

THE DEVELOPMENT AND VALIDATION OF EFFECT SIZE MEASURES FOR IRT AND
CFA STUDIES OF MEASUREMENT EQUIVALENCE

BY

CHRISTOPHER DAVID NYE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Fritz Drasgow, Chair
Professor Brent W. Roberts
Professor James Rounds
Assistant Professor Daniel A. Newman
Assistant Professor Sungjin Hong

ABSTRACT

Evaluating measurement equivalence is a necessary first step before comparisons can be made across groups or over time. As a result, techniques for evaluating equivalence have received much attention in the literature. Given the many benefits of these approaches, measurement equivalence is most appropriately assessed using item response theory (IRT) or confirmatory factor analytic (CFA) techniques. For both methods, the identification of biased items typically involves statistical significance tests based on the chi-square distribution or empirically derived rules of thumb for determining nonequivalence. However, because of the disadvantages of these criteria, it may be informative to use effect size estimates to judge the magnitude of the observed effects as well. As such, the present work proposed the development and evaluation of effect size measures for CFA and IRT studies of measurement equivalence. First, simulation research was used to illustrate the advantages of effect size measures and to develop guidelines for interpreting the magnitude of an effect. Next, these indices were used to evaluate nonequivalence in both cognitive and noncognitive data. In sum, the results show the benefits of evaluating the effect size of DIF in addition to assessing its statistical significance.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
CHAPTER 2: A SIMULATION STUDY OF AN EFFECT SIZE MEASURE FOR CFA ANALYSES OF MEASUREMENT EQUIVALENCE.....	26
CHAPTER 3: A FIELD STUDY OF THE CFA EFFECT SIZE MEASURE	32
CHAPTER 4: A SIMULATION STUDY OF AN EFFECT SIZE MEASURE FOR IRT ANALYSES OF DIFFERENTIAL ITEM FUNCTIONING	34
CHAPTER 5: A FIELD STUDY OF THE IRT EFFECT SIZE MEASURE	38
CHAPTER 6: GENERAL DISCUSSION	44
TABLES AND FIGURES	49
REFERENCES	74
APPENDIX A: DERIVATIONS OF Δ VARIANCE AND Δ COVARIANCE	85
APPENDIX B: DERIVATIONS OF $\Delta r_{xy(CFA)}$	87

CHAPTER 1: INTRODUCTION

The measurement of psychological constructs is a fundamental issue for understanding and predicting behavior. Given the unobservable nature of the constructs examined in psychological research, the measures used are the primary method for testing theories and studying relations. As a result, the quality of a measure and the accuracy with which it assesses its latent construct are important considerations when drawing conclusions from a study. Without appropriate measures, the interpretation of results and the theories that are developed are likely to provide inaccurate descriptions of human behavior. Because of their importance, factors that may affect psychological measurement have received a substantial amount of attention.

A quality measure must have a number of essential properties in order to measure a construct effectively. It is well-known that tests must demonstrate reliability and validity to justify their use. However, other properties are also crucial for the accurate interpretation of empirical results. When comparing groups, measurement equivalence is an essential property. Measurement equivalence occurs when individuals with the same standing on the latent trait have equal expected observed scores on the test (Drasgow, 1987). Without first establishing measurement equivalence, studies comparing scores across cultures, sexes, races, or over time may be misleading. Therefore, measurement equivalence is a necessary first step before drawing substantive conclusions from group-level comparisons.

The issue of measurement equivalence (or differential item functioning [DIF] in IRT terminology) is particularly salient in domains where mean-level differences have been observed between groups. Vandenberg and Lance (2000) noted that tests of measurement equivalence have been primarily conducted to inform group differences across gender (e.g., Stacy, MacKinnon, & Penz, 1993), age (e.g., Marsh, 1993), race (e.g., Chan, 1997), and culture (e.g.,

Nye, Roberts, Saucier, & Zhou, 2008). However, the issue of measurement equivalence has arguably gained the most attention in the cognitive ability domain where mean differences between majority and minority groups are well documented (Kuncel & Hezlet, 2007). In a recent report, Camara and Schmidt (1999) showed mean-level differences on the SAT, ACT, National Assessment of Educational Progress (NAEP), and the Advanced Placement (AP) Examinations across demographic groups. As a result of these differences, the use of such tests for college admissions decisions and in educational settings has been questioned using the arguments that such tests are unfair and biased against subpopulations of test-takers. Although fairness is inherently subjective, issues of bias are partially addressed via statistical analyses of measurement equivalence. Within the standardized testing domain, most items display equivalence but some content has been found to function differently across groups (Kuncel & Hezlett, 2007). Specifically, men tend to perform better than women on science-related questions (Lawrence, Curly, & McHale, 1988) and some verbal stimuli tend to favor European Americans over Hispanic individuals (Schmitt & Dorans, 1988).

In contrast to the results of DIF studies on cognitive predictors within a single country, research has shown that measurement nonequivalence may be pervasive in personality scales when administered internationally. For example, Nye et al. (2008) found a lack of invariance for the majority of items in a common measure of personality when compared across three independent languages. Similarly, Oishi (2007) demonstrated the nonequivalence of a subjective well-being scale when comparing U. S. and Chinese respondents and Chirkov, Ryan, Kim, and Kaplan (2003) found substantial nonequivalence in a measure of individualism and collectivism.

Although the issue of measurement equivalence has been particularly important for comparing demographic groups, other research has used these techniques for examining changes

in the psychometric properties of a measure over time. These changes, known as item drift, may occur as items become obsolete or outdated as a result of educational, technological, and/or cultural changes (Bock, Muraki, & Pfeifferberger, 1988). For example, Chan, Drasgow, and Sawin (1999) examined the Army's ASVAB test and found that items with greater semantic/knowledge components exhibited the most DIF over time. In addition, Drasgow, Nye, and Guo (2008) found significant levels of item drift on the National Council Licensure Examination for Registered Nurses (NCLEX-RN). However, in this study, DIF cancelled out at the test level suggesting that these items would not have a substantial impact on overall test scores.

Measurement equivalence is most appropriately examined using item response theory (IRT) differential functioning techniques or confirmatory factor analytic (CFA) mean and covariance structure (MACS) analyses. Although several articles have proposed various decision rules for determining if measurement nonequivalence exists (Cheung & Rensvold, 2002; Meade, Lautenschlager, & Johnson, 2007; Hu & Bentler, 1999), these rules generally involve cutoffs or statistical significance tests. As such, these criteria do not address the practical importance of observed differences between groups. In the broader psychological literature, many suggest that empirical work should be evaluated relative to effect sizes rather than tests for significance (Cohen, 1990, 1994; Kirk, 2006, Schmidt, 1996). However, empirically validated effect size indices for CFA and IRT evaluations of measurement equivalence at the item level do not exist. Thus, the goal of this research is to develop and evaluate such measures using a combination of simulation and empirical research to explore their boundaries and examine their use with real-world data.

Measurement Equivalence Techniques

Drasgow (1984) suggested that techniques for examining bias in psychological measures can be distinguished by the type of bias examined. Specifically, Drasgow described the differences between measurement and relational equivalence. As described above, measurement equivalence occurs when individuals with the same standing on the latent trait have an equivalent expected observed score (Drasgow, 1984). In contrast, relational equivalence is concerned with the relations between test scores and some external criteria. Relational equivalence has a long history in psychology and a number of indices have been proposed. Probably the most widely known model of relational equivalence is the regression model, suggested by Cleary (1968). This model defines bias as group level differences between the predicted criterion scores. In other words, bias occurs when the regression line for one group predicts criterion scores that are consistently higher or lower than those for another group. Other methods of assessing relational equivalence have also been suggested. These methods include, for example, assessing differential validity, the Constant Ratio Model (Thorndike, 1971), the Conditional Probability Model (Cole, 1973), and the Equal Probability Model (Linn, 1973), among others. However, these approaches have been criticized and, therefore, are not commonly used (see Drasgow, 1982 and Petersen & Novick, 1976 for a description of these approaches and their criticisms).

Drasgow (1984) suggested a sequential process for assessing bias. In the first step, measurement equivalence is assessed. If the measurement properties are shown to be invariant, the second step is to examine relational equivalence with an external criterion. The theory behind this sequential procedure is that if measurement equivalence does not hold, the interpretation of equivalent relations would be ambiguous. Specifically, if relational nonequivalence is found, these differences may be attributable to variance in the measurement rather than true differences

in the relationship. Furthermore, finding equivalent relations despite nonequivalent measurement only calls into question the invariance of the criterion. As additional support for this sequential process, Drasgow and Kang (1984) showed that the power of relational invariance analyses was significantly reduced in the presence of measurement nonequivalence such that bias was not identified in many cases and the effects were substantially reduced when it was. These findings are particularly problematic given the general tendency to assess equivalent relations without first establishing measurement invariance. As a result, the common finding that equivalent relations hold for most tests may be confounded by nonequivalent measurement. Thus, establishing measurement equivalence is of primary importance in the study of bias.

A variety of methods have been developed for examining measurement equivalence. Some have suggested using t-tests, analysis of variance, or other statistical tests of mean-level differences to evaluate bias in psychological measures. However, these methods are inappropriate for this purpose because they confound measurement bias with true mean differences (often referred to as impact; Stark, Chernyshenko, & Drasgow, 2004) between the groups. Stated differently, these methods assume equal ability distributions across groups; an assumption that does not necessarily hold in practice (Hulin, Drasgow, & Parsons, 1983). Showing the inadequacy of these tests, Camilli and Shepard (1987) demonstrated that when true mean differences exist between groups, ANOVA comparisons were incapable of detecting bias. In fact, even when bias accounted for 35% of the observed mean difference, the effects suggested by ANOVA were negligible. More importantly, these authors found that the presence of true group differences can result in higher Type I error rates when group means are compared. In addition, Linn and Drasgow (1987) suggested that highly discriminating items (those that

differentiate between high and low ability individuals well) could result in significant mean level differences across groups and \hat{p} values that suggest bias.

Others have used exploratory factor analyses to assess measurement equivalence. This practice is commonly used in personality research where the factor structure of personality measures is determined in one culture and then confirmed in another by evaluating factor congruencies or using target rotation (van de Vijver & Leung, 2001). However, their exploratory nature and reliance on heuristic decision rules for judging equivalence make these analyses more difficult to use for evaluating measurement invariance. Moreover, this methodology only assesses a single type of nonequivalence that results from varying factor structures across groups. Thus, even if equivalent factor structures are found, EFA techniques do not rule out the possibility of biased means or differences across groups.

Given the limitations of mean-level comparisons and EFA, CFA methodology, known as mean and covariance structure (MACS) analysis, and IRT DIF analysis have been suggested as more appropriate alternatives for examining measurement equivalence (Cheung & Rensvold, 2000; Little, 1997; Stark, Chernyshenko, & Drasgow, 2006; Vandenberg & Lance, 2000). These techniques have a number of advantages over the alternative methods described above. First, they do not assume equal distributions of the latent trait across groups (Drasgow & Kanfer, 1985). Thus, unlike mean-level comparisons, individuals in one group may generally have higher (or lower) scores on the latent trait than the comparison group. Second, these methods allow a more advanced assessment of nonequivalence than either EFA or mean-level comparisons can provide. As will be discussed next, equivalence can be evaluated relative to the factor structure, the ability of the item to discriminate against high or low levels of the trait, and/or the item-level

difficulty/intercept. Therefore, these approaches apply a more comprehensive definition of nonequivalence to the data.

If DIF is identified using these methods, IRT and CFA techniques will also be useful for examining mean-level differences. Although the observed differences will be affected by nonequivalence, latent mean differences will not be when nonequivalent items are allowed to vary across groups. In other words, when the parameters for the DIF items are freely estimated in each of the groups and only the invariant items are constrained to be equivalent, the latent means can be compared (Ployhart & Oswald, 2004). Thus, even in the presence of DIF, these techniques can be used to evaluate hypothesized differences between groups or over time.

Mean and Covariance Structure Analysis (MACS)

Recent review articles have detailed several steps to MACS analyses. Although the number and order of these steps vary across studies (Vandenberg & Lance, 2000), researchers are generally interested in tests of configural, metric, and scalar invariance. However, a number of additional tests for invariance are available and the exact forms of invariance that are assessed should be linked to the purposes of the study (Steenkamp & Baumgartner, 1998). Nevertheless, these additional tests should be preceded by a confirmation of configural, metric, and scalar equivalence.

Configural invariance is generally considered to be the first step in assessing measurement equivalence (Vandenberg & Lance, 2000). The equivalence of these models confirms the null hypothesis that the pattern of fixed and free loadings is equivalent across groups. Essentially, this test assesses the extent to which individuals in both groups employ the same conceptualization of the focal constructs and confirms that the factor structure holds across groups. This type of invariance has been particularly important in the personality literature where

there have been substantial debates over the factor structure of individual differences (see Saucier & Goldberg, 2001). If the pattern of zero and nonzero loadings differ across groups, the process stops and no further tests are required; constructs that are conceptualized differently are not comparable across groups. In contrast, if the configural invariance of the measure is confirmed, assessments of metric invariance should proceed.

Metric invariance is tested by constraining the factor loadings to be equivalent across groups (i.e., $\Lambda^g = \Lambda^{g'}$). The factor loadings index the relationships between the item responses and the latent variables they assess (Bollen, 1989, Vandenberg, 2002). This relationship is clearly represented in the MACS model by

$$\mathbf{x}^g = \boldsymbol{\tau}_x + \Lambda_x \boldsymbol{\xi}^g + \boldsymbol{\delta}^g$$

where \mathbf{x}^g is the vector of observed variables for the g th group, $\boldsymbol{\tau}_x$ is the vector of intercepts of the regressions of the observed variables on the latent factors $\boldsymbol{\xi}$, and $\boldsymbol{\delta}$ is the vector of measurement errors. Thus, tests of metric invariance are implicitly assessing the conceptual scaling of the observed responses (Vandenberg, 2002).

In contrast to tests of configural invariance, failure to support metric invariance does not preclude further tests of DIF. Instead, partial metric invariance can be assessed in the next step. In other words, parameters that are found to vary across groups are freely estimated while the rest of the parameters remain constrained. Vandenberg and Lance (2000) suggest that a disadvantage of this approach is the exploratory nature of the post hoc process used to identify biased items and the possibility that items may be identified as nonequivalent as a result of chance (i.e., due to the number of comparisons). However, recent research by Stark et al. (2006) suggested that constraining single items at a time (known as the free-baseline approach) provides higher power and lower Type I error rates than the more common approach of constraining all of

the items (the constrained-baseline approach) and systematically freeing parameters to diagnose nonequivalence at the item level. Using this approach, post hoc analyses would not be required to identify nonequivalent items.

Items that are found invariant at the metric level can then be assessed for scalar equivalence. Here, the item intercepts are constrained in addition to the factor loadings (i.e., $\tau^g = \tau^{g'}$ and $\Lambda^g = \Lambda^{g'}$). As such, this test assesses the comparability of scores across groups. A failure to support the null hypothesis suggests that the scores, and hence group means, are not directly comparable. Therefore, tests of scalar invariance have substantial import for drawing conclusions about group differences. Ployhart and Oswald (2004) suggest that the decision to constrain indicator intercepts to equality should be considered carefully. Specifically, one must determine whether item intercepts are expected to be equal. Therefore, the examination of scalar invariance should be based on substantive hypotheses rather than being tested blindly.

Because tests for scalar invariance will be confounded with metric nonequivalence when it exists, these tests are generally assessed sequentially. However, Stark et al. (2006) suggest that it may be more appropriate to evaluate these forms of invariance simultaneously. The authors offer several reasons for this recommendation. First, MACS analyses were as sensitive to DIF when these forms of equivalence were examined simultaneously as when they were treated sequentially. Second, examining metric and scalar equivalence separately increases the number of comparisons and, therefore, also increases the risk of Type I errors. Finally, the sequential process may propagate errors from one step (e.g., metric invariance) to another (i.e., scalar invariance).

Differential Item Functioning (DIF)

Nonequivalence can also be assessed using IRT DIF analysis. In contrast to the linear relationship between the underlying construct and the observed score proposed in the CFA framework, IRT proposes a non-linear relationship in the form of a logistic function. This function relates the probability of a correct response on an item to the ability of an individual and the characteristics of the item. For a range of ability levels, these relations are represented by an item characteristic curve (ICC). However, a number of IRT models are available for defining this curve. One of the more widely used models for dichotomous items is the 3-parameter logistic (3pl) model defined by

$$P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

where $P_i(\theta)$ is the probability of a correct response to item i at ability level θ , D is the scaling constant 1.702, a_i is the discrimination parameter that represents the slope of the ICC, b_i is the difficulty parameter or location of the item, and c_i is the guessing parameter or the lower asymptote of the ICC.

Measurement equivalence, or DIF, occurs when the parameters of the item are invariant across groups. Using the 3pl model shown above, this means that

$$a_i^R = a_i^F, b_i^R = b_i^F, c_i^R = c_i^F$$

where R represents the reference group (i.e., the majority group or the group used for comparison) and F represents the focal group (i.e., the minority group or the group that DIF is expected in) in IRT terminology. The equivalence of item parameters can also be expressed by the invariance of the ICCs across groups. Here, DIF occurs when the area between the two curves is significantly different from zero. Although the typical MACS approach generally follows a sequential procedure for testing the equivalence of item intercepts and loadings, IRT

DIF analyses have traditionally taken a simultaneous approach to assessing differences in the item location (b-parameter) and discrimination parameters (a-parameters, respectively). Again, Stark et al. (2006) provide several theoretical justifications for this approach. The *c*-parameter is frequently ignored because it is often very difficult to estimate accurately.

A number of procedures have been developed to test for significant differences between the item parameters or ICCs. For example, Lord (1980) developed a chi-square index of differences between parameters, Thissen, Steinberg, and Wainer (1988) suggested a likelihood ratio test, Linn, Levine, Hastings, and Wardrop (1981) directly compared ICCs across groups, Raju (1988, 1990) derived alternative formulas for calculating the significance of the area between two ICCs, and Raju, van der Linden, and Fleer (1995) developed a method they call differential functioning for items and tests (DFIT). Although each takes a slightly different approach, these methods each assess DIF in either the item parameters or the ICCs. For example, Lord's (1980) chi-square focused on a statistical test of the differences between item parameters. In contrast, Raju et al.'s (1995) approach defines DIF as

$$ES_{iR} \neq ES_{iF},$$

where ES_{iR} is the expected score on item *i* for an examinee in the reference group and ES_{iF} is the expected item score for an examinee in the focal group with the same θ .

Differential functioning at the item level can also be summed to assess the effects of bias at the test level. Thus, an important property of DIF is the potential for compensatory effects when aggregated to the test level. In other words, DIF towards the focal group on one item may cancel out items with DIF against the reference group at the test level. Therefore, despite nonequivalence at the item level, the test as a whole may not function differently. This form of nonequivalence, termed differential test functioning (DTF), is reflected by the differences in the

test characteristic curves. If DTF is observed, a common practice is to systematically remove items with significant DIF until DTF becomes non-significant (Raju et al., 1995). Thus, if significant DTF exists, items making a substantial contribution to the overall effect will be identified for removal, resulting in a test that is invariant across groups.

