813
11	0.6098332	0.5586117	-1.03621	77.677	0.00000
17	0.5029484	0.5707301	0.42500	13.084	0.00030
37	0.7022912	0.7791310	0.71575	29.431	0.00000
53	0.6092860	0.5493005	-0.98202	75.812	0.00000
64	0.8103028	0.7601840	-1.11781	73.360	0.00000
70	0.6983208	0.7177366	-0.21267	2.615	0.10585
73	0.8049650	0.8349546	-0.10832	0.390	0.53216
76	0.4099343	0.4028590	-0.83521	37.303	0.00000
78	0.4762925	0.4965920	-0.21722	3.294	0.06952
88	0.7554984	0.7766959	-0.42281	7.389	0.00656
90	0.6128249	0.7100130	0.83775	42.633	0.00000
98	0.6119275	0.7545396	1.64603	142.448	0.00000
102	0.6270896	0.7320791	1.08585	56.032	0.00000
104	0.3750536	0.4402209	0.39761	6.755	0.00935
107	0.7498573	0.8126015	0.50692	10.050	0.00152
112	0.2727403	0.3594108	0.82292	44.687	0.00000
115	0.1790661	0.1198441	-1.96418	123.207	0.00000
117	0.5241439	0.6020840	0.53313	20.882	0.00000
119	0.4679817	0.5364583	0.41764	11.114	0.00086
121	0.7003290	0.6393762	-1.10608	90.305	0.00000
128	0.4886185	0.4840702	-0.90851	41.758	0.00000
140	0.8479038	0.8946341	0.66318	15.048	0.00010
145	0.5693431	0.6097163	-0.07288	0.279	0.59711
149	0.5032923	0.4995096	-0.49299	17.548	0.00003
165	0.5330756	0.5666775	-0.02241	0.028	0.86792

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 12
Mantel-Haenszel Analyses: T-F White Male vs. White Female

	Group Means				
Item	WM	WF	Δ	X^2	p-value
4	0.6295389	0.7258649	-0.29918	5.3313	0.02095
26	0.1805516	0.3763618	0.77700	28.1294	0.00000
29	0.4328487	0.5341431	-0.67256	28.6695	0.00000
72	0.6662382	0.8226697	0.20857	1.4724	0.22496
79	0.1444682	0.2968496	0.28728	3.2600	0.07099
81	0.3876268	0.5337443	0.18214	2.3967	0.12159
84	0.4218795	0.5622764	0.27519	5.7279	0.01670
86	0.2400057	0.4609883	0.75544	28.7526	0.00000
89	0.5699928	0.7821685	1.03171	51.7713	0.00000
91	0.7571000	0.8261434	-0.01807	0.0092	0.92346
93	0.4979971	0.6418105	0.42005	13.1418	0.00029
100	0.3698630	0.4792343	-0.44088	13.1612	0.00029
103	0.3471865	0.5081169	-0.17131	1.6888	0.19375
105	0.2386461	0.3151220	-0.72546	29.7506	0.00000
108	0.9074286	0.9322309	-0.14136	0.4177	0.51811
111	0.4869267	0.6482803	-0.22042	2.6291	0.10492
114	0.2866333	0.4776363	-0.18223	1.3515	0.24501
120	0.7273247	0.8249432	-0.05784	0.1330	0.71539
122	0.2502142	0.2128488	-0.87714	45.5712	0.00000
133	0.8219708	0.8804950	-0.09757	0.3103	0.57747
147	0.0412386	0.0623174	-0.40854	2.5253	0.11204
154	0.3001004	0.4341718	-0.19214	2.2139	0.13677
158	0.1647817	0.2173772	-0.53052	13.4297	0.00025

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 13
Mantel-Haenszel Analyses: J-P White Male vs. White Female

	Group Means				
Item	WM	WF	Δ	X^2	p-value
1	0.2926515	0.2587891	-0.93660	40.913	0.000000
9	0.2905251	0.3322518	0.40430	12.051	0.000518
13	0.4350527	0.3839869	-0.93388	56.089	0.000000
20	0.5695326	0.5350066	-0.57750	24.839	0.000001
27	0.4953485	0.4138044	-1.54389	141.571	0.000000
35	0.3311911	0.3598571	0.21807	3.390	0.065612
42	0.6582315	0.5927380	-1.21744	90.210	0.000000
49	0.1637067	0.1334421	-0.95737	31.389	0.000000
55	0.4180726	0.4195964	-0.30120	5.340	0.020845
60	0.3271711	0.2661894	-1.10111	71.185	0.000000
74	0.4590351	0.6426485	2.41329	341.664	0.000000
85	0.2724153	0.3413524	1.13405	49.639	0.000000
94	0.1933229	0.2618430	0.95198	52.097	0.000000
97	0.2822489	0.4232890	1.67084	197.217	0.000000
99	0.7831360	0.7647059	-0.40613	9.063	0.002608
109	0.5662479	0.6480519	0.84783	49.107	0.000000
113	0.4448571	0.4140955	-0.43019	15.609	0.000078
118	0.3557555	0.4793898	1.39531	136.952	0.000000
124	0.4523980	0.3915171	-0.72887	44.272	0.000000
132	0.3571633	0.3427734	-0.46500	12.959	0.000318
142	0.3728717	0.4167481	0.36553	9.501	0.002054
151	0.1986262	0.2255125	0.29432	4.034	0.044591
153	0.2630376	0.1876623	-1.26684	89.676	0.000000

Note. A single underline denotes a Δ value indicative of intermediate DIF. A double underline highlights large DIF.

Table 14
Item Parameters for Subgroups on the E-I Scale

	a-parameter				b-parameter				c-parameter			
Item	WM	BF	BM	WF	WM	BF	BM	WF	WM	BF	BM	WF
6	.881	1.011	.900	.917	-.223	.267	.254	-.135	.021	.015	.047	.003
15	.639	.535	.645	.635	-.313	-.027	-.738	.416	.021	.024	.069	.002
19	.727	.711	.737	.748	-.662	-.503	-.489	-.617	.010	.031	.047	.005
25	1.032	1.260	1.159	1.305	1.012	.988	1.309	.901	.000	.011	.022	.003
33	.893	.806	.787	.953	.306	.713	.517	.583	.001	.014	.048	.001
41	.472	.526	.581	.458	-.622	-.458	-.380	-.303	.004	.029	.064	.006
47	.371	.448	.841	.401	1.877	.642	1.533	1.389	.002	.025	.169	.029
50	1.962	2.152	1.722	2.068	.204	.406	.408	.468	.002	.006	.022	.000
58	.478	.616	.415	.575	-.330	-.737	-.734	-.840	.002	.027	.066	.007
66	.633	.571	.756	.685	.888	.687	.558	1.382	.001	.019	.037	.001
77	.406	.442	.411	.479	.432	-.660	.275	-.369	.001	.032	.054	.005
87	.781	.815	1.082	.761	-.165	-.037	-.544	.310	.001	.022	.044	.001
92	.736	.635	.850	.708	.831	.476	.475	.645	.000	.021	.029	.008
95	.474	.434	.354	.440	1.255	.967	2.011	.812	.001	.041	.079	.004
106	.618	.475	.741	.628	.403	.534	.190	.604	.001	.024	.050	.002
116	.950	.848	.874	.909	1.910	2.223	1.915	2.157	.001	.011	.010	.000
126	.996	1.246	.839	1.184	.093	.372	.630	.299	.005	.016	.040	.017
129	.345	.429	.290	.269	-2.966	-1.613	-2.659	-2.300	.005	.032	.060	.007
134	.769	.836	.718	.715	-.438	-.148	-.194	-.280	.001	.019	.049	.006
138	.714	.582	.606	.779	-.266	-.681	-.686	-.060	.002	.030	.050	.003
148	.834	.777	.867	.901	.283	.166	.311	.686	.001	.038	.080	.019
160	.572	.388	.469	.559	.842	1.439	.916	1.394	.001	.030	.052	.001

