January 2012

Detection and Classification of DIF Types Using Parametric and Nonparametric Methods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST, and Logistic Regression Procedures

Gabriel E. Lopez
University of South Florida, gabriel_e_lopez@yahoo.com

Detection and Classification of DIF Types Using Parametric and Nonparametric Methods: A comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST, and Logistic Regression Procedures

by

Gabriel E. Lopez Rivas

A dissertation submitted in partial fulfillment
of the requirements for the degree of Doctor of Philosophy
in Industrial/Organizational Psychology
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Stephen Stark, Ph.D.
Michael T. Brannick, Ph.D.
Michael D. Coovert, Ph.D.
Bill N. Kinder, Ph.D.
Yi-Hsin Chen, Ph.D.

Date of Approval:
January 19, 2012

Keywords: differential item functioning, DIF, item bias, crossing simultaneous item bias test, item response theory likelihood ratio test, logistic regression

Table of Contents

List of Tables

List of Figures

Abstract

The purpose of this investigation was to compare the efficacy of three methods for detecting differential item functioning (DIF). The performance of the crossing simultaneous item bias test (CSIBTEST), the item response theory likelihood ratio test (IRT-LR), and logistic regression (LOGREG) was examined across a range of experimental conditions including different test lengths, sample sizes, DIF and differential test functioning (DTF) magnitudes, and mean differences in the underlying trait distributions of comparison groups, herein referred to as the reference and focal groups. In addition, each procedure was implemented using both an all-other anchor approach, in which the IRT-LR baseline model, CSIBEST matching subtest, and LOGREG trait estimate were based on all test items except for the one under study, and a constant anchor approach, in which the baseline model, matching subtest, and trait estimate were based on a predefined subset of DIF-free items. Response data for the reference and focal groups were generated using known item parameters based on the three-parameter logistic item response theory model (3-PLM). Various types of DIF were simulated by shifting the generating item parameters of select items to achieve desired DIF and DTF magnitudes based on the area between the groups' item response functions. Power, Type I error, and Type III error rates were computed for each experimental condition based on 100 replications and effects analyzed via ANOVA. Results indicated that the procedures varied in efficacy, with LOGREG when implemented using an all-other approach providing the best balance of power and Type I

error rate. However, none of the procedures were effective at identifying the type of DIF that was simulated.

CHAPTER 1

Introduction

Assessment is prevalent in organizational settings because studies have shown that the use of valid tests for selection and promotion greatly enhances decision making and therefore productivity (Schmidt & Hunter, 1998). Unfortunately, evidence of mean test score differences across demographic groups, particularly between a majority group and a minority group identified as "protected" under the Civil Rights Act of 1964 (Sackett, Schmitt, Ellingson, & Kabin, 2001), raises concerns that tests are biased and thus increase the likelihood of litigation, which can reduce the anticipated utility of assessment-based selection programs. In recognition of these concerns, the American Psychological Association (APA) and similar organizations have, over the years, commissioned scientific task forces or review panels, consisting of psychologists, educators, and measurement specialists. The objective was to examine the issue of test bias, develop precise statistical and psychometric definitions that distinguish bias from group mean differences in scores that relate to actual differences in performance, and identify powerful and up-to-date methods for detecting and revising problematic items or instruments. The results of these efforts have been codified in documents, such as the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 1999) and the Society for Industrial and Organizational Psychologists (SIOP) Principles for the Validation and Use of Personnel Selection Procedures (2003), which provide

recommendations, or "best practices", intended to guide organizational professionals who are involved directly with test development, validation, and use.

These documents and, more generally, the psychometric literature discuss two broad forms of test bias: external and internal (Drasgow, 1984). External or *predictive bias* refers to differences across comparison groups in a test's relationships with external criteria. Since the 1960s, the recommended way of testing for predictive bias is to compare regression lines for test score-criterion relationships across groups, with and without the inclusion of group membership and, possibly, interaction terms as additional predictors. Internal bias, on the other hand, is more precisely referred to as differential functioning. *Differential item functioning* (DIF) is said to occur when individuals from different groups have unequal expected item scores, after conditioning, or matching, on the primary trait, attribute, or ability the test is designed to measure. Similarly, *differential test functioning* (DTF) is said to occur when individuals from different groups have unequal expected test (i.e., number correct or total) scores, after conditioning on trait level. Importantly, both item response theory (IRT) and confirmatory factor analysis (CFA) methods for detecting differential functioning, which are advocated by the testing standards, are capable of distinguishing DIF and DTF, internal problems with a measuring instrument, from *impact*, defined as a true difference in the distribution of the latent trait or attribute measured by a test across comparison groups.

To assess whether an instrument shows bias in a general sense, it is therefore necessary to conduct tests for both internal and external bias using appropriate statistical and psychometric methods. The presence or absence of mean differences across groups in item or test scores does not provide any meaningful information concerning internal

bias, because DIF/DTF and impact can work in opposite directions. An unbiased instrument can show substantial impact if the comparison groups differ markedly in terms of their actual skills. Alternately, an instrument that exhibits no impact can contain a large number of DIF items, resulting in significant DTF or none at all. The latter is possible when items vary in the direction of DIF so that cancellation occurs when forming total scores. Finally, a lack of internal bias does not guarantee that a test will have equivalent relationships with external criteria, or vice versa. A test could have internal bias while showing similar regressions and correlations with external measures.

The Present Study

This study focuses exclusively on the effectiveness of IRT-based DIF detection methods. However, because IRT methods may be new to some readers, this presentation begins with a brief summary of the recommended approach for examining predictive bias involving moderated linear regression. Graphical illustrations of relational equivalence and various forms of predictive bias are then used as springboards for introducing IRT functions that relate expected item scores to trait scores via nonlinear regression lines. Following that are more detailed definitions of DIF, visual illustrations of its various manifestations, and a review of parametric and nonparametric DIF detection methods, including those that are the focus of this investigation. This presentation continues with a summary of the Monte Carlo simulation, methods of data analysis, and hypotheses concerning the results, and concludes with a discussion of the implications of the study findings upon practice and future research.

CHAPTER 2

Predictive Bias

Predictive bias, also known as *differential prediction*, refers to a difference across examinee subgroups in the relationship between test scores and an external criterion measure, such as job performance (Cleary, 1968; Drasgow, 1984; Humphreys, 1952). Figure 1 presents some hypothetical scenarios. Panel (a) presents a test that shows no predictive bias; that is, the same regression line can be used to describe the predictor-performance relationship for the comparison groups, generally referred to as the *reference* and *focal* groups (Holland & Thayer, 1988). The other panels of Figure 1 present scenarios in which a test shows bias due to differences in intercepts (b), slopes (c), or both intercepts and slopes (d and e). First, note that in the presence of slope differences (c), one group may be uniformly favored across the entire test score range, because an ordinal interaction occurs; specifically, the regression lines converge somewhat, but do not cross. On the other hand, a disordinal interaction (d) can occur, wherein the reference group is favored at some trait levels and the focal group at others. The scenario involving intercept differences that typically raises concern among testing critics is one in which the focal group has a higher intercept than the reference group, but the reference group regression line is used for selection decisions, thus leading to underprediction of focal group member performance. Such is the case shown in panels (b) and (e).

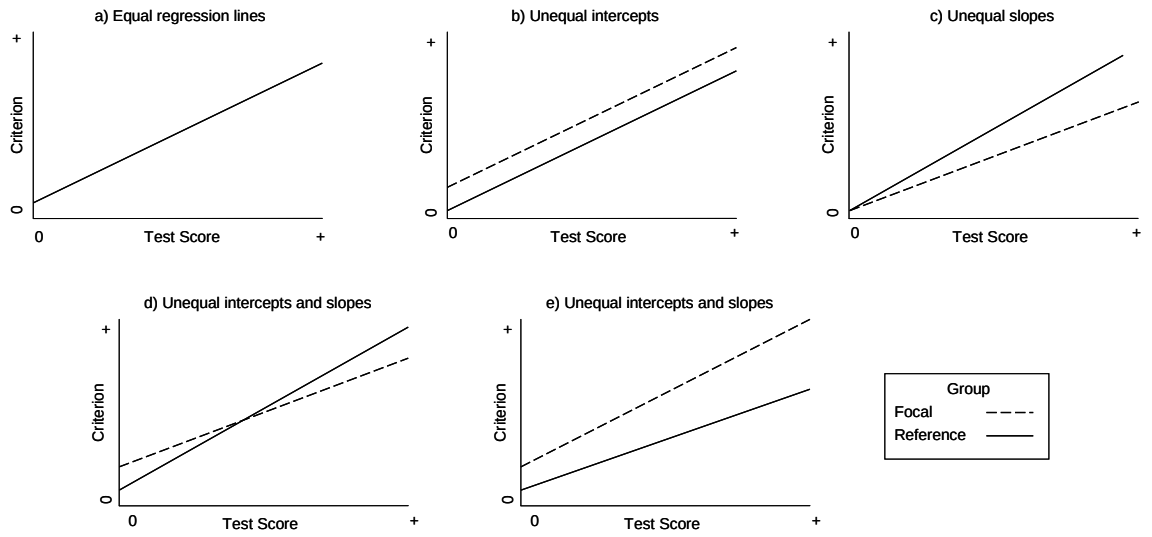

Figure 1. Examples of predictive bias.

According to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) and the Principles for the Validation and Use of Personnel Selection Procedures (SIOP, 2003), hypotheses about predictive bias can be tested using a moderated linear regression approach that examines either changes in the multiple correlation coefficient or the statistical significance of regression weights for nested models involving terms for test scores, group membership, and their interaction. Specifically, a compact (reduced) model is formed first, in which a dependent variable (y) is chosen, such as job performance, and the test score (x) is used as a predictor. Next, an augmented (full) model is created by adding two terms: one for group membership (d) and another for the interaction of group membership with test scores (dx). The differential prediction analysis is conducted by entering the predictors for the compact and augmented models in separate blocks. A significant change from Block 1 to Block 2 in the multiple correlation indicates that differential prediction is present. More precisely, a significant regression coefficient for the interaction term indicates inequality

of slopes, whereas a significant coefficient for the group membership term indicates inequality of intercepts (Stark, Chernyshenko, & Drasgow, 2004).

Research in applied settings has not shown much support for hypotheses involving differential prediction. In the few studies where differences have been found, the most common source of nonequivalence has been intercept differences that have been attributed to measurement or sampling error, and, contrary to intuition, these differences would have resulted in slight *over*prediction of focal group performance if a common regression line had been used (Jensen, 1980; Linn, 1982). Therefore, the more pressing concern seems to be group mean differences that affect selection rates directly, via proportions of correct (true positive and true negative) and incorrect (false positive and false negative) selection decisions within comparison groups. An unfortunate fact is that higher scoring groups not only show more true positives, but also benefit more from false positive decisions, and, conversely, lower scoring groups are disproportionately impacted by false negatives. (Recognition of this issue prompted discussion in the late 1970s of procedures to adjust test scores based on group membership; interested readers should see Hartigan and Widgor [1989] and Sackett and Wilk [1994] for details.) It is therefore important that testing professionals not only evaluate their instruments for predictive bias but that they also use complimentary methods to detect internal bias (Drasgow, 1984), which could exacerbate mean differences due to impact, and thus, the undesirable secondary effects of testing on society.

In summary, external bias does not seem to be a pervasive phenomenon that adversely affects selection rates for protected group members. Yet, evidence of equivalent relationships between test scores and criterion measures across groups does

not eliminate the possibility that a test exhibits internal bias (Drasgow, 1984). The internal psychometric properties of instruments must be specifically examined using methods rooted in IRT or CFA to determine if the items measure equally well for comparison groups. The next chapter focuses on the issue of internal, or measurement, bias, its manifestations, and its effects on item responses.

CHAPTER 3

Measurement Bias (a.k.a., Differential Functioning)

Unlike predictive bias, which addresses differences in the relationship between an assessment and an external criterion, measurement bias concerns how a test's internal psychometric properties vary across comparison groups. For example, are the items equally difficult and do they discriminate equally well for members of reference and focal groups *after controlling for differences in ability (a.k.a. trait level)*? If so, then measurement equivalence is said to obtain (Drasgow, 1984); otherwise, measurement bias, properly referred to as *differential functioning*, may be present.

A key issue in identifying differential functioning lies in its distinction from impact. As mentioned previously, *impact* refers to a "true" difference across comparison groups in the distribution of the trait a test is designed to measure; that is, the groups differ in a substantive or meaningful way in terms of the skill that is assessed, rather than due to artifacts or spurious factors associated with problems in the measuring instrument. Impact is accounted for in an IRT framework through mathematical transformations, or processes broadly referred to as linking methods, that put quantities estimated in different groups on a common scale for comparison purposes. On the other hand, differential functioning occurs when comparison groups differ in their expected item or test scores after accounting for impact. In particular, d*ifferential item functioning* (DIF) is said to occur when comparison groups differ in their probability of correctly answering an item after conditioning on, or controlling for, trait level (Hambleton & Swaminathan, 1985;

Hulin, Drasgow, & Parsons, 1983; Lord, 1980; Shealy & Stout, 1993). And, *differential test functioning* (DTF) is said to occur when DIF at the item level accumulates to produce differences in expected test scores.

Because IRT methods do not confound differential functioning with impact, mean differences in observed test scores across comparison groups can be broken down into components, as shown, Mean Difference = DTF + IMPACT, and the relative effects of these two sources of variation on test scores and selection rates can be examined to determine if test revision is warranted (Stark et al., 2004). (Note that DTF can be further decomposed into individual DIF results.) For example, consider the two hypothetical observed score distributions shown in Figure 2. For convenience, these distributions were chosen to be normal and equal in variance, but, in practice, this need not be the case.