Raju, Laffitte, and Byrne (2002) point out several differences between MACS and DIF approaches to examining equivalence. As described above, one of the primary differences is how the relationship between the latent construct and the observed score is modeled. This relationship is linear in CFA whereas in IRT it is expressed as a nonlinear logistic function. Second, because a logistic regression model is more appropriate for describing the relationship between a continuous underlying trait and a dichotomous outcome, IRT may be more appropriate for modeling dichotomous data. Indeed, two recent studies (Meade & Lautenschlager, 2004; Stark et al., 2006) comparing the Type I error rates and power of CFA and IRT assessments of invariance have shown that IRT methodology does provide slightly superior results when dichotomous data are analyzed. In contrast, a key advantage of the CFA approach is that this model is well-suited for tests of invariance with multidimensional models. Drasgow and Parsons (1983) showed that unidimensional IRT models are robust to moderate violations of unidimensionality. However, if the dimensions are substantively important and/or researchers are interested in the relationships among them, it will be important to model the full factor structure. Although multidimensional IRT models are available, most DIF analyses have been developed solely for the dichotomous case and have yet to be generalized to a multidimensional model. Finally, another difference between CFA and IRT methodology is that CFA models may be estimated more accurately in small samples (Stark et al., 2006).

Interpreting the Results of IRT and CFA Studies of Measurement Equivalence

Due to their importance for evaluating the quality of a scale or test, analyses of measurement equivalence have received much attention in the literature and several studies have provided empirical recommendations for evaluating IRT and CFA results to determine if measurement equivalence exists. For both IRT and CFA, the most commonly used indices of fit appear to be statistical significance tests based on a χ^2 distribution. In MACS research, chi-square difference tests are generally used to compare nested models. Although frequently used for this purpose, it is well-known that χ^2 tests are severely affected by sample size. Thus, in large sample analyses, even small differences will be detected as statistically significant (Meade, Johnson, & Braddy, 2008). A similar dependence on sample size has been demonstrated for various DIF detection methods (e.g., McLaughlin & Drasgow, 1987; Budgell, Raju, & Quartetti, 1995; Raju et al., 1995; Stark et al., 2006).

As a result, many have suggested using other indices (i.e., changes in fit statistics) to evaluate equivalence. This practice has resulted in a variety of empirically derived rules of thumb for identifying nonequivalence in CFA or IRT studies. For example, Cheung and Rensvold (2002) showed that ΔCFI is the most accurate indicator of nonequivalence and suggested that a ΔCFI greater than .01 be used as a cutoff for determining substantial changes in fit. More recently, Meade et al. (2008) showed that this cutoff was too liberal and did not detect some forms of nonequivalence. Instead, these authors recommended that a $\Delta\text{CFI} > .002$ is suggestive of DIF.

A similar trend has occurred in IRT DIF studies. Because of the sensitivity of the chi-square test to sample size, Fler (1993) suggested establishing empirically derived cutoff values for assessing DIF with the noncompensatory DIF index (NCDIF) in the DFIT framework. As

such, Raju, van der Linden, and Fler (1995) suggested a cutoff of .016 and subsequent Monte Carlo simulations have provided support for it (Flowers, Oshima, & Raju, 1999). More recently, Meade et al. (2007) showed that an NCDIF value of .009 is a more accurate indicator of DIF.

Notice that in both cases the trend was moving towards more conservative criteria for evaluating nonequivalence. Thus, Meade et al. (2007) differentiated between detectable and practically significant DIF, suggesting that these two goals were independent issues. These authors state that the development of conservative cutoffs is primarily focused on the range of detectable DIF whereas the practical significance of these observations is inherently subjective. Nevertheless, the use of cutoff values only suggests that some degree of nonequivalence may be present in the scale and does not address the extent of this bias.

The Need for Effect Size Indices

This emphasis on statistical significance tests and empirically derived cutoff values has been criticized in the broader psychological literature (Cohen, 1990, 1994; Harlow, Mulaik, & Steiger, 1997; Kirk, 1996; Schmidt, 1996). Although a number of criticisms have been raised, four primary limitations are frequently discussed. First, many have suggested that null hypothesis significance testing (NHST) does not tell researchers what they want to know (Cohen, 1994). Although researchers are interested in the probability that the null hypothesis is true given the data, significance tests describe the probability of the data given the null hypothesis. Despite this difference, many continue to believe that null hypotheses describe the former. This belief has led to the ultimately false conclusions that 1) the p value represents the probability that the null hypothesis is correct and 2) $1 - p$ is the probability of replicating the present result in future research (Kirk, 1996).

Another important criticism is that NHST is a trivial exercise because the null hypothesis can always be shown to be false to some degree (Cohen, 1990). The null hypothesis states that the magnitude of the effect is zero but, when measured with enough accuracy, an effect can always be found between two variables. As a result, a significant effect can be found for any two variables given a large enough sample (Tukey, 1991). Thus, a more important question may be: How large is the effect? The answer to this question cannot be addressed using NHST. As a result, Tukey (1991) suggested that a significance test may be more useful for demonstrating the direction rather than the existence of an effect.

Given the falsity of the null hypothesis, the general focus on Type I errors in psychology is unfortunate because these errors can never occur (i.e., if the null can never be true, one can never falsely reject it). Moreover, the focus on Type I errors has led to a general neglect of power and Type II error rates (see Cohen, 1962), the only type of error that can occur.

The third criticism of NHST can also be applied to the use of cutoff values or rules of thumb. Kirk (2006) critiques these criteria because they force researchers to turn a decision continuum into a dichotomous reject/do not reject decision. As Kirk points out, this practice treats a p value that is only slightly larger than the cutoff (e.g. .055) the same as a much larger difference. The result is that some researchers treat statistical significance as an indicator of the importance or significance of the finding.

Another criticism of NHST is that it does not demonstrate the practical significance of a finding. In other words, NHST does not reflect the magnitude, value, or importance of an effect (Kirk, 2006). Nevertheless, statistically significant results are often treated as “important” in the psychological literature (Cohen, 1990). Again, this criticism can be applied to cutoff values as well. As noted by Meade et al. (2008), the cutoff criteria used in CFA research were developed

to detect a range of nonequivalence from trivial to substantial. Thus, observing a $\Delta\text{CFI} > .002$ indicates that DIF exists but does not reflect the importance or magnitude of these differences.

As a result of these criticisms (and others), it is widely believed that the interpretation of empirical results should be based on an evaluation of effect sizes rather than tests of statistical significance. Therefore, a number of effect size statistics have been developed for analyses of variance (e.g., Hays, 1963), t-tests (e.g., Cohen, 1988), and other traditional statistical tests. However, no such indices exist for CFA analyses and few alternatives are available for item-level IRT analyses of measurement equivalence. Stark, Chernyshenko, and Drasgow (2004) proposed an effect size for analyses of differential test functioning (DTF) that they refer to as d_{dtf} . Despite the usefulness of this measure, it does not address bias at the item-level which can be informative for test development. By assessing the magnitude of DIF, one can more easily detect items and/or specific content that may be problematic for a specific group. Moreover, d_{dtf} is not applicable to CFA methodology where no viable alternatives exist for these techniques.

Effect Size Indices for DIF and MACS Analyses

Some effect size indices have been proposed for IRT DIF studies. For example, Dorans and Kulick (1986) suggested using the standardized difference between $P_i(\theta)$ in the focal and reference groups as an effect size measure. This index is calculated by summing the differences over ability levels and then standardizing the sum using a common weight across groups. However, this index is based on ad hoc decisions (see Dorans & Kulick, p. 360) and does not have a consistent interpretation across research studies. Instead, an effect size measure in a standardized metric—similar to those developed by Cohen (1969) and Glass (1976)—may be more informative for evaluating the magnitude of an effect.

Other effect size indices have been proposed by Rudas and Zwick (1997), Holland and Thayer (1988), and Zumbo (1999). However, because these indices are not based on IRT parameters, they have trouble detecting some forms of non-uniform DIF (see Hidalgo & López-Pina, 2004; Holland & Thayer, 1988; Swaminathan & Rogers, 1990). Therefore, parametric indices may be more useful in some situations.

As suggested by Stark, et al. (2004), an index of practically important nonequivalence can be defined as the contribution that biases make to expected score differences for each item. Whereas Stark et al. examined bias at the test level, the present work proposes that differences at the item level (i.e. DIF) should be examined. In IRT, one DIF measure computes the area between the item characteristic curves (ICCs) for the reference and focal groups. In other words, this measure compares the probability of a correct response in each group [i.e. $P_{iR}(\theta)$ and $P_{iF}(\theta)$, respectively] while controlling the level of the latent trait. Thus, by averaging these differences over the levels of ability (θ), one can obtain an estimate of the contribution made by DIF to differences in the expected item scores between the two groups. To put this difference in a standardized metric comparable to other effect size measures (e.g. Cohen's d), this value can be divided by the pooled standard deviation of item i in the reference and focal groups (SD_{iP} ; cf. Stark et al., 2004). Thus, an effect size for DIF, d_{DIF} , can be defined as

$$d_{DIF} = \frac{1}{SD_{iP}} \sqrt{\int [P_{iR}(\theta) - P_{iF}(\theta)]^2 f_F(\theta) d\theta}$$

where $f_F(\theta)$ is the ability density of the focal group with the mean and variance estimated from the transformed $\hat{\theta}$ distribution.

A similar conceptualization of the magnitude of an effect can be derived for the MACS analyses in a CFA framework. Again, an effect is defined as the contribution of measurement

nonequivalence to expected mean score differences at the item level. Although CFA analyses are most often conducted by placing simultaneous constraints on all of the items in a scale (i.e., the constrained-baseline approach), single-item constraints (i.e., the free-baseline approach) are frequently recommended to diagnose the source of any nonequivalence (Vandenberg & Lance, 2000). More importantly, recent research has shown that assessing invariance using the free-baseline approach provides lower Type I and II error rates than the more common constrained-baseline analyses (Stark, et al., 2006). Given these results, an effect size measure for MACS analyses will be most beneficial at the item level.

In CFA methodology, the mean predicted response \hat{X}_{iR} to item i for an individual in group R with a score of ξ on the latent variable is given by

$$\hat{X}_{iR} = \tau_{iR} + \lambda_{iR}\xi$$

where τ_{iR} is the intercept and λ_{iR} is the item's loading. Here, nonequivalence is reflected in the area between this regression line and that for the same item in group F. Thus, an effect size for MACS analysis is defined as

$$d_{MACS} = \frac{1}{SD_{iP}} \sqrt{\int (\hat{X}_{iR} - \hat{X}_{iF} | \xi)^2 f_F(\xi) d\xi}$$

where SD_{iP} is the pooled standard deviation of item i in groups R and F and $f(\xi)$ is the ability density of the latent factor (Nye & Drasgow, 2011). Again, the latent factor is assumed to be normally distributed.

Practical Consequences of Measurement Nonequivalence

Although these effect size measures can be used to describe the magnitude of an effect, they still provide little information about the observed consequences of measurement nonequivalence. For example, what effects does DIF have on the mean and the variance of the

scale? How will DIF affect the correlation between a scale and an external variable? To address these issues, equations were derived to calculate these effects. These equations will help researchers and practitioners to further understand the effects of nonequivalence.

In group-level comparisons, observed mean differences can be defined as

$$\text{Observed Differences} = \text{DIF} + \text{impact}.$$

To quantify the effects of DIF on the mean of a scale in the IRT framework, one can calculate

$$\Delta \text{Mean}(X_{DIF}) = \sum_1^n \int [P_{iR}(\theta) - P_{iF}(\theta)] f_F(\theta) d\theta$$

where X_{DIF} is the scale score. Notice that the integral in this equation is similar to that for d_{DIF} except that the differences between the mean predicted responses in groups R and F at each ability level are not squared so that DIF in opposite directions can cancel. In addition, because we are interested in the change in the mean of the *scale*, item-level differences are summed across all n items to obtain the overall mean difference in raw score points. In sum, $\Delta \text{Mean}(X_{DIF})$ refers to the amount of the observed difference that can be attributed to DIF; impact is not a factor in this calculation. Using similar logic, the effects of nonequivalence on the mean of the scale in the CFA framework (cf. Nye & Drasgow, 2011) is calculated by

$$\Delta \text{Mean}(X_{MACS}) = \sum_1^n \int (\hat{X}_{iR} - \hat{X}_{iF} | \xi) f_F(\xi) d\xi.$$

Differences between the variances of a scale in the reference and focal groups due to DIF can also be calculated. For a given θ , the variance of item i in the reference group is

$$\text{Var}_{iR}(\theta) = P_{iR}(\theta)[1 - P_{iR}(\theta)].$$

Using this equation, Lord (1980) provided the formula for calculating the variance of a scale in the total group (see p. 42)

$$Var_R(X) = \int \sum_1^n P_{iR}(\theta)[1 - P_{iR}(\theta)]f_F(\theta)d\theta + \int [\sum_1^n P_{iR}(\theta)]^2 f_F(\theta)d\theta - [\int \sum_1^n P_{iR}(\theta)f_F(\theta)d\theta]^2 .$$

Thus, the effects of DIF on the variance of a scale can be defined as

$$\Delta Var(X_{DIF}) = Var_R(X) - Var_F(X) .$$

In contrast, the formula for calculating the variance of item i in the reference group using CFA methodology is

$$Var(x_{iR}) = \lambda_{iR}^2 \phi_R + Var(\delta_{iR})$$

where λ_{iR}^2 is the squared factor loading of item i , δ is the measurement error of the item, and ϕ_R is the variance of the latent factor in the reference group. Based on this formula, the effects of nonequivalence on the variance of an item are

$$\Delta Var(x_i) = 2C_i \lambda_{iR} \phi_F + C_i^2 \phi_F ,$$

where C_i is the difference between the factor loadings for item i in the reference and focal groups (Nye & Drasgow, 2011). Two key assumptions were made in this derivation. First, because we are only interested in identifying differences due to DIF, we assumed $\phi_R = \phi_F$. When this is the case, ΔVar is not influenced by true group level differences in the latent construct. Instead, only metric nonequivalence (i.e., differences in the factor loadings) can result in $\Delta Var > 0$. The second simplifying assumption is $Var(\delta_{iR}) = Var(\delta_{iF})$. Several authors have suggested that requiring equivalent error variances is the least important hypothesis to test and is generally unnecessary for analyses of measurement equivalence (Bentler, 1995; Byrne, 1998; Joreskog, 1971). Speaking about constraining the error variances and covariances to equality in multi-group tests, Byrne noted that “it is now widely accepted that to do so represents an overly restrictive test of the data” (p. 261). Therefore, differences in δ are not included when evaluating the effects of DIF.

To estimate the total effect of DIF on the variance of a scale, the $\Delta Var(x_i)$ can be aggregated across all n items in the scale using the formula for calculating the variance of a composite. In other words,

$$\Delta Var(X_{MACS}) = \Delta Var(x_1) + \Delta Var(x_2) \dots + \Delta Var(x_n) \\ + 2\Delta Cov(x_1, x_2) \dots + 2\Delta Cov(x_{n-1}, x_n)$$

where

$$\Delta Cov(x_i, x_j) = \lambda_{jR} C_i \phi_F + C_j \lambda_{iR} \phi_F + C_i C_j \phi_F$$

is the covariance of items i and j . Appendix A shows the derivations of ΔVar and ΔCov .

Although the effects of DIF on scale-level properties are important for evaluating the quality of a scale, researchers may also be interested in how an observed level of DIF will affect research results. In this regard, one can also calculate the effects of nonequivalence on the correlation between a scale and an external criterion. The correlation between a scale score X and criterion Y in CFA methodology is

$$r_{XY(CFA)} = \frac{\sum_{iR}^n \lambda_{iR} Cov(\xi_R, Y_R)}{SD_{XR} SD_{YR}}$$

in the reference group. Here, $Cov(\xi_R, Y_R)$ is the covariance of the latent trait and the criterion Y and SD_{XR} and SD_{YR} are the standard deviations of the scale and the criterion, respectively. Using this equation, the effects of nonequivalence on the correlation can be defined as

$$\Delta r_{XY(CFA)} = \frac{Cov(\xi_R, Y_R) \Delta SD(X) - Cov(\xi_R, Y_R) SD_{XR}}{SD_{XR}^2 SD_{YR} + SD_{XR} SD_{YR} \Delta SD(X)}.$$

where $\Delta SD(X)$ is the change in the standard deviation of the reference group due to nonequivalence [i.e., $\Delta SD(X) = \sqrt{\Delta Var(X)}$]. The derivation of $\Delta r_{XY(CFA)}$ is provided in

Appendix B.

In IRT, Lord (1980, p. 9) provided the formula for estimating the correlation of a scale with an external criterion

$$r_{XYR(IRT)} = \frac{\sum_{iR}^n SD_{iR} r_{iYR}}{SD_{XR}}$$

where r_{iYR} is the correlation between item i and criterion Y in the reference group. Unlike the formula for $\Delta r_{XY(CFA)}$, this equation cannot be reduced to a simpler formula for calculating the effects of DIF on the correlation. Therefore, $\Delta r_{XY(IRT)}$ is defined as

$$\Delta r_{XYR(IRT)} = r_{XYR(IRT)} - r_{XYF(IRT)}.$$

Overview of the Present Research

For the current research, the effect size measures described above were evaluated using a combination of simulation and empirical research. The goals of the simulations were 1) to illustrate the benefits of effect size values for DIF analysis by demonstrating their robustness to various data characteristics and 2) to develop empirically derived operational definitions of small, medium, and large effects for CFA and IRT studies of measurement equivalence. To accomplish these goals, item-level data were generated with a range of DIF levels (e. g., small, medium, and large amounts of DIF). These data were then used to evaluate the effect size measures relative to traditional indicators of DIF. In terms of IRT, comparisons were made with Lord's Chi-Square (Lord, 1980) and Raju, et al.'s (1995) NCDIF index. For CFA methodology, the effect size measure d_{MACS} was compared to the χ^2 difference tests between nested MACS models and the ΔCFI recommended by Cheung and Rensvold (2002) and Meade, et al. (2008). To demonstrate the robustness of effect size measures to various data characteristics, these comparisons were made under simulated conditions that varied the sample size, number of items,

and magnitude of DIF in the population. A review of relevant research on each of these characteristics is provided next.

Previous research suggests that both the sample size and the number of items in a test will have strong effects on CFA and IRT invariance research. Stark, et al. (2006) found that both techniques were affected by sample size. When $N = 500$, IRT tests resulted in high Type I error rates. In contrast, when $N = 250$, MACS analyses exhibited low power. Similar results were also found by Meade et al. (2008) for MACS analyses and are well established for a variety of IRT DIF analyses (e.g., McLaughlin & Drasgow, 1987; Budgell, Raju, & Quartetti, 1995; Raju et al., 1995; Stark et al., 2006).

Sample size also has the potential to affect the magnitude of the effect size measures proposed here by influencing the parameter estimates used to calculate \hat{X}_i in the CFA model and $P_i(\theta)$ in IRT. Although a commonly cited rule of thumb for the sample sizes required in CFA models is to have ten times as many subjects as variables (Nunnally, 1967), several studies have suggested that this heuristic is insufficient for providing adequate solutions (Marsh, Hau, Balla, & Grayson, 1998; Velicer & Fava, 1987). Other studies have demonstrated the effects of sample size on the quality of CFA model estimates. For example, Boomsma (1982) found that sample size affected the percentage of proper solutions, the accuracy of parameter estimates, and the appropriateness of the chi-square statistic. In addition, Velicer and Fava (1987) showed that sample size influenced model convergence and goodness-of-fit indices. The effects of sample size on the accuracy of IRT parameter estimates are also widely known (Hulin, Drasgow, & Parsons, 1983). Thus, sample size may affect both IRT and CFA estimates and the corresponding accuracy of predicted values (i.e., \hat{X}_i and $P_i[\theta]$).