Table 15
Item Parameters for Subgroups on the S-N Scale

Item	a-parameter				b-parameter				c-parameter			
	WM	BF	BM	WF	WM	BF	BM	WF	WM	BF	BM	WF
2	.946	1.057	1.210	1.002	.601	.713	.494	.382	.002	.063	.027	.010
11	.590	.570	.664	.720	-.539	-.427	-.588	-.137	.012	.103	.063	.067
17	.605	.610	.862	.844	.116	-.061	.121	-.047	.059	.088	.056	.128
37	.596	.425	.605	.617	-.951	-1.254	-.655	-1.354	.054	.118	.065	.083
53	.559	.409	.725	.635	-.382	.153	-.051	.413	.086	.117	.102	.236
64	.458	.462	.563	.461	-2.083	-1.421	-1.714	-1.552	.020	.112	.065	.068
70	.697	.578	.727	.728	-.891	-.833	-.792	-.914	.025	.101	.061	.059
73	.955	1.191	.823	1.085	-1.258	-1.142	-1.255	-1.340	.030	.088	.060	.042
76	1.088	.882	1.265	1.132	.311	.909	.574	.368	.006	.058	.030	.024
78	.744	.635	.586	.679	.089	-.153	-.038	.027	.002	.067	.055	.014
88	1.079	.984	.800	1.031	-.969	-.845	-1.060	-1.082	.013	.132	.073	.021
90	.848	.776	.685	.729	-.451	-.574	-.434	-.909	.008	.086	.066	.038
98	.861	.837	.758	.968	-.449	-.122	.136	-.973	.005	.169	.057	.048
102	1.076	1.060	1.252	1.141	-.463	-.378	-.092	-.831	.003	.073	.052	.028
104	1.445	1.086	1.424	1.389	.387	.403	.583	.187	.002	.050	.043	.010
107	1.036	.805	.712	1.060	-.967	-1.727	-1.487	-1.238	.007	.103	.065	.033
112	.645	.872	.593	.650	1.117	1.858	2.567	.761	.005	.039	.033	.036
115	.840	1.042	.996	1.058	1.448	2.471	1.501	1.851	.002	.049	.031	.028
117	.658	.851	.501	.605	-.018	-.160	-.010	-.450	.047	.186	.078	.031
119	.938	1.188	.814	1.026	.336	.060	.430	-.041	.104	.079	.102	.062
121	.505	.426	.517	.456	-1.150	-.721	-1.274	-.764	.010	.097	.069	.041
128	1.220	.937	1.209	1.254	.026	.173	.096	.069	.003	.069	.037	.022
140	.597	.582	.447	.633	-2.011	-1.913	-2.252	-2.339	.027	.104	.069	.060
145	.953	.616	.864	1.056	-.266	.221	.212	-.372	.005	.132	.035	.028
149	.765	.602	.747	.758	.028	1.031	.501	.216	.027	.181	.051	.101
165	.647	.570	.559	.750	-.124	-.629	-.738	-.176	.021	.109	.063	.064

Table 16
Item Parameters for Subgroups on the T-F Scale

	a-parameter				b-parameter				c-parameter			
Item	WM	BF	BM	WF	WM	BF	BM	WF	WM	BF	BM	WF
4	.553	.692	.542	.663	-.685	-.567	-.523	-1.076	.007	.079	.019	.012
26	.969	.805	.707	1.046	1.333	.739	1.660	.413	.011	.047	.029	.002
29	.665	.907	.781	.795	.302	.125	.618	-.151	.011	.070	.030	.005
72	1.054	1.008	.896	1.065	-.611	-.838	-.520	-1.297	.005	.119	.021	.005
79	1.038	1.334	1.378	1.021	1.476	.851	1.223	.763	.004	.028	.003	.014
81	.573	.604	.545	.530	.551	-1.094	-.365	-.175	.006	.108	.019	.010
84	.458	.530	.589	.501	.451	.315	.506	-.342	.006	.092	.018	.009
86	1.100	1.220	1.241	1.050	.950	.473	.966	.120	.007	.048	.014	.006
89	.811	.843	1.068	.865	-.297	-.910	-.051	-1.205	.003	.086	.013	.006
91	.454	.485	.335	.433	-1.667	-1.757	-2.222	-2.354	.009	.097	.021	.009
93	.433	.533	.356	.449	.011	.377	.374	-.864	.009	.115	.031	.009
100	.652	1.062	.746	.667	.591	.614	.692	.097	.007	.189	.014	.012
103	.863	.791	1.018	.919	.588	.251	.481	.000	.007	.059	.040	.028
105	.708	.760	.811	.635	1.198	.742	.811	.871	.002	.060	.020	.003
108	.423	.435	.555	.420	-3.468	-2.629	-2.435	-3.982	.014	.097	.020	.011
111	.934	.789	1.163	.882	.026	-.093	.232	-.590	.002	.066	.011	.004
114	1.237	1.340	1.232	1.431	.720	.371	.751	.064	.008	.045	.023	.012
120	.610	.736	.698	.749	-1.165	-1.434	-1.112	-1.577	.007	.085	.020	.009
122	.364	.855	.709	.272	4.252	1.981	******	3.453	.188	.052	.020	.044
133	.507	.478	.476	.568	-2.037	-2.143	-1.450	-2.417	.013	.098	.022	.010
147	.584	.657	.802	.659	3.657	******	3.100	2.895	.003	.087	.011	.003
154	.785	.619	.896	.731	.943	.866	.530	.268	.037	.064	.014	.005
158	.531	.856	.796	.554	2.081	1.847	1.445	1.595	.006	.089	.021	.004

Table 17
Item Parameters for Subgroups on the J-P Scale

	a-parameter				b-parameter				c-parameter			
Item	WM	BF	BM	WF	WM	BF	BM	WF	WM	BF	BM	WF
1	1.025	1.284	1.018	1.108	.761	.843	1.037	.851	.009	.031	.010	.004
9	.597	.597	.710	.614	1.435	1.124	1.083	1.351	.085	.030	.021	.129
13	.734	.958	.946	.843	.256	.188	.448	.435	.005	.012	.010	.005
20	.538	.425	.548	.637	-.369	-.375	-.379	-.158	.009	.033	.024	.013
27	.991	1.226	1.302	.920	.019	.275	.266	.313	.018	.015	.036	.010
35	.611	.725	.611	.627	.973	.627	1.050	.840	.047	.024	.022	.060
42	.777	.695	.717	.892	-.646	-.382	-.352	-.363	.028	.025	.022	.011
49	.869	.839	.897	.907	1.535	1.858	1.485	1.651	.011	.010	.006	.003
55	.992	1.052	1.037	.936	.306	.360	.492	.279	.019	.020	.035	.006
60	.791	.857	.854	.822	.818	1.242	.978	1.067	.041	.018	.029	.032
74	.938	.595	.782	.901	.137	-.096	.502	-.561	.008	.029	.020	.009
85	1.634	1.501	2.018	1.355	.685	.760	1.029	.483	.001	.008	.013	.002
94	.687	.567	.817	.569	1.512	1.732	1.946	1.256	.002	.019	.012	.004
97	.658	.810	.970	.610	1.030	.455	1.120	.352	.005	.021	.013	.007
99	.499	.463	.548	.557	-1.730	-1.720	-1.778	-1.476	.020	.026	.024	.014
109	.747	.638	.608	.633	-.268	-.233	.336	-.701	.021	.028	.020	.017
113	.406	.428	.401	.376	.362	1.758	1.677	.707	.010	.048	.044	.034
118	.762	.668	.752	.662	.591	.495	1.057	.075	.005	.019	.030	.006
124	.560	.522	.405	.393	1.142	.903	.785	.741	.238	.036	.026	.011
132	.804	1.030	1.089	.873	.560	.875	1.021	.598	.004	.014	.013	.006
142	.642	.768	.748	.714	.614	.449	.879	.445	.017	.046	.025	.044
151	.841	.953	.916	.843	1.306	1.556	1.710	1.156	.003	.012	.005	.005
153	.498	.505	.683	.514	1.514	1.675	1.121	1.945	.027	.029	.010	.008