In the figure, it can be seen that Group 2 has a substantially higher mean than Group 1, which would result in disproportionately more members from Group 2 being hired, licensed, or admitted into an organization, regardless of the cut score used for decision making. If this mean difference in test scores occurred using an unbiased test (i.e., it is attributable solely to impact), then modifying or replacing items would not be necessary from a psychometric standpoint and would not necessarily cure the disparity in selection rates. On the other hand, if the groups had identical trait distributions, yet the test scores differed as shown due to internal problems with the instrument, then test revision would be required and the disparity in selection rates would be mitigated. Finally, if the difference occurred because of both DTF and impact, then one would need

to judge whether modifying the test would have any practical effect on decision making and act on the basis of ethical and legal grounds.

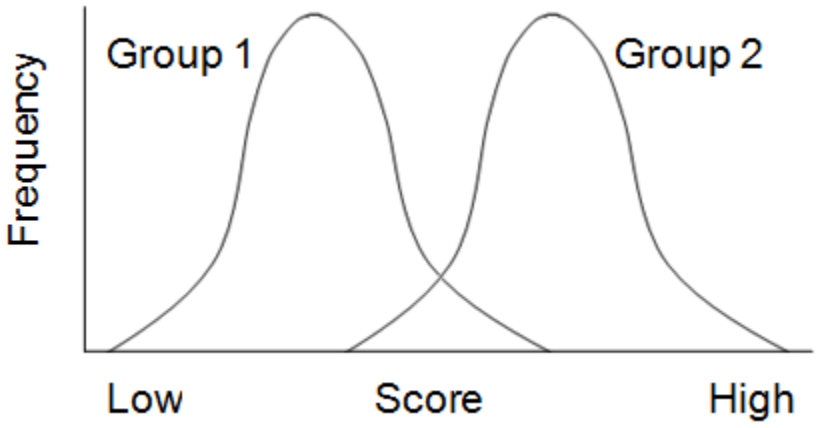

Figure 2. Hypothetical distributions of observed test scores for two comparison groups.

Item Response Functions

In the IRT framework, it is customary to illustrate the relationship between the probability of a correct response and examinee trait level graphically, using what is referred to as an *item response function* (IRF). Basically, an IRF represents the nonlinear regression of an item's expected score on examinee trait level (Hambleton & Swaminathan, 1985) with the horizontal axis representing examinee trait level, θ, and the vertical axis the probability of a correct or positive response. IRFs are useful for visually examining the quality of test items because the steepness of a curve indicates how well an item differentiates among examinees of different trait levels and its lateral position along the trait axis indicates its difficulty. Example IRFs for items differing in discrimination and difficulty are presented in Figure 3.

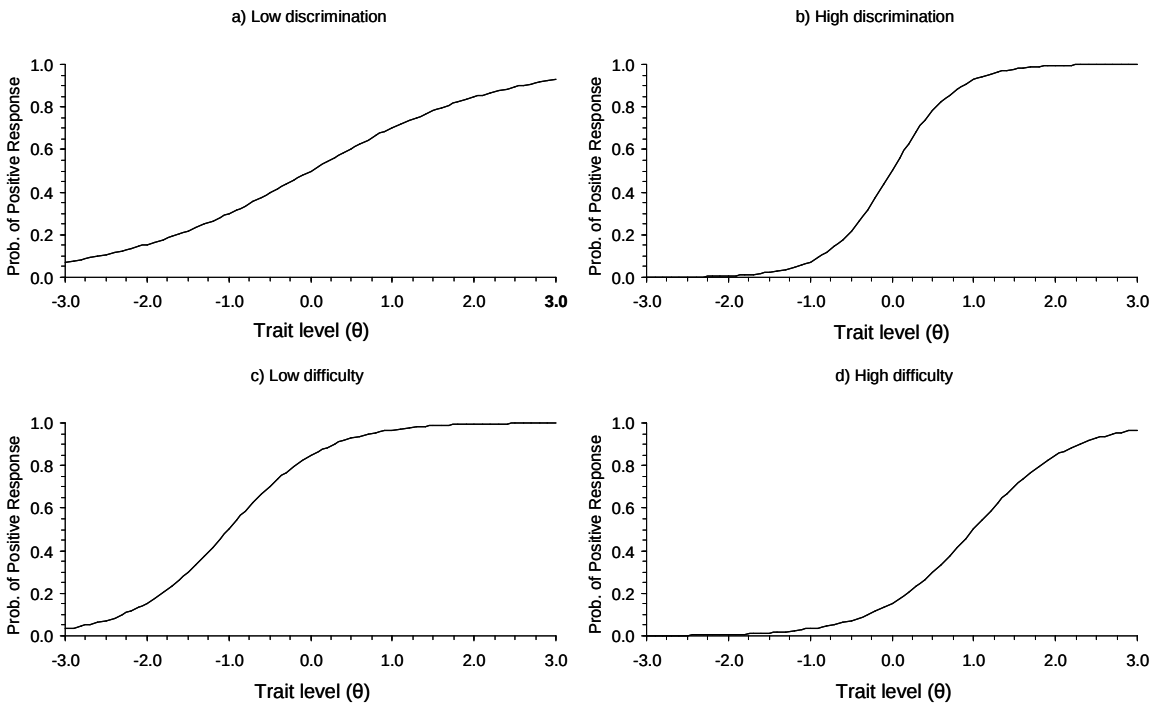

Figure 3. Example item response functions (IRFs).

Figure 3a presents an IRF having a shallow or relatively flat slope, which indicates that the probability of a positive response does not change rapidly as a function of trait level. This item exhibits low discrimination and is less informative for examinees, on average, than an item having a steep IRF, such as the one shown in Figure 3b. The item represented in Figure 3b exhibits good discrimination for most examinees, because the probability of answering it correctly varies markedly in the middle regions of the trait continuum, even over relatively narrow ranges of θ. It does not, however, discriminate well for examinees having trait levels beyond -/+1.5, where the IRF is relatively flat. Note also that the items represented in panels (a) and (b) are equally difficult, with both showing response probabilities to .5 at $\theta = 0$.

Panels (c) and (d) of Figure 3 present IRFs for items that are equally discriminating, but different in difficulty. The difficulty of an item is indicated by its lateral position along the horizontal axis. The item presented in panel (c) is relatively easy, because the probability of a correct response is high even at a low trait levels (e.g., θ = -1), whereas the item presented in panel (d) is relatively difficult because even examinees having high trait level (e.g., θ = +1) have only a .5 probability of answering it correctly.

If one works with parametric models that formally specify the relationship between trait level and item response probabilities using a mathematical model, then to compute an IRF, one must first estimate item parameters from the response data, using a procedure, such as marginal maximum likelihood estimation (MML; Bock & Lieberman, 1970). These item parameter estimates can then be substituted into the equation for the model to compute response probabilities at various trait levels for plotting. With nonparametric methods, on the other hand, there are usually no formal models for item responding. Instead, only general assumptions are made about the shapes of IRFs (e.g., monotonicity) and the IRFs are estimated directly from the observed response data. The purported advantages of nonparametric methods are that sample size requirements are smaller than with most parametric methods and less restrictive assumptions about the shape of IRFs may make them applicable to a wider variety of item types (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). The lack of parameters, however, may reduce interpretability when psychometric problems are suspected and item revision is warranted.

In the context of DIF analysis, where one wishes to determine if an item is biased, an IRF must be estimated for each comparison group. The differences between the curves for each item must then be tested for statistical significance by comparing the groups' item parameters, response probabilities, or the area between their IRFs. If the results of these tests are nonsignificant, measurement equivalence is said to hold; otherwise, DIF is present. Graphical examples of items showing measurement equivalence and various types of DIF across reference and focal groups are shown in Figure 4.

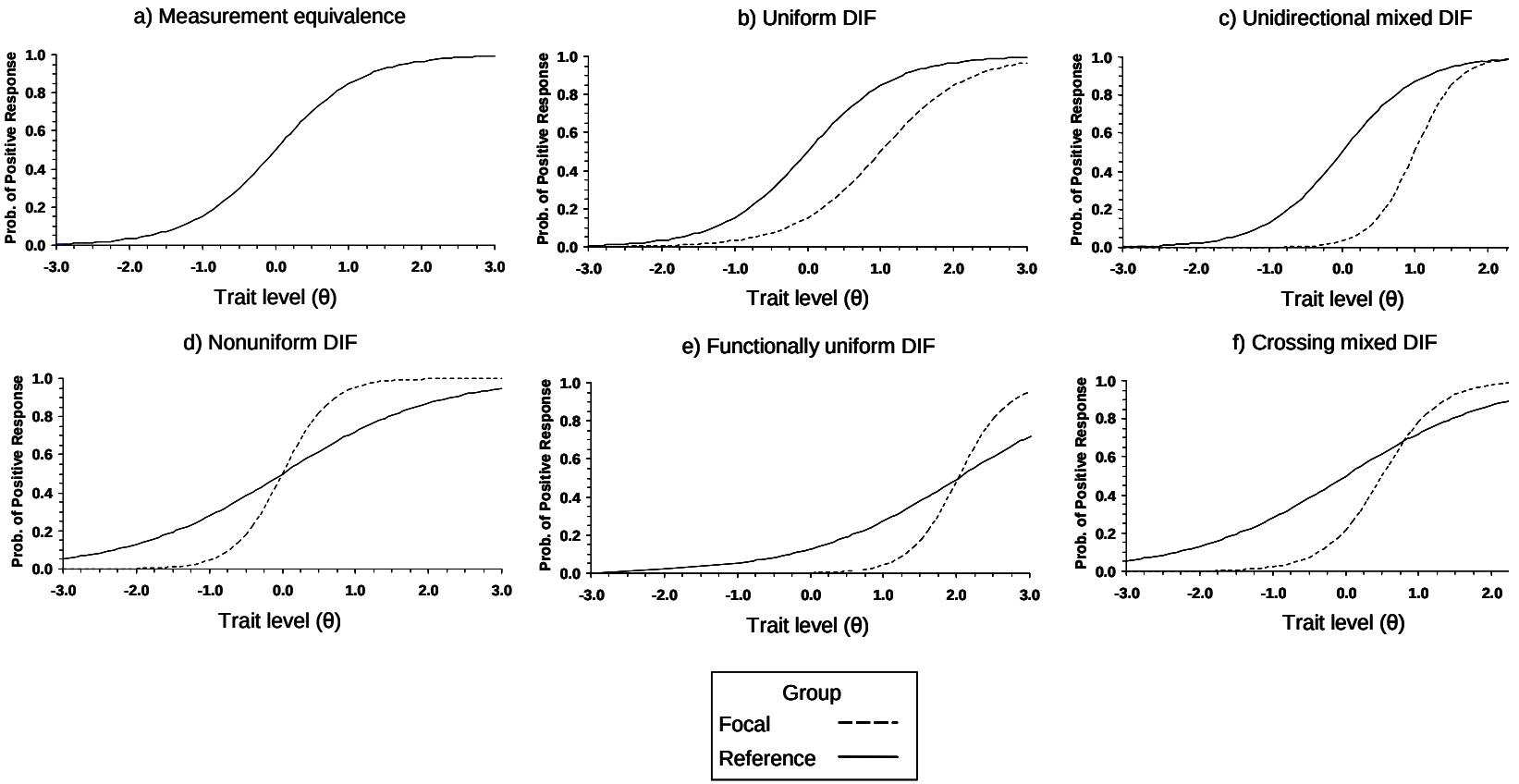

Figure 4. Example of a non-DIF item (a) and items exhibiting various types of DIF against the focal group (panels b through f). (*Note*: DIF = differential item functioning.).

Figure 4a presents the case for an item showing measurement equivalence (i.e., a non-DIF item). After accounting for impact, the reference and focal group IRFs are virtually identical, thus yielding curves that overlap. This is the ideal scenario for every test item because examinees having the same trait level have equal probabilities of responding correctly, regardless of the group to which they belong. In practice, however, small differences in the IRFs across groups are likely to occur, not only because of estimation error, but possibly due to violations of model assumptions, such as multidimensionality, which by some accounts is the underlying source of DIF. Shealy and Stout (1993) and Camilli (1992), for example, postulate that DIF occurs when comparison groups differ along a secondary dimension that influences item responding on the primary attribute measured by a test (interested readers should refer to those papers for details). Panels (b) through (e) of Figure 4 present examples of items showing DIF in reference and focal group comparisons.

Figure 4b shows the case for an item that is equally discriminating across comparison groups but exhibits a difference in difficulty. Note that the slopes of the IRFs are identical, but the focal group IRF is shifted toward higher trait levels even after accounting for impact. This item is therefore said to exhibit *unidirectional DIF* (Li & Stout, 1996) against the focal group, because the focal group response probabilities are lower across the trait continuum, except at extremes where the IRFs converge. Because this unidirectional DIF results only from differences in difficulty, the term *uniform DIF* is also used to describe this situation (Mellenbergh, 1982). This type of DIF is analogous to the predictive bias scenario in Figure 1b where the focal group regression line is higher than that of the reference group because of an intercept difference.

Figure 4c presents another type of unidirectional DIF, which results from differences in both difficulty and discrimination. As can be seen, the focal and reference group IRFs are neither parallel nor crossing, resulting in a situation referred to as *unidirectional mixed DIF* (Li & Stout, 1996; Rogers & Swaminathan, 1993). This scenario is analogous to Figure 1c where the regression lines differ in both intercept and slope, but do not intersect, thus indicating an ordinal interaction.