The length of a test can also have similar effects on parameter estimates in IRT and CFA research (Hulin, Drasgow, & Parsons, 1983; Marsh et al., 1998). Cheung and Rensvold (2002) showed that many of the goodness of fit indices for CFA models are affected by the number of items in the test. In addition, Boomsma (1982) found that the accuracy of the parameter estimates and their sampling variability improved with the number of indicators and Marsh et al. (1998) noted that more items resulted in greater interpretability and more reliable factors. As such, the parametric models of effect size described here may be affected by the length of a test.

The magnitude of the effect in the population should also influence d_{DIF} and d_{MACS} . Therefore, the present study will examine the magnitude of these indices with respect to various levels of DIF in the population models. An added benefit of this approach is that it allows an evaluation of the conventions recommended by Cohen (1988) for interpreting the magnitude of an effect and, if necessary, new guidelines can be developed for d_{DIF} and d_{MACS} . A major contribution of Cohen's (1969) work was that he provided guidelines for interpreting the magnitude of his d index (Kirk, 1996). Specifically, values near .20 are considered small, near .50 are medium, and .80 or greater are large (Cohen, 1992). However, Cohen (1988, 1992) has noted that these definitions are arbitrary and based on subjective judgments. As such, although these guidelines are commonly used and have been linked to the observed effect sizes in a number of fields (e.g., Sedlemeier & Gigerenzer, 1989), these conventions have not been empirically tested. Therefore, another contribution of the present work is to develop guidelines that facilitate the identification of small, medium, and large amounts of measurement nonequivalence using d_{DIF} and d_{MACS} .

After evaluating these indices with simulated data and developing guidelines for interpreting the size of an effect, the magnitude of DIF was examined for both cognitive and

noncognitive measures. Specifically, DIF was assessed in the SAT, AP Spanish Language, and AP World History exams across gender and ethnic groups. Using IRT to calibrate items, the d_{DIF} index was applied to calculate the magnitude of the effect and was compared to Lord's Chi-Square and the NCDIF index. Effect sizes in the CFA framework were also explored using a widely researched measure of personality. In a re-analysis of data from Nye et al. (2008), the Mini-Markers Scale (Saucier, 1994) was examined to determine the extent of nonequivalence across American-English, Greek, and Chinese languages. The effect size of nonequivalence was then compared to $\Delta\chi^2$ and ΔCFI . For both IRT and CFA, the effects of nonequivalence on observed scale properties were also calculated using the equations described above. These analyses provide additional information about the influence of nonequivalence on research findings and conclusions.

CHAPTER 2: A SIMULATION STUDY OF AN EFFECT SIZE MEASURE FOR CFA ANALYSES OF MEASUREMENT EQUIVALENCE

Study 1: Methods

For Study 1, Monte Carlo computer simulations were used to study the effects of sample size, the number of items, and the magnitude of DIF in the population on estimates of d_{MACS} .

The study consisted of a 3 (sample size) \times 3 (number of items) \times 4 (magnitude of DIF) research design. In each design cell, values of d_{MACS} were compared to the Type I error rates and power of traditional indicators of nonequivalence over 200 replications.

The MACS Model

To enhance the generalizability of the results, measurement models were simulated to represent the models that are frequently found in psychological research. DiStefano (2002) reviewed issues of *Educational and Psychological Measurement* and *Psychological Bulletin* between 1992 and 1999 to ascertain the characteristics of commonly observed psychological models. This author found that two-dimensional models with 12 to 16 indicators and factor loadings ranging from .30 to .70 were typical. Therefore, the current study used these guidelines to generate the population models.

Two dimensional CFA models were simulated in samples of 250, 500, and 1000. To examine the influence of the number of items, each model was comprised of 6, 12, or 16 normally distributed variables. In other words, each factor had 3, 6, or 8 indicators. As suggested by DiStefano, the factor loadings ranged from .30 to .70 and item uniquenesses were generated to create indicators with unit variance. Half of the indicators were nonequivalent across groups. Therefore, 3, 6, and 8 items contained DIF in the 6, 12, and 16 item conditions, respectively, and

items with DIF were distributed evenly across the two factors whenever possible. For simplicity, the two factors were orthogonal and the manifest indicators only loaded on a single dimension.

Operationalizations of Small, Medium, and Large DIF

Models were also generated with varying magnitudes of DIF in the factor loadings and intercepts. Specifically, four DIF conditions were simulated: a no DIF condition and small, medium, and large DIF conditions. To operationalize small, medium, and large effects, a literature review was conducted to determine the magnitudes of DIF that have been observed in organizational research. After a careful search, studies were included in this review if they met three primary criteria. First, the study must have been published in one of the top organizational journals. For the purposes of the present analysis, the top journals were identified as the *Journal of Applied Psychology*, *Personnel Psychology*, *Academy of Management Journal*, *Organizational Research Methods*, the *Journal of Management*, and *Organizational Behavior and Human Decision Processes*. Second, the article must have examined a substantive topic—simulation studies were excluded from the review. Finally, the authors must have reported standardized parameter estimates or differences between these parameters in the reference and focal groups. Using these criteria, searches were conducted in the American Psychological Association's PsycINFO database (1887-2010) and Google Scholar. From this search, 522 comparisons were found in 16 separate studies. Unfortunately, item intercepts were reported in only one of these studies and, therefore, were excluded from this review. This finding is consistent with other reviews of the literature showing that scalar invariance is not commonly tested in organizational research (cf. Vandenberg & Lance, 2000).

Figure 1 shows the distribution of differences between the standardized factor loadings in the reference and focal groups. As shown here, the majority of the differences were below .10

and few were greater than .50. Given these results, the lower bounds for the small, medium, and large DIF conditions were operationalized as .10, .20, and .30, respectively. In other words, the factor loadings in the focal group were increased by .10, .20, or .30 depending on the simulated condition. These definitions were selected so that the majority of the items would display only negligible differences while substantially fewer items exhibit medium or large effects. With these values, nearly 60% of the items reported in previous organizational research would be considered negligible (i.e., $< .10$), 20% small, 10% medium, and 10% large. Thus, the lower bound for a medium effect corresponds roughly to the median of the substantively important effects (i.e., small, medium, or large effects). Cutoffs for small and large effects were then chosen to be equidistant from the medium value. Interestingly, these results are consistent with previous simulation research on MACS analyses (e.g., Meade et al., 2008; Stark et al., 2006).

Because intercepts were not reported in the articles examined above, operationalizations of small, medium, and large differences between these parameters in the reference and focal groups were based on previous simulation studies in this area (Meade et al., 2008; Stark et al., 2006). Specifically, DIF was simulated by increasing the intercepts in the focal group by .20 in the small DIF condition, .30 in the medium DIF condition, and .40 in the large DIF condition. Note that the original variables were generated as standard normal variables so DIF in the largest condition accounts for slightly less than half of a standard deviation. Again, the intercepts in the small and large DIF conditions were selected to be equidistant from intercepts in the medium condition.

Analyses

Using the simulated data, values of d_{MACS} were calculated for every item and across all 200 replications in each of the 3 (sample sizes) $\times 3$ (number of items) $\times 4$ (degrees of bias) = 36

conditions and compared to traditional indicators of nonequivalence. Specifically, the effect size of each item was compared to the widely used chi-square difference test and the ΔCFI (cf. Meade et al., 2008). To make these comparisons, we first fit the two-factor configural model to the simulated data and used the results as the baseline for comparing subsequent models. Next, metric and scalar models were fit simultaneously for a single item at a time in each of the 200 replicated samples and the chi-square tests and ΔCFI were calculated. The $\Delta\chi^2$ was evaluated for significance and the ΔCFI was compared to the rule of thumb proposed by Meade et al. (i.e., .002; 2008). Type I error rates for these indices were reflected by the number of items in the no DIF condition that were inaccurately identified as functioning differently across the 200 replicated samples. In contrast, power was reflected by the number of items correctly identified as nonequivalent.

Study 1: Results

Results from the computer simulations are presented in Tables 1 and 2. In Table 1, frequencies for the d_{MACS} index are shown. Cutoffs for interpreting the magnitude of an effect were selected to maximize the number of items with effect sizes that accurately reflected the simulated level of DIF and minimize the number of invariant items with effects incorrectly classified as small, medium, or large. For example, the cutoff for a small effect was selected so that most items in the small DIF condition and few items in the no DIF condition would have effect sizes exceeding this value. Consequently, small DIF consisted of d_{MACS} values between .15 and .30, medium DIF was the interval between .30 and .45, and large DIF was defined as values of .45 and greater.

As illustrated in Table 1, values of d_{MACS} were consistently below the cutoff for a small effect size when no DIF was simulated. Therefore, a cutoff of .15 appears useful for

differentiating between negligible effects and small differences. However, there was some variation in effect sizes when the sample size was small (i.e., $N = 250$) and/or there were few indicators in the model. For example, with a sample size of 250 and only six indicators, 30% of the non-DIF items had effect sizes greater than .15 across the 200 replications. These results are most likely due to poorly estimated model parameters and are reflected in the corresponding conditions in Table 2.

The shaded areas in Table 1 identify the proportion of items with effect sizes that accurately reflected their simulated level of DIF. For example, the shaded area in the medium DIF condition denotes the proportion of items with effect size values greater than .30. As shown in these shaded areas, the d_{MACS} index accurately reflected the magnitude of nonequivalence for the majority of items. In nearly all of the cases, more than 80% of the nonequivalent items were correctly classified as small, medium, or large. Therefore, the guidelines for interpreting the magnitude of an effect appear to adequately differentiate between small, medium, and large effects.

Figures 2–4 illustrate the frequency distributions of d_{MACS} for several selected conditions. With 6, 12, or 16 items and $N = 500$, these figures show that d_{MACS} consistently fell within the range of values that correspond to the simulated level of DIF. In the large DIF condition, effect sizes for most items were greater than .45. In the medium DIF condition, the majority of items were between .30 and .45. Comparable results were obtained in the small DIF condition and when no DIF was present.

The Type I error rates and power for the chi-square difference tests and the ΔCFI are provided in Table 2. As shown here, the chi-square tests performed moderately well. Although power was nearly always greater than .80, the Type I error rates were also higher than the desired

.05 level. In contrast, the Δ CFI displayed good Type I error rates under most conditions. Similar to the results presented in Table 1, Δ CFI displayed high Type I error rates when the sample size was small and/or there were few indicators. In addition, the power of Δ CFI was low in the small DIF condition. Nevertheless, power was consistently high when medium or large levels of DIF were simulated.

CHAPTER 3: A FIELD STUDY OF THE CFA EFFECT SIZE MEASURE

Study 2: Methods and Results

In Study 2, a noncognitive measure was examined to determine the magnitude of nonequivalence in cross-cultural personality data. Using the guidelines developed in Study 1, effect sizes in this data can be interpreted as trivial, small, medium, or large.

The methods and results from this study have been published by the American Psychological Association (see Nye & Drasgow, 2011) and cannot be reproduced here. Therefore, interested readers are referred to the full article in the *Journal of Applied Psychology* available from <http://psycnet.apa.org/index.cfm?fa=browsePA.ofp&jcode=apl>. However, the results for $\Delta r_{XY(CFA)}$ have not been previously published and, therefore, are reported below.

The Effects of Nonequivalence on Δr_{XY}

The effects of nonequivalence were examined across American, Greek, and Chinese samples. Participants in all samples responded to Saucier's Mini-Marker scale (Saucier, 1994) and analyses were conducted on the Extraversion, Agreeableness, and Conscientiousness scales (see Nye & Drasgow, 2011 for further details). The effects of nonequivalence on correlations with external criteria are shown in Table 3. Although two latent factors were estimated for each of the scales examined here, these constructs reflected 2 methodological dimensions rather than substantive differences. Therefore, the effects of nonequivalence on the correlations were calculated by estimating a single factor model for each of the scales. This approach was necessary to obtain estimates of λ_i that could be used to calculate the scale level effects. As described above, calculating the effects of DIF on a correlation requires knowledge of the covariance between the latent trait and the criterion [i.e., $Cov(\xi_R, Y_R)$]. However, because these values were unknown for the present data, the effects of DIF were tested at three different values

of $Cov(\xi_R, Y_R)$. These assumed values are provided in the first column of Table 3. The next column shows the correlation between the personality scale (X) and the criterion (Y) that corresponds to each covariance level in the reference group and the final column illustrates the effects of DIF on this correlation. Negative values suggest that DIF would lower the correlation and positive values indicate a larger relationship between the predictor and criterion in the focal group. As shown in Table 3, DIF had only small effects on observed correlations in most cases. In fact, for the Extraversion scale, DIF did not change the correlation under any of the conditions examined here. In contrast, the largest effects were $-.06$ and $-.07$ for the Agreeableness scale in the Greek sample and the positive Conscientiousness items in the Chinese sample, respectively. Moreover, it appears that the effects of DIF increased as $Cov(\xi_R, Y_R)$ became larger. Nevertheless, only 3 of the 18 effects shown in Table 3 were greater than $.04$, suggesting that DIF in a scale will not have a substantial effect on its correlations with external criteria.

CHAPTER 4: A SIMULATION STUDY OF AN EFFECT SIZE MEASURE FOR IRT ANALYSES OF DIFFERENTIAL ITEM FUNCTIONING

Study 3: Methods

In Study 3, Monte Carlo computer simulations were used to study the effects of sample size, the number of items, and the magnitude of DIF in the population on IRT DIF analyses and values of d_{DIF} . This study consisted of a 3 (sample size) \times 2 (number of items) \times 4 (magnitude of DIF) research design and values of d_{DIF} were compared to traditional IRT indicators of DIF across 100 replications.

Simulated Data

Because of the larger samples required for accurate parameter estimates in IRT analyses, samples of 500, 1000, and 2000 were simulated. In each sample, 20 and 40 dichotomous items were generated from the 3pl IRT model using the 3PLGEN computer program (Stark, 2000). The 3pl parameters for generating these data were obtained from the 2004 Preliminary Scholastic Achievement Test (PSAT). This exam contains approximately 115 total items that assess academic skills and parameters were selected from these items. To examine the effects due to the magnitude of DIF, four levels of nonequivalence were simulated. In one condition, the item parameters used to generate the data were the same in both groups. In other conditions, parameters with small, medium, or large differences between groups were used to generate DIF in a quarter of the items.

Although a literature review similar to the review conducted in Study 1 was also attempted here to operationalize small, medium, and large DIF, only five studies were identified with sufficient information for calculating parameter differences between the reference and focal groups. Therefore, instead of basing the definitions of the magnitude of an effect on a small

sample of studies, we used operationalizations of small, medium, and large effects that were consistent with previous simulation research in this area (cf. Flowers, Oshima, & Raju, 1999; Meade & Lautenschlager, 2004; Meade et al., 2007; Stark et al., 2006). Specifically, small, medium, and large effects were defined by differences in the a-parameter of .20, .30, and .40, respectively. In addition, DIF was also generated by adding .40, .70, or 1.00 to the b-parameters of the reference group for the small, medium, or large DIF conditions, respectively. All nonequivalent items contained differences in both the a- and b-parameters.

Analyses

To evaluate DIF, item parameters for the simulated data from the reference and focal groups were estimated separately using marginal maximum likelihood estimation in BILOG (Mislevy & Bock, 1991). Next, these parameters were equated across groups and DIF was estimated using the ITERLINK computer program (Stark, 2006). This program identifies DIF using the iterative process suggested by Candell and Drasgow (1988). The first step in this approach was to link item parameters using a modified version of the linear equating procedure developed by Stocking and Lord (1983). Next, Lord's (1980) chi-square with a Bonferroni correction was used to identify DIF. Because linking on DIF items can result in errors, items were then relinked using only those items identified as unbiased. Finally, bias was assessed again. The final two steps (i.e., relinking and estimating DIF) were repeated until the same items were identified as nonequivalent in two successive steps.

Because of its popularity, we also assessed DIF using Raju, van der Linden, and Fleer's (1995) DFIT framework. Here, the equated parameters from ITERLINK were used as input to Raju's (1999) DFITD4 computer program for dichotomous items. DIF was identified if values of the NCDIF index were larger than the .006 cutoff suggested by Meade et al. (2007).

Study 3: Results

The results of the Monte Carlo simulations are shown in Tables 4 and 5. Table 4 provides the percent of d_{DIF} values at several cut points and across all conditions and Table 5 shows the Type I error rates and power of the chi-square and NCDIF indices. Consistent with the simulation results presented in Study 1, cutoffs were selected to maximize the number of items that accurately reflected the simulated levels of DIF. As such, a small effect was defined by values of d_{DIF} that were greater than .10 but less than .20, medium effects were represented by values between .20 and .30, and large effects were greater than .30.

Table 4 shows that these cutoff values adequately identified the simulated level of DIF in the majority of cases. For example, when no DIF was present, 80% or more of the simulated items had effect sizes less than .10. In addition, between 75% and 88% of effect sizes were above this level when small amounts of DIF were simulated. In contrast, many of the items in the medium and large DIF conditions had effect sizes greater than .20. In the medium DIF conditions, the majority of items had an effect size between .20 and .30. In the largest DIF conditions, more than 90% of the effect size values were greater than .30 in samples of 1000 or 2000. Although slightly fewer effect sizes were in this range when the sample size was 500, 79% or more were still above .30. Figures 5 and 6 also illustrate the distributions of effect sizes for the 20 and 40 item conditions when $N = 500$. Although the distributions were not as pronounced as those for the CFA simulations, the majority of effect sizes were within the appropriate range of magnitude in the 20-item condition. In addition, d_{DIF} differentiated between effect sizes more clearly in the 40-item condition. As such, values of .10, .20, and .30 appear to be appropriate guidelines for evaluating the magnitude of an effect.

The Type I error rates and power of the traditional indicators of DIF are provided in Table 5. Again, Type I error rates represent the percent of unbiased items that were incorrectly identified as nonequivalent and power was indicated by the percent of DIF items that were correctly identified. As shown in Table 5, both Lord's chi-square and the NCDIF index exhibited low Type I error rates under all conditions. In every condition, fewer than 5% of the equivalent items were identified as nonequivalent. Not surprisingly, results also suggested that power increases with the magnitude of the effect. In the small DIF condition, power was as low as .16. However, power increased and was as high as .97 for the chi-square index in the large DIF condition.

Overall, the NCDIF index displayed relatively low power for detecting an effect. Across all conditions, power for this index ranged from .26 in the small DIF condition to .56 when large amounts of DIF were simulated. These results are likely due to the number of items with simulated DIF and are consistent with previous research in this area. Flowers et al. (1999) found power as low as .50 in some simulated conditions when differences were generated in 20% of the items in a scale. In contrast, these authors found that power was 1.00 when DIF was simulated in only 10% of the items. Thus, DFIT may be most effective when DIF is expected in few of the items being evaluated.

As expected, the chi-square index was substantially affected by sample size. In the small DIF condition, the chi-square index only detected 16% of the nonequivalent items in samples of 500 but detected 74% of the problematic items when $N = 2000$. Although the range of differences between small and large samples decreased as the size of the effect increased, these results suggest that the chi-square index may be problematic for detecting DIF in samples smaller than 1000.