Table 18
BILOG results: Classical Item Statistics For Subtest E-I White Males

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
6	6986.0	3971.0	.568	.16	.482	.607
15	6984.0	4025.0	.576	.18	.399	.504
19	6998.0	4584.0	.655	.38	.425	.549
25	6997.0	1650.0	.236	-.69	.456	.628
33	6995.0	2941.0	.420	-.19	.499	.630
41	6853.0	4172.0	.609	.26	.325	.414
47	6967.0	1759.0	.252	-.64	.234	.318
50	6993.0	3049.0	.436	-.15	.660	.831
58	6966.0	3887.0	.558	.14	.315	.397
66	6961.0	2195.0	.315	-.46	.388	.507
77	7003.0	3035.0	.433	-.16	.286	.361
87	7001.0	3789.0	.541	.10	.466	.586
92	7007.0	2175.0	.310	-.47	.424	.556
95	7007.0	2042.0	.291	-.52	.305	.403
106	7007.0	2908.0	.415	-.20	.402	.508
116	7004.0	688.0	.098	-1.30	.340	.585
126	7006.0	3342.0	.477	-.05	.513	.643
129	6978.0	5835.0	.836	.96	.192	.288
134	6991.0	4234.0	.606	.25	.454	.577
138	6994.0	3936.0	.563	.15	.441	.555
148	6941.0	2976.0	.429	-.17	.469	.591
160	7001.0	2349.0	.336	-.40	.361	.468

Table 19
BILOG results: Classical Item Statistics For Subtest E-I Black Females

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
6	329.0	142.0	.432	-.16	.514	.648
15	329.0	169.0	.514	.03	.338	.424
19	329.0	206.0	.626	.30	.424	.542
25	328.0	74.0	.226	-.73	.449	.625
33	330.0	110.0	.333	-.41	.436	.565
41	321.0	191.0	.595	.23	.333	.422
47	327.0	133.0	.407	-.22	.281	.355
50	330.0	121.0	.367	-.32	.654	.837
58	329.0	217.0	.660	.39	.378	.488
66	327.0	122.0	.373	-.31	.341	.435
77	330.0	204.0	.618	.28	.275	.351
87	330.0	171.0	.518	.04	.480	.602
92	329.0	134.0	.407	-.22	.395	.500
95	329.0	122.0	.371	-.31	.246	.314
106	330.0	138.0	.418	-.19	.304	.384
116	330.0	28.0	.085	-1.40	.261	.468
126	329.0	130.0	.395	-.25	.546	.693
129	329.0	246.0	.748	.64	.262	.356
134	329.0	179.0	.544	.10	.480	.603
138	329.0	212.0	.644	.35	.370	.475
148	326.0	156.0	.479	-.05	.438	.549
160	330.0	104.0	.315	-.46	.221	.289

Table 20
BILOG results: Classical Item Statistics For Subtest E-I Black Males

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
6	348.0	156.0	.448	-.12	.471	.592
15	346.0	233.0	.673	.43	.370	.481
19	346.0	215.0	.621	.29	.438	.558
25	347.0	59.0	.170	-.93	.416	.618
33	348.0	137.0	.394	-.25	.424	.539
41	341.0	203.0	.595	.23	.345	.437
47	348.0	106.0	.305	-.49	.206	.270
50	347.0	127.0	.366	-.32	.598	.766
58	346.0	219.0	.633	.32	.248	.318
66	346.0	131.0	.379	-.29	.417	.532
77	347.0	166.0	.478	-.05	.258	.324
87	347.0	228.0	.657	.38	.507	.655
92	348.0	134.0	.385	-.28	.439	.559
95	348.0	105.0	.302	-.49	.151	.199
106	348.0	164.0	.471	-.07	.409	.514
116	347.0	37.0	.107	-1.25	.320	.537
126	347.0	124.0	.357	-.35	.421	.541
129	346.0	272.0	.786	.77	.110	.155
134	348.0	194.0	.557	.14	.412	.518
138	348.0	226.0	.649	.36	.365	.470
148	348.0	159.0	.457	-.10	.435	.547
160	348.0	129.0	.371	-.31	.275	.352

Table 21
BILOG results: Classical Item Statistics For Subtest E-I White Females

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
6	3068.0	1624.0	.529	.07	.497	.624
15	3066.0	1240.0	.404	-.23	.406	.515
19	3069.0	1961.0	.639	.34	.422	.541
25	3072.0	718.0	.234	-.70	.524	.723
33	3068.0	1031.0	.336	-.40	.505	.653
41	3028.0	1658.0	.548	.11	.316	.397
47	3059.0	959.0	.314	-.46	.244	.319
50	3074.0	1014.0	.330	-.42	.657	.854
58	3055.0	2020.0	.661	.39	.345	.446
66	3047.0	642.0	.211	-.78	.375	.529
77	3079.0	1727.0	.561	.14	.321	.405
87	3074.0	1285.0	.418	-.19	.457	.577
92	3079.0	1081.0	.351	-.36	.430	.554
95	3075.0	1123.0	.365	-.33	.299	.383
106	3079.0	1128.0	.366	-.32	.403	.516
116	3082.0	229.0	.074	-1.48	.313	.585
126	3082.0	1266.0	.411	-.21	.543	.686
129	3064.0	2239.0	.731	.59	.176	.237
134	3076.0	1722.0	.560	.14	.430	.542
138	3073.0	1561.0	.508	.02	.458	.574
148	3055.0	1003.0	.328	-.42	.461	.599
160	3077.0	742.0	.241	-.67	.328	.450

Table 22
BILOG results: Classical Item Statistics For Subtest S-N White Males

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
2	6938.0	2345.0	.338	-.40	.505	.653
11	6956.0	4242.0	.610	.26	.407	.517
17	6953.0	3497.0	.503	.01	.383	.481
37	6896.0	4843.0	.702	.50	.380	.502
53	6849.0	4173.0	.609	.26	.353	.448
64	6969.0	5647.0	.810	.85	.281	.406
70	6908.0	4824.0	.698	.49	.421	.555
73	7009.0	5642.0	.805	.83	.450	.647
76	7006.0	2872.0	.410	-.21	.549	.694
78	7002.0	3335.0	.476	-.06	.468	.587
88	7002.0	5290.0	.755	.66	.513	.703
90	7002.0	4291.0	.613	.27	.489	.623
98	7009.0	4289.0	.612	.27	.504	.641
102	6999.0	4389.0	.627	.31	.553	.707
104	6999.0	2625.0	.375	-.30	.604	.771
107	7008.0	5255.0	.750	.65	.514	.700
112	6959.0	1898.0	.273	-.58	.374	.501
115	7003.0	1254.0	.179	-.90	.378	.554
117	6979.0	3658.0	.524	.06	.407	.511
119	6996.0	3274.0	.468	-.08	.452	.567
121	6991.0	4896.0	.700	.50	.343	.452
128	6985.0	3413.0	.489	-.03	.598	.750
140	6989.0	5926.0	.848	1.01	.317	.484
145	6987.0	3978.0	.569	.16	.542	.683
149	6986.0	3516.0	.503	.01	.468	.587
165	6984.0	3723.0	.533	.08	.433	.544