Panels (d) through (f) of Figure 4 exhibit three types of *nonuniform* DIF (Mellenbergh, 1982), in which neither group is favored consistently across trait levels. Panel (d) presents a situation in which the reference group is favored at low trait levels and the focal group is favored at high trait levels, solely because of differences in item discrimination. Note that the IRFs, in this case, cross to produce a disordinal interaction (Swaminathan & Rogers, 1990), like the one shown in Figure 1d.

On the other hand, Figure 4e illustrates another type of nonuniform DIF item (i.e., group IRFs exhibit equal difficulty but unequal discrimination) in which the IRFs cross at such high trait levels that its manifestation is indistinguishable from unidirectional DIF (Figure 4b), thereby making identification of DIF type inherently problematic (Li & Stout, 1996). Going forward, this special case is therefore referred to as *functionally uniform* DIF.

Finally, Figure 4f demonstrates a type of nonuniform DIF referred to as *crossing mixed DIF* (Li & Stout, 1996; Rogers & Swaminathan, 1993). In this scenario, the focal and reference group IRFs have unequal difficulty and discrimination to an extent that is sufficient to produce IRFs that cross, thus differentiating this case from the previously

described unidirectional mixed DIF. This scenario is reflective of the predictive bias example shown in Figure 1e in which an ordinal interaction is present.

The examples of DIF described above illustrate some of the most common manifestations of this phenomenon, but are by no means comprehensive. For example, the direction of DIF could have been reversed in each case, producing a bias that, on average, favored the focal group over the reference group. Although such findings might seem surprising and defy conventional wisdom, they are not uncommon with real test data (e.g., Drasgow, 1987; Stark et al., 2004).

CHAPTER 4

DIF Detection Using IRT

The Standards for Educational and Psychological Testing (AERA et al., 1999) state that all assessments should be screened for items that may exhibit DIF in order to ensure fairness in testing. Simple statistical or classical test theory approaches, such as comparisons of proportion correct scores (*p*-values) or analysis of variance (ANOVA) tests for group by item interactions, are ineffective in this regard, because they confound differential functioning with impact (Drasgow, 1987). At present, there are two broad classes of methods capable of DIF detection: those based on CFA and those based on IRT (Glöckner-Rist & Hoijtink, 2003; Meade & Lautenschlager, 2004, 2004b; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006). CFA methods, such as mean and covariance structure analysis (MACS, Sörbom, 1974), involve a linear model that is well suited for polytomous data and may be extended to tests designed intentionally to be multidimensional. However, because most tests in educational and organizational settings are designed to measure one dominant dimension at a time, IRT methods involving nonlinear models that are suitable for both polytomous and dichotomous data are more often recommended for DIF detection (e.g., Drasgow & Hulin, 1990). Furthermore, because IRT offers a wide array of item response models and DIF detection methods, it is possible to identify those that are most consistent with theoretical and practical considerations in each application and, if necessary, to conduct comparative examinations of appropriateness.

Parametric vs. Nonparametric DIF Detection

In the context of IRT-based DIF detection, the term *parametric* refers to whether parameters characterizing items and persons are estimated explicitly during the course of data analysis. Parametric methods require one to specify a formal model for item responding and to estimate its associated parameters as a way of comparing item properties across reference and focal groups. *Nonparametric* methods, on the other hand, detect DIF by comparing item and test scores obtained directly from examinee responses, and thus circumvent some of the steps and difficulties that may be encountered during parameter estimation (Raju & Ellis, 2002). When DIF is found, nonparametric methods arguably provide little insight into the potential causes of DIF whereas parameters that have substantive interpretations might provide some guidance as to how an item or test could be revised. For this reason, parametric DIF detection methods are often preferred. Nonetheless, whether one chooses a parametric or nonparametric method, the distinction between differential functioning and impact remains clear (Drasgow & Hulin, 1990; Hulin et al., 1983; Stark et al., 2006). The way impact is handled by parametric and nonparametric DIF detection procedures differs, however, as do the mechanics of their implementation.

Parametric DIF Detection

In general, parametric DIF detection methods involve the following sequence of steps. First, a model for item responding is chosen based on theoretical and practical considerations. Theoretical considerations could include the nature of the construct being assessed (cognitive or noncognitive), the response format (e.g., dichotomous or polytomous with ordered/unordered categories), and the possibility of response sets (e.g.,

guessing or impression management), whereas practical considerations could include the available sample size (models having more parameters require larger samples) and test length (longer tests yield better parameter estimates). Once an IRT model is chosen, its item parameters (e.g., discrimination and difficulty) must be estimated for both the reference and focal groups (see Hambleton & Swaminathan, 1985). To account for impact, parameters can be estimated simultaneously, using a procedure known as *concurrent calibration,* or estimated separately for each group and then placed on a common scale by a linear transformation through a process known as *linking*. In either case, model-data fit is examined and, if reasonably good fit is observed, DIF detection proceeds by comparing the item parameters or IRFs of the reference and focal groups using a statistical test with an *a priori* specified significance level. If an item shows a statistically significant difference between the reference and focal groups, then the null hypothesis of "no DIF" is rejected and the item is flagged for further inspection and revision or removal.

Within the parametric framework, DIF is said to be present when an item exhibits IRFs or parameters that differ across comparison groups beyond what is expected due to sampling and estimation error (Hulin et al., 1983; Lord, 1980). Although it is advisable to form hypotheses in advance about which test items might show DIF and why, this is difficult, if not impossible, to do; so, every item is usually examined in practice. DIF findings are often unintuitive but, in some cases, parametric methods provide insights for item revision through post hoc inspections of content. For example, if a mathematical reasoning item shows a higher difficulty parameter for an English-as-second-language focal group relative to a native English speaker reference group, its content could be

reviewed to determine if abstruse vocabulary or ambiguous wording could have unintentionally affected what the item was designed to measure. Yet, explanations for differences in discrimination, or differences in difficulty that run counter to all reasonable expectations, typically remain elusive.

The benefits of parametric approaches come with a cost; namely, some models involve strong assumptions about the nature of the response data. For example, with most parametric models, one must verify that the response data are *essentially unidimensional* (Stout, 1990); that is, one dominant or prepotent dimension underlies item responding. With dichotomous data, tests for unidimensionality typically involve linear principal axis factoring of item tetrachoric correlations or procedures, such as modified parallel analysis, designed specifically to determine whether the data are sufficiently unidimensional for the application of a unidimensional IRT model (Drasgow & Lissak, 1983; Drasgow & Parsons, 1983; Hulin et al., 1983). In practice, the unidimensionality assumption is not a severe limitation to model application, because most parameter estimation procedures have been shown to be robust to weak to moderate violations (Drasgow & Lissak, 1983; Kirisci, Hsu, & Yu, 2001). However, even if the dimensionality of the response data and the model chosen for parameter estimation are in agreement, one must still examine model-data fit at the item level using graphical and/or statistical methods (Drasgow, Levine, Tsien, Williams, & Mead, 1995). If the model does not fit the data well, then the benefits of using IRT for DIF detection might be diminished or even negated. So, one must make an informed choice, based on previous simulation research, as to whether one should proceed with fitting a theoretically

appropriate model, as planned, or consider alternative models that provide better fit empirically as the basis for DIF detection.

Parametric DIF Detection Methods

Area Methods. A wide variety of parametric methods are available for detecting DIF. Some focus on comparing the area between IRFs estimated for reference and focal groups after establishing a common or base metric (Raju, 1988, 1990; Raju, van der Linden, & Fleer, 1995). These are known as *area methods*. An advantage of area methods is that they can be adapted for use with a wide variety of models, but the disadvantage is that the distributions of the test statistics are usually unknown and, thus, require computationally intensive resampling methods to obtain critical values for hypothesis testing. Methods involving the direct comparisons of item parameters are therefore more common.

Lord's Chi-Square. A popular item parameter comparison method is Lord's (1980) chi-square, which can be used to test for differences in one or more item parameters simultaneously across reference and focal groups. Vectors of item parameter differences and the inverse of the variance-covariance matrix for these differences are used to calculate a chi-square statistic that is compared to a critical value based on an *a priori* specified level of significance, with degrees of freedom (df) corresponding to the number of parameters examined for each item. If the observed chi-square exceeds the critical value, then the null hypothesis of no DIF is rejected. The advantages of this method are that it is readily adapted to any parametric model, critical values are easily obtained for different df and levels of significance, and the index is sensitive to both uniform and nonuniform DIF. The disadvantage, however, is that it requires a complete

variance-covariance matrix, elements of which are not readily available from some common MML estimation programs for polytomous models. Baker (1992) suggested setting off-diagonal elements of the variance-covariance matrices to zero in such situations, but the extent to which this simplifying assumption affects DIF detection is an open question. Thus, polytomous test data are often examined using an alternative DIF test, such as the likelihood ratio test, which uses a model-comparison approach similar to what is done in CFA investigations.

IRT Likelihood Ratio Test. The item response theory likelihood ratio test for DIF (IRT-LR; Thissen et al., 1988) can be used with both polytomous and dichotomous data, and since its advent has proven to be one of the most effective methods for detecting uniform and nonuniform DIF and DTF. Essentially, the test involves comparing the goodness of fit statistics for a series of nested models in which parameters for items suspected of DIF are constrained, or alternatively allowed to vary, relative to a baseline model. If the change in the goodness of fit statistic, which is distributed roughly as a chi-square, exceeds the critical value with df equal to the number of parameters in question, then the item is flagged for DIF; otherwise, the null hypothesis of no DIF is retained.

As discussed by Stark, Chernyshenko, and Drasgow (2006), implementations of the IRT-LR test vary widely in the literature. However, some implementations are more effective for DIF detection and control of Type I error than others, with Type I error rates varying from the expected level of .05 to levels higher than .90 in some cases. In particular, the approach suggested originally by Thissen et al. (1988), referred to here as the *constant* anchor item (or free-baseline) method, has been shown to outperform the alternative approach, known as the *all-other* anchor (or constrained-baseline) method,

when multiple DIF items are present (Stark et al. 2006; Wang & Yeh, 2003). This is because the constant anchor method begins by specifying a baseline model that has the best chance of fitting the reference and focal groups' response data - namely, one in which the parameters for all items are free to vary, except for a presumably DIF free anchor item, or subset of items, that is needed to identify the latent metric. Comparison models for DIF analyses are formed by constraining one at a time in addition to those in the anchor subset, which remains constant across comparisons. In contrast, the all-other approach begins with a baseline model in which the parameters for all items are constrained across reference and focal groups, and comparison models for DIF analyses are formed by freeing one item at a time in succession. The baseline model thus changes by necessity across iterations and the fit is certain to be adversely affected if a test contains any DIF items at all. Research to date suggests that the greater the number of DIF items, the more "contaminated" is the baseline model, and the higher is the likelihood of Type I errors (e.g., Finch, 2005; Wang & Yeh, 2003).

Stark et al. (2006) proposed constraining one item at a time, in addition to the anchor subset, and testing for DIF on all item parameters simultaneously (an omnibus test) to avoid issues, such as partial invariance that sometimes arise with CFA. The results of their simulation indicated that, for both IRT and CFA applications, the constant anchor implementation of this omnibus test was more effective for uniform and nonuniform DIF detection and control of Type I error than the traditional all-other implementation. These results supported and extended the findings of Wang and Yeh (2003), which showed that the constant anchor implementation was more effective than the all-other method under a variety of realistic testing conditions. In addition, a recent

study by Lopez Rivas, Stark, and Chernyshenko (2009) demonstrated that high power for DIF detection could be achieved with the constant anchor method by using just one well discriminating unbiased referent. Furthermore, like other recent studies (Wang, 2004; Wang & Yeh, 2003), it was found that power could be increased by using a group of unbiased anchor items (three), but improvements were generally small when the anchor group was expanded from three items to five.

In summary, although IRT-LR procedures for detecting DIF may be seen by some as cumbersome, they have been shown overall to be effective and versatile. A preponderance of studies have examined the efficacy of the all-other IRT-LR implementation and found that it works well under many conditions. However, recently proposed constant anchor item implementations are beginning to garner more attention in the literature because of their higher power to detect DIF with better Type I error rates in realistic testing situations. Consequently, both the constant and all-other implementations of the IRT-LR test were explored in the Monte Carlo study that is described later in this presentation.

Nonparametric DIF Detection

Nonparametric DIF detection methods include a variety of approaches revolving around contingency tables and regression. Typically, these methods assume only that the underlying IRFs are monotonic, so that item responses can be summed to obtain a total or number correct score for each examinee that serves as an estimator of the latent trait, θ, often discussed in conjunction with parametric models. Essentially, proportion correct scores for items, or bundles of items, are compared across reference and focal groups

after conditioning on number correct or total test scores, and various statistical corrections (e.g., Jiang & Stout, 1998) are used to distinguish DIF and DTF from impact.

Besides making fewer assumptions about the data, another key advantage of nonparametric methods is that they do not require item parameter estimation. These methods may therefore provide better DIF detection than parametric methods for analyses involving small samples. However, the lack of parameters may make it more difficult to understand the source(s) of DIF when items are flagged. In addition, because number correct score is only a good estimator of theta for long tests (e.g., 20 items or more), the performance of these methods with shorter tests is open to question.

Nonparametric DIF Detection Methods

Mantel-Haenszel. Numerous nonparametric DIF detection methods have been developed, and one of the most widely used and extensively researched is the Mantel-Haenszel procedure (MH; Holland & Thayer, 1988). MH detects DIF by comparing the odds ratios of item endorsement frequencies across reference and focal groups after matching examinees on the trait measured by the test using total test scores. Specifically, the reference and focal groups are split into *K* subgroups, representing different levels of the total test score, and at each score level, a 2 x 2 contingency table is constructed showing group membership as a function of item response frequency. The odds of correctly answering an item at each score level is obtained for the reference and focal groups, and the results are aggregated across score levels to compute the MH statistic, which is distributed as a chi-square with 1 df. If the observed MH exceeds the critical MH value (3.84), then the item is flagged as exhibiting DIF, and the process is repeated

for the remaining items. (For details on the computations, interested readers are referred to Holland and Thayer [1988] and Hidalgo and López-Pina [2004].)