CHAPTER 5: A FIELD STUDY OF THE IRT EFFECT SIZE MEASURE

Study 4: Methods

In Study 4, cognitive ability data from the SAT and AP testing programs were examined. The dichotomous data from each test was fit with the 3pl IRT model and the extent of DIF was evaluated with d_{DIF} . Finally, the effects of DIF on the means and variances of the scales, and their correlations with an external variable, were also examined using the equations described above.

Samples

First, DIF was examined in a sample of 20,000 examinees from the 2007 SAT. Approximately 58% ($N = 11,534$) of the sample was female and 58% ($N = 11,671$) was White. When sample sizes were large enough for accurate parameter estimates, DIF was explored across gender and ethnic groups. Specifically, we examined DIF in the African American ($N = 1,686$), Hispanic (individuals that identified themselves as Mexican or Mexican American, Puerto Rican, or another Hispanic/Latino group; $N = 1,987$), Asian ($N = 3,281$), and female samples.

We also assessed DIF in the 2006 AP World History and Spanish Language exams. Samples of 20,000 examinees were analyzed for both tests. For the World History exam, 56% of the sample was female ($N = 11,147$), 7% African American ($N = 1,427$), 13% Asian ($N = 2,638$), and 14% Hispanic ($N = 2,701$). The sample for the Spanish Language exam was 65% female ($N = 12,930$), 53% Hispanic ($N = 10,690$), and 6% Asian ($N = 1,228$). African Americans were not examined for the Spanish Language test because of the small sample size for this group ($N = 443$).

Analyses

Single factor CFA models were fit to the multiple choice (MC) items in each sample to confirm the unidimensionality of the tests. Models for the AP exams were estimated with maximum likelihood (ML) estimation in LISREL 8.7 (Jöreskog & Sörbom, 1993). In contrast, because of the large number of missing responses to the SAT (the valid N for CFA estimation was 1,769 after listwise deletion), full information maximum likelihood (FIML) estimation was used to model the SAT data. Rather than discarding information due to missing responses, FIML uses all of the available data for estimating item parameters. Therefore, this method is preferable to ML estimation when a substantial amount of data is missing (Newman, 2003).

Next, the 3pl IRT model was fit to the multiple choice items in each data set and parameters were linked using the ITERLINK computer program (Stark, 2006). DIF was then examined across gender and ethnicity using the male and White samples as the reference groups. For each test, values of d_{DIF} were compared to Lord's (1980) chi-square and Raju et al.'s (1995) DFIT framework. Again, significant observed chi-squares and values of NCDIF greater than .006 reflected nonequivalence.

Study 4: Results

CFA analyses confirmed that all three tests were unidimensional (SAT: RMSEA = 0.05, SRMR = 0.044; AP World History: RMSEA = 0.020, SRMR = 0.017; AP Spanish Language: RMSEA = 0.022, SRMR = 0.021) and results from the IRT analyses of the SAT and AP tests are presented for each of the focal groups in Tables 6–8. For each group, the first column identifies the items that displayed significant DIF using Lord's chi-square. The next column shows the NCDIF index and the last column presents effect sizes. Items with NCDIF values greater than

.006 are indicated by asterisks and the shaded cells show the DIF items that were identified by both the chi-square index and NCDIF.

As shown in Table 6, a number of SAT items functioned differently across groups. Although the chi-square identified the largest number of nonequivalent items, the NCDIF measure also suggested that many items displayed DIF. Items from the Mathematics subtests (i.e., M1 to M54) displayed the most DIF and these differences were particularly salient in the female and Asian samples. For example, 18 items showed significant DIF in the Asian sample using both Lord's chi-square and the NCDIF index.

Despite significant DIF, the magnitudes of these differences were generally negligible or small. For example, the majority of the items had effect size values less than .10. Moreover, even when DIF was flagged by both the chi-square and NCDIF measures, the magnitude of the effect was still between .20 and .30 in many conditions. However, several items displayed larger differences between groups. For instance, the effect sizes for items M28 and M52 were .52 and .54, respectively, in the Asian sample.

For the AP World History test in Table 7, few items showed statistically significant DIF across sex and/or ethnic groups. Although Lord's chi-square was significant for a number of items, there was little overlap with the NCDIF measure. Moreover, effect sizes were either negligible or small for all of the items. Overall, the AP World History test displayed the least amount of DIF of the tests examined here.

In contrast, substantial DIF was identified for the items in the AP Spanish Language exam. As with the SAT and World History tests, most of the statistically significant DIF represented either small or negligible effects in the female and Asian samples. However, large effects were observed in the Hispanic group. As shown in Table 8, the largest effect observed in

this sample was .95 for items 11 and 47 and 9 items had effect sizes larger than .60. Thus, the Spanish Language test exhibited the largest differences between groups.

The Effects of IRT DIF on Observed Scale-Level Properties

Tables 9–11 also show the effects of DIF on scale-level properties. In Table 9, the effects of DIF on the mean and variance of the scale are illustrated. Table 11 shows the impact of DIF on the correlation with an external variable. As shown in Table 9, DIF can have substantial effects on the observed mean and variance of a scale. Not surprisingly, the largest effects on the mean were observed in the 2007 SAT when comparing men and women or White and Asian test-takers and in the AP Spanish Language exam when comparing White and Hispanic individuals. In all three cases, DIF influenced the observed mean by more than three points. Consistent with Table 7, the AP World History exam appears to have the smallest effect on the mean of the scale.

The second column in Table 9 provides the observed mean differences between the number-right scores in each of the groups and the third column is the percentage of the observed difference that can be attributed to nonequivalence. Although DIF accounted for less than 5% of the observed difference in several cases, DIF accounted for 192% of the observed difference between men and women on the AP Spanish Language exam. This indicates that the effects of DIF actually reversed the direction of the observed mean difference. Although the reference group (i.e., men) had a higher mean on the latent construct, the effects of DIF caused the observed mean in the female sample to be higher. Thus, DIF would have led to the false conclusion that women performed better on this exam. Nevertheless, the observed mean difference and the impact of DIF were both small.

The fourth column in Table 9 provides the effects of DIF on the variance of the scale. Here, DIF had the largest absolute effect on the variance of the SAT in samples of women and

Asian respondents. However, DIF accounted for the largest percentage of the observed difference in the female samples for the AP exams. For both the World History and Spanish Language tests, DIF accounted for over 100% of the observed difference. In contrast to the results for the mean-level differences, the signs of the Δ Variance and the observed differences for these tests were in opposite directions. Therefore, the observed differences were actually lower than the true differences in the variances of the scales.

For the 2007 SAT, differences between the effect sizes of the items in each of the sub-tests (i.e., Writing, Mathematics, and Reading) were apparent in Table 6. For example, the largest effect sizes for each of the groups were estimated for the items in the Mathematics test (i.e., M1 to M54). Because of these differences, it is possible that the effects of DIF on the means and variances will vary across these sub-tests. Therefore, Table 10 shows the effects of DIF on the sub-test level properties in the SAT. As shown in this table, the largest effects on the mean (i.e., Δ Mean) were observed for the Mathematics sub-test. When comparing men and women, DIF resulted in a higher mean in the male sample. In contrast, Δ Mean was -4.65 when comparing White and Asian test-takers and this difference accounted for 92% of the observed difference. In contrast, DIF had much smaller effects on the means of the Reading and Writing sub-tests. DIF also had substantial effects on the variance of the Mathematics test. However, the largest effects were observed for the Reading test where DIF increased the variances in the African American, Asian, and Hispanic groups by more than 4.

Note that in several cases, DIF accounted for more than 100% of the observed differences in the means and variances of the tests. For example, DIF accounted for 787% of the observed differences between the variances of the Mathematics test in the White and African American group. Consistent with the findings presented in Table 9, the effects of DIF actually reversed the

signs of some differences and resulted in observed differences that were in the opposite direction of the true effect. In those cases, the observed differences were smaller than the effects of DIF making the percentage of the difference accounted for appear substantial.

Finally, Table 11 shows the effects of DIF on the correlation between the scale (X) and an external variable (Y). The first column shows the assumed correlation between item i and variable Y in the reference group. Because this value is unknown, the effects of DIF were tested at three separate correlations: .05, .10, and .15. For simplicity, all items were assumed to have the same relationship to the external variable. The second and third columns show the standard deviation of the scale and the observed correlation in the reference group. The last column provides the effects of DIF on the observed correlation. For nearly all of the tests examined here, the effects of DIF were near zero. The one notable exception was for the AP Spanish Language exam. Although the effects were still small, when r_{iYR} was .10 or .15, the correlation was .05 or .07 higher in the reference group. Thus, the correlation in the focal group would be .45 compared with .52 in the reference group when $r_{iYR} = .15$. These results suggest that DIF can result in moderate changes in the correlation between a measure and an external criterion.

CHAPTER 6: GENERAL DISCUSSION

DIF can have important consequences for both research and practice. It can affect employee selection decisions, academic admissions, or may lead to inaccurate conclusions about group differences more generally. Consequently, methods for examining DIF have garnered a substantial amount of attention in psychological research. Still, a number of questions remain. How large are the effects? What are the consequences for conclusions about group differences? How will DIF affect correlations with external variables? The present study attempts to answer some of these questions and to provide researchers and practitioners with the tools to further understand measurement nonequivalence.

Although most methods for identifying DIF rely on statistical significance tests or empirically derived cutoff values, these methods have a number of disadvantages. One way to compensate for these limitations is to use effect size measures (Cohen, 1988). However, measures of effect magnitude have not been available for IRT or CFA studies of measurement equivalence at the item level. Therefore, one of the primary contributions of the present research was the development of the d_{DIF} and d_{MACS} indices. As shown in the studies presented here, these indices provide researchers with additional information about their results when used in combination with more traditional indicators of DIF. Most importantly, d_{DIF} and d_{MACS} are useful for differentiating between statistical significance and practical importance.

The current work also provides researchers and practitioners with formulas for evaluating the consequences of DIF on the total scale or test scores. Current recommendations for dealing with nonequivalence suggest that problematic items be removed from the scale in order to justify group-level comparisons (e.g., Raju et al., 1995). However, this practice has important disadvantages. First, this approach will be problematic when all or most of the items in a scale

are nonequivalent. In this situation, too few items may remain to justify comparisons. Concomitantly, removing items may sacrifice content coverage and limit the validity of a scale. Therefore, another approach is needed that allows researchers to make comparisons without removing a substantial number of items. Because observed differences between groups are equal to the sum of DIF and impact, differences between means can be corrected when the effects of DIF on the mean of the scale are known. Once these effects are eliminated, only impact remains and groups can be compared to identify substantively meaningful differences. In this regard, ΔMean can be used to calculate these effects and the necessary corrections can be made. This is important because, as shown in Table 9, the effects of DIF may account for a substantial portion of the observed group-level differences.

The effect size index proposed here might also be used to identify the items with the largest effects for removal. The results of the field studies showed that a range of DIF can be found within a single scale. For example, the Extraversion scale in Study 2 contained items with both small ($d_{MACS} = .26$) and substantial ($d_{MACS} = 1.11$) nonequivalence (see Nye & Drasgow, 2011). In the IRT analyses presented in Study 4, substantial DIF was identified when White and Hispanic test-takers were compared on a subset of the items in the AP Spanish Language exam (e.g., $d_{DIF} = .95$). By removing only those items with the largest effects, enough items may remain for accurate comparisons. Indeed, this was the case for the AP Spanish Language exam. Nevertheless, the utility of this approach is predicated on the magnitude of DIF for the items in a scale. If the majority of the items display substantial DIF, this approach may be unsatisfactory.

The present work also makes a substantial contribution to the literature by using simulation research to suggest guidelines for interpreting effects as small, medium, or large. Cohen's (1988) guidelines are useful for interpreting the magnitude of an effect for mean-level

differences. Thus, the present work advocates similar guidelines that can be used to evaluate the magnitude of nonequivalence and will facilitate the interpretation of this artifact. However, in contrast to Cohen's guidelines, which were developed based on an arbitrary decision process (Cohen, 1988, 1992), the present research developed guidelines empirically using simulation studies and operationalizations of effect magnitudes that were based on a review of previous research. Although subjective judgments were still necessary in this process, the guidelines developed from this research provide rough criteria that evaluate effects relative to the magnitudes of DIF encountered in past organizational research.

It is important to emphasize that these guidelines should not be used in the same way that p-values have been used with significance testing. As described by Cortina and Landis (2011), one disadvantage of effect size measures is that they are often used in this way. In other words, effect sizes greater than .20 are often considered important while anything less than that is generally ignored as inconsequential. Again, this practice translates a decision continuum into a dichotomous outcome. Instead, effect size measures should be used to compliment significance testing and provide additional information about the magnitude of a particular finding on a continuum. With this approach, effect size indices can mitigate many of the limitations of significance testing and provide a more comprehensive picture of research outcomes.

Limitations

As with all empirical work, the current research has limitations. First, the effect size index proposed for MACS analyses is only applicable to items with a single loading on one latent factor (i.e., models with simple structure). Items with secondary loadings on another latent factor cannot be accurately evaluated using d_{MACS} because additional loadings will influence the mean predicted response and necessitate a more complicated formula for estimating the

magnitude of an effect. Similarly, as with most IRT models, the d_{DIF} index is limited to unidimensional scales. Although multidimensional IRT models are available, these techniques are less well-developed and, similar to the d_{MACS} index, would require more complicated formulas for estimating effect sizes. However, note that these limitations do *not* preclude analyses on multidimensional *scales* in the CFA framework when each item loads on only one factor. The limitation described for d_{MACS} only applies to single *items* that load on multiple latent factors. In the present work, the scales examined in Study 2 were all multidimensional. Nevertheless, CFA models were estimated so that each of the items in the scale only loaded on one of the latent factors. As a result, the CFA effect size measure proposed is still applicable and Study 2 demonstrates its use with complex survey data.

Both indices described here are also limited to tests and scales with at least one unbiased referent item. Scales that are completely nonequivalent do not have an unbiased baseline. For this reason, we could not obtain accurate estimates of the effect sizes for all of the scales in Study 2. Instead, we provide an example to illustrate the importance of using an item that is invariant across groups (see Nye & Drasgow, 2011). Although this was not an issue for any of the tests that were examined in the IRT analyses, the application of the d_{DIF} index also requires at least one non-DIF item for linking parameters between the two groups. Nevertheless, this is a problem for all methods of assessing DIF (Candell & Drasgow, 1988; Vandenberg & Lance, 2000). As a result, some research has examined techniques for identifying an equivalent referent item (Candell & Drasgow, 1988; Cheung & Rensvold, 1999; Hernandez, Chernyshenko, Stark, & Drasgow, 2008). However, if these methods are used and an equivalent referent item cannot be found, DIF analyses cannot proceed.

Finally, the effect size indices identified here will be severely affected by the accuracy of the parameter estimates. Because item parameters are used as input to calculate the differences between groups, inaccurate estimates will confound interpretations of the effect magnitude. Consequently, any factors that affects the IRT or CFA parameter estimates will also influence the calculation of effect sizes. For example, parameter estimates can be inaccurate and have large standard errors when the sample size is small (Boomsma, 1982; Hoogland & Boomsma, 1998; Lei & Lomax, 2005; Raju et al., 1995; Stark et al., 2006). As a result, the indices presented here should be evaluated critically whenever the sample size is small and/or the standard errors of the parameters are large. However, it should be noted that this is a limitation of SEM and IRT models more generally and is not specific to the indices developed here.

Conclusion

Scales and tests are used as measures of the latent constructs fundamental to applied psychology. Therefore, analyses of measurement equivalence are necessary for establishing the quality of a scale and for drawing accurate conclusions when data from two or more groups are collected. The present research proposes several new tools for examining these measurement properties and providing researchers and practitioners with important information. Using this information, decisions can be made about the appropriate course of action when nonequivalence is found on a scale or test. Thus, the effect size measures described here can complement the statistical significance tests and empirically derived rules of thumb that are commonly used and yet provide only limited information about measurement instruments.

TABLES AND FIGURES

Table 1

Frequencies of d_{MACS} in the Small, Medium, Large, and No DIF MACS Conditions

No DIF Condition					
# of Items	N	Cutoff Values ^a			
			Small	Medium	Large
		$\leq .15$	$\geq .15$	$\geq .30$	$\geq .45$
6	250	0.70	0.30	0.08	0.04
	500	0.87	0.14	0.02	0.01
	1000	0.94	0.06	0.01	0.00
12	250	0.85	0.15	0.01	0.00
	500	0.97	0.03	0.00	0.00
	1000	1.00	0.00	0.00	0.00
16	250	0.92	0.08	0.00	0.00
	500	0.98	0.02	0.00	0.00
	1000	1.00	0.00	0.00	0.00
Small DIF Condition ($\Delta\lambda = .10, \Delta\tau = .20$)					
# of Items	N	Cutoff Values ^a			
			Small	Medium	Large
		$\leq .15$	$\geq .15$	$\geq .30$	$\geq .45$
6	250	0.20	0.80	0.30	0.08
	500	0.18	0.82	0.20	0.01
	1000	0.12	0.88	0.14	0.00
12	250	0.21	0.79	0.23	0.01
	500	0.17	0.83	0.14	0.00
	1000	0.14	0.86	0.07	0.00
16	250	0.17	0.83	0.23	0.01
	500	0.10	0.90	0.14	0.00
	1000	0.04	0.96	0.07	0.00
Medium DIF Condition ($\Delta\lambda = .20, \Delta\tau = .30$)					
# of Items	N	Cutoff Values ^a			
			Small	Medium	Large
		$\leq .15$	$\geq .15$	$\geq .30$	$\geq .45$
6	250	0.03	0.97	0.76	0.29
	500	0.02	0.98	0.81	0.22
	1000	0.00	1.00	0.91	0.15
12	250	0.01	0.99	0.84	0.32
	500	0.00	1.00	0.91	0.25
	1000	0.00	1.00	0.95	0.16

Table 1 (cont.)

Medium DIF Condition ($\Delta\lambda = .20, \Delta\tau = .30$)					
		Cutoff Values ^a			
			Small	Medium	Large
# of Items	N	$\leq .15$	$\geq .15$	$\geq .30$	$\geq .45$
16	250	0.01	0.99	0.88	0.30
	500	0.00	1.00	0.93	0.22
	1000	0.00	1.00	0.97	0.18
Large DIF Condition ($\Delta\lambda = .30, \Delta\tau = .40$)					
		Cutoff Values ^a			
			Small	Medium	Large
# of Items	N	$\leq .15$	$\geq .15$	$\geq .30$	$\geq .45$
6	250	0.01	0.99	0.94	0.77
	500	0.00	1.00	0.91	0.83
	1000	0.00	1.00	1.00	0.93
12	250	0.00	1.00	0.99	0.87
	500	0.00	1.00	1.00	0.93
	1000	0.00	1.00	1.00	0.98
16	250	0.00	1.00	1.00	0.90
	500	0.00	1.00	1.00	0.96
	1000	0.00	1.00	1.00	0.98

Note: N = sample size. ^a Cutoff values are provided for small, medium, and large effects.