Table 23
BILOG results: Classical Item Statistics For Subtest S-N Black Females

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
2	326.0	108.0	.331	-.41	.465	.604
11	327.0	202.0	.618	.28	.357	.455
17	330.0	179.0	.542	.10	.353	.444
37	324.0	234.0	.722	.56	.242	.324
53	320.0	170.0	.531	.07	.230	.289
64	325.0	244.0	.751	.65	.244	.332
70	319.0	219.0	.687	.46	.343	.449
73	330.0	267.0	.809	.85	.473	.682
76	328.0	99.0	.302	-.49	.407	.536
78	329.0	181.0	.550	.12	.399	.502
88	328.0	245.0	.747	.64	.444	.603
90	329.0	215.0	.653	.37	.423	.546
98	329.0	198.0	.602	.24	.366	.464
102	330.0	204.0	.618	.28	.499	.636
104	329.0	130.0	.395	-.25	.505	.641
107	330.0	285.0	.864	1.09	.341	.535
112	326.0	45.0	.138	-1.08	.293	.458
115	329.0	27.0	.082	-1.42	.163	.296
117	328.0	203.0	.619	.29	.357	.455
119	327.0	164.0	.502	.00	.513	.643
121	330.0	212.0	.642	.34	.275	.353
128	330.0	156.0	.473	-.06	.466	.584
140	329.0	277.0	.842	.98	.285	.430
145	326.0	168.0	.515	.04	.335	.420
149	328.0	137.0	.418	-.20	.250	.316
165	330.0	216.0	.655	.38	.345	.444

Table 24
BILOG results: Classical Item Statistics For Subtest S-N Black Males

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
2	343.0	120.0	.350	-.36	.549	.707
11	348.0	223.0	.641	.34	.395	.507
17	348.0	168.0	.483	-.04	.454	.569
37	343.0	222.0	.647	.36	.358	.460
53	343.0	190.0	.554	.13	.360	.453
64	348.0	281.0	.807	.84	.282	.406
70	344.0	237.0	.689	.47	.401	.526
73	348.0	275.0	.790	.78	.386	.546
76	348.0	114.0	.328	-.42	.550	.716
78	346.0	181.0	.523	.05	.351	.440
88	347.0	262.0	.755	.66	.394	.539
90	346.0	212.0	.613	.27	.403	.513
98	348.0	168.0	.483	-.04	.433	.542
102	348.0	186.0	.534	.08	.565	.710
104	348.0	114.0	.328	-.42	.552	.717
107	348.0	282.0	.810	.85	.354	.512
112	344.0	42.0	.122	-1.16	.232	.376
115	347.0	57.0	.164	-.96	.371	.556
117	344.0	183.0	.532	.08	.290	.364
119	347.0	154.0	.444	-.13	.392	.493
121	348.0	256.0	.736	.60	.288	.388
128	346.0	163.0	.471	-.07	.568	.712
140	348.0	289.0	.830	.93	.224	.333
145	347.0	155.0	.447	-.13	.486	.611
149	346.0	138.0	.399	-.24	.414	.525
165	346.0	227.0	.656	.38	.340	.439

Table 25
BILOG results: Classical Item Statistics For Subtest S-N White Female

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
2	3043.0	1202.0	.395	-.25	.524	.665
11	3054.0	1706.0	.559	.14	.444	.559
17	3068.0	1751.0	.571	.17	.444	.560
37	3038.0	2367.0	.779	.74	.358	.500
53	3002.0	1649.0	.549	.12	.304	.382
64	3044.0	2314.0	.760	.68	.297	.409
70	3022.0	2169.0	.718	.55	.417	.556
73	3084.0	2575.0	.835	.95	.455	.680
76	3078.0	1240.0	.403	-.23	.547	.693
78	3081.0	1530.0	.497	-.01	.439	.550
88	3081.0	2393.0	.777	.73	.490	.683
90	3076.0	2184.0	.710	.53	.414	.549
98	3084.0	2327.0	.755	.66	.478	.654
102	3083.0	2257.0	.732	.59	.524	.704
104	3078.0	1355.0	.440	-.14	.607	.764
107	3079.0	2502.0	.813	.86	.479	.694
112	3055.0	1098.0	.359	-.34	.383	.492
115	3079.0	369.0	.120	-1.17	.299	.486
117	3071.0	1849.0	.602	.24	.388	.492
119	3072.0	1648.0	.536	.09	.515	.646
121	3078.0	1968.0	.639	.34	.323	.415
128	3076.0	1489.0	.484	-.04	.593	.744
140	3075.0	2751.0	.895	1.26	.284	.478
145	3067.0	1870.0	.610	.26	.550	.699
149	3059.0	1528.0	.500	.00	.415	.521
165	3067.0	1738.0	.567	.16	.453	.570

Table 26
BILOG results: Classical Item Statistics For Subtest T-F White Male

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
4	6940.0	4369.0	.630	.31	.337	.431
26	6962.0	1257.0	.181	-.89	.402	.588
29	6947.0	3007.0	.433	-.16	.394	.496
72	7002.0	4665.0	.666	.41	.486	.631
79	7005.0	1012.0	.144	-1.05	.402	.621
81	6999.0	2713.0	.388	-.27	.363	.462
84	7002.0	2954.0	.422	-.19	.306	.386
86	7004.0	1681.0	.240	-.68	.471	.646
89	6965.0	3970.0	.570	.17	.444	.560
91	7007.0	5305.0	.757	.67	.271	.371
93	6990.0	3481.0	.498	.00	.295	.369
100	7008.0	2592.0	.370	-.31	.385	.493
103	7002.0	2431.0	.347	-.37	.458	.591
105	7002.0	1671.0	.239	-.68	.384	.528
108	7000.0	6352.0	.907	1.34	.177	.309
111	6999.0	3408.0	.487	-.03	.494	.619
114	6995.0	2005.0	.287	-.54	.517	.687
120	7001.0	5092.0	.727	.58	.339	.454
122	7002.0	1752.0	.250	-.65	.070	.096
133	6982.0	5739.0	.822	.90	.259	.380
147	7008.0	289.0	.041	-1.85	.164	.368
154	6971.0	2092.0	.300	-.50	.379	.499
158	6985.0	1151.0	.165	-.95	.264	.395

Table 27
BILOG results: Classical Item Statistics For Subtest T-F Black Female

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
4	327.0	213.0	.651	.37	.379	.489
26	323.0	111.0	.344	-.38	.412	.531
29	329.0	164.0	.498	.00	.461	.578
72	330.0	249.0	.755	.66	.422	.578
79	329.0	85.0	.258	-.62	.520	.704
81	329.0	245.0	.745	.63	.311	.422
84	329.0	160.0	.486	-.03	.287	.360
86	329.0	125.0	.380	-.29	.527	.672
89	329.0	244.0	.742	.62	.409	.554
91	330.0	264.0	.800	.82	.249	.356
93	330.0	162.0	.491	-.02	.292	.366
100	329.0	150.0	.456	-.10	.362	.454
103	330.0	153.0	.464	-.09	.433	.543
105	329.0	118.0	.359	-.34	.407	.523
108	328.0	284.0	.866	1.10	.171	.269
111	330.0	181.0	.548	.11	.420	.528
114	330.0	133.0	.403	-.23	.537	.680
120	328.0	268.0	.817	.88	.336	.490
122	330.0	47.0	.142	-1.06	.272	.422
133	330.0	277.0	.839	.97	.212	.320
147	330.0	16.0	.048	-1.75	.110	.234
154	326.0	117.0	.359	-.34	.327	.419
158	327.0	63.0	.193	-.84	.259	.374

Table 28
BILOG results: Classical Item Statistics For Subtest T-F Black Males

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
4	344.0	207.0	.602	.24	.327	.415
26	348.0	65.0	.187	-.87	.320	.464
29	347.0	125.0	.360	-.34	.428	.549
72	348.0	221.0	.635	.33	.425	.544
79	348.0	53.0	.152	-1.01	.508	.775
81	348.0	199.0	.572	.17	.329	.415
84	346.0	138.0	.399	-.24	.327	.415
86	348.0	78.0	.224	-.73	.506	.705
89	348.0	177.0	.509	.02	.520	.652
91	348.0	267.0	.767	.70	.145	.201
93	345.0	159.0	.461	-.09	.224	.281
100	348.0	117.0	.336	-.40	.398	.515
103	348.0	132.0	.379	-.29	.489	.624
105	347.0	106.0	.305	-.48	.421	.553
108	348.0	306.0	.879	1.17	.224	.364
111	348.0	147.0	.422	-.18	.534	.673
114	347.0	98.0	.282	-.55	.503	.670
120	348.0	257.0	.739	.61	.342	.463
122	347.0	78.0	.225	-.73	.111	.154
133	348.0	257.0	.739	.61	.269	.363
147	348.0	14.0	.040	-1.87	.170	.385
154	347.0	124.0	.357	-.35	.479	.615
158	346.0	67.0	.194	-.84	.378	.543