Unlike parametric methods that involve concurrent calibration or linear transformations of parameter estimates to account for impact, impact is handled with MH by matching examinees on the ability, or latent trait, estimate prior to comparing response frequencies across groups. Moreover, to aid in interpretation of MH results, a logarithmic transformation (Holland & Thayer, 1988) is sometimes applied to produce a scale that is symmetric about an origin of zero. In that situation, negative values indicate that an item exhibits DIF against the focal group, whereas positive values indicate DIF against the reference group. Some researchers have proposed using the magnitude of the transformed values as a way of gauging the practical importance of DIF (e.g., Zwick & Erickan, 1989), but whether or not such log odds results provide meaningful indications of effect size is left to readers to determine.

In summary, the MH procedure is a straightforward and adaptable method for detecting DIF but it has two strong limitations. First, although it can be extended to handle polytomous and even multidimensional data by expanding contingency tables and using more than one test score for matching examinees (e.g., Mazor, Kanjee, & Clauser, 1995), the number of score categories quickly becomes large and problems with low cell frequencies can arise. Second, research has shown that the MH procedure is generally ineffective at detecting nonuniform or crossing DIF. This limitation has led to the development of other procedures such as the crossing simultaneous item bias test (CSIBTEST; Li & Stout, 1996) and the application of logistic regression to DIF detection (LOGREG; Swaminathan & Rogers, 1990). These procedures have become increasingly

popular alternatives in applied measurement settings because both can detect unidirectional and crossing DIF and they are easy to implement.

Crossing Simultaneous Item Bias Test. CSIBTEST was developed by Li and Stout (1996) as an extension of the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) for DIF, which is rooted conceptually in multidimensional IRT. According to Shealy and Stout, DIF occurs when examinees from comparison groups differ in their standing on a secondary dimension, which may or may not be meaningfully related to performance on the primary attribute measured by a test. If the secondary dimension is considered unrelated to the purpose of testing, it is then referred to as a nuisance dimension or nuisance determinant, and the items showing DIF are said to be "biased;" otherwise, the items are said to exhibit benign DIF against the affected group. (In educational measurement, a distinction is sometimes drawn between DIF and internal or measurement bias, which is overlooked in other fields, and which was not explicated earlier in this presentation for simplicity.)

CSIBTEST builds on the foundational work of SIBTEST, which was designed to detect unidirectional DIF by comparing reference and focal group responses after conditioning on a *matching subtest* score. The matching subtest can be specified by a test constructor if a subset of non-DIF items is available *a priori*, or it can be derived in a manner analogous to what was suggested by Thissen et al. (1988) for obtaining a non-DIF set of *constant anchor items* to begin the free-baseline implementation of the IRT-LR procedure. However, because a priori information about DIF is usually not available and because the length of the matching subtest is an important concern, the matching subtest is often taken to be *all items except the one under study, just as with the all-other*

implementation of the IRT-LR test. The "automatic" option of the SIBTEST program (Stout, 1999) uses this all-other approach to test successively each item in a measure for DIF against the focal group, reference group, or either group by using one-, one-, or two-tailed significance tests, respectively. Examinees in the reference and focal groups are matched on levels of the matching subtest score and a regression-based correction is used to adjust for bias in the estimation of trait level as well as impact (Jiang & Stout, 1998; Shealy & Stout, 1993). The null hypothesis of no DIF is tested for each item by computing a statistic called B_{uni}, which has a mean of zero and is approximately normally distributed in large samples. If the observed B_{uni} exceeds the critical value for a standard normal distribution at the desired level of statistical significance, or alternatively the observed *p*-value is less than the critical *p*-value, then a studied item is flagged as DIF.

CSIBTEST (Li & Stout, 1996) extended the SIBTEST methodology to include the capability to detect crossing DIF as well as unidirectional. The approaches to DIF detection are similar in that reference and focal group examinees are sorted into score categories based on matching subtest scores and the proportions of correct responses within these categories are compared. However, unlike SIBTEST, CSIBTEST also attempts to estimate the point on the trait range at which the reference and focal group IRFs cross. The computations for CSIBTEST and SIBTEST are almost identical, except that CSIBTEST partitions the sum of the group differences over the levels of the matching subtest scores into two components centered on the estimated crossing point, k_c. In other words, rather than one summation over all the levels of matching subtest scores,

$k = 0,1,...,n$, two summations are required: 0 to $k_c - 1$ and k_c.+ 1 to n. The CSIBTEST test statistic, B_{cro}, for the null hypothesis of no DIF is thus given by:

$$B_{cro} = \frac{\hat{\beta}_{cro}}{\hat{\sigma}(\hat{\beta}_{cro})}, \text{ where} \tag{1}$$

$$\hat{\beta}_{cro} = \sum_{k=0}^{k_c-1} \hat{p}_k \left(\bar{Y}^*_{Fk} - \bar{Y}^*_{Rk}\right) + \sum_{k=k_c+1}^{n} \hat{p}_k \left(\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}\right), \tag{2}$$

$$\hat{\sigma}\left(\hat{\beta}_{cro}\right) = \left(\sum_{k=0}^{k_c-1} + \sum_{k=k_c+1}^{n}\right)\left(\hat{p}_k^2\left(\frac{1}{J_{Rk}}\hat{\sigma}^2\left(Y|k,R\right) + \frac{1}{J_{Fk}}\hat{\sigma}^2\left(Y|k,F\right)\right)\right)^{1/2}, \text{ and} \tag{3}$$

$$\hat{p}_k = \left(J_{Rk} + J_{Fk}\right) \Big/ \sum_{j=0}^{n}\left(J_{Rj} + J_{Fj}\right). \tag{4}$$

In the equations above, *R* represents the reference group, *F* represents the focal group, $k = 0,1,...,n$ represents levels of matching subtest (*Y*) scores, and k_c represents the estimated crossing point for the comparison IRFs. If *G* is allowed to represent either the reference or focal group, then J_{Gk} represents the number of examinees in group *G* with matching subtest score level *k* and J_{Gj} is the number of examinees in each group thus $\hat{p}_k$ is the proportion of examinees at *k*, $\bar{Y}^*_{Gk}$ is the impact-adjusted mean for examinees on the studied item (or subtest/bundle of items, in the case of DTF analysis) having matching subtest score level *k*, $\hat{\beta}_{cro}$ is the average weighted difference between the two marginal IRFs or, in other words, the sum of the unidirectional DIF against the reference group in the lower trait level and the unidirectional DIF against the focal group in the

higher trait level, $\hat{\sigma}^2(Y|k,G)$ is the sample variance of the studied item (or bundle) among examinees having matching subtest score level *k*, and $\hat{\sigma}(\hat{\beta}_{cro})$ is the standard error of the B_{cro} test statistic. Note that because B_{cro} depends on k_c and there is "no easily derived distribution when no DIF exists," a randomization test (Edgington, 1987) is used to determine statistical significance (Li & Stout, 1996, p. 654).

According to Li and Stout (1996), an item exhibits crossing DIF if one group shows a significantly higher probability of a correct response in one trait range and the other group shows it in another. Therefore, the identification of DIF type depends upon where the group IRFs cross. This crossing point is estimated by regression and can consequently be outside of the matching subtest score range. Specifically, for a matching subtest of length *n*, if the crossing point occurs between zero and *n*, that is, between none correct (or endorsed) and all correct, then crossing DIF is implied. On the other hand, if the crossing point occurs at a value less than zero or greater than *n*, unidirectional DIF is implied.

Li and Stout (1996) conducted a series of simulations and found that CSIBTEST demonstrated better Type I error (i.e., .05 or less) and power rates for nonuniform DIF detection than both the SIBTEST and MH procedures. Furthermore, they illustrated CSIBTEST's capability to distinguish unidirectional DIF from crossing DIF. However, it is unknown how well this method works with tests that are relatively short, as well as with tests that contain a substantial proportion of DIF items. In addition, given that the matching subtest typically consists of all items except the one under study (an all-other anchor implementation), questions arise as to whether power and Type I error rates could be improved by using a constant anchor approach to DIF detection – namely a matching

subtest comprising only a subset of DIF-free items that remains the same across all reference and focal group comparisons. Doing so would allow one to examine the efficacy of the CSIBTEST methodology independently of the issue of contamination, but, in turn, raises questions as to the number of items needed to get reliable matching subtest score estimates; in IRT-LR terms, the issue is the length of the baseline model. It is also unknown how well this method performs in comparison with other methods capable of detecting nonuniform or crossing DIF, such as the IRT-LR test and logistic regression. Moreover, as Finch and French (2008) noted, CSIBTEST's ability to detect unidirectional DIF has been frequently overlooked.

Logistic Regression. The LOGREG approach to DIF detection was proposed by Swaminathan and Rogers (1990) to overcome the limitations of the MH procedure. Their aim was to develop a method that was capable of detecting both uniform and nonuniform DIF within dichotomous data, yet was computationally simpler than approaches requiring the explicit estimation of item parameters, such as Lord's chi-square (1980) or Thissen et al.'s IRT-LR test (1988). A key virtue of the LOGREG approach is that DIF analysis can be conducted easily using common statistical packages, such as SAS and SPSS. One must merely specify a series of comparison models involving examinee trait estimates (X), which are typically total test scores or total test scores computed without the item under study, a dummy variable for group membership (D), and the interaction of total test score with group membership (DX) as predictors. Dichotomous item scores for each studied item serve as the dependent variables. By statistically testing the change in the goodness of fit or R-square for compact models involving only total test score as a predictor versus augmented models that include the additional term(s) involving group

membership, one can determine if uniform or nonuniform DIF is present. (One can also examine the statistical significance of the beta weights associated with the group membership terms in the augmented models.)

To illustrate this process for detecting DIF, consider the logistic regression model, $P(u=1|z)=\frac{e^z}{1+e^z}$. Here, $P(u=1|z)$ represents the probability of a correct response to the studied item (the item suspected of DIF) and z represents a linear function of predictor variables. For the baseline model that includes only total score as a predictor, $z=\beta_0+\beta_1 X$. For the augmented model used to test for DIF, $z=\beta_0+\beta_1 X+\beta_2 D+\beta_3 DX$. If the goodness of fit improves or *R*-square increases significantly when adding the additional predictor terms (*D* and/or *DX*), then the item exhibits DIF. Specifically, if β_3 is significant, then the item is flagged for nonuniform DIF. If only β_2 is significant, then the item is flagged for uniform DIF. If neither β_2 nor β_3 are significant, then the null hypothesis of no DIF is retained.

Although similar in appearance to the IRT two-parameter logistic model (2-PLM; Birnbaum, 1968), the logistic regression DIF detection method is often referred to as nonparametric, because, like MH and CSIBTEST, it does not involve estimation of item or latent trait parameters. Instead, the LOGREG procedure uses total test score as a trait estimate (Millsap & Everson, 1993) and response probabilities are compared for reference and focal groups after taking trait differences into account.

Overall, simulation studies have shown that the LOGREG procedure provides as good or better uniform DIF detection than MH (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). Because it is also capable of detecting nonuniform DIF

and can be extended for use with tests designed intentionally to be multidimensional by the inclusion of additional predictor variables (Mazor et al., 1995), it is certainly worthy of consideration as a DIF detection tool. Its ease of implementation in almost any statistical software package broadens its appeal among non-psychometricians. However, studies of LOGREG power and Type I error rates have shown mixed results, with the effects of impact, for example, remaining unclear (e.g., Finch & French, 2007; Li & Stout, 1996; Narayanan & Swaminathan, 1996). Also unknown is whether LOGREG performance can be improved by using a constant subset of non-DIF items, rather than total test score, as a predictor in comparison models when DIF items are present. The issue of contamination is just as relevant here as with CSIBTEST and IRT-LR, so research is clearly needed to explore the effects of matching subtest length and implementation methods (all other vs. constant anchor) on LOGREG DIF detection.

Study Rationale and Objectives

This chapter has provided a brief review of parametric and nonparametric DIF detection methods. Some of these methods have been studied and their efficacy documented under a variety of conditions, while others are still relatively new and underutilized. Arguably, even so-called industry standards differ widely enough in terms of implementation that researchers are still discovering issues that may have substantial effects on power and Type I error rates. This has led to myths and confusion about which methods should be preferred and when. Moreover, what is considered realistic in the context of large scale testing programs is sometimes quite different from what is seen in typical industrial-organizational settings and in academic research outside of

measurement circles; so studies comparing multiple methods under conditions likely to be encountered in these situations are needed.

This paper proposes such a study. Specifically, the efficacy of the IRT-LR (Thissen et al., 1988), the LOGREG (Swaminathan & Rogers, 1990), and the CSIBTEST (Li & Stout, 1996) DIF detection methods was compared under a wide range of experimental conditions using a Monte Carlo simulation. The constant anchor implementation of the IRT-LR test has been shown to perform well in three recent investigations (Lopez Rivas et al., 2009; Stark et al., 2006; Wang & Yeh, 2003), so it served as a benchmark for examining the efficacy of the conventional all-other implementation of IRT-LR. In addition, both the popular LOGREG and the understudied CSIBTEST methods was examined using both a constant set of anchor items and an all-other implementation, in which all test items but the one under study serve as anchors.