Table 2

Type I Error Rates and Power for $\Delta\chi^2$ and ΔCFI in the MACS Simulations

# of Items	N	Simulated Levels of DIF							
		No DIF		Small		Medium		Large	
		$\Delta\chi^2$	ΔCFI^a	$\Delta\chi^2$	ΔCFI^a	$\Delta\chi^2$	ΔCFI^a	$\Delta\chi^2$	ΔCFI^a
6	250	0.08	0.17	0.58	0.63	0.82	0.80	0.88	0.88
	500	0.08	0.10	0.81	0.79	0.97	0.96	1.00	0.97
	1000	0.10	0.06	0.97	0.91	0.96	0.99	1.00	1.00
12	250	0.12	0.06	0.74	0.49	0.95	0.78	0.99	0.91
	500	0.13	0.01	0.91	0.53	1.00	0.88	1.00	0.98
	1000	0.12	0.00	0.99	0.59	1.00	0.96	1.00	1.00
16	250	0.14	0.02	0.83	0.39	0.99	0.73	1.00	0.93
	500	0.18	0.01	0.96	0.37	1.00	0.81	1.00	0.98
	1000	0.15	0.00	1.00	0.41	1.00	0.93	1.00	1.00

Note: N = sample size. ^a ΔCFI greater than .002 were used to identify nonequivalence.

Table 3

Effects of CFA Nonequivalence on Scale-Level Correlations with External Criteria

	$Cov(\xi_R, Y_R)$	$r_{XYR(CFA)}$	$\Delta r_{XY(CFA)}$
Extraversion (Greek)	.10	.12	.00
	.15	.19	.00
	.25	.31	.00
Extraversion (Chinese)	.10	.12	.00
	.15	.19	.00
	.25	.31	.00
Agreeableness (Greek)	.05	.10	-.02
	.10	.20	-.04
	.15	.30	-.06
Agreeableness (Chinese)	.05	.10	-.01
	.10	.20	-.03
	.15	.30	-.04
Positive Conscientiousness Items (Greek)	.05	.13	.01
	.10	.25	.03
	.15	.37	.04
Positive Conscientiousness Items (Chinese)	.05	.13	-.02
	.10	.25	-.05
	.15	.37	-.07

Note: $Cov(\xi_R, Y_R)$ = the assumed covariance of the latent trait and the criterion Y in the reference group. $r_{XYR(CFA)}$ = the correlation between the scale and an external criterion given $Cov(\xi_R, Y_R)$. $\Delta r_{XY(CFA)}$ = the change in the correlation due to nonequivalence.

Table 4

Frequencies of d_{DIF} in the Small, Medium, Large, and No DIF IRT Conditions

No DIF Condition					
# of Items	N	Cutoff Values ^a			
			Small	Medium	Large
		$\leq .10$	$\geq .10$	$\geq .20$	$\geq .30$
20	500	0.81	0.19	0.01	0.00
	1000	0.93	0.07	0.00	0.00
	2000	0.98	0.02	0.00	0.00
40	500	0.80	0.20	0.02	0.00
	1000	0.94	0.06	0.00	0.00
	2000	0.98	0.02	0.00	0.00
Small DIF Condition ($\Delta a = .20, \Delta b = .40$)					
# of Items	N	Cutoff Values ^a			
			Small	Medium	Large
		$\leq .10$	$\geq .10$	$\geq .20$	$\geq .30$
20	500	0.20	0.80	0.36	0.12
	1000	0.20	0.80	0.41	0.11
	2000	0.25	0.75	0.42	0.12
40	500	0.21	0.79	0.37	0.10
	1000	0.18	0.82	0.41	0.11
	2000	0.12	0.88	0.45	0.07
Medium DIF Condition ($\Delta a = .30, \Delta b = .70$)					
# of Items	N	Cutoff Values ^a			
			Small	Medium	Large
		$\leq .10$	$\geq .10$	$\geq .20$	$\geq .30$
20	500	0.05	0.95	0.76	0.49
	1000	0.02	0.98	0.80	0.55
	2000	0.00	1.00	0.88	0.52
40	500	0.01	0.99	0.82	0.48
	1000	0.01	0.99	0.88	0.56
	2000	0.00	1.00	0.95	0.57

Table 4 (cont.)

Large DIF Condition ($\Delta a = .40$, $\Delta b = 1.00$)					
# of Items	N	Cutoff Values ^a			
			Small	Medium	Large
		$\leq .10$	$\geq .10$	$\geq .20$	$\geq .30$
20	500	0.00	1.00	0.97	0.79
	1000	0.00	1.00	1.00	0.96
	2000	0.00	1.00	1.00	0.95
40	500	0.00	1.00	0.99	0.86
	1000	0.00	1.00	1.00	0.94
	2000	0.00	1.00	1.00	0.99

Note: N = sample size. ^a Cutoff values are provided for small, medium, and large effects.

Table 5

Type I Error Rates and Power for χ^2 and NCDIF in the IRT Simulations

# of Items	N	Simulated Levels of DIF							
		No DIF		Small		Medium		Large	
		χ^2	NCDIF ^a	χ^2	NCDIF ^a	χ^2	NCDIF ^a	χ^2	NCDIF ^a
20	500	0.01	0.00	0.20	0.26	0.45	0.42	0.68	0.49
	1000	0.02	0.00	0.37	0.31	0.69	0.41	0.97	0.56
	2000	0.03	0.00	0.54	0.35	0.81	0.40	0.91	0.55
40	500	0.01	0.00	0.16	0.26	0.61	0.39	0.85	0.42
	1000	0.02	0.00	0.47	0.30	0.85	0.40	0.96	0.47
	2000	0.02	0.00	0.74	0.35	0.90	0.40	0.94	0.49

Note: All chi-square values are Lord's (1980) chi-square. N = sample size. ^a A cutoff of .006 was used to identify nonequivalence with the NCDIF index.

Table 6
DIF Results for the 2007 SAT

Items	χ^2	Women		African American			χ^2	Asian		χ^2	Hispanic	
		NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}		NCDIF	d_{DIF}		NCDIF	d_{DIF}
W1	NO	0.000	0.04	NO	0.000	0.10	YES	0.000	0.11	NO	0.000	0.02
W2	YES	0.002	0.14	NO	0.000	0.01	NO	0.000	0.05	YES	0.001	0.06
W3	NO	0.000	0.05	NO	0.000	0.08	YES	0.005	0.26	NO	0.000	0.02
W4	YES	0.001	0.17	YES	0.000	0.13	NO	0.000	0.01	NO	0.000	0.05
W5	YES	0.000	0.06	YES	0.000	0.08	YES	0.001	0.10	NO	0.001	0.04
W6	YES	0.001	0.12	NO	0.000	0.03	YES	0.006	0.20	YES	0.000	0.07
W7	YES	0.001	0.11	NO	0.000	0.04	NO	0.000	0.06	YES	0.000	0.08
W8	YES	0.001	0.08	YES	0.000	0.07	YES	0.003	0.19	YES	0.001	0.11
W9	NO	0.000	0.07	NO	0.000	0.04	NO	0.001	0.07	NO	0.000	0.04
W10	YES	0.000	0.05	YES	0.001	0.07	NO	0.000	0.03	YES	0.001	0.09
W11	NO	0.000	0.02	NO	0.000	0.03	YES	0.005	0.19	YES	0.001	0.09
W12	NO	0.000	0.03	NO	0.000	0.09	NO	0.000	0.03	NO	0.000	0.08
W13	NO	0.000	0.04	NO	0.000	0.02	YES	0.003	0.23	NO	0.000	0.02
W14	NO	0.000	0.03	NO	0.002	0.06	NO	0.003	0.10	YES	0.001	0.09
W15	NO	0.000	0.05	YES	0.001	0.06	YES	0.000	0.12	YES	0.002	0.05
W16	YES	0.001	0.05	YES	0.002	0.13	YES	0.003	0.13	NO	0.000	0.05
W17	NO	0.000	0.04	YES	0.007*	0.11	YES	0.013*	0.47	YES	0.011*	0.27
W18	NO	0.000	0.02	YES	0.001	0.12	YES	0.005	0.21	YES	0.001	0.09
W19	YES	0.001	0.07	YES	0.000	0.10	YES	0.002	0.15	YES	0.001	0.11
W20	YES	0.002	0.15	NO	0.001	0.08	YES	0.007*	0.20	YES	0.001	0.12
W21	YES	0.001	0.06	NO	0.003	0.10	YES	0.014*	0.27	YES	0.001	0.07
W22	NO	0.001	0.11	NO	0.001	0.09	NO	0.008*	0.24	NO	0.005	0.18
W23	NO	0.000	0.01	NO	0.001	0.09	NO	0.002	0.10	NO	0.000	0.04
W24	YES	0.001	0.08	YES	0.001	0.09	YES	0.003	0.13	YES	0.001	0.07
W25	YES	0.001	0.06	NO	0.001	0.05	YES	0.010	0.26	YES	0.001	0.08
W26	NO	0.000	0.03	YES	0.002	0.12	YES	0.000	0.05	YES	0.000	0.07
W27	NO	0.000	0.04	NO	0.001	0.09	NO	0.000	0.06	NO	0.001	0.10
W28	NO	0.000	0.03	NO	0.001	0.04	NO	0.013*	0.23	YES	0.001	0.11
W29	YES	0.000	0.05	YES	0.000	0.15	YES	0.001	0.12	YES	0.000	0.09
W30	YES	0.000	0.05	YES	0.001	0.11	NO	0.000	0.06	NO	0.000	0.05
W31	YES	0.000	0.04	YES	0.002	0.12	YES	0.002	0.11	YES	0.000	0.05
W32	NO	0.000	0.04	NO	0.001	0.06	NO	0.001	0.05	YES	0.001	0.12
W33	NO	0.001	0.05	NO	0.000	0.04	YES	0.001	0.08	YES	0.001	0.04
W34	NO	0.001	0.13	NO	0.000	0.06	NO	0.004	0.20	NO	0.001	0.10
W35	YES	0.000	0.09	NO	0.000	0.01	NO	0.000	0.08	NO	0.000	0.08
W36	NO	0.001	0.03	NO	0.000	0.09	YES	0.002	0.14	YES	0.000	0.13
W37	NO	0.000	0.01	YES	0.000	0.12	YES	0.000	0.08	NO	0.000	0.05
W38	NO	0.000	0.02	NO	0.000	0.08	YES	0.001	0.12	NO	0.000	0.04
W39	YES	0.001	0.04	YES	0.000	0.08	YES	0.006	0.22	NO	0.000	0.04
W40	NO	0.000	0.01	YES	0.000	0.08	YES	0.002	0.12	NO	0.000	0.03
W41	YES	0.001	0.07	NO	0.001	0.05	YES	0.001	0.10	NO	0.001	0.03
W42	YES	0.002	0.11	YES	0.001	0.14	NO	0.001	0.10	YES	0.001	0.12

Table 6 (cont.)

Items	<u>Women</u>			<u>African American</u>			<u>Asian</u>			<u>Hispanic</u>		
	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}
W43	YES	0.001	0.14	NO	0.001	0.04	YES	0.002	0.11	NO	0.000	0.04
W44	NO	0.000	0.03	NO	0.002	0.17	NO	0.002	0.09	NO	0.000	0.02
W45	YES	0.004	0.13	NO	0.000	0.06	YES	0.000	0.05	NO	0.000	0.06
W46	NO	0.001	0.07	NO	0.000	0.03	NO	0.002	0.10	NO	0.001	0.06
W47	YES	0.000	0.03	NO	0.001	0.05	YES	0.001	0.06	NO	0.000	0.03
W48	NO	0.000	0.06	YES	0.005	0.15	YES	0.003	0.14	NO	0.000	0.06
W49	YES	0.000	0.07	NO	0.001	0.06	NO	0.000	0.03	NO	0.000	0.03
M1	NO	0.000	0.02	NO	0.000	0.02	NO	0.000	0.02	NO	0.000	0.03
M2	NO	0.000	0.06	NO	0.000	0.03	NO	0.000	0.05	YES	0.000	0.08
M3	NO	0.000	0.08	NO	0.000	0.02	NO	0.000	0.02	NO	0.000	0.02
M4	YES	0.002	0.19	YES	0.002	0.15	NO	0.000	0.05	NO	0.000	0.05
M5	YES	0.001	0.12	YES	0.003	0.14	NO	0.000	0.06	NO	0.001	0.05
M6	YES	0.005	0.23	NO	0.001	0.07	YES	0.014*	0.34	NO	0.001	0.04
M7	YES	0.008	0.26	YES	0.005	0.16	YES	0.001	0.13	YES	0.006	0.19
M8	YES	0.001	0.14	NO	0.000	0.05	YES	0.010*	0.27	NO	0.001	0.11
M9	YES	0.003	0.20	NO	0.000	0.02	YES	0.012*	0.30	NO	0.001	0.06
M10	YES	0.002	0.14	NO	0.000	0.06	YES	0.002	0.12	NO	0.000	0.02
M11	YES	0.004	0.21	NO	0.000	0.02	YES	0.004	0.18	YES	0.001	0.11
M12	YES	0.006	0.23	YES	0.003	0.11	YES	0.005	0.20	YES	0.000	0.06
M13	YES	0.005	0.21	YES	0.001	0.11	YES	0.001	0.09	YES	0.001	0.07
M14	YES	0.008*	0.26	NO	0.000	0.05	YES	0.014*	0.34	NO	0.000	0.04
M15	YES	0.002	0.17	NO	0.000	0.03	YES	0.018*	0.37	YES	0.000	0.04
M16	YES	0.007*	0.23	NO	0.007*	0.17	YES	0.020*	0.38	NO	0.000	0.04
M17	YES	0.007*	0.23	NO	0.001	0.06	YES	0.016*	0.36	YES	0.001	0.08
M18	YES	0.005	0.20	NO	0.006	0.17	YES	0.003	0.20	YES	0.002	0.10
M19	NO	0.006	0.21	NO	0.002	0.09	YES	0.012*	0.31	NO	0.000	0.04
M20	NO	0.001	0.09	YES	0.009*	0.18	NO	0.006	0.22	NO	0.000	0.05
M21	NO	0.000	0.06	YES	0.000	0.08	YES	0.002	0.16	YES	0.001	0.08
M22	YES	0.002	0.16	NO	0.000	0.10	YES	0.001	0.14	NO	0.001	0.10
M23	YES	0.001	0.10	NO	0.001	0.06	YES	0.003	0.16	NO	0.000	0.02
M24	YES	0.006	0.28	NO	0.001	0.07	YES	0.001	0.14	YES	0.001	0.08
M25	YES	0.004	0.18	YES	0.001	0.10	YES	0.004	0.19	NO	0.000	0.02
M26	YES	0.001	0.11	NO	0.000	0.03	YES	0.004	0.23	NO	0.001	0.04
M27	YES	0.005	0.20	NO	0.005	0.17	YES	0.001	0.14	NO	0.001	0.07
M28	YES	0.008*	0.23	NO	0.001	0.10	YES	0.038*	0.52	NO	0.001	0.09
M29	YES	0.001	0.11	YES	0.001	0.07	YES	0.005	0.21	NO	0.000	0.06
M30	YES	0.001	0.08	NO	0.000	0.05	YES	0.002	0.14	NO	0.002	0.07
M31	YES	0.004	0.20	YES	0.005	0.25	YES	0.006	0.24	NO	0.000	0.06
M32	YES	0.010*	0.32	YES	0.002	0.15	YES	0.004	0.18	YES	0.002	0.13
M33	YES	0.003	0.15	NO	0.001	0.07	YES	0.016*	0.36	NO	0.000	0.01
M34	YES	0.001	0.09	NO	0.000	0.01	YES	0.012*	0.31	NO	0.000	0.03
M35	YES	0.009*	0.29	YES	0.002	0.11	YES	0.006	0.20	YES	0.002	0.11
M36	YES	0.016*	0.37	YES	0.012*	0.26	YES	0.012*	0.30	YES	0.001	0.09
M37	YES	0.001	0.09	NO	0.000	0.05	YES	0.005	0.20	NO	0.000	0.03

Table 6 (cont.)

Items	<u>Women</u>			<u>African American</u>			<u>Asian</u>			<u>Hispanic</u>		
	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}
M38	YES	0.007*	0.25	NO	0.003	0.13	YES	0.015*	0.32	NO	0.000	0.05
M39	YES	0.003	0.25	YES	0.001	0.19	YES	0.001	0.09	NO	0.000	0.05
M40	YES	0.002	0.14	YES	0.001	0.18	YES	0.003	0.19	NO	0.000	0.01
M41	YES	0.002	0.17	NO	0.000	0.03	YES	0.001	0.17	NO	0.000	0.05
M42	YES	0.001	0.08	YES	0.000	0.06	YES	0.001	0.07	NO	0.000	0.06
M43	YES	0.002	0.17	NO	0.001	0.05	YES	0.002	0.15	NO	0.000	0.07
M44	YES	0.001	0.14	NO	0.000	0.06	NO	0.000	0.07	YES	0.000	0.17
M45	YES	0.002	0.10	YES	0.001	0.10	NO	0.000	0.01	YES	0.001	0.13
M46	YES	0.003	0.23	YES	0.004	0.15	YES	0.002	0.19	YES	0.001	0.08
M47	YES	0.000	0.07	NO	0.000	0.03	YES	0.007*	0.26	NO	0.000	0.02
M48	YES	0.005	0.20	NO	0.000	0.04	YES	0.018*	0.38	NO	0.000	0.02
M49	YES	0.001	0.13	YES	0.000	0.08	YES	0.003	0.13	NO	0.000	0.07
M50	YES	0.005	0.21	NO	0.001	0.08	YES	0.012*	0.32	NO	0.000	0.04
M51	YES	0.010*	0.27	YES	0.003	0.10	YES	0.001	0.09	YES	0.002	0.09
M52	YES	0.003	0.17	NO	0.002	0.09	YES	0.034*	0.54	NO	0.000	0.05
M53	YES	0.003	0.15	NO	0.000	0.05	YES	0.025*	0.43	NO	0.000	0.05
M54	NO	0.002	0.10	NO	0.000	0.07	NO	0.029*	0.49	YES	0.001	0.10
CR1	NO	0.000	0.07	NO	0.000	0.09	NO	0.000	0.08	NO	0.000	0.08
CR2	YES	0.001	0.14	YES	0.000	0.07	YES	0.005	0.31	NO	0.000	0.04
CR3	YES	0.000	0.05	NO	0.000	0.05	NO	0.000	0.10	NO	0.000	0.03
CR4	YES	0.001	0.10	YES	0.001	0.14	YES	0.003	0.15	NO	0.000	0.01
CR5	YES	0.001	0.10	YES	0.002	0.13	YES	0.017*	0.32	YES	0.014*	0.33
CR6	YES	0.001	0.06	NO	0.001	0.04	YES	0.001	0.07	NO	0.000	0.03
CR7	YES	0.001	0.08	NO	0.001	0.07	YES	0.002	0.10	YES	0.003	0.14
CR8	NO	0.001	0.09	NO	0.000	0.03	YES	0.005	0.23	NO	0.001	0.09
CR9	NO	0.001	0.06	NO	0.000	0.08	NO	0.001	0.06	NO	0.001	0.15
CR10	YES	0.001	0.10	NO	0.000	0.05	YES	0.004	0.19	YES	0.000	0.07
CR11	YES	0.001	0.11	YES	0.000	0.08	YES	0.007*	0.25	YES	0.016*	0.41
CR12	NO	0.001	0.12	NO	0.000	0.02	NO	0.001	0.11	NO	0.001	0.07
CR13	NO	0.000	0.02	NO	0.000	0.06	NO	0.001	0.07	YES	0.000	0.08
CR14	NO	0.000	0.02	NO	0.001	0.09	NO	0.000	0.03	NO	0.001	0.06
CR15	NO	0.000	0.02	NO	0.000	0.05	NO	0.001	0.04	NO	0.000	0.04
CR16	YES	0.000	0.05	YES	0.000	0.10	YES	0.004	0.15	YES	0.001	0.12
CR17	NO	0.000	0.03	NO	0.000	0.04	NO	0.000	0.04	NO	0.000	0.04
CR18	NO	0.000	0.03	YES	0.000	0.06	YES	0.000	0.06	NO	0.000	0.02
CR19	YES	0.000	0.06	NO	0.000	0.02	YES	0.000	0.09	YES	0.000	0.05
CR20	NO	0.000	0.07	NO	0.000	0.05	YES	0.002	0.19	NO	0.001	0.05
CR21	NO	0.000	0.04	YES	0.001	0.10	YES	0.000	0.06	YES	0.001	0.11
CR22	NO	0.000	0.02	NO	0.000	0.04	YES	0.000	0.09	NO	0.000	0.02
CR23	YES	0.001	0.10	NO	0.000	0.05	YES	0.002	0.16	NO	0.000	0.01
CR24	NO	0.000	0.03	NO	0.000	0.02	NO	0.000	0.05	NO	0.000	0.03
CR25	NO	0.000	0.04	NO	0.000	0.04	NO	0.000	0.06	NO	0.001	0.02
CR26	YES	0.000	0.03	NO	0.000	0.02	YES	0.003	0.21	YES	0.005	0.29
CR27	YES	0.002	0.10	NO	0.000	0.04	YES	0.002	0.13	NO	0.000	0.05

Table 6 (cont.)