Table 29
BILOG results: Classical Item Statistics For Subtest T-F White Females

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
4	3035.0	2203.0	.726	.57	.374	.500
26	3029.0	1140.0	.376	-.30	.508	.648
29	3046.0	1627.0	.534	.08	.454	.569
72	3079.0	2533.0	.823	.90	.444	.652
79	3079.0	914.0	.297	-.51	.468	.618
81	3082.0	1645.0	.534	.08	.348	.437
84	3075.0	1729.0	.562	.15	.331	.417
86	3076.0	1418.0	.461	-.09	.523	.657
89	3062.0	2395.0	.782	.75	.428	.601
91	3083.0	2547.0	.826	.92	.238	.351
93	3071.0	1971.0	.642	.34	.297	.381
100	3082.0	1477.0	.479	-.05	.409	.513
103	3080.0	1565.0	.508	.02	.485	.608
105	3075.0	969.0	.315	-.46	.375	.491
108	3084.0	2875.0	.932	1.54	.162	.311
111	3082.0	1998.0	.648	.36	.481	.620
114	3063.0	1463.0	.478	-.05	.586	.735
120	3079.0	2540.0	.825	.91	.373	.549
122	3082.0	656.0	.213	-.77	.125	.176
133	3071.0	2704.0	.880	1.17	.263	.428
147	3081.0	192.0	.062	-1.59	.211	.416
154	3061.0	1329.0	.434	-.16	.419	.528
158	3073.0	668.0	.217	-.75	.296	.415

Table 30
BILOG results: Classical Item Statistics For Subtest J-P White Males

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
1	6981.0	2043.0	.293	-.52	.506	.670
9	7008.0	2036.0	.291	-.53	.295	.391
13	6921.0	3011.0	.435	-.15	.453	.570
20	6932.0	3948.0	.570	.16	.352	.444
27	6987.0	3461.0	.495	-.01	.501	.628
35	7002.0	2319.0	.331	-.41	.349	.453
42	6955.0	4578.0	.658	.39	.421	.544
49	6982.0	1143.0	.164	-.96	.403	.603
55	6994.0	2924.0	.418	-.19	.500	.631
60	6978.0	2283.0	.327	-.42	.414	.538
74	7006.0	3216.0	.459	-.10	.499	.626
85	6993.0	1905.0	.272	-.58	.605	.810
94	7009.0	1355.0	.193	-.84	.358	.516
97	7008.0	1978.0	.282	-.55	.376	.501
99	7009.0	5489.0	.783	.76	.269	.377
109	7004.0	3966.0	.566	.16	.419	.528
113	7000.0	3114.0	.445	-.13	.289	.363
118	7002.0	2491.0	.356	-.35	.432	.555
124	6985.0	3160.0	.452	-.11	.245	.308
132	6994.0	2498.0	.357	-.35	.474	.608
142	6989.0	2606.0	.373	-.31	.403	.515
151	6988.0	1388.0	.199	-.82	.421	.603
153	6999.0	1841.0	.263	-.61	.294	.397

Table 31
BILOG results: Classical Item Statistics For Subtest J-P Black Females

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
1	328.0	83.0	.253	-.64	.525	.714
9	330.0	94.0	.285	-.54	.329	.437
13	325.0	139.0	.428	-.17	.503	.634
20	326.0	183.0	.561	.15	.277	.349
27	329.0	130.0	.395	-.25	.544	.691
35	329.0	115.0	.350	-.37	.402	.518
42	329.0	190.0	.578	.18	.390	.493
49	329.0	38.0	.116	-1.20	.368	.605
55	329.0	125.0	.380	-.29	.493	.629
60	330.0	68.0	.206	-.79	.409	.580
74	330.0	170.0	.515	.04	.374	.468
85	329.0	79.0	.240	-.68	.599	.823
94	330.0	65.0	.197	-.83	.300	.430
97	330.0	124.0	.376	-.30	.441	.563
99	330.0	252.0	.764	.69	.236	.325
109	329.0	179.0	.544	.10	.382	.479
113	330.0	88.0	.267	-.60	.230	.310
118	329.0	125.0	.380	-.29	.402	.512
124	324.0	111.0	.343	-.38	.315	.407
132	329.0	83.0	.252	-.64	.498	.677
142	329.0	131.0	.398	-.24	.414	.525
151	329.0	46.0	.140	-1.07	.398	.621
153	327.0	76.0	.232	-.70	.263	.363

Table 32
BILOG results: Classical Item Statistics For Subtest J-P Black Males

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
1	346.0	84.0	.243	-.67	.499	.684
9	347.0	98.0	.282	-.55	.385	.512
13	346.0	135.0	.390	-.26	.511	.649
20	344.0	202.0	.587	.21	.332	.420
27	347.0	155.0	.447	-.13	.556	.699
35	348.0	107.0	.307	-.48	.351	.461
42	347.0	206.0	.594	.22	.419	.531
49	348.0	59.0	.170	-.93	.412	.612
55	348.0	136.0	.391	-.26	.518	.658
60	346.0	100.0	.289	-.53	.433	.575
74	348.0	137.0	.394	-.25	.452	.573
85	348.0	69.0	.198	-.82	.601	.860
94	348.0	43.0	.124	-1.15	.329	.531
97	348.0	81.0	.233	-.70	.473	.653
99	346.0	280.0	.809	.85	.272	.393
109	347.0	154.0	.444	-.13	.367	.462
113	348.0	103.0	.296	-.51	.218	.288
118	348.0	100.0	.287	-.53	.396	.526
124	347.0	138.0	.398	-.24	.230	.292
132	347.0	84.0	.242	-.67	.515	.706
142	348.0	111.0	.319	-.45	.420	.547
151	348.0	46.0	.132	-1.11	.387	.612
153	346.0	94.0	.272	-.58	.398	.533

Table 33
BILOG results: Classical Item Statistics For Subtest J-P White Females

Item	Number Tried	Number Right	Percent	LOGIT/1.7	Item*Test Pearson	Correlation Biserial
1	3072.0	795.0	.259	-.62	.510	.690
9	3082.0	1024.0	.332	-.41	.270	.350
13	3060.0	1175.0	.384	-.28	.478	.609
20	3028.0	1620.0	.535	.08	.397	.498
27	3057.0	1265.0	.414	-.20	.477	.603
35	3079.0	1108.0	.360	-.34	.352	.451
42	3057.0	1812.0	.593	.22	.476	.603
49	3065.0	409.0	.133	-1.10	.389	.614
55	3072.0	1289.0	.420	-.19	.485	.612
60	3073.0	818.0	.266	-.60	.404	.544
74	3081.0	1980.0	.643	.35	.462	.593
85	3076.0	1050.0	.341	-.39	.574	.742
94	3082.0	807.0	.262	-.61	.338	.457
97	3083.0	1305.0	.423	-.18	.380	.479
99	3077.0	2353.0	.765	.69	.300	.413
109	3080.0	1996.0	.648	.36	.372	.478
113	3079.0	1275.0	.414	-.20	.257	.325
118	3081.0	1477.0	.479	-.05	.406	.509
124	3065.0	1200.0	.392	-.26	.268	.341
132	3072.0	1053.0	.343	-.38	.485	.626
142	3069.0	1279.0	.417	-.20	.411	.519
151	3073.0	693.0	.226	-.73	.423	.589
153	3080.0	578.0	.188	-.86	.279	.405