The objectives of this study are twofold. The first objective is to compare the power and Type I error rates of CSIBTEST, LOGREG, and IRT-LR using dichotomous data exhibiting a wide array of DIF types – specifically those defined in Chapter 3 and illustrated in panels (b) through (f) of Figure 4: uniform, unidirectional mixed, nonuniform, functionally uniform, and crossing mixed. To the author's knowledge, this is the first study to examine the capacity of these procedures to detect multiple forms of DIF and will provide the most comprehensive assessment of their power to date.
The second objective of this study is to examine the accuracy of DIF classifications made by the selected methods. That is, can CSIBTEST, LOGREG, and IRT-LR reliably distinguish between the different types of DIF being simulated? In other words, when and why is the null hypothesis of no DIF rejected for the wrong reason (i.e., a Type III

error; Mosteller, 1948)? A recently study by Finch and French (2008) touched on this issue; the authors reported the frequency with which DIF items were erroneously detected, referring to them collectively as "anomalous Type I errors". This study examined the DIF classifications produced by CSIBTEST, LOGREG, and IRT-LR.

CHAPTER 5

Method

Study Design

A Monte Carlo simulation was conducted to examine the power, Type I, and Type III error rates of three DIF detection methods with independent variables manipulated as follows:

1. Method for detecting DIF: (a) IRT-LR, (Thissen et al., 1988), (b) LOGREG (Swaminathan & Rogers, 1990), and (c) CSIBTEST (Li & Stout, 1996);
2. Test length: (a) 15 items and (b) 30 items (to reduce noise in the Monte Carlo study, the 15-item test was nested within the 30-item test);
3. Sample size per group: (a) 250, (b) 500, and (c) 1000;
4. Impact: (a) None (both reference and focal group trait distributions were standard normal, $N[0,1]$), and (b) one half standard deviation against the focal group (the reference group distribution was standard normal and the focal group distribution was $N[-0.5, 1]$); and
5. Implementation of DIF test: (a) All-other (the baseline model, matching subtest, or trait estimates needed for the respective comparisons were created using all items except the one under study), and (b) Constant (the baseline model, matching subtest, or trait estimates needed for all DIF analyses were based on a preselected, fixed subset of non-DIF items).

In addition to the variables listed above, nested factors of particular interest were explored:

- Number of items in the anchor/matching subtest: (a) 5 items in the 15-item test length conditions and (b) 10 items in the 30-item test length conditions (the 10-item anchor group contained the five items from the 15-item test plus an additional five);
- Magnitude of DIF, defined as the area between comparison IRFs calculated using a formula provided by Raju (1988): (a) None, (b) 0.4 per DIF item, and (c) 0.8 per DIF item, with 5 DIF items specified in the 15-item test length conditions and 10 DIF items specified in the 30-item test length conditions; and
- DTF in the 0.4 and 0.8 DIF conditions: (a) DTF (DIF simulated to favor only one group and accumulates across items), and (b) No DTF (DIF simulated to favor both groups and cancels across items to produce negligible DTF).

In summary, this study involved 360 conditions representing unique combinations of the independent variables, as shown in Table 1. In each condition, 100 separate analyses, or *replications*, were conducted. Unique data sets were generated for both the reference and focal groups on each replication (details on data generation are presented later in this chapter), the DIF detection methods applied, and the power, Type I, and Type III error rates computed over replications.

Table 1.

Summary of Simulation Conditions for CSIBTEST, IRT-LR, and LOGREG

Test length	Sample size per group	Impact	DIF magnitude	DTF	Test length	Sample size per group	Impact	DIF magnitude	DTF
15	250	.0	.0	.0	30	250	.0	.0	.0
15	500	.0	.0	.0	30	500	.0	.0	.0
15	1000	.0	.0	.0	30	1000	.0	.0	.0
15	250	-.5	.0	.0	30	250	-.5	.0	.0
15	500	-.5	.0	.0	30	500	-.5	.0	.0
15	1000	-.5	.0	.0	30	1000	-.5	.0	.0
15	250	.0	.4	DTF	30	250	.0	.4	DTF
15	500	.0	.4	DTF	30	500	.0	.4	DTF
15	1000	.0	.4	DTF	30	1000	.0	.4	DTF
15	250	-.5	.4	DTF	30	250	-.5	.4	DTF
15	500	-.5	.4	DTF	30	500	-.5	.4	DTF
15	1000	-.5	.4	DTF	30	1000	-.5	.4	DTF
15	250	.0	.4	No DTF	30	250	.0	.4	No DTF
15	500	.0	.4	No DTF	30	500	.0	.4	No DTF
15	1000	.0	.4	No DTF	30	1000	.0	.4	No DTF
15	250	-.5	.4	No DTF	30	250	-.5	.4	No DTF
15	500	-.5	.4	No DTF	30	500	-.5	.4	No DTF
15	1000	-.5	.4	No DTF	30	1000	-.5	.4	No DTF
15	250	.0	.8	DTF	30	250	.0	.8	DTF
15	500	.0	.8	DTF	30	500	.0	.8	DTF
15	1000	.0	.8	DTF	30	1000	.0	.8	DTF
15	250	-.5	.8	DTF	30	250	-.5	.8	DTF
15	500	-.5	.8	DTF	30	500	-.5	.8	DTF
15	1000	-.5	.8	DTF	30	1000	-.5	.8	DTF
15	250	.0	.8	No DTF	30	250	.0	.8	No DTF
15	500	.0	.8	No DTF	30	500	.0	.8	No DTF
15	1000	.0	.8	No DTF	30	1000	.0	.8	No DTF
15	250	-.5	.8	No DTF	30	250	-.5	.8	No DTF
15	500	-.5	.8	No DTF	30	500	-.5	.8	No DTF
15	1000	-.5	.8	No DTF	30	1000	-.5	.8	No DTF

Notes. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. For each procedure, both the constant and all-other implementations were used. The number of DIF items and the number of items in the constant baseline conditions was equal to a third of test lenght condition (i.e., 5 for each in the 15-item test and 10 for each in the 30-item test).

In this Monte Carlo study, *Type I error* represented the proportion of times a non-DIF item was flagged incorrectly as a DIF item across replications; in other words, it is the number of false positives divided by the number of replications. On the other hand, *power* was defined as the number of times an item known to exhibit DIF is flagged by a DIF detection method; therefore, it is the number of true positives (a.k.a., hits) divided by the number of replications. When a hit occurs, the DIF type indicated by a detection

method can be checked for accuracy. If the DIF type was mischaracterized (e.g., an item known to exhibit nonuniform DIF is misclassified as a uniform DIF item), then a classification error, or *Type III error* (Mosteller, 1948), is said to occur. Type III errors were thus calculated as the number of misclassifications divided by the number of hits.

Data Generation

This simulation study focused on the efficacy of DIF detection with unidimensional dichotomous data. Therefore, data were simulated in accordance with a unidimensional model that reasonably characterizes the process of answering dichotomously scored test items. In this study, the three-parameter logistic IRT model (3-PLM; Birnbaum, 1968) was used for response data generation because it has been shown to fit cognitive ability data well in a multitude of studies over the last 30 years and, though its role in personality data remains unclear (Reise & Waller, 2002), it has been used to calibrate personality data when response sets, such as faking or impression management, were suspected (e.g., Stark et al., 2004).

In essence, an IRT model relates the psychometric properties of items and a respondent's standing on the underlying construct measured by a test to the probability of correctly answering (or endorsing) an item. The equation for the 3-PLM is shown below,

$$P\left(u_{ij}=1\middle|\theta_i,a_j,b_j,c_j\right)=c_j+\frac{1-c_j}{1+\exp[-1.7a_j(\theta_i-b_j)]}, \text{ where} \qquad (5)$$

i is an index for examinees, j is an index for items, u_{ij} represents the i^{th} examinee's scored response to the j^{th} item (u_{ij} = 1 if correct; 0 otherwise), $P\left(u_{ij}=1\middle|\theta_i\right)$ represents the

probability that an examinee will answer an item correctly given his or her trait level θ_i and the item's parameters (a_j, b_j, c_j), which reflect discrimination, difficulty, and the probability of obtaining the correct answer just by guessing, and 1.7 is a scaling constant that relates the logistic model item parameters to the metric of a normal ogive model (Reise & Waller, 2002, p. 91).

To generate item response data using the 3-PLM, trait scores and item parameters are needed. In this study, trait scores for examinees in reference and focal groups of various sizes were obtained by sampling values from independent normal distributions. For each examinee, a trait score, θ_i, is substituted into Equation 5 along with an item's parameters, (a_j, b_j, c_j), and the probability of a correct response is computed. A random number is then sampled from a uniform distribution and compared to this response probability. If the response probability is greater than the random number, then the item response, u_{ij}, is scored as 1; otherwise it is scored as 0. This process is repeated for all test items to obtain a response pattern for each examinee, and the method is repeated for other examinees until the desired number of response patterns has been created. In this study, 3-PLM response data were generated using the 3PLGEN computer program (Stark, 2000).

Constructing Tests for Simulations. Tests consisting of 15 and 30 items were created for the Monte Carlo study by randomly selecting item parameters from tables published by Narayanan and Swaminathan (1996), which showed item calibration results for an administration of the Graduate Management Admissions Test. One exception to this random process was the choice of Item 1, which was chosen to provide an item with reasonably high discrimination and moderate difficulty in the anchor set, as per the

findings of Lopez Rivas et al. (2009). Following the procedure used by Narayanan and Swaminathan (1996), the *c*-parameters for all items were set to 0.2, which is near the upper end of that parameter's typically observed range (Reise & Waller, 2002). Importantly, because manipulations involving *c*-parameters are difficult to interpret in the context of DIF studies and guessing is a realistic possibility in many testing applications, fixing the *c*-parameters reduces Monte Carlo noise.

Table 2 presents the item discrimination and item difficulty parameters for the resulting 15- and 30-item tests that were used to generate item responses for both the reference and focal groups. Trait scores for the respective groups were sampled from independent normal distributions, which may differ depending on whether impact is desired. Note that the items marked by an asterisk are the DIF-free subset that were used in the constant method conditions and items marked by two asterisks had their parameters shifted to produce DIF having magnitudes of 0.4 or 0.8 in those respective DIF conditions.

Table 2.

Generating Parameters for 15- and 30-Item Tests for No DIF Conditions

Item	*a*	*b*	Item	*a*	*b*
1*	1.05	0.10	16*	0.73	0.61
2*	0.44	-0.30	17*	1.11	-0.35
3*	0.55	-1.06	18*	1.32	0.57
4*	0.82	1.02	19*	0.55	1.09
5*	1.02	1.28	20*	0.92	1.13
6	0.82	0.61	21	0.64	-1.55
7	0.92	0.42	22	1.01	0.81
8	0.65	1.68	23	0.61	-0.53
9	0.29	-1.39	24	0.70	1.05
10	0.51	-0.09	25	1.02	0.64
11**	1.00	0.00	26**	1.00	-0.50
12**	0.56	0.00	27**	0.60	0.00
13**	0.56	2.00	28**	0.60	2.00
14**	0.56	0.00	29**	0.60	-0.28
15**	1.12	0.00	30**	1.20	-0.50

Note. * DIF-free item to be used in constant method conditions. ** DIF item in .4 and .8 DIF conditions. *a* = item discrimination. *b* = item difficulty. For 15-item test, items 1 to 15 were used. All item *c* -parameters (lower asymptotes) were set to .2.

Simulating Impact. To simulate impact against the focal group in selected conditions, trait scores for focal group examinees were sampled from normal distributions having means of -0.5 and variances of 1.0; reference group trait scores were sampled from standard normal distributions. In the "no impact" conditions, both the reference and focal group trait scores were sampled from independent standard normal distributions.

Simulating DIF and DTF. The most common way of simulating differential functioning is to construct tests that include a designated number or percentage of DIF items. DIF is then simulated on a particular item by shifting a comparison group's item parameters higher or lower by designated amounts prior to response data generation. (With the 3-PLM, it is conventional to manipulate only the discrimination, *a*, and

difficulty, *b*, parameters.) The magnitude and direction of these shifts can be held constant or allowed to vary across items, depending on the purpose of the study. Because this study aims to examine the power, Type I, and Type III error rates associated with different DIF types and magnitudes, the latter approach to simulating DIF was adopted. That is, each test included five DIF item types (one of each in the 15-item test, for a total of 5 DIF items, and two of each in the 30-item test, for a total of 10) and each of these reflected one of the DIF prototypes shown in Figure 4, panels b through f; thus, each DIF item in this study exhibited parameter shifts that achieved the desired DIF prototype and magnitude.

Fixed shifts can produce discrepant effects on the area between focal and reference group IRFs, depending on an item's discrimination and difficulty parameters, so the size of the shifts for items 11 through 15 and 26 through 30 were chosen empirically using an unsigned area DIF equation for 3-PLM, provided by Raju (1988):

$$(1-c)\left|\frac{2(a_F-a_R)}{1.7a_Ra_F}\ln\left(1+\exp\left(\frac{1.7a_Ra_F(b_F-b_R)}{a_F-a_R}\right)\right)-(b_F-b_R)\right|, \text{ where} \qquad (6)$$

the subscripts *R* and *F* represent the reference and focal groups, respectively, and *a*, *b*, and *c* represent the 3-PLM item parameters. By substituting different values of *a* and *b* into Equation 6, along with *c* = 0.2, the author identified shifts in the focal group *a*- and *b*-parameters that yielded the desired DIF magnitudes of 0.4 and 0.8.

The resulting parameters for the DIF items are shown in Tables 3 and 4 by DIF and test length condition. For the 0.4 and 0.8 DIF with maximum DTF conditions (Table 3), the

reference group item responses were generated using the parameters for the 15- and 30-item tests shown in Table 2, whereas focal group item responses were generated by substituting the appropriate values for items 11 through 15 and 26 to 30 given in Table 3. Note that the parameters in Table 3 can be compared to those marked with a double asterisk in Table 2 to observe the shifts that were required to achieve the desired DIF magnitudes.