Items	<u>Women</u>			<u>African American</u>			<u>Asian</u>			<u>Hispanic</u>		
	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	Items	χ^2
CR28	NO	0.000	0.03	YES	0.001	0.10	YES	0.005	0.19	YES	0.003	0.13
CR29	YES	0.001	0.06	YES	0.000	0.06	YES	0.002	0.11	NO	0.001	0.14
CR30	YES	0.001	0.11	NO	0.001	0.08	YES	0.006	0.27	YES	0.001	0.12
CR31	YES	0.001	0.09	NO	0.000	0.02	NO	0.000	0.05	NO	0.000	0.05
CR32	YES	0.003	0.15	NO	0.000	0.05	NO	0.000	0.03	YES	0.003	0.14
CR33	NO	0.001	0.07	YES	0.001	0.07	NO	0.000	0.03	NO	0.000	0.05
CR34	YES	0.001	0.09	NO	0.000	0.09	YES	0.001	0.15	NO	0.000	0.05
CR35	NO	0.000	0.04	NO	0.000	0.11	YES	0.002	0.17	YES	0.001	0.18
CR36	YES	0.001	0.06	NO	0.001	0.08	NO	0.000	0.06	YES	0.001	0.09
CR37	NO	0.000	0.01	NO	0.000	0.04	NO	0.000	0.10	NO	0.000	0.04
CR38	YES	0.000	0.07	YES	0.001	0.15	YES	0.004	0.23	YES	0.000	0.13
CR39	YES	0.000	0.07	NO	0.000	0.02	YES	0.004	0.16	NO	0.001	0.04
CR40	NO	0.001	0.12	NO	0.000	0.05	NO	0.001	0.08	NO	0.000	0.04
CR41	YES	0.001	0.10	NO	0.001	0.05	YES	0.004	0.17	YES	0.001	0.08
CR42	YES	0.001	0.11	YES	0.001	0.10	YES	0.010*	0.28	YES	0.005	0.24
CR43	YES	0.001	0.09	YES	0.001	0.09	YES	0.001	0.10	NO	0.000	0.04
CR44	NO	0.000	0.02	NO	0.000	0.01	YES	0.002	0.17	NO	0.000	0.01
CR45	YES	0.002	0.17	NO	0.000	0.06	YES	0.003	0.11	YES	0.000	0.06
CR46	YES	0.001	0.13	NO	0.000	0.07	YES	0.002	0.12	NO	0.000	0.03
CR47	YES	0.001	0.11	YES	0.001	0.12	NO	0.000	0.04	NO	0.000	0.05
CR48	YES	0.001	0.07	YES	0.000	0.11	NO	0.001	0.09	YES	0.000	0.08
CR49	NO	0.001	0.03	YES	0.003	0.19	NO	0.000	0.03	NO	0.000	0.05
CR50	NO	0.000	0.03	NO	0.000	0.06	YES	0.002	0.18	NO	0.000	0.04
CR51	YES	0.001	0.08	YES	0.000	0.12	NO	0.000	0.04	YES	0.002	0.19
CR52	YES	0.000	0.09	NO	0.000	0.05	YES	0.001	0.19	NO	0.000	0.01
CR53	YES	0.002	0.10	YES	0.006	0.17	YES	0.007*	0.21	YES	0.004	0.16
CR54	YES	0.003	0.20	NO	0.000	0.04	YES	0.007*	0.21	YES	0.002	0.13
CR55	NO	0.006	0.23	YES	0.002	0.14	NO	0.000	0.10	NO	0.000	0.04
CR56	NO	0.000	0.01	NO	0.000	0.02	NO	0.001	0.08	NO	0.001	0.05
CR57	YES	0.000	0.06	NO	0.000	0.05	NO	0.000	0.04	YES	0.001	0.07
CR58	YES	0.000	0.07	NO	0.001	0.07	YES	0.001	0.08	NO	0.000	0.01
CR59	NO	0.000	0.03	NO	0.000	0.02	YES	0.001	0.18	YES	0.001	0.21
CR60	NO	0.000	0.05	NO	0.001	0.05	NO	0.001	0.06	NO	0.001	0.05
CR61	NO	0.001	0.07	NO	0.000	0.03	NO	0.000	0.03	YES	0.000	0.03
CR62	NO	0.000	0.02	NO	0.001	0.06	NO	0.000	0.03	NO	0.001	0.03
CR63	YES	0.000	0.08	NO	0.000	0.01	NO	0.000	0.03	NO	0.000	0.03
CR64	YES	0.000	0.06	NO	0.000	0.03	YES	0.000	0.04	NO	0.001	0.05
CR65	YES	0.000	0.04	YES	0.001	0.10	YES	0.001	0.05	NO	0.000	0.02
CR67	YES	0.000	0.03	NO	0.000	0.05	YES	0.001	0.08	NO	0.000	0.03
CR68	YES	0.000	0.04	YES	0.005	0.20	YES	0.007*	0.18	YES	0.006	0.20

Table 7
DIF Results for the 2006 AP World History Exam

Items	Women			African American			Asian			Hispanic		
	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}
1	YES	0.000	0.08	NO	0.001	0.04	NO	0.000	0.01	NO	0.001	0.09
2	YES	0.000	0.05	YES	0.001	0.09	YES	0.001	0.10	NO	0.000	0.03
3	NO	0.000	0.04	NO	0.000	0.09	YES	0.004	0.18	YES	0.001	0.09
4	NO	0.000	0.03	YES	0.001	0.14	NO	0.000	0.04	NO	0.000	0.02
5	YES	0.001	0.07	YES	0.002	0.16	YES	0.004	0.16	YES	0.001	0.13
6	NO	0.000	0.09	NO	0.001	0.09	NO	0.001	0.05	NO	0.002	0.13
7	YES	0.001	0.09	NO	0.000	0.02	YES	0.002	0.14	NO	0.001	0.07
8	NO	0.000	0.02	YES	0.004	0.18	NO	0.003	0.13	YES	0.000	0.06
9	NO	0.000	0.05	NO	0.000	0.02	YES	0.001	0.10	NO	0.001	0.08
10	NO	0.000	0.06	NO	0.001	0.10	YES	0.004	0.17	NO	0.000	0.03
11	YES	0.001	0.09	NO	0.001	0.09	NO	0.002	0.09	YES	0.001	0.08
12	YES	0.001	0.08	NO	0.002	0.05	NO	0.000	0.04	NO	0.000	0.04
13	YES	0.000	0.11	YES	0.001	0.12	NO	0.001	0.05	NO	0.000	0.02
14	YES	0.001	0.13	YES	0.002	0.12	YES	0.000	0.07	YES	0.005	0.23
15	NO	0.000	0.03	NO	0.001	0.07	YES	0.002	0.09	NO	0.002	0.13
16	NO	0.001	0.03	YES	0.001	0.08	YES	0.001	0.09	YES	0.000	0.12
17	YES	0.001	0.09	NO	0.001	0.08	YES	0.002	0.14	YES	0.002	0.14
18	NO	0.000	0.03	NO	0.000	0.08	YES	0.001	0.07	NO	0.000	0.07
19	YES	0.003	0.16	YES	0.001	0.06	YES	0.002	0.17	YES	0.001	0.10
20	NO	0.001	0.06	YES	0.001	0.08	NO	0.000	0.04	NO	0.001	0.10
21	YES	0.000	0.04	YES	0.001	0.15	NO	0.005	0.21	YES	0.000	0.10
22	NO	0.000	0.08	YES	0.001	0.04	YES	0.001	0.06	YES	0.001	0.17
23	YES	0.000	0.07	NO	0.000	0.04	YES	0.002	0.12	NO	0.000	0.04
24	NO	0.005	0.21	NO	0.000	0.03	YES	0.001	0.06	NO	0.001	0.05
25	YES	0.000	0.04	NO	0.003	0.07	NO	0.000	0.02	YES	0.001	0.08
26	NO	0.004	0.18	NO	0.003	0.11	NO	0.003	0.11	NO	0.000	0.03
27	YES	0.001	0.07	YES	0.009*	0.24	NO	0.000	0.05	YES	0.004	0.16
28	NO	0.000	0.02	NO	0.000	0.04	NO	0.000	0.03	YES	0.001	0.12
29	NO	0.000	0.05	NO	0.000	0.04	NO	0.000	0.03	NO	0.000	0.03
30	NO	0.000	0.06	NO	0.000	0.07	NO	0.000	0.03	NO	0.000	0.05
31	NO	0.000	0.05	YES	0.000	0.08	YES	0.000	0.08	YES	0.001	0.16
32	YES	0.003	0.15	NO	0.001	0.07	NO	0.001	0.05	NO	0.001	0.09
33	YES	0.002	0.11	YES	0.002	0.11	NO	0.000	0.04	NO	0.000	0.04
34	NO	0.000	0.12	NO	0.000	0.04	YES	0.003	0.10	NO	0.001	0.10
35	NO	0.002	0.11	NO	0.000	0.03	NO	0.000	0.01	NO	0.001	0.06
36	YES	0.001	0.08	YES	0.002	0.07	NO	0.000	0.04	NO	0.000	0.02
37	YES	0.003	0.09	YES	0.002	0.12	YES	0.000	0.07	YES	0.006	0.12
38	YES	0.000	0.04	NO	0.001	0.06	YES	0.000	0.07	NO	0.000	0.04
39	YES	0.001	0.10	YES	0.001	0.11	NO	0.000	0.02	NO	0.000	0.06
40	YES	0.000	0.04	NO	0.000	0.05	NO	0.000	0.04	NO	0.000	0.06
41	NO	0.001	0.08	YES	0.002	0.13	NO	0.005	0.19	NO	0.000	0.01
42	NO	0.001	0.08	NO	0.000	0.05	NO	0.003	0.12	NO	0.000	0.04

Table 7 (cont.)

Items	<u>Women</u>			<u>African American</u>			<u>Asian</u>			<u>Hispanic</u>		
	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	Items	χ^2
43	NO	0.000	0.06	YES	0.001	0.09	NO	0.001	0.08	NO	0.001	0.08
44	YES	0.003	0.13	YES	0.002	0.11	YES	0.002	0.12	YES	0.001	0.07
45	YES	0.001	0.08	YES	0.004	0.15	YES	0.003	0.14	YES	0.003	0.09
46	YES	0.000	0.05	YES	0.001	0.09	YES	0.001	0.08	NO	0.000	0.02
47	NO	0.000	0.04	YES	0.000	0.08	YES	0.001	0.08	YES	0.001	0.06
48	YES	0.001	0.13	NO	0.000	0.02	YES	0.001	0.10	NO	0.000	0.04
49	NO	0.000	0.02	NO	0.000	0.03	YES	0.001	0.06	YES	0.002	0.04
50	YES	0.000	0.05	YES	0.004	0.16	YES	0.003	0.13	NO	0.000	0.07
51	YES	0.000	0.04	YES	0.001	0.06	YES	0.003	0.12	NO	0.000	0.03
52	YES	0.001	0.10	NO	0.001	0.04	NO	0.000	0.07	NO	0.000	0.02
53	YES	0.001	0.08	YES	0.002	0.14	YES	0.001	0.07	YES	0.001	0.05
54	NO	0.000	0.05	NO	0.001	0.07	NO	0.000	0.08	YES	0.001	0.07
55	YES	0.001	0.10	NO	0.000	0.04	YES	0.000	0.06	YES	0.002	0.13
56	YES	0.001	0.07	YES	0.001	0.13	NO	0.000	0.03	YES	0.001	0.06
57	YES	0.000	0.07	NO	0.001	0.08	YES	0.002	0.12	NO	0.000	0.04
58	YES	0.001	0.07	NO	0.000	0.06	YES	0.001	0.12	NO	0.000	0.04
59	YES	0.000	0.06	NO	0.000	0.09	YES	0.001	0.11	NO	0.000	0.03
60	YES	0.002	0.11	NO	0.001	0.10	YES	0.001	0.08	NO	0.001	0.07
61	YES	0.004	0.15	YES	0.001	0.10	NO	0.000	0.03	NO	0.000	0.02
62	YES	0.000	0.05	NO	0.000	0.05	YES	0.001	0.09	NO	0.000	0.02
63	YES	0.000	0.05	NO	0.001	0.06	NO	0.001	0.08	NO	0.000	0.04
64	YES	0.002	0.14	NO	0.000	0.04	YES	0.002	0.13	YES	0.001	0.08
65	YES	0.001	0.16	NO	0.000	0.03	YES	0.002	0.14	YES	0.002	0.09
66	YES	0.000	0.08	YES	0.001	0.15	YES	0.003	0.21	NO	0.001	0.06
67	YES	0.000	0.04	YES	0.001	0.07	NO	0.000	0.06	YES	0.001	0.09
68	NO	0.001	0.10	NO	0.001	0.10	NO	0.002	0.14	NO	0.000	0.04
69	NO	0.000	0.03	NO	0.000	0.08	YES	0.000	0.09	NO	0.000	0.05
70	YES	0.002	0.15	NO	0.004	0.18	NO	0.001	0.10	NO	0.001	0.08

Table 8
DIF Results for the 2006 AP Spanish Language Exam

Items	Women			Asian			Hispanic		
	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}	χ^2	NCDIF	d_{DIF}
1	YES	0.001	0.08	YES	0.001	0.11	YES	0.025*	0.61
2	YES	0.002	0.09	NO	0.000	0.04	YES	0.023*	0.58
3	YES	0.001	0.08	NO	0.000	0.02	YES	0.011*	0.56
4	YES	0.000	0.06	NO	0.001	0.12	YES	0.025*	0.64
5	YES	0.001	0.07	NO	0.000	0.07	YES	0.039*	0.89
6	YES	0.000	0.04	NO	0.000	0.11	YES	0.011*	0.49
7	YES	0.001	0.07	NO	0.000	0.04	YES	0.022*	0.80
8	YES	0.003	0.10	NO	0.001	0.06	YES	0.057*	0.90
9	YES	0.001	0.09	NO	0.000	0.01	YES	0.002	0.20
10	NO	0.000	0.02	NO	0.001	0.08	YES	0.001	0.11
11	YES	0.001	0.08	YES	0.001	0.09	YES	0.054*	0.95
12	NO	0.002	0.11	NO	0.000	0.01	YES	0.006	0.16
13	NO	0.000	0.03	NO	0.001	0.05	YES	0.003	0.20
14	NO	0.000	0.02	NO	0.001	0.07	NO	0.001	0.04
15	NO	0.001	0.11	NO	0.000	0.03	YES	0.002	0.10
16	YES	0.003	0.14	YES	0.001	0.13	YES	0.001	0.11
17	NO	0.000	0.02	NO	0.000	0.03	YES	0.003	0.18
18	YES	0.001	0.08	NO	0.003	0.10	YES	0.063*	0.82
19	NO	0.000	0.03	NO	0.002	0.10	YES	0.026*	0.48
20	NO	0.000	0.01	YES	0.001	0.09	YES	0.003	0.23
21	NO	0.000	0.03	NO	0.000	0.03	NO	0.000	0.05
22	YES	0.000	0.05	NO	0.000	0.03	YES	0.005	0.22
23	NO	0.000	0.04	NO	0.000	0.05	YES	0.005	0.26
24	YES	0.000	0.05	NO	0.001	0.05	YES	0.002	0.13
25	YES	0.000	0.04	NO	0.000	0.04	YES	0.011*	0.45
26	YES	0.002	0.11	NO	0.001	0.08	YES	0.005	0.16
27	NO	0.000	0.01	NO	0.001	0.14	YES	0.000	0.08
28	YES	0.002	0.10	NO	0.002	0.08	YES	0.046*	0.84
29	NO	0.001	0.06	NO	0.000	0.05	NO	0.000	0.06
30	NO	0.002	0.08	YES	0.001	0.09	YES	0.001	0.11
31	YES	0.002	0.13	NO	0.001	0.07	YES	0.006	0.16
32	YES	0.001	0.09	NO	0.000	0.03	YES	0.013*	0.40
33	YES	0.002	0.12	NO	0.000	0.06	YES	0.004	0.20
34	YES	0.001	0.06	NO	0.000	0.04	YES	0.014*	0.52
35	NO	0.000	0.01	NO	0.000	0.02	YES	0.002	0.12
36	NO	0.001	0.07	NO	0.000	0.05	NO	0.002	0.13
37	NO	0.000	0.03	NO	0.000	0.06	YES	0.008*	0.42
38	YES	0.000	0.05	NO	0.001	0.06	YES	0.005	0.29
39	NO	0.000	0.04	NO	0.000	0.06	YES	0.001	0.07
40	YES	0.000	0.08	NO	0.000	0.03	YES	0.011*	0.35
41	NO	0.000	0.03	NO	0.000	0.09	YES	0.005	0.24
42	YES	0.001	0.05	NO	0.000	0.06	YES	0.001	0.04

Table 8 (cont.)