Table 34
Transformed Item Parameters for Subgroups on the E-I Scale

	a-parameters				b-parameters			
Item	WM*	WF	BF	BM	WM*	WF	BF	BM
6	0.881	0.807	0.861	0.914	-0.223	-0.387	0.120	0.229
15	0.639	0.559	0.455	0.655	-0.312	0.238	-0.225	-0.747
19	0.726	0.658	0.606	0.748	-0.662	-0.934	-0.783	-0.502
25	1.031	1.149	1.073	1.176	1.011	0.788	0.965	1.268
33	0.893	0.838	0.686	0.799	0.305	0.428	0.643	0.488
41	0.471	0.403	0.448	0.589	-0.622	-0.577	-0.730	-0.394
47	0.370	0.353	0.381	0.853	1.877	1.342	0.560	1.489
50	1.961	1.821	1.833	1.748	0.204	0.297	0.283	0.381
58	0.477	0.506	0.524	0.421	-0.329	-1.187	-1.057	-0.743
66	0.632	0.603	0.487	0.767	0.888	1.334	0.613	0.528
77	0.406	0.422	0.376	0.417	0.431	-0.652	-0.967	0.250
87	0.780	0.670	0.694	1.098	-0.165	0.118	-0.236	-0.556
92	0.736	0.623	0.541	0.863	0.830	0.497	0.365	0.447
95	0.474	0.387	0.369	0.359	1.255	0.687	0.941	1.960
106	0.618	0.553	0.404	0.752	0.403	0.451	0.433	0.166
116	0.949	0.800	0.723	0.887	1.910	2.215	2.414	1.865
126	0.995	1.042	1.061	0.851	0.093	0.105	0.243	0.600
129	0.345	0.236	0.365	0.294	-2.965	-2.845	-2.084	-2.639
134	0.769	0.629	0.712	0.728	-0.437	-0.551	-0.366	-0.211
138	0.714	0.685	0.496	0.615	-0.265	-0.302	-0.992	-0.696
148	0.833	0.793	0.661	0.880	0.283	0.545	0.001	0.285
160	0.572	0.492	0.330	0.476	0.842	1.348	1.495	0.881

Note. * = the WM metric onto which the other parameters were transformed.

Table 35
Transformed Item Parameters for Subgroups on the S-N Scale

	a-parameters				b-parameters			
Item	WM*	WF	BF	BM	WM*	WF	BF	BM
2	0.945	1.040	1.269	1.209	0.601	0.402	0.290	0.494
11	0.589	0.747	0.684	0.664	-0.538	-0.097	-0.658	-0.587
17	0.605	0.876	0.732	0.862	0.116	-0.010	-0.354	0.120
37	0.595	0.640	0.510	0.605	-0.951	-1.268	-1.346	-0.655
53	0.559	0.659	0.491	0.724	-0.382	0.431	-0.175	-0.051
64	0.457	0.478	0.555	0.563	-2.083	-1.459	-1.485	-1.713
70	0.696	0.755	0.693	0.727	-0.890	-0.845	-0.996	-0.792
73	0.955	1.127	1.430	0.823	-1.258	-1.255	-1.254	-1.254
76	1.088	1.175	1.058	1.264	0.311	0.388	0.453	0.574
78	0.743	0.705	0.762	0.586	0.088	0.060	-0.430	-0.038
88	1.078	1.070	1.181	0.800	-0.968	-1.007	-1.006	-1.059
90	0.847	0.757	0.931	0.684	-0.450	-0.840	-0.780	-0.434
98	0.860	1.005	1.005	0.758	-0.449	-0.901	-0.404	0.136
102	1.075	1.185	1.272	1.252	-0.463	-0.765	-0.618	-0.092
104	1.445	1.442	1.303	1.424	0.386	0.214	0.032	0.583
107	1.035	1.100	0.966	0.712	-0.966	-1.157	-1.740	-1.486
112	0.644	0.674	1.046	0.593	1.116	0.767	1.243	2.567
115	0.839	1.098	1.251	0.996	1.448	1.817	1.754	1.500
117	0.657	0.628	1.022	0.500	-0.018	-0.398	-0.436	-0.010
119	0.937	1.065	1.426	0.814	0.336	-0.004	-0.253	0.430
121	0.504	0.473	0.511	0.517	-1.150	-0.701	-0.903	-1.273
128	1.219	1.302	1.125	1.209	0.026	0.101	-0.159	0.095
140	0.597	0.657	0.699	0.447	-2.011	-2.217	-1.895	-2.251
145	0.953	1.096	0.739	0.863	-0.265	-0.323	-0.119	0.211
149	0.765	0.786	0.722	0.747	0.027	0.242	0.555	0.501
165	0.646	0.779	0.685	0.558	-0.124	-0.134	-0.827	-0.738

Note. * = the WM metric onto which the other parameters were transformed.

Table 36
Transformed Item Parameters for Subgroups on the T-F Scale

	a-parameters				b-parameters			
Item	WM*	WF	BF	BM	WM*	WF	BF	BM
4	0.552	0.641	0.516	0.407	-0.684	-0.481	-0.392	-0.625
26	0.968	1.011	0.601	0.530	1.332	1.058	1.357	2.282
29	0.665	0.768	0.676	0.586	0.302	0.474	0.535	0.894
72	1.054	1.029	0.752	0.672	-0.610	-0.710	-0.755	-0.621
79	1.037	0.986	0.996	1.034	1.475	1.419	1.507	1.700
81	0.572	0.512	0.451	0.408	0.550	0.449	-1.097	-0.415
84	0.457	0.484	0.395	0.442	0.450	0.277	0.790	0.745
86	1.099	1.015	0.911	0.931	0.950	0.754	1.001	1.357
89	0.810	0.836	0.629	0.801	-0.296	-0.615	-0.851	0.003
91	0.453	0.418	0.362	0.251	-1.667	-1.803	-1.984	-2.888
93	0.433	0.434	0.397	0.267	0.010	-0.262	0.872	0.569
100	0.651	0.645	0.793	0.560	0.590	0.731	1.190	0.992
103	0.862	0.888	0.590	0.764	0.588	0.631	0.704	0.711
105	0.707	0.614	0.567	0.608	1.198	1.531	1.361	1.151
108	0.422	0.406	0.324	0.416	-3.467	-3.487	-3.152	-3.172
111	0.934	0.852	0.589	0.873	0.025	0.020	0.243	0.380
114	1.236	1.383	1.000	0.924	0.720	0.697	0.864	1.071
120	0.609	0.724	0.549	0.523	-1.165	-1.000	-1.552	-1.409
122	0.363	0.262	0.638	0.532	4.252	4.202	3.020	.
133	0.507	0.549	0.357	0.357	-2.037	-1.868	-2.502	-1.859
147	0.584	0.637	0.490	0.602	3.657	3.624	.	4.200
154	0.785	0.706	0.462	0.672	0.942	0.907	1.528	0.776
158	0.531	0.535	0.639	0.597	2.080	2.280	2.840	1.995

Note. * = the WM metric onto which the other parameters were transformed.