Table 3.

Discrimination and Difficulty Parameters for Creating Desired Magnitudes of DIF with DTF

	Focal group (.4 DIF)		Focal group (.8 DIF)	
Item	*a*	*b*	*a*	*b*
11	1.00	0.50	1.00	1.00
12	0.85	0.00	1.79	0.00
13	0.85	2.00	1.79	2.00
14	0.81	0.25	1.50	0.50
15	1.40	0.50	2.01	0.99
26	1.00	0.00	1.00	0.50
27	0.95	0.00	2.28	0.00
28	0.95	2.00	2.28	2.00
29	0.89	0.00	1.81	0.22
30	1.48	0.00	2.09	0.50

Note. a = item discrimination. b = item difficulty. DIF values correspond to the area between the group item response functions (Raju, 1988). All item c- parameters (lower asymptotes) were set to .2.

For the 0.4 and 0.8 DIF with no DTF conditions (Table 4), two of the DIF item types: Unidirectional mixed (items 15 and 30) and functionally uniform (items 13 and 28), favored the focal group and the remaining DIF items favored the reference. Thus, the generating parameters for both groups are presented in Table 4. These DIF item types

were selected for this purpose because they generated the least DTF possible given the DIF item types employed in this study (see Appendix B for details).

Table 4.

Discrimination and Difficulty Parameters for Creating Desired Magnitudes of DIF without DTF

	.4 DIF				.8 DIF			
	Reference group		Focal group		Reference group		Focal group	
Item	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
11	1.00	0.50	1.00	0.00	1.00	1.00	1.00	0.00
12	0.56	0.00	0.85	0.00	0.56	0.00	1.79	0.00
13	0.85	2.00	0.56	2.00	1.79	2.00	0.56	2.00
14	0.56	0.00	0.81	0.25	0.56	0.00	1.50	0.50
15	1.12	0.00	1.40	0.50	1.12	0.00	2.01	0.99
26	1.00	0.00	1.00	-0.50	1.00	0.50	1.00	-0.50
27	0.60	0.00	0.95	0.00	0.60	0.00	2.28	0.00
28	0.95	2.00	0.60	2.00	2.28	2.00	0.60	2.00
29	0.60	-0.28	0.89	0.00	0.60	-0.28	1.81	0.22
30	1.20	-0.50	1.48	0.00	1.20	-0.50	2.09	0.50

Note. a = item discrimination. b = item difficulty. DIF values correspond to the area between the group item response functions (Raju, 1988). All item c- parameters (lower asymptotes) were set to .2.

DIF and DTF Manipulation Check. To ensure that the parameter shifts for the DIF items produced effects consistent with their condition, a series of manipulation checks were conducted. First, as can be seen from Figures 5 through 8, the primary difference between the DIF magnitude conditions (Figures 5 and 7 for 0.4 DIF and Figures 6 and 8 for 0.8 DIF) lies in the total area between the reference and focal group IRFs, with the areas being somewhat larger in the 0.8 condition; it should, therefore, be easier to detect DIF in those cases.

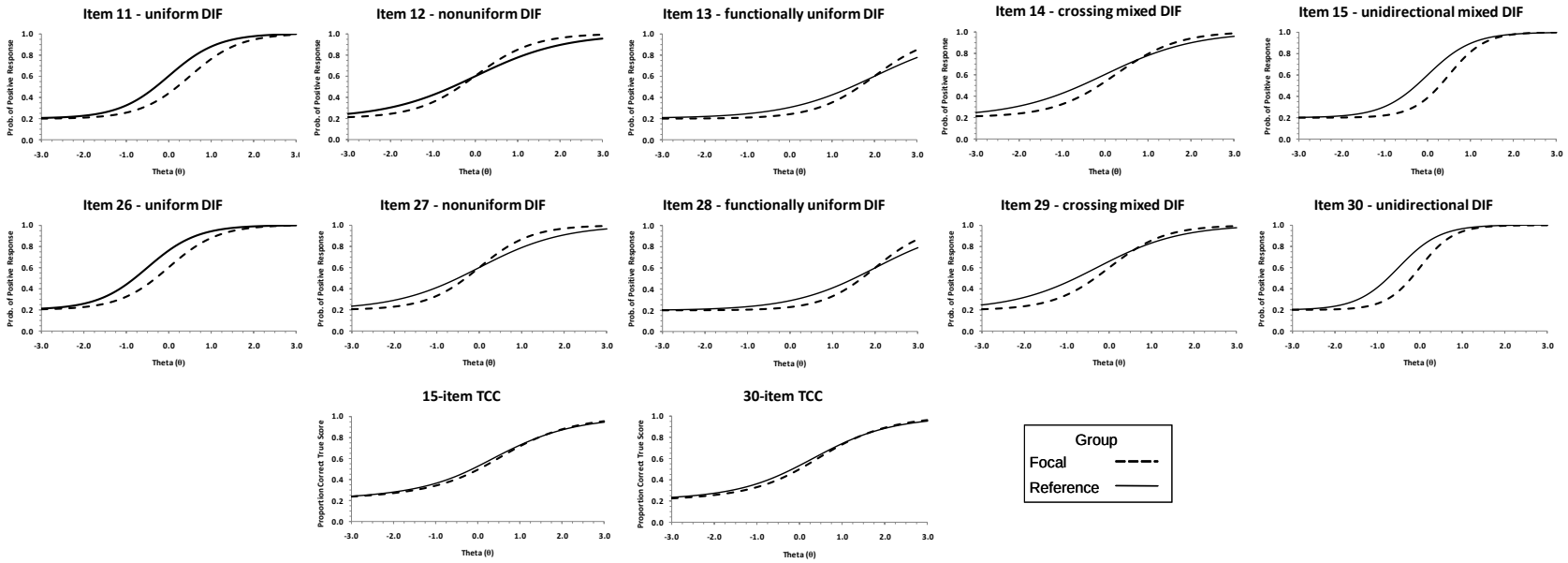

Figure 5. Reference and focal group item response functions (IRFs) and test characteristic curves (TCCs) for the 0.4 DIF magnitude per item with DTF conditions. (*Note*: DIF = differential item functioning and DTF = differential test functioning).

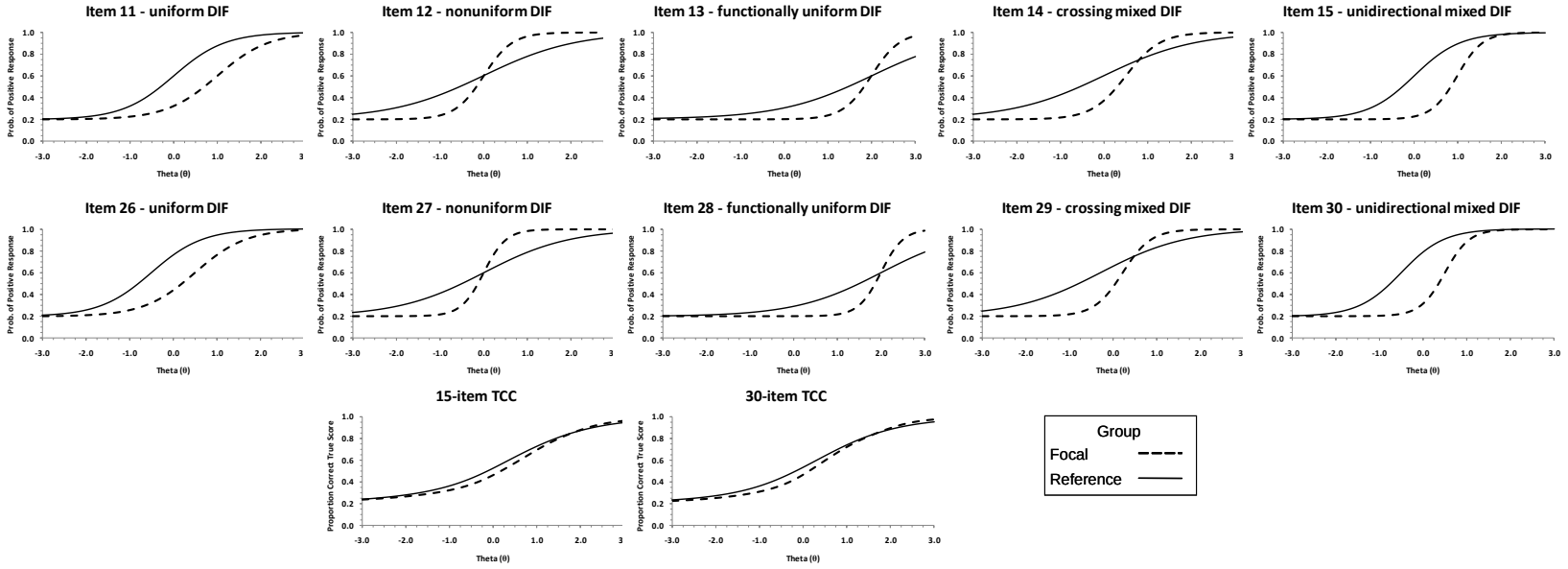

Figure 6. Reference and focal group item response functions (IRFs) and test characteristic curves (TCCs) for the 0.8 DIF magnitude per item with DTF conditions. (*Note*: DIF = differential item functioning and DTF = differential test functioning).

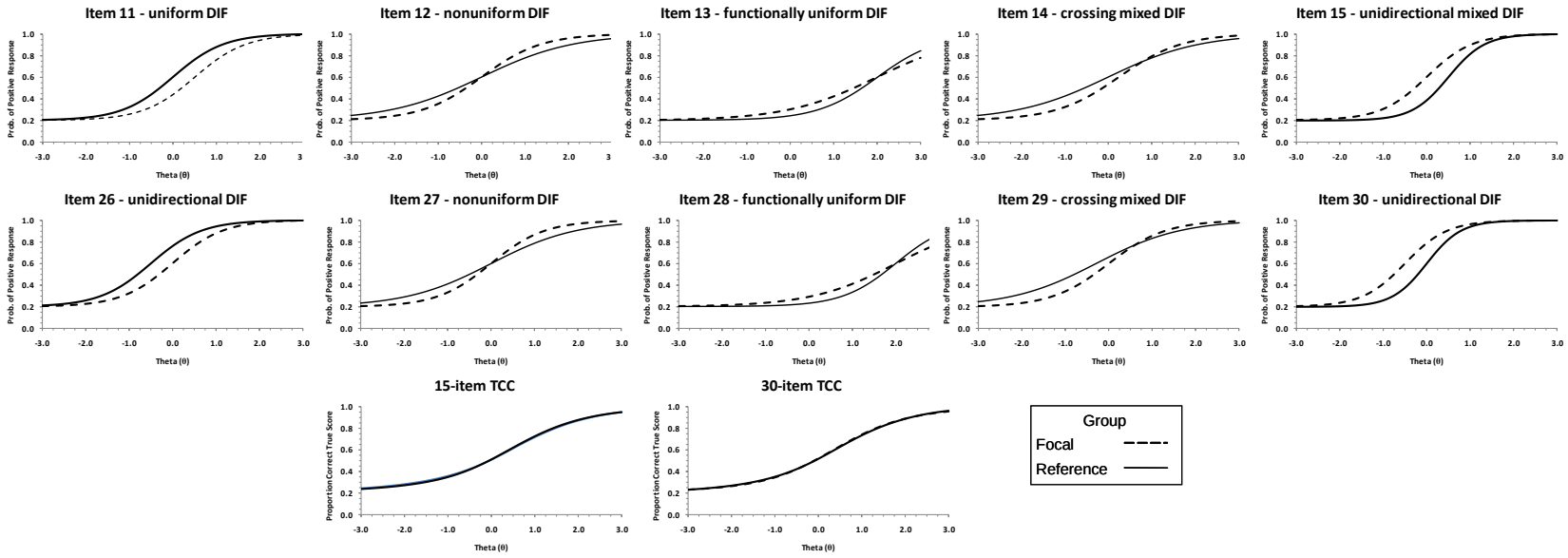

Figure 7. Reference and focal group item response functions (IRFs) and test characteristic curves (TCCs) for the 0.4 DIF magnitude per item with no DTF conditions. (*Note*: DIF = differential item functioning and DTF = differential test functioning).