Items	χ^2	<u>Women</u> NCDIF	d_{DIF}	χ^2	<u>Asian</u> NCDIF	d_{DIF}	χ^2	<u>Hispanic</u> NCDIF	d_{DIF}
43	NO	0.000	0.05	NO	0.000	0.06	YES	0.001	0.10
44	YES	0.001	0.07	NO	0.000	0.09	YES	0.010*	0.36
45	NO	0.004	0.17	NO	0.000	0.03	YES	0.001	0.09
46	YES	0.002	0.07	NO	0.001	0.06	NO	0.017*	0.28
47	YES	0.001	0.05	YES	0.001	0.15	YES	0.061*	0.95
48	NO	0.001	0.07	NO	0.000	0.03	NO	0.002	0.11
49	NO	0.000	0.05	NO	0.000	0.03	YES	0.002	0.16
50	NO	0.001	0.06	NO	0.000	0.03	YES	0.014*	0.43
51	YES	0.001	0.06	NO	0.001	0.06	YES	0.006	0.26
52	YES	0.001	0.07	NO	0.000	0.06	YES	0.009*	0.35
53	YES	0.000	0.07	NO	0.000	0.05	YES	0.003	0.14
54	YES	0.001	0.05	NO	0.000	0.04	YES	0.014*	0.43
55	YES	0.001	0.04	NO	0.001	0.05	YES	0.012*	0.37
56	NO	0.001	0.05	NO	0.003	0.15	YES	0.004	0.21
57	NO	0.001	0.09	NO	0.000	0.04	YES	0.008*	0.25
58	YES	0.001	0.09	NO	0.000	0.05	YES	0.008*	0.31
59	NO	0.001	0.09	NO	0.000	0.04	YES	0.007*	0.30
60	YES	0.002	0.11	NO	0.000	0.05	YES	0.013*	0.36
61	YES	0.000	0.07	NO	0.000	0.08	YES	0.018*	0.49
62	YES	0.002	0.11	NO	0.000	0.04	YES	0.004	0.29
63	YES	0.001	0.12	NO	0.001	0.05	YES	0.001	0.16
64	YES	0.001	0.10	NO	0.000	0.05	YES	0.008*	0.27
65	NO	0.000	0.04	NO	0.000	0.02	NO	0.012*	0.29
66	YES	0.001	0.08	NO	0.000	0.07	YES	0.002	0.15
67	NO	0.001	0.05	NO	0.000	0.03	YES	0.008*	0.39
68	YES	0.001	0.05	NO	0.000	0.04	YES	0.019*	0.50
69	YES	0.001	0.10	NO	0.002	0.09	YES	0.015*	0.49
70	NO	0.000	0.03	NO	0.002	0.14	YES	0.025*	0.43
71	YES	0.000	0.05	NO	0.000	0.04	YES	0.014*	0.33
72	NO	0.000	0.02	NO	0.000	0.04	YES	0.001	0.07
73	YES	0.002	0.12	NO	0.000	0.05	YES	0.025*	0.46
74	YES	0.000	0.06	NO	0.001	0.06	YES	0.011*	0.32
75	NO	0.003	0.15	NO	0.001	0.11	YES	0.005	0.26

Table 9
Effects of IRT DIF on Scale-Level Means and Variances

	Δ Mean Due to DIF	Observed Mean Difference	% of Observed Mean Difference	Δ Variance	Differences in Observed Var.	% of Observed Differences	d_{DIF} Min – Max
<u>2007 SAT</u>							
Women	3.38	5.30	64%	–20.32	35.71	57%	0.01 – 0.47
African American	–0.07	26.77	0%	2.64	–70.06	4%	0.01 – 0.31
Asian	–3.57	–4.62	77%	–21.35	–154.24	14%	0.01 – 0.54
Hispanic	0.54	18.60	3%	–4.81	–90.01	5%	0.01 – 0.43
<u>AP World History</u>							
Women	0.67	4.34	15%	–1.92	1.84	104%	0.02 – 0.21
African American	0.13	9.95	1%	–12.80	–17.45	73%	0.02 – 0.24
Asian	0.15	–0.69	22%	4.41	–14.32	31%	0.01 – 0.21
Hispanic	–0.05	9.85	1%	3.06	–18.68	16%	0.01 – 0.23
<u>AP Spanish</u>							
Women	–0.27	–0.14	192%	–11.25	10.54	107%	0.01 – 0.18
Asian	0.02	–.32	6%	–2.47	9.05	27%	0.01 – 0.15
Hispanic	–3.62	–8.13	45%	–14.35	30.32	47%	0.04 – 1.02

Table 10
Effects of IRT DIF on Scale-Level Means and Variances for the SAT Sub-Scales

	Δ Mean Due to DIF	Observed Mean Difference	% of Observed Mean Difference	Δ Variance	Differences in Observed Var.	% of Observed Differences	d_{DIF} Min – Max
<u>2007 SAT-Writing</u>							
Women	-0.47	0.00	--	-2.46	2.54	97%	0.01 – 0.17
African American	-0.71	6.81	10%	-1.01	-3.10	33%	0.01 – 0.17
Asian	0.26	-0.02	1300%	-0.99	-11.66	8%	0.01 – 0.47
Hispanic	0.12	5.21	2%	3.56	-1.80	197%	0.02 – 0.27
<u>2007 SAT-Math</u>							
Women	3.64	4.36	83%	-2.63	2.05	128%	0.02 – 0.37
African American	1.26	9.92	13%	5.35	0.68	787%	0.02 – 0.26
Asian	-4.65	-5.07	92%	3.93	-12.67	31%	0.01 – 0.54
Hispanic	0.66	6.55	10%	0.47	-7.21	7%	0.01 – 0.19
<u>2007 SAT-Reading</u>							
Women	0.20	0.93	22%	-0.79	9.88	8%	0.01 – 0.23
African American	-0.62	10.03	6%	-4.29	-19.87	22%	0.01–0.20
Asian	0.81	0.47	172%	-13.66	-40.66	34%	0.03 – 0.32
Hispanic	-0.24	6.83	4%	-8.01	-25.21	32%	0.01 – 0.41

Table 11
Effects of IRT DIF on Scale-Level Correlations with External Criteria

	r_{iYR}	SD_{XR}	$r_{XYR(IRT)}$	$\Delta r_{XY(IRT)}$
2007 SAT (Women)	.05	28.09	.13	.00
	.10	28.09	.26	.00
	.15	28.09	.39	.00
2007 SAT (African American)	.05	32.74	.12	.00
	.10	32.74	.23	.00
	.15	32.74	.35	.00
2007 SAT (Asian)	.05	28.76	.12	.00
	.10	28.76	.25	.00
	.15	28.76	.37	.01
2007 SAT (Hispanic)	.05	30.13	.13	.00
	.10	30.13	.25	.00
	.15	30.13	.38	.00
AP World History (Women)	.05	12.19	.13	.00
	.10	12.19	.27	.00
	.15	12.19	.40	.00
AP World History (African American)	.05	12.57	.13	.00
	.10	12.57	.26	.01
	.15	12.57	.39	.01
AP World History (Asian)	.05	12.40	.13	.00
	.10	12.40	.26	.00
	.15	12.40	.39	-.01

Table 11 (cont.)

	r_{iYR}	SD_{XR}	$r_{XYR(IRT)}$	$\Delta r_{XY(IRT)}$
AP World History (Hispanic)	.05	13.75	.12	.00
	.10	13.75	.24	.00
	.15	13.75	.36	.00
AP Spanish (Women)	.05	11.05	.15	.01
	.10	11.05	.30	.02
	.15	11.05	.45	.02
AP Spanish (Asian)	.05	11.41	.15	.00
	.10	11.41	.30	.00
	.15	11.41	.45	.01
AP Spanish (Hispanic)	.05	9.30	.17	.02
	.10	9.30	.35	.05
	.15	9.30	.52	.07

Note: r_{iYR} = the assumed correlation between item i and the criterion Y in the reference group. SD_{XR} = the standard deviation of the scale in the reference group. $r_{XYR(IRT)}$ = the correlation between the scale and an external criterion in the reference group given r_{iYR} . $\Delta r_{XY(IRT)}$ = the change in the correlation between the scale and an external criterion due to DIF.

Figure 1
Frequency Distribution of Empirical Differences between Standardized Factor Loadings in the Reference and Focal Groups

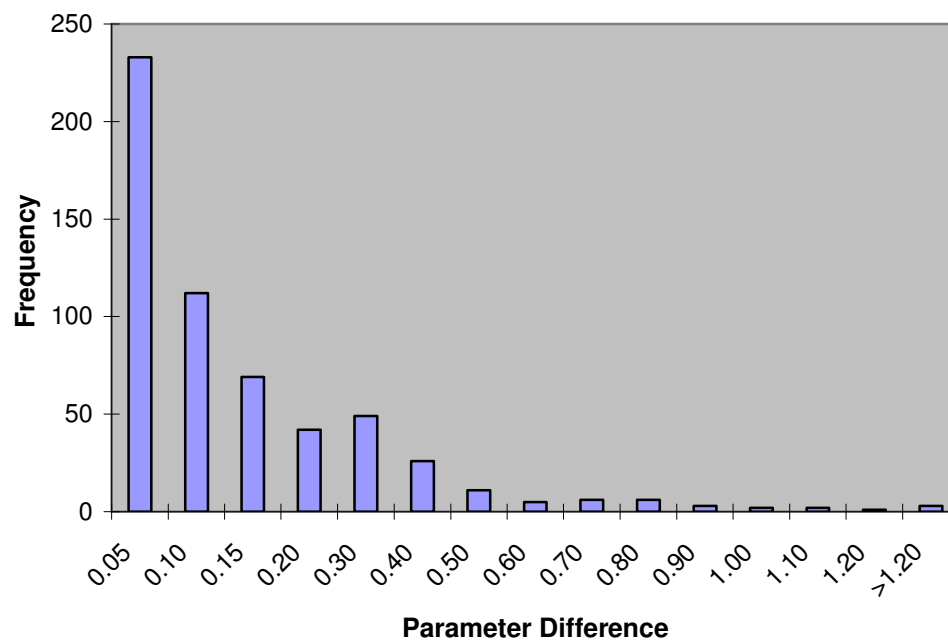


Figure 2

Frequency Distributions of d_{MACS} for 6 Items and $N = 500$ in the Small, Medium, Large, and No DIF Conditions

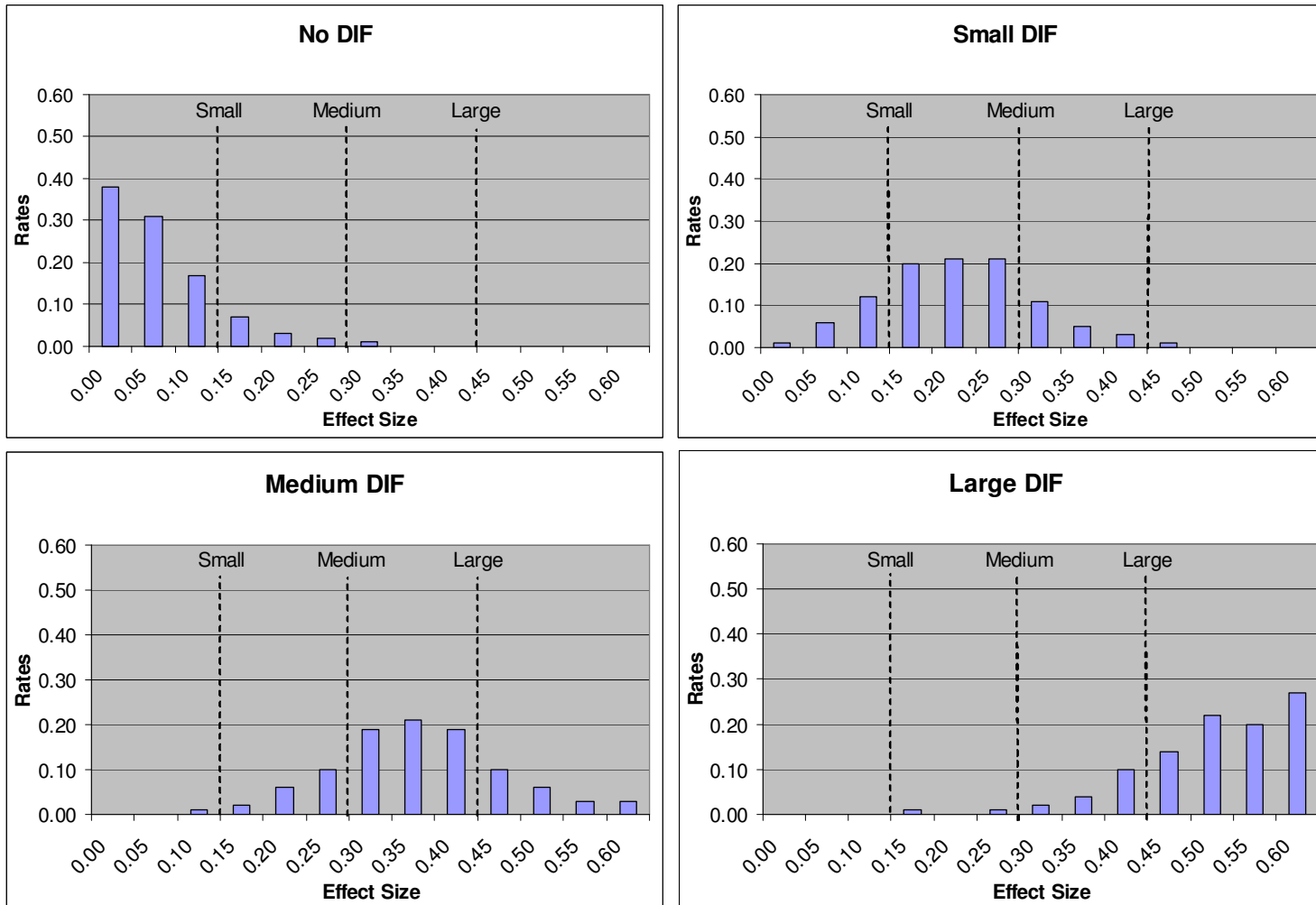


Figure 3

Frequency Distributions of d_{MACS} for 12 Items and $N = 500$ in the Small, Medium, Large, and No DIF Conditions

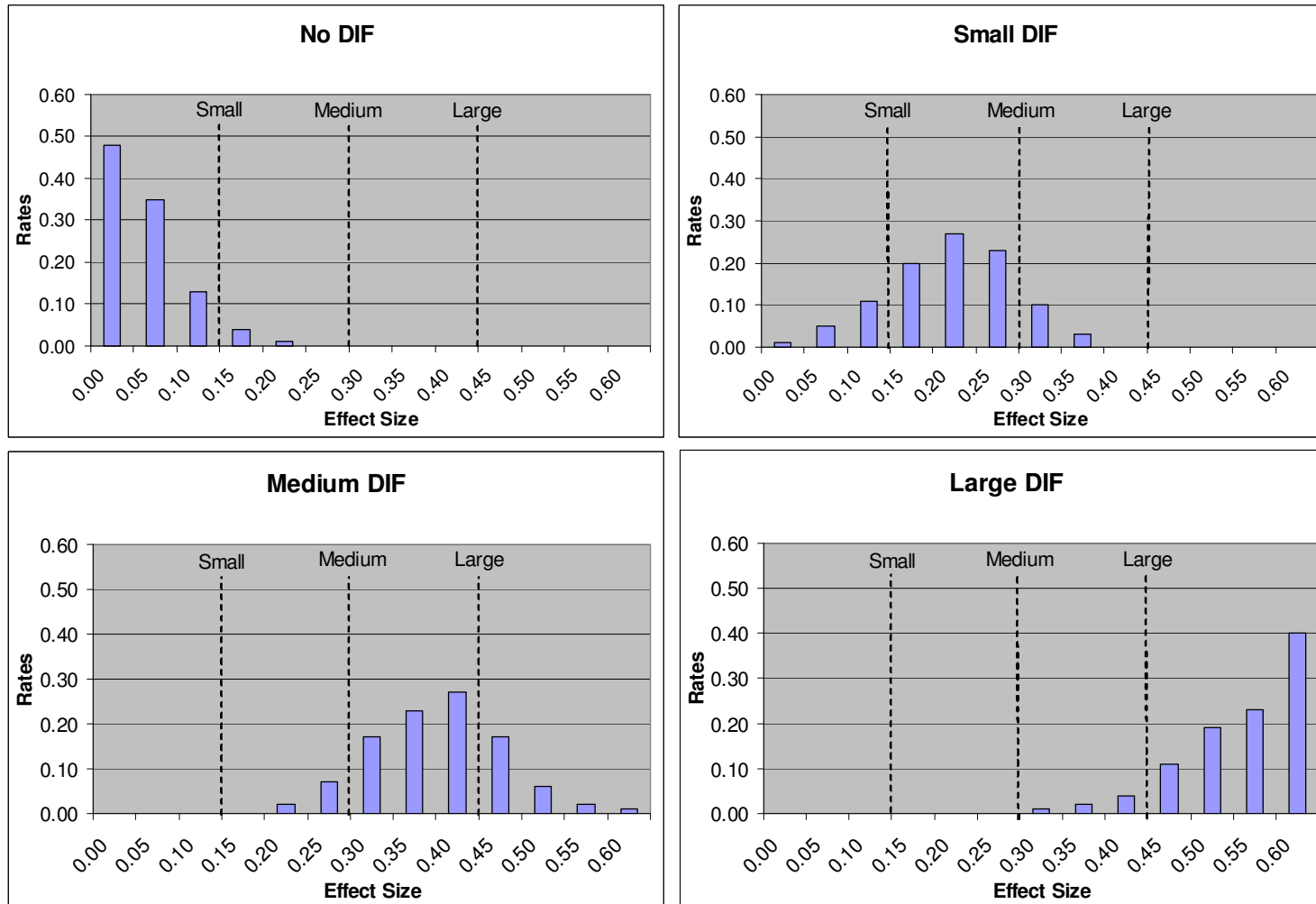


Figure 4

Frequency Distributions of d_{MACS} for 16 Items and $N = 500$ in the Small, Medium, Large, and No DIF Conditions

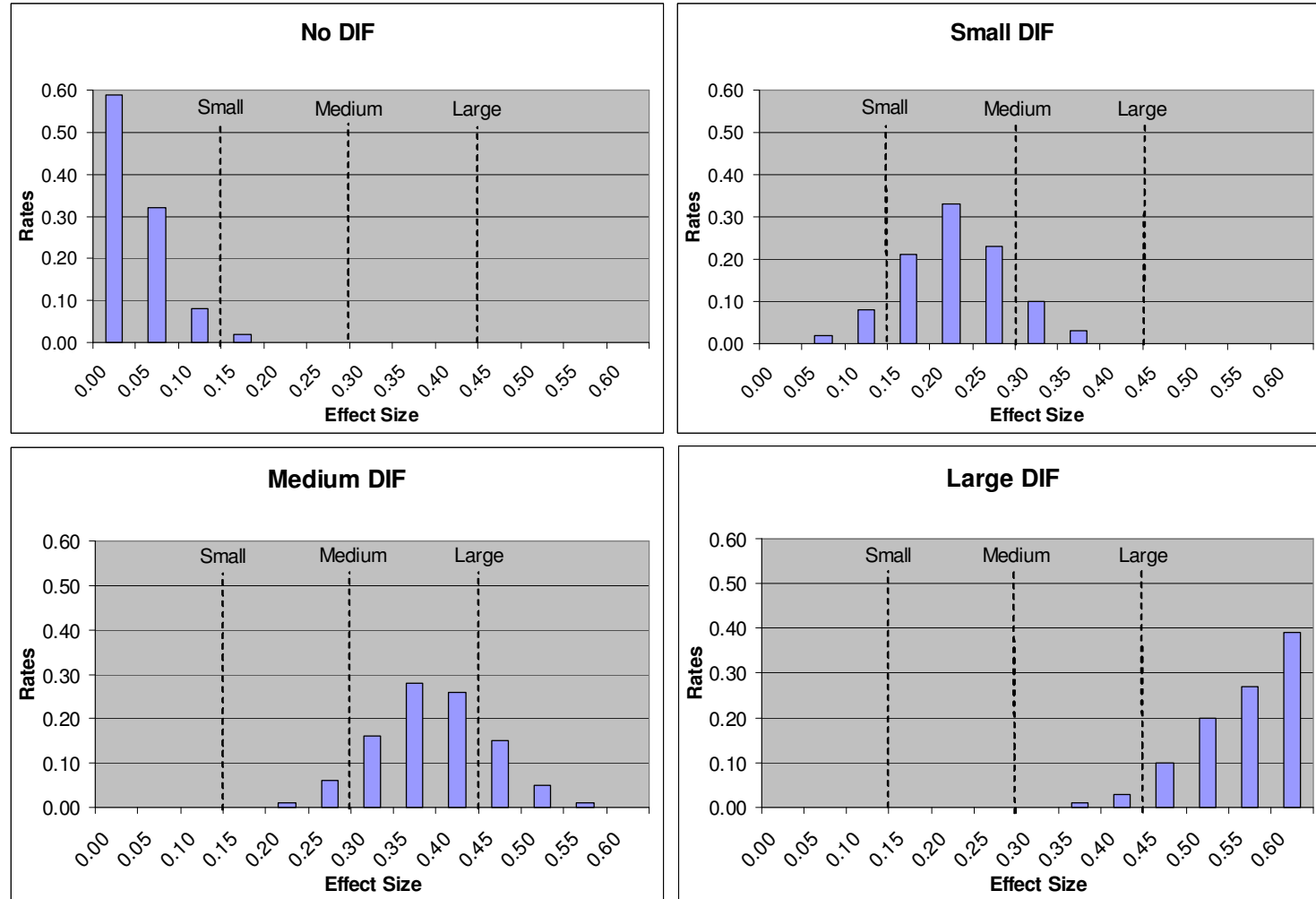


Figure 5

Frequency Distributions of d_{DIF} for 20 Items and $N = 500$ in the Small, Medium, Large, and No DIF Conditions