Table 37
Transformed Item Parameters for Subgroups on the J-P Scale

	a-parameters				b-parameters			
Item	WM*	WF	BF	BM	WM*	WF	BF	BM
1	1.025	1.060	1.336	0.996	0.761	0.831	0.784	0.877
9	0.597	0.587	0.621	0.695	1.435	1.353	1.055	0.924
13	0.734	0.807	0.997	0.926	0.256	0.397	0.155	0.276
20	0.537	0.609	0.442	0.536	-0.369	-0.222	-0.385	-0.568
27	0.990	0.880	1.275	1.275	0.018	0.269	0.238	0.089
35	0.610	0.599	0.754	0.597	0.972	0.820	0.577	0.891
42	0.777	0.854	0.723	0.702	-0.646	-0.436	-0.392	-0.540
49	0.868	0.867	0.872	0.877	1.535	1.667	1.760	1.335
55	0.992	0.895	1.094	1.015	0.305	0.233	0.320	0.320
60	0.790	0.786	0.891	0.835	0.818	1.057	1.168	0.817
74	0.937	0.862	0.618	0.766	0.136	-0.643	-0.117	0.331
85	1.633	1.297	1.562	1.975	0.685	0.447	0.704	0.869
94	0.686	0.544	0.590	0.800	1.512	1.255	1.639	1.806
97	0.658	0.584	0.843	0.949	1.030	0.310	0.412	0.962
99	0.499	0.533	0.481	0.536	-1.729	-1.599	-1.678	-1.997
109	0.746	0.605	0.664	0.595	-0.267	-0.789	-0.249	0.161
113	0.405	0.359	0.445	0.392	0.361	0.681	1.664	1.531
118	0.762	0.634	0.694	0.736	0.591	0.021	0.450	0.898
124	0.559	0.375	0.543	0.396	1.142	0.716	0.842	0.620
132	0.804	0.835	1.072	1.065	0.559	0.567	0.816	0.861
142	0.641	0.683	0.799	0.732	0.613	0.407	0.406	0.716
151	0.840	0.807	0.991	0.896	1.306	1.150	1.469	1.565
153	0.498	0.491	0.525	0.668	1.513	1.974	1.584	0.963

Note. * = the WM metric onto which the other parameters were transformed.

Table 38
Rank Ordered Area Statistics and Corresponding M-H Values: E-I WM-BF

Item	BF Area	E-I BF Δ
77	0.83760	2.03972
47	0.70661	1.90980
58	0.53935	1.25536
138	0.48459	1.15989
92	0.47123	1.44500
95	0.38722	1.01635
148	0.32960	0.84355
66	0.31708	0.86568
160	0.31616	-0.12801
6	0.31560	-1.81452
106	0.28307	0.17675
33	0.27158	-1.08476
129	0.24939	-1.30089
15	0.22048	-0.64686
19	0.13643	-0.33730
116	0.13144	-0.06970
126	0.11842	-1.07900
87	0.11834	-0.17931
41	0.09189	-0.10118
50	0.07929	-1.54429
134	0.07212	-0.72732
25	0.06324	0.13104

Table 39
Rank Ordered Area Statistics and Corresponding M-H Values: S-N WM-BF

Item	BF Area	S-N BF Δ
165	0.67149	1.46561
107	0.60434	1.93943
117	0.59624	1.20006
78	0.54930	0.94515
119	0.50241	0.72773
104	0.44505	0.93326
17	0.41684	0.55932
2	0.40531	0.29369
90	0.40154	0.48763
128	0.29998	-0.05532
145	0.27983	-0.64150
149	0.26983	-0.82999
11	0.26004	0.04850
102	0.25823	-0.27313
115	0.25320	-1.86906
37	0.24659	0.17637
98	0.24331	-0.21566
112	0.21243	-2.07533
88	0.17555	-0.57766
70	0.16859	-0.26190
73	0.16693	-0.42780
76	0.12372	-1.25847
53	0.11969	-0.82612
64	0.10698	-1.06332
140	0.07960	-0.41169
121	0.03742	-0.81897

Table 40
Rank Ordered Area Statistics and Corresponding M-H Values: T-F WM-BF

Item	BF Area	T-F BF Δ
81	1.22502	3.27872
122	0.56240	-1.94439
89	0.50648	0.96145
111	0.37518	-0.97095
26	0.33888	1.11757
154	0.32676	-0.56011
72	0.31152	-0.54453
103	0.30923	-0.02271
120	0.26066	0.26571
100	0.25703	-0.18766
93	0.25201	-0.89393
158	0.22837	-0.49208
105	0.22123	0.41044
108	0.21988	-1.74689
114	0.19832	-0.36517
86	0.18676	0.36530
84	0.15508	-0.17026
133	0.12888	-0.64195
91	0.12873	-0.17175
4	0.11628	-0.80231
29	0.09815	-0.51334
79	0.07777	0.57297
147	.	-0.49157

Table 41
Rank Ordered Area Statistics and Corresponding M-H Values: J-P WM-BF

Item	BF Area	J-P BF Δ
113	0.65173	-1.80759
97	0.46925	1.74341
60	0.36974	-1.51771
124	0.36517	-1.00354
74	0.35569	1.29960
132	0.25363	-1.14600
27	0.24179	-0.89910
35	0.23552	0.57176
42	0.22218	-0.68716
13	0.20438	0.39044
142	0.19516	0.73859
118	0.16365	0.80054
9	0.15366	0.18926
151	0.14817	-0.80265
49	0.14299	-0.69905
20	0.13007	0.17794
1	0.11183	-0.04344
94	0.09944	0.51573
109	0.08325	0.15667
153	0.06357	-0.16514
55	0.06128	0.10855
99	0.04119	-0.02405
85	0.03814	0.55046

Table 42

Rank Ordered Area Statistics and Corresponding M-H Values: E-I WM-BM

Item	BM Area	E-I BM Δ
87	0.44564	1.65718
15	0.40260	1.17511
126	0.38261	-1.63102
138	0.38175	1.07955
92	0.35745	1.17345
6	0.34531	-1.64302
58	0.33008	0.79824
66	0.32053	0.93525
47	0.28618	0.70748
106	0.27598	0.73042
95	0.22035	0.17368
77	0.21743	0.47615
25	0.20077	-1.08289
160	0.19646	0.49323
50	0.17879	-1.38652
129	0.17688	-0.87370
148	0.17067	0.49899
33	0.15780	-0.24485
134	0.12972	-0.66184
41	0.10013	-0.18820
116	0.08581	0.66729
19	0.06454	-0.50427

Table 43
Rank Ordered Area Statistics and Corresponding M-H Values: S-N WM-BM

Item	BM Area	S-N BM Δ
112	0.61128	-2.27457
165	0.48981	1.68618
98	0.40996	-1.38261
145	0.36969	-1.24669
107	0.34509	1.22657
149	0.33229	-0.93593
78	0.25517	0.86906
102	0.25450	-1.02970
17	0.22863	0.07598
53	0.21422	-0.42521
76	0.20557	-0.62466
117	0.19530	0.38336
88	0.18945	0.11357
90	0.17969	0.28956
37	0.17926	-0.52065
104	0.17753	0.21407
121	0.15445	0.52458
2	0.14580	0.87092
11	0.14140	0.57195
140	0.10376	-0.32302
119	0.09693	0.18777
73	0.07836	-0.28347
64	0.07663	-0.03436
115	0.06314	0.24926
128	0.06312	0.44986
70	0.01782	0.04171

Table 44
Rank Ordered Area Statistics and Corresponding M-H Values: T-F WM-BM

Item	BM Area	T-F BM Δ
81	0.67981	2.19689
29	0.36217	-0.88474
26	0.33987	0.02227
133	0.29906	-1.28944
111	0.29522	-0.74639
72	0.28725	-0.18854
93	0.28681	-0.34622
114	0.28589	0.05024
86	0.27611	-0.25517
89	0.24251	-0.66858
100	0.22789	-0.36217
91	0.22309	0.17864
4	0.19216	-0.26301
79	0.15457	0.06594
105	0.15234	1.08717
84	0.14498	-0.18534
103	0.13221	0.60554
120	0.11513	0.26930
154	0.10845	0.89903
108	0.08357	-0.82015
147	0.05844	-0.14566
158	0.02321	0.57420
122	.	-0.33339

Table 45

Rank Ordered Area Statistics and Corresponding M-H Values: J-P WM-BM

Item	BM Area	J-P BM Δ
113	0.52776	-1.25104
109	0.34638	-0.78154
132	0.28774	-1.00833
153	0.24221	0.61264
9	0.23724	0.35249
94	0.21578	-0.82464
151	0.19836	-0.71555
74	0.19061	0.05058
97	0.19002	-0.10223
118	0.18560	-0.21875
99	0.17805	1.03474
85	0.17277	-0.47779
20	0.16249	0.80526
124	0.15990	-0.22855
13	0.14804	0.25795
27	0.13116	0.32922
49	0.12131	0.93354
42	0.11247	-0.01505
142	0.10045	-0.04247
1	0.09340	0.08651
60	0.05661	0.16379
35	0.03023	0.23193
55	0.02198	0.61143

Table 46
Rank Ordered Area Statistics and Corresponding M-H Values: E-I WM-WF

Item	WF Area	E-I WF Δ
77	0.66073	1.76150
58	0.57786	1.67889
15	0.45753	-1.54182
95	0.41549	1.29464
47	0.34852	1.09026
129	0.32616	-1.34677
92	0.31790	1.21346
66	0.28432	-1.02150
160	0.25558	-0.82818
87	0.23925	-0.93128
25	0.18463	0.98983
148	0.18348	-0.69059
19	0.17356	0.45040
134	0.14422	0.08422
6	0.10737	0.29394
33	0.10365	-0.43194
41	0.09922	-0.27230
50	0.09729	-0.80130
116	0.08122	-0.20885
106	0.07367	-0.06190
138	0.03643	-0.02942
126	0.01964	-0.02280

Table 47
Rank Ordered Area Statistics and Corresponding M-H Values: S-N WM-WF

Item	WF Area	S-N WF Δ
98	0.47112	1.64603
90	0.33940	0.83775
102	0.32234	1.08585
112	0.32071	0.82292
115	0.27324	-1.96418
37	0.26409	0.71575
117	0.25712	0.53313
121	0.25509	-1.10608
119	0.25000	0.41764
53	0.23808	-0.98202
17	0.23209	0.42500
64	0.22579	-1.11781
11	0.22331	-1.03621
107	0.19781	0.50692
2	0.18998	0.29662
104	0.18626	0.39761
140	0.16465	0.66318
165	0.10196	-0.02241
145	0.10055	-0.07288
149	0.09421	-0.49299
73	0.07554	-0.10832
78	0.05947	-0.21722
76	0.05185	-0.83521
128	0.04846	-0.90851
88	0.03974	-0.42281
70	0.03491	-0.21267

Table 48
Rank Ordered Area Statistics and Corresponding M-H Values: T-F WM-WF

Item	WF Area	T-F WF Δ
122	0.29119	-0.87714
89	0.28415	1.03171
26	0.18754	0.77700
86	0.18378	0.75544
29	0.17891	-0.67256
93	0.16993	0.42005
105	0.15449	-0.72546
4	0.12674	-0.29918
81	0.11072	0.18214
84	0.10958	0.27519
120	0.10578	-0.05784
158	0.09807	-0.53052
100	0.09018	-0.44088
72	0.08682	0.20857
79	0.08667	0.28728
111	0.06337	-0.22042
114	0.05382	-0.18223
133	0.04387	-0.09757
154	0.03886	-0.19214
91	0.03804	-0.01807
147	0.03166	-0.40854
103	0.02367	-0.17131
108	0.02326	-0.14136

Table 49

Rank Ordered Area Statistics and Corresponding M-H Values: J-P WM-WF

Item	WF Area	J-P WF Δ
74	0.69133	2.41329
97	0.56642	1.67084
118	0.48539	1.39531
109	0.35721	0.84783
153	0.28430	-1.26684
94	0.27954	0.95198
85	0.24806	1.13405
27	0.23821	-1.54389
124	0.23316	-0.72887
142	0.21116	0.36553
60	0.20554	-1.10111
42	0.18613	-1.21744
9	0.18534	0.40430
35	0.14267	0.21807
151	0.13346	0.29432
13	0.12968	-0.93388
113	0.11516	-0.43019
49	0.11082	-0.95737
20	0.10559	-0.57750
1	0.08218	-0.93660
55	0.06115	-0.30120
99	0.04380	-0.40613
132	0.02044	-0.46500

Table 50
Across-Comparison Findings of Differential Item Functioning on the E-I Scale

	WM-BF		WM-BM		WM-WF	
Item	Δ	area	Δ	area	Δ	area
6	large (-)	√	large (-)	√		
15			intermediate (+)	√	large (-)	√
19						
25			intermediate (-)			
33	intermediate (-)					
41						
47	large (+)	√			intermediate (+)	√
50	large (-)		intermediate (-)			
58	intermediate (+)	√		√	large (+)	√
66		√		√	intermediate (-)	
77	large (+)	√			large (+)	√
87			large (+)	√		
92	intermediate (+)	√	intermediate (-)	√	intermediate (+)	√
95	intermediate (+)	√			intermediate (+)	√
106						
116						
126	intermediate (-)		large (-)	√		
129	intermediate (-)				intermediate (-)	√
134						
138	intermediate (+)	√	intermediate (+)	√		
148		√				
160		√				

Note. (+) = DIF that favors the focal group (BF, BM, or WF); (-) = DIF that favors the reference group (WM); √ = an area over 0.30 in size; (D) = an item that is differentially scored by gender.

Table 51
Across-Comparison Findings of Differential Item Functioning on the S-N Scale

Item	WM-BF Δ	WM-BF area	WM-BM Δ	WM-BM area	WM-WF Δ	WM-WF area
2		√				
11					intermediate (-)	
17		√				
37						
53						
64	intermediate (-)				intermediate (-)	
70						
73						
76	intermediate (-)					
78		√				
88						
90		√				√
98			intermediate (-)	√	large (+)	√
102			intermediate (-)		intermediate (+)	√
104		√				
107	large (+)	√	intermediate (+)	√		
112	large (-)		large (-)	√		√
115	large (-)				large (-)	
117	intermediate (+)	√				
119		√				
121					intermediate (-)	
128						
140						
145			intermediate (-)	√		
149				√		
165	intermediate (+)	√	large (+)	√		

Note. (+) = DIF that favors the focal group (BF, BM, or WF); (-) = DIF that favors the reference group (WM); √ = an area over 0.30 in size; (D) = an item that is differentially scored by gender.

Table 52
Across-Comparison Findings of Differential Item Functioning on the T-F Scale

Item	WM-BF Δ	WM-BF area	WM-BM Δ	WM-BM area	WM-WF Δ	WM-WF area
4 (D)						
26	intermediate (+)	√		√		
29				√		
72		√				
79 (D)						
81	large (+)	√	large (+)	√		
84						
86 (D)						
89		√			intermediate (+)	
91						
93						
100 (D)						
103		√				
105			intermediate (+)			
108	intermediate (-)					
111 (D)		√				
114 (D)						
120						
122 (D)	large (-)	√				
133			intermediate (+)			
147 (D)						
154 (D)		√				
158 (D)						

Note. (+) = DIF that favors the focal group (BF, BM, or WF); (-) = DIF that favors the reference group (WM); √ = an area over 0.30 in size; (D) = an item that is differentially scored by gender.

Table 53
Across-Comparison Findings of Differential Item Functioning on the J-P Scale

	WM-BF		WM-BM		WM-WF	
Item	Δ	area	Δ	area	Δ	area
1						
9						
13						
20						
27					large (-)	
35						
42					intermediate (-)	
49						
55						
60	large (-)	√			intermediate (-)	
74	intermediate (+)	√			large (+)	√
85					intermediate (+)	
94						
97	large (+)	√			large (+)	√
99			intermediate (+)			
109				√		√
113	large (-)	√	intermediate (-)	√		
118					intermediate (+)	√
124	intermediate (-)	√				
132	intermediate (-)		intermediate (-)			
142						
151						
153					intermediate (-)	

Note. (+) = DIF that favors the focal group (BF, BM, or WF); (-) = DIF that favors the reference group (WM); √ = an area over 0.30 in size; (D) = an item that is differentially scored by gender.