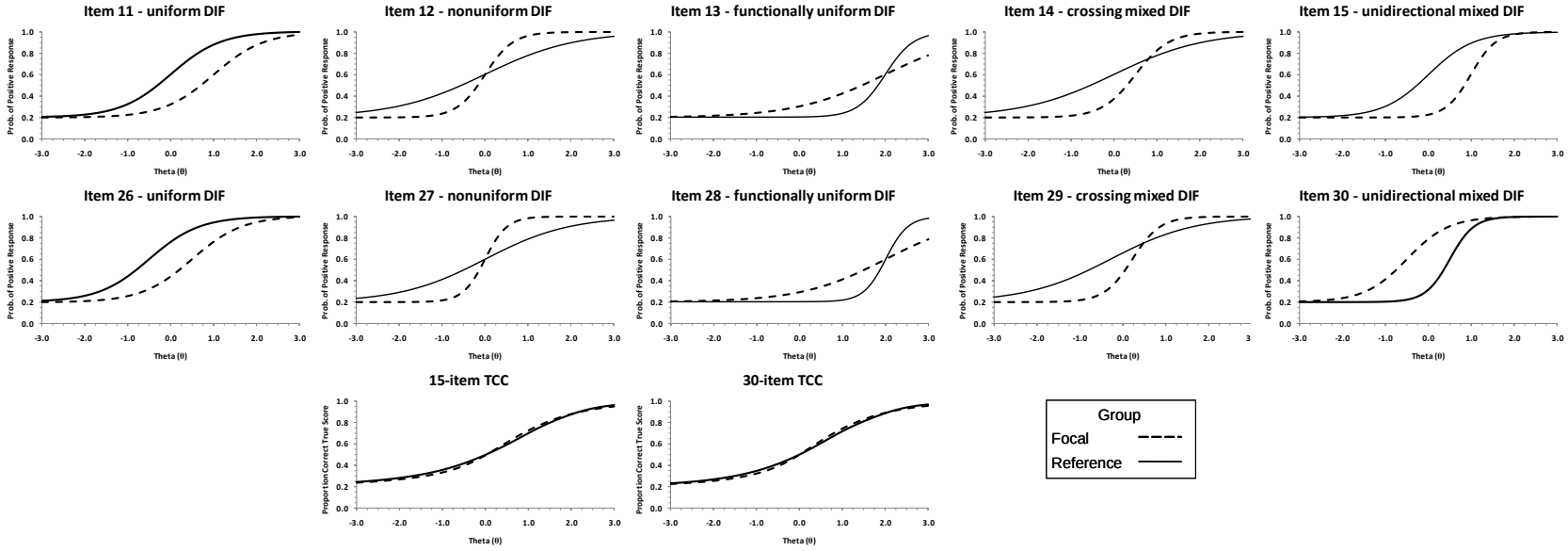

Figure 8. Reference and focal group item response functions (IRFs) and test characteristic curves (TCCs) for the 0.8 DIF magnitude per item with no DTF conditions. (*Note*: DIF = differential item functioning and DTF = differential test functioning).

Next, as can be seen from Figures 5 through 8, the DIF item generating parameters for all four DIF conditions exhibit a range of DIF types that coincided with the examples shown in Figure 4. Specifically, items 11 and 26 represent uniform DIF (unequal *b*-parameters and equal *a*-parameters) yielding reference and focal group IRFs that are nearly parallel over almost the entire trait range. Items 12 and 27 illustrate nonuniform DIF (unequal *a*'s and equal *b*'s) resulting in IRFs that cross in the middle of the trait range. Items 13 and 28 represent functionally uniform DIF (unequal *a*'s and equal, but extreme, *b*'s) exhibiting IRFs that cross outside the middle of the trait range. Item 14 and 29 show crossing mixed DIF (unequal *a*'s and unequal *b*'s) having IRFs that cross near the middle of the trait range. Item 15 and 30 represent unidirectional mixed DIF (unequal *a*'s and unequal *b*'s) resulting in IRFs that are not quite parallel; yet they do not cross until relatively low and/or high trait levels. Finally, in relation to the DTF manipulation, note how the group TCCs in Figures 5 and 6, representing the DIF with maximum DTF conditions, show that only the reference group benefits from DTF, whereas in Figures 7 and 8, representing the DIF with minimal DTF conditions, the TCCs are nearly identical or cross indicating that both groups benefit.

As an additional check of the DTF manipulation, a small simulation was conducted in which data sets of 1000 reference and 1000 focal group examinees were generated by sampling trait scores from independent $N(0, 1)$ distributions using the respective item parameters for the 15- and 30-item tests in the no impact, DIF conditions. The data were then analyzed using the differential functioning of items and tests computer program (DFITD4; Raju, 1995).

The focal and reference group TCCs calculated using the generating parameters are presented in Figure 9 by condition. As expected, the TCCs in the no DTF conditions overlap almost perfectly (panels c and g) or they cross (panels d and h) so that differences in expected scores cancel when averaging across trait levels. On the other hand, the TCCs for the 0.4 DTF (panels a and b) and 0.8 DTF (panels e and f) conditions show systematic, non-cancelling discrepancies favoring the reference group. These results, in conjunction with the remarkable similarity of the group TCCs shown in Figures 5 through 9, indicate that the objective of minimizing DTF was met.

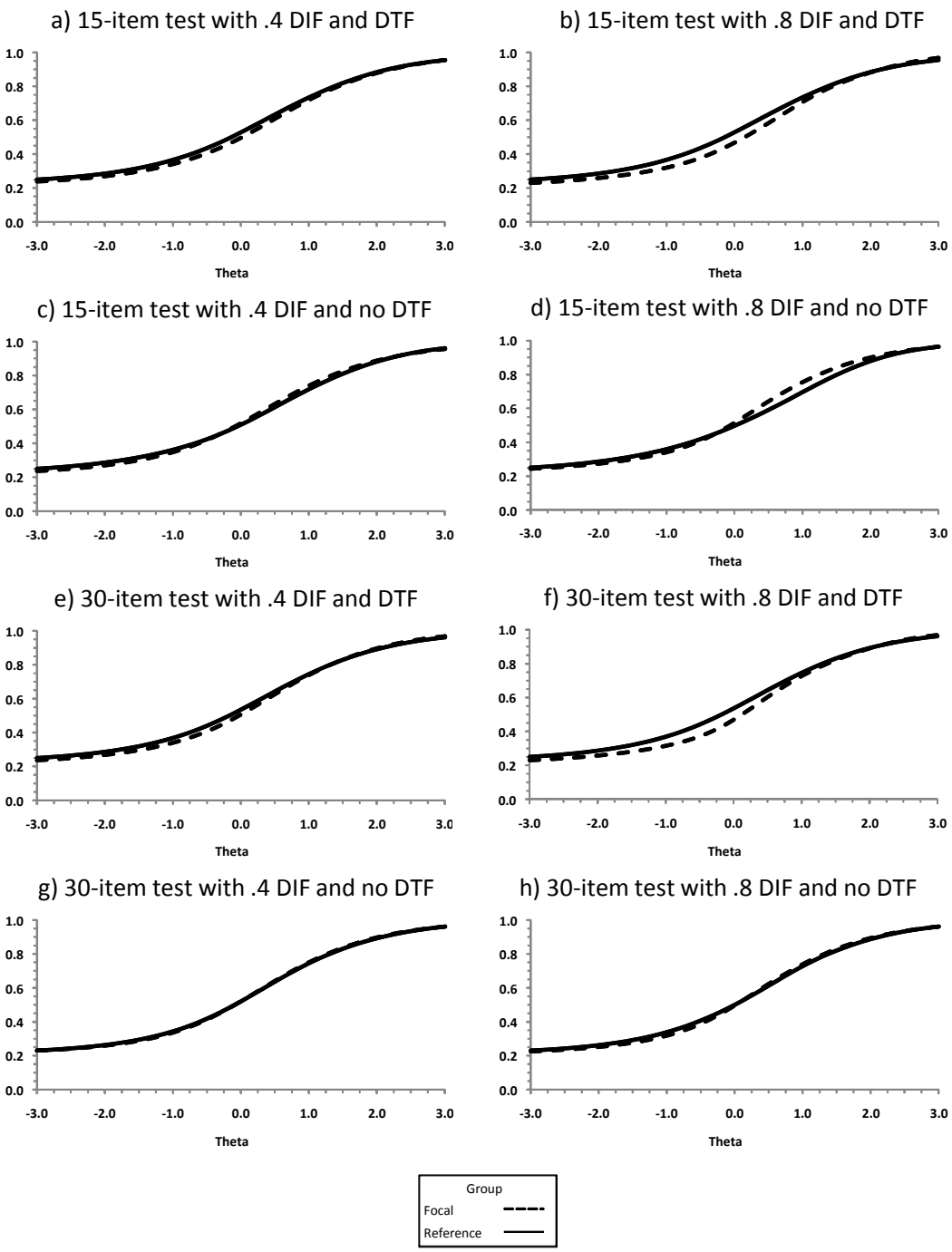

Figure 9. Estimated reference and focal group test characteristic curves (TCCs) by test length, DIF and DTF magnitude conditions. (*Note*: DIF = differential item functioning and DTF = differential test functioning.).

Implementation of DIF Detection Methods

IRT-LR analyses were conducted using MULTILOG 7.0 (Thissen, 2003), LOGREG analyses using SPSS 17.0 (SPSS Inc., 2008), and CSIBTEST analyses using the CSIBTEST software (Stout, 1999). Details of how each analysis was conducted are provided below. All significance tests were based on a critical *p*-value of .05 for rejecting the null hypothesis of no DIF.

Table 5 illustrates how classification accuracy was examined for the three DIF detection methods under investigation. Column 2 shows the broad DIF types, or categories, that each procedure is designed to detect. Column 3 shows how DIF prototypes (Figure 4), as well as the specific DIF items simulated in tests examined here (Figures 5 through 8), are subsumed within these broad DIF categories. Finally, Column 4 denotes the key statistical criteria that were used with the respective detection methods to make broad DIF classification decisions when hits occur.

For example, if a hit occurred using the IRT-LR test, the DIF would be labeled as "Nonuniform" (Column 2) if there was a statistically significant difference in the *a*-parameters across the reference and focal groups, but a nonsignificant difference in the *b*-parameters (Column 4). The classification decision would be deemed correct if the type of DIF simulated by shifting focal group parameters matched the prototypical form indicated in the adjacent cell in Column 3 (i.e., "nonuniform" or "functionally uniform"). Otherwise, the outcome would be viewed as a Type III error. The same process would be used to examine classification accuracy with the CSIBTEST and LOGREG methods, except that crossing point location and the significance of group and interaction terms, respectively, would be used in lieu of item parameter comparisons. Additional examples

of accurate classifications and Type III errors for each procedure are presented in the following sections.

Table 5.

Classification Criteria for Items Flagged as Exhibiting DIF by each DIF Detection Procedure

Procedure	Broad DIF Type	DIF Prototype Falling within Broad DIF Type	Key Criteria for Classifying Flagged Item
CSIBTEST	Unidirectional	Uniform, functionally uniform, and unidirectional mixed	Presence of extreme crossing point
	Crossing	Nonuniform and crossing mixed	Presence of non-extreme crossing point
IRT-LR	Uniform	Uniform	Unequal *b* -parameter only
	Nonuniform	Nonuniform and functionally uniform	Unequal *a* -parameter only
	Mixed	Crossing mixed and unidirectional mixed	Unequal *a* - and *b* -parameters
LOGREG	Uniform	Uniform	Only group term is statistically significant
	Nonuniform	Nonuniform and functionally uniform	Only trait estimate by group interaction term is statistically significant
	Mixed	Crossing mixed and unidirectional mixed	Both group term and trait estimate by group interaction term are statistically significant

Note . CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. *a* = item discrimination. *b* = item difficulty.

In addition, item-level identification results were presented in the form of a "confusion matrix." A confusion matrix concisely shows the concordance between the type of DIF identified by a procedure and the broad DIF type that was simulated. Thus, for all items flagged as DIF (false positives as well as hits), the known and observed DIF types could be readily compared to determine whether each DIF detection method was prone to a particular classification error (e.g., systematic misidentification of uniform DIF as nonuniform or mixed).

IRT Likelihood Ratio Test. The first step of the constant anchor IRT-LR test (Lopez Rivas et al., 2009; Stark et al., 2006) was to construct a baseline model that allowed all items parameters to vary across reference and focal groups, with the exception of an unbiased anchor set. In this case, the anchor set included items 1 through 5 for the 15-item test and, in addition, items 16 through 20 for the 30-item test. Next, compact comparison models, one for each item under investigation, were formed by simultaneously constraining a studied item's parameters (i.e., requiring them to be equal) across the reference and focal groups. For the all-other IRT-LR implementation, a baseline model was formed by constraining all items across groups, and comparison models were formed by freeing parameters for just one item at a time.

For both implementations, the change in goodness of fit between a baseline and a comparison model was compared to a chi-square having degrees of freedom equal to the difference in model parameters. Thus, when testing for DIF on *a*- and *b*-parameters simultaneously, the observed chi-square was compared to a critical chi-square having 2 df, and if the observed chi-square exceeded the critical chi-square, then the studied item was flagged as DIF. Note that impact is addressed in MULTILOG by fixing the latent trait distribution for the reference group to standard normal (i.e., mean = 0 and standard deviation = 1) and allowing the focal group mean and standard deviation to vary (Thissen, 2003).

As an example, if investigating DIF with a 15-item test, a constant baseline model would be formed by allowing parameters for all items to vary across the reference and focal groups, except for an anchor subset – say, items 1 through 5. DIF analyses would then be conducted for items 6 through 15. To test for DIF on Item 6, the parameters for

Item 6 would be constrained in addition to the anchor group and the change in the goodness of fit for this model relative to the baseline model would be compared to a critical chi-square having df equal to the difference in the number of parameters estimated. If both the discrimination and difficulty parameters for Item 6 were constrained simultaneously (i.e., an omnibus test for DIF due to difficulty and discrimination), then the critical chi-square would be based on 2 df $\left(\chi^2_{2;.05} = 5.99\right)$. If, alternatively, one wished to test for DIF on just one parameter at a time (e.g., discrimination or difficulty), the critical chi-square would be based on 1 df $\left(\chi^2_{1;.05} = 3.84\right)$. In any case, the process would be repeated for the remaining studied items, 7 through 15.

For the all-other implementation, the process would be similar except for the model construction aspects. Specifically, the baseline model would be formed by constraining all 15 items across reference and focal groups. Then, a comparison model for, say, Item 6 would be formed by freeing just the discrimination and/or difficulty parameters for Item 6 and comparing the change in goodness of fit to a critical chi-square with the appropriate degrees of freedom. If the improvement in fit was statistically significant, then Item 6 would be labeled a DIF item. The process would then be repeated for items 7 through 15.

Because MULTILOG does not explicitly report information about DIF type in connection with a statistically significant omnibus test, additional analyses were needed to examine DIF classification accuracy in a manner consistent with the other methods explored in this investigation. Specifically, each time a known DIF item was flagged by an omnibus IRT-LR test, follow-up 1 df tests were performed to determine whether the apparent cause of DIF accorded with the parameter(s) that were manipulated. For

example, if an item known to have uniform DIF was flagged by an omnibus test, two follow-up 1 df tests on the individual *a*- and *b*-parameters of that item were conducted. If only the follow-up test constraining or freeing the *b*-parameter was significant, then the detection was considered an accurate identification of uniform DIF. However, if both follow-up tests were significant, then the result would be counted as a Type III error, because it was known a priori that uniform DIF was simulated by shifting only the focal group *b*-parameter. Of course, a Type III error would also result if uniform DIF were simulated and only the follow-up test on the *a*-parameter was significant.

Logistic Regression. Logistic regression DIF analysis can be conducted using SPSS or other popular statistical packages. The process begins by obtaining a trait estimate for each examinee. This can be done by computing a number correct score based on all items or by computing a series of number correct scores that exclude the item under study in each step of the DIF analysis. (There is still debate as to whether the studied item should be included or excluded in the total score computation. However, I chose to exclude the studied item here for consistency with the all-other IRT-LR implementation and with the way CSIBTEST creates matching subtests under the "automatic" DIF option described in the next section). Conversely, a subset of items could be used as the trait estimate, namely, items 1 to 5 in the 15-item test and items 1 through 5 as well as items 15 through 20 in the 30-item test, as was done in the constant anchor conditions.

Next, for each suspect item, a logistic regression analysis is conducted in which the dependent variable is the studied item response and the predictor variables are examinee trait estimate, group membership, and a trait by group interaction. Trait

estimate can be entered as a predictor in Block 1, and the group membership and interaction terms entered in Block 2. If the goodness of fit improves significantly when the group membership and interaction terms are added, as indicated by the change in log likelihood values based on a 2 df chi-square test, then the studied item is flagged as DIF.

In this simulation, power and Type I error rates were computed in the usual manner and, when an item known to exhibit DIF was correctly flagged by the LOGREG procedure, the statistical significance of the regression weights for the group and interaction terms were examined to determine which type of DIF was identified. (The statistical significance of the individual parameters was tested using a Wald chi-square test for which the beta estimate is divided by its squared standard error and the resulting value compared to a critical chi-square with 1 df). Specifically, a significant group term suggests uniform DIF, whereas a significant interaction term suggests nonuniform DIF. If both terms are significant, then mixed DIF is implied. For example, if an item was known to exhibit uniform DIF, then it would be correctly classified as uniform if only the group term was significant. Alternatively, if both terms were significant or if only the interaction term was significant, then the classification would be considered a Type III error.

Crossing Simultaneous Item Bias Test. Stout's (1999b) software was used to conduct CSIBTEST analyses using the "automatic single-item" option, which tests for DIF one item at a time, in succession. This is what has been referred to previously as the all-other anchor approach, because the respective matching subtests are composed of all items except the one under study. To examine the efficacy of a constant-anchor approach with CSIBTEST, a predefined set of DIF-free items was provided for the matching

subtest – items 1 through 5 for the 15-item test, and items 1 through 5 plus 16 through 20 for the 30-item test. Note that CSIBTEST requires a user to provide an estimate of the proportion of correct responding due to guessing; this was set at 0.2 in this simulation, as per the recommendation in the user manual.

Unlike the IRT-LR and LOGREG methods, when CSIBTEST detects DIF, the result is labeled as either "*Crossing*" or "*Unidirectional*" in the program output, depending on the estimated location of the crossing point (the trait level at which the reference and focal group IRFs cross). In this study, a Type III error was recorded if the label appearing in the program output (Table 3, Column 2) was discordant with the type of DIF simulated (Table 3, Column 4). In other words, if the label was *Unidirectional* and the DIF type simulated was nonuniform or crossing mixed DIF, then a Type III error would be recorded; *Crossing* on the other hand, would be considered a correct classification.

Because CSIBTEST cannot identify items as exhibiting both crossing and unidirectional DIF (these definitions are mutually exclusive), the accurate classification of mixed DIF items was based on whether the presence of a crossing point was correctly indicated. That is, CSIBTEST had to identify crossing mixed DIF as *Crossing* and unidirectional mixed DIF as *Unidirectional*.

Analyses of Monte Carlo Results

In each condition of the simulation, 100 reference and 100 focal group data sets were generated using a different seed on each replication. Power, Type I, and Type III error rates were computed and compared at the condition level for each procedure.

Furthermore, the power and Type III error rates associated with the identification of different DIF types were assessed by procedure.

Power and Type III error were calculated based the detection of items known to exhibit DIF: items 11 to 15 and 26 to 30. Type I error rates were calculated based on the detection of items known to *not* exhibit DIF: items 6 to 10 and 21 to 25; the items that constituted the DIF-free subset in the constant anchor implementations were not included in the Type I error calculation to ensure the comparability of the all-other and constant implementation results. (Under the constant implementation, the anchor items were DIF-free and were therefore not investigated for DIF in the actual analysis of the test items.) Hypotheses regarding the effects of the experimental manipulations upon the power, Type I, and Type III error rates of the three DIF detection procedures are presented below.

Hypotheses

Power and Type I Error. Numerous studies (e.g., Narayanan & Swaminathan, 1996) have shown that greater magnitudes of DIF are easier to detect; however, in this study, results were expected to vary according to how a procedure was implemented. Specifically, it was anticipated that the constant anchor implementation of all three procedures would demonstrate lower Type I error rates and higher power than the all-other implementation when DTF was present and as DIF magnitude increased. This is because using a constant subset of DIF-free anchor items precludes the possibility of contamination in the IRT-LR baseline model and in the test scores used as trait estimates in the CSIBTEST and LOGREG analyses. In contrast, the all-other anchor implementation uses all items but the one under study in the anchor, including any DIF

items present in the test. This introduces contamination into the baseline model/matching subtest, which should adversely affect power and Type I; an effect that should be magnified as the amount of DIF increases. Consequently, it was anticipated that the effects of contamination would be most apparent in the large DIF with DTF conditions. Additionally, it was expected that all of the study procedures would exhibit greater power in the longer test conditions of this study. This stems from the fact that longer tests were expected to provide more accurate and reliable estimates of trait scores for matching focal and reference group examinees.

Hypothesis 1: Higher Type I error rates will be observed in the all-other implementation conditions than in the constant for all procedures when DTF is present and as DIF increases.

Hypothesis 2: Higher power will be observed in the constant anchor implementation conditions than in the all-other for all procedures when DTF is present and as DIF increases.

Hypothesis 3: Higher power will be observed in the longer test conditions for all procedures.

There is a large literature examining the effects of sample size on DIF detection. And, research has consistently shown that power to detect DIF with IRT methods is higher with large samples (e.g., Rogers & Swaminathan, 1993) due to better item

parameter estimation, or in the case of nonparametric methods, perhaps due to smoother, more representative (i.e., diverse or complete) distributions of test scores used for trait matching. Higher power was therefore expected here for the IRT-LR, CSIBTEST, and LOGREG methods in the large sample conditions than in the small sample conditions.

Hypothesis 4: Higher power will be observed in the larger sample conditions for all procedures.

Few studies have examined all of the DIF types included in this investigation. However, some expectations can be formed in relation to the power of these procedures to detect uniform and nonuniform DIF items. It was expected that the power of CSIBTEST to detect nonuniform DIF would be greater than that of LOGREG (Finch & French, 2007). In addition, studies have suggested that LOGREG's power to detect uniform DIF exceeds its ability to detect nonuniform DIF (e.g., Swaminathan & Rogers, 1990). In regard to DIF type, it was expected that functionally uniform DIF would be the most difficult to detect for all three procedures. This is because the effects of DIF become manifest only at high trait levels (above +2.0), as shown in panel (c) of Figures 5 and 6.

Hypothesis 5: Higher power to detect nonuniform DIF will be observed for CSIBEST than for LOGREG.

Hypothesis 6: Higher power to detect uniform DIF than nonuniform DIF will be observed for LOGREG.

Hypothesis 7: Lower power to detect functionally uniform DIF than the other DIF prototypes will be observed for all procedures.

Type III Error. In relation to the effects of the study manipulations upon Type III error rates, there has been little to no research on which to base hypotheses. However, because Type III error is dependent on power (i.e., a misclassification can only occur after an accurate DIF detection), it is likely that the effects of the manipulated variables upon Type III error will mirror those for power. That is, factors which affect a procedure's power - such as magnitude of DIF, sample size, contamination in the anchor set, and test length - ought to also affect its Type III error rate. This assumption extends to DIF type; namely, as expected for power, it was anticipated that all three procedures would be ineffectual for classifying functionally uniform DIF.

Hypothesis 8: Lower Type III error rates will be observed in the larger DIF magnitude conditions for all procedures.

Hypothesis 9: Lower Type III error rates will be observed in the larger sample size conditions for all procedures.

Hypothesis 10: Lower Type III error rates will be observed in the constant anchor implementation conditions than in the all-other for all procedures when DTF is present and as DIF increases.

Hypothesis 11: Lower Type III error rates will be observed in the longer test length conditions for all procedures.

Hypothesis 12: Higher Type III error rates for functionally uniform DIF detection than the other DIF prototypes will be observed all procedures.

CHAPTER 6

Results

In keeping with the guidelines suggested by Harwell, Stone, Hsu, and Kirisci (1996) for Monte Carlo investigations, results were analyzed using analysis of variance (ANOVA) to facilitate interpretation. Separate analyses were conducted for the different study criteria: Type I error, overall power, power to detect the five DIF prototypes, and Type III error. As previously stated, Type I error was defined as the proportion of times across the 100 replications a non-DIF item was incorrectly identified as a DIF item, power as the proportion of times an item known to exhibit DIF was correctly detected, and Type III error as the proportion of times the type of DIF attributed to a correctly detected item was different from the type of DIF that was simulated (see Table 5 for each procedure's classification criteria). Overall power and Type III error rates were computed using all of the items known to exhibit DIF. Power to detect the DIF prototypes shown in Figure 4 was computed using the item(s) known to exhibit that given DIF type: Items 11 and 26 for uniform, 12 and 27 for nonuniform, 13 and 28 for functionally uniform, 14 and 29 for crossing mixed, and 15 and 30 for unidirectional mixed.

To compare the study procedures' performance (CSIBTEST, IRT-LR, and LOGREG), a one-way ANOVA was conducted for each of the study criteria. Additionally, a one-way ANOVA was conducted for each procedure comparing their ability to detect the different types of DIF. Hypotheses pertaining to the comparative

efficacy of the procedures or differences in their ability to detect the DIF prototypes were checked by reviewing the statistical significance findings of the respective analyses and via post hoc comparisons.

To determine the effects of the study manipulations, separate analyses were conducted for each procedure. A full factorial ANOVA model was conducted including implementation, DIF magnitude, DTF, impact, sample size, and test length; however, the analysis could not be run as the error degrees of freedom were zero. Instead a model for main effects, 2-, 3-, and 4-way interactions was run (error df = 15). Due to the large number of significance test that were performed, a Bonferroni correction was used to control the family-wise error rate (resulting significance level was .00089); additionally, the eta squared value for each term was reported to illustrate its effect upon performance. To ensure significant main effects were not attributable to an interaction, mean cell differences were examined. When a significant interaction was ordinal and did not reverse the trend observed for the corresponding main effects, the main effect was reported; otherwise, the interaction was reported. Hypotheses pertaining to the study manipulations were investigated by reviewing the statistical significance of their respective main effect or interaction terms.

Note that, when interpreting observed power rates, Type I error rate must be considered because an inflated value indicates that the procedure detected DIF regardless of its presence, which renders its associated power spurious. Also, note that the varying degrees of freedom across the comparisons are attributable to whether the analysis was conducted at the condition level (comparisons of overall power and Type I error rates) or at the item level (comparisons of power and Type III error rates for the specific DIF

prototypes) and the criteria being analyzed (for power results, the No DIF conditions were omitted because there were no DIF items to detect).

Type I Error Rates

It was found that the average Type I error rate of the procedures significantly differed, $F(2,357) = 92.32$, $p < .05$. Post hoc tests showed that IRT-LR demonstrated significantly higher error rates than any of the other procedures (.39). In turn, it was found that CSIBTEST (.13) produced a significantly higher error rate than LOGREG (.05). The effects of the manipulations upon Type I error varied by procedure; Table 6 presents the ANOVA results by procedure and Table 7 the mean error rates for each procedure by manipulation.

A review of Table 6 shows that a number of factors produced a significant effect. Consistent with their recommended implementation in the literature, a lower error rate was attained by CSIBTEST and LOGREG in the all-other conditions and IRT-LR in the constant. Furthermore, a number of implementation-related interactions were observed. Specially, a significant implementation x test length interaction was found, with CSIBTEST and LOGREG showing a marked increase in the 15-item, constant conditions and IRT-LR an increase in the 15-item, all-other conditions.

It was also found that CSIBTEST was adversely affected by the presence of contamination within the anchor subtest. That is, a significant interaction was found for implementation x DIF magnitude that revealed greater error rates in the all-other conditions as DIF increased whereas the constant conditions demonstrated a similar, albeit inflated, error rate across DIF levels. Relatedly, it was hypothesized that higher Type I errors would be observed for the all-other implementations of the three methods in

the 0.8 DIF conditions when DTF was present than when DTF was absent (H1). This was investigated by reviewing the significance of the implementation x DIF x DTF interaction for each procedure and their corresponding cell means. This hypothesis was only partially supported. It was found that this factor was significant only for CSIBTEST and that, although the highest Type I error rate for the all-other approach occurred in the 0.8 DIF magnitude conditions when DTF was present, the error rates for the constant approach were greater across all levels of DIF and DTF.

Additionally, it was observed that LOGREG was the only procedure to