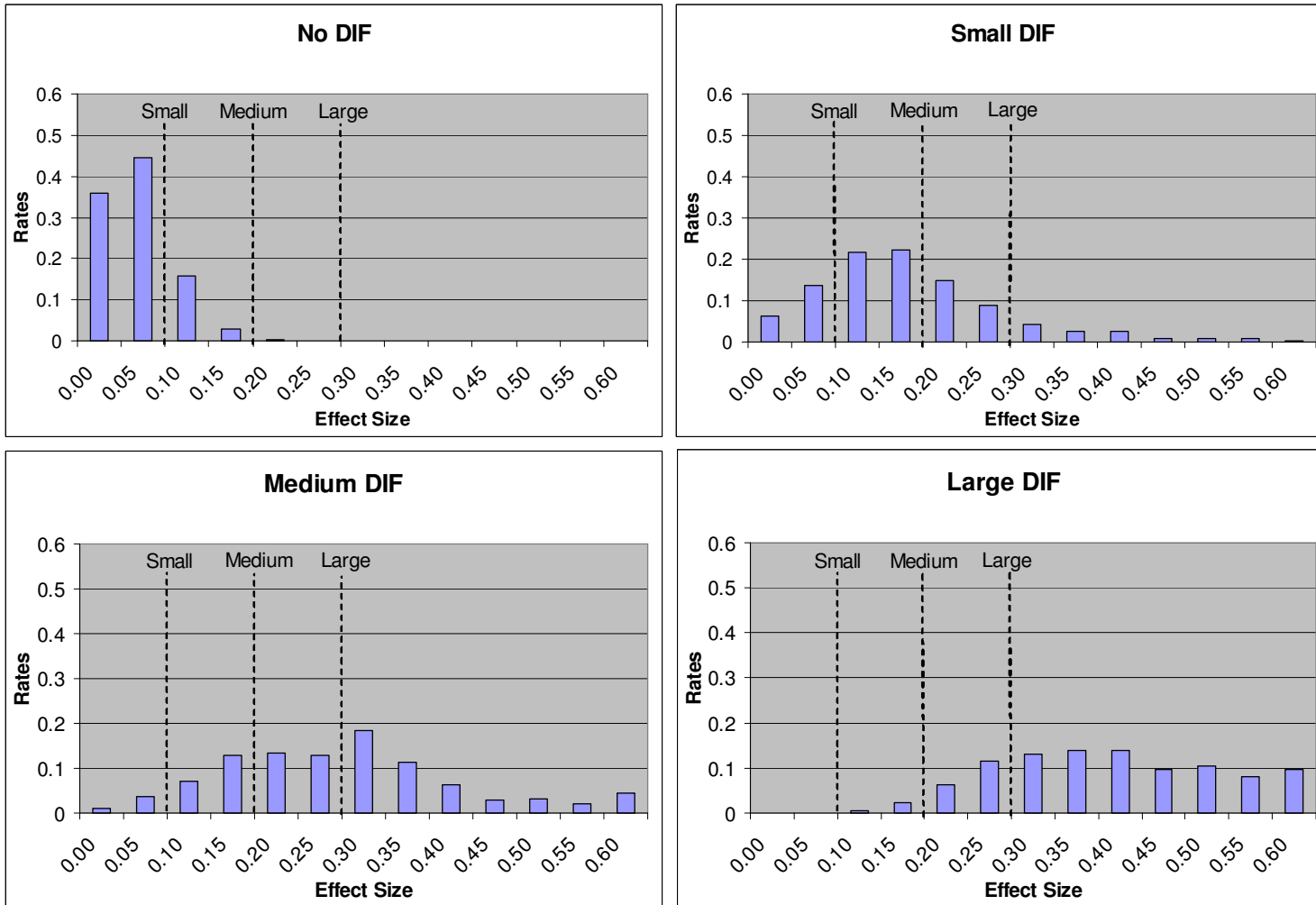
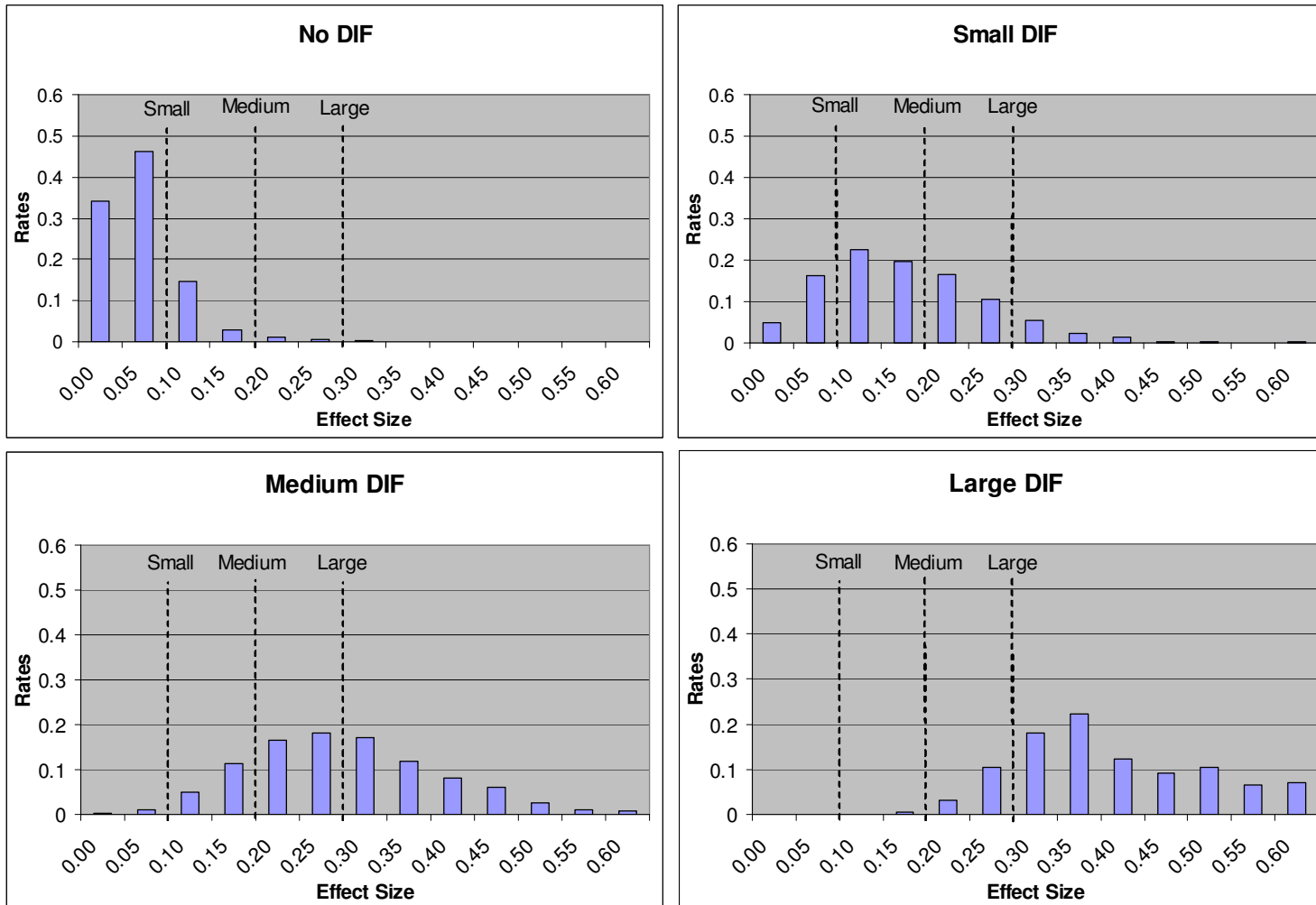


Figure 6

Frequency Distributions of d_{DIF} for 40 Items and $N = 500$ in the Small, Medium, Large, and No DIF Conditions



REFERENCES

- Bentler, P. M. (1995). *EQS: Structural equations program manual*. Encino, CA: Multivariate Software, Inc.
- Bock, R. D., Muraki, E., & Pfeifenger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, 25, 275–285.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Boomsma, A. (1982). Robustness of LISREL against small sample sizes in factor analysis models. In K. G. Joreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Part I) (pp. 149–173). Amsterdam: North Holland.
- Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, 19, 309–324.
- Byrne, B. M. (1998). *Structural equation modeling in LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Camara, W. J., & Schmidt, A. E. (1999). *Group differences in standardized testing and social stratification* (College Board Rep. No. 99-5). New York, NY: College Board Publications.
- Camilli, G., & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics*, 12, 87–99.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, 12, 253–260.
- Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology*, 82, 311–320.

- Chan, K.-Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery over 16 years. *Journal of Applied Psychology, 84*, 610–619.
- Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1–27.
- Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 187–212.
- Cheung, G. W., & Rensvold, R. B. (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.
- Chirkov, V., Ryan, R. M., Kim, Y., & Kaplan, U. (2003). Differentiating autonomy from individualism and independence: A self-determination theory perspective on internalization of cultural orientations and well-being. *Journal of Personality and Social Psychology, 84*, 97–110.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312.

- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, 49, 997–1003.
- Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement*, 10, 237–255.
- Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, 12, 381–398.
- Cortina, J. M., & Landis, R. S. (2011). The earth is not round ($p = .00$). *Organizational Research Methods*, 14, 332–349.
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, 9, 327–346.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, 23, 355–368.
- Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin*, 92, 526–531.
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, 95, 134–135.
- Drasgow, F. (1987). Study of measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, 19–29.
- Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, 70, 662–680.

- Drasgow, F., & Kang, T. (1984). Statistical power of differential validity and differential prediction for detecting measurement nonequivalence. *Journal of Applied Psychology*, 69, 498–508.
- Drasgow, F., Nye, C. D., & Guo, J. (2008). *The stability of item parameters over time for the NCLEX-RN exam: A report to the National Council of State Boards of Nursing*. The National Council of State Boards of Nursing.
- Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189–199.
- Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias. (Doctoral dissertation, Illinois Institute of Technology). *Dissertation Abstracts International*, 54 (04), 2266B.
- Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*, 23, 309–326.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3–8.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Hays, W. L., (1963). *Statistics for psychologists*. New York: Holt, Rinehart, and Winston.
- Hernandez, A., Chernyshenko, O., Stark, S., & Drasgow, F. (April, 2008). *DIF detection with MACS: Effectiveness and efficiency of two approaches*. Paper presented at the 23rd annual conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

- Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, 64, 903–915.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 139–145). Hillsdale, NJ: Lawrence Erlbaum.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, 26, 329–367.
- Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, 14, 332–346.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 3, 424–453.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones Irwin.
- Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling*, 16, 642–657.
- Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426.
- Jöreskog, K., & Sörbom, D. (1993). *New features in LISREL 8*. Chicago: Scientific Software International.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746–759.

- Kirk, R. E. (2006). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference*. DOI: 10.1016/j.jspi.2006.09.011.
- Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, 315, 1080–1081.
- Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). *Differential item functioning for males and females on SAT-verbal reading subscore items* (College Board Rep. No. 88-4). New York, NY: College Board Publications.
- Lei, M., & Lomax, R. G. (2005). The effects of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, 12, 1–27.
- Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139–161.
- Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational measurement: Issues and Practice*, 6, 13–17.
- Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5, 159–173.
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal*, 30, 841–860.

- Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220.
- McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161–173.
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568–592.
- Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361–388.
- Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with likert data. *Applied Psychological Measurement, 31*, 430–455.
- Mislevy, R. J., & Bock, R. D. (1991). *BILOG User's Guide*. Chicago, IL: Scientific Software.
- Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6*, 328–362.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
- Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, Advance online publication. doi: [10.1037/a0022955](https://doi.org/10.1037/a0022955).

- Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality traits across cultures. *Journal of Research in Personality*, 42, 1524–1536.
- Oishi, S. (2007). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, 40, 411–423.
- Peterson, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 13, 3–29.
- Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, 7, 27–65.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495–502.
- Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197–207.
- Raju, N. S. (1999). *DFITD4: A Fortran program for calculating DIF/DTF* [computer program]. Chicago: Illinois Institute of Technology.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measure of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353–368.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87, 517–529.

- Rudas, T., & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics*, 22, 31–45.
- Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, 63, 506–516.
- Saucier, G., & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality*, 69, 847–879.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115–129.
- Schmitt, A. P., & Dorans, N. J. (1988). *Differential item functioning for minority examinees on the SAT* (Educational Testing Service Rep. No. 88-32). Princeton, NJ: Educational Testing Service.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.
- Stacy, A. W., MacKinnon, D. P., & Pentz, M. A. (1993). Generality and specificity in health behavior: Application to warning-label and social influence expectancies. *Journal of Applied Psychology*, 78, 611–627.
- Stark, S. (2000). *3PLGEN* [computer program]. Department of Psychology, University of Illinois at Urbana-Champaign.
- Stark, S. (2006). *ITERLINK: Iterative linking and pairwise DIF detection for the 3PL IRT model using Lord's chi-square* [computer program]. Department of Psychology, University of Illinois at Urbana-Champaign.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are

- statistically significant effects practically important? *Journal of Applied Psychology*, 89, 497–508.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology*, 91, 1292–1306.
- Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–90.
- 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., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.
- Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement*, 8, 63–70.
- Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6, 100–116.
- van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality*, 69, 1007–1031.
- Vandenburg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, 5, 139–158.

- Vandenburg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70.
- Velicer, W. F., & Fava, L. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivariate Behavioral Research*, 22, 193–209.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

APPENDIX A: DERIVATIONS OF Δ VARIANCE AND Δ COVARIANCE

Derivation of $\Delta Var(x_i)$

The variances of item i in the reference and focal groups can be defined as

$$Var(x_{iR}) = \lambda_{iR}^2 \phi_R + Var(\delta_{iR})$$

and

$$Var(x_{iF}) = (\lambda_{iR} + C_i)^2 \phi_F + Var(\delta_{iF}),$$

respectively, where $Var(\delta_{iR})$ and $Var(\delta_{iF})$ are the variances of the error terms for the item and C_i

$$= \lambda_{iF} - \lambda_{iR}.$$

Consequently,

$$Var(x_{iF}) = \lambda_{iR}^2 \phi_F + 2C_i \lambda_{iR} \phi_F + C_i^2 \phi_F + Var(\delta_{iF})$$

and $\Delta Var(x_i)$ is given by

$$Var(x_{iR}) - Var(x_{iF}) = \Delta Var(x_i) = 2C_i \lambda_{iR} \phi_F + C_i^2 \phi_F$$

when $Var(\delta_{iR}) = Var(\delta_{iF})$. Because we are only interested in the effects of DIF on the variance of the scale (rather than the combined effects of DIF and true difference in the variance), ϕ_R and ϕ_F are also treated as equal for the purposes of this calculation.

Derivation of $\Delta Cov(x_i, x_j)$

The covariance of items i and j is

$$Cov(x_{iR}, x_{jR}) = \lambda_{iR} \lambda_{jR} \phi_R$$

and

$$Cov(x_{iF}, x_{jF}) = (\lambda_{iR} + C_i)(\lambda_{jR} + C_j) \phi_F,$$

for the reference and focal group, respectively. Then

$$Cov(x_{iF}, x_{jF}) = (\lambda_{iR}\lambda_{jR} + \lambda_{jR}C_i + \lambda_{iR}C_j + C_iC_j)\phi_F$$

and subtracting this from $Cov(x_{iR}, x_{jR})$ gives

$$\Delta Cov(x_i, x_j) = \lambda_{jR}C_i\phi_F + \lambda_{iR}C_j\phi_F + C_iC_j\phi_F.$$

APPENDIX B: DERIVATIONS OF $\Delta r_{xy(CFA)}$

Derivation of $\Delta r_{XY(CFA)}$

The correlation of scale X with criterion Y in the reference group can be defined as

$$r_{XYR(CFA)} = \frac{\sum_{iR}^n \lambda_{iR} Cov(\xi_R, Y_R)}{SD_{XR} SD_{YR}}$$

where $Cov(\xi_R, Y_R)$ is the covariance of the latent trait and the criterion. In addition, SD_{XR} and SD_{YR} are the standard deviations of the scale and the criterion, respectively, in the reference group. In contrast, the correlation in the focal group is

$$r_{XYF(CFA)} = \frac{\sum_{iR}^n (\lambda_{iR} + C_i) Cov(\xi_F, Y_F)}{[SD_{XR} + \Delta SD(X_S)] SD_{YF}}$$

where $\Delta SD(X_S)$ is the change in the standard deviation of the scale due to nonequivalence.

To obtain $\Delta r_{XY(CFA)}$, one can calculate the difference between the correlations in the reference and focal groups. This difference can be defined by

$$\Delta r_{XY(CFA)} = r_{XYR(CFA)} - r_{XYF(CFA)} = \frac{\sum_{iR}^n \lambda_{iR} Cov(\xi_R, Y_R)}{SD_{XR} SD_{YR}} - \frac{\sum_{iR}^n (\lambda_{iR} + C_i) Cov(\xi_F, Y_F)}{[SD_{XR} + \Delta SD(X_S)] SD_{YF}}.$$

Again, because we are interested in the effects of nonequivalence in X on the correlation, it is assumed that $Cov(\xi_R, Y_R) = Cov(\xi_F, Y_F)$ and $SD_{YR} = SD_{YF}$. Thus, differences identified here are only the result of nonequivalence and true differences between the correlations will not be a factor. Calculating the result of this equation gives

$$\Delta r_{XY(CFA)} = \frac{Cov(\xi_R, Y_R) \Delta SD(X_S) - Cov(\xi_R, Y_R) SD_{XR}}{SD_{XR}^2 SD_{YR} + SD_{XR} SD_{YR} \Delta SD(X_S)}.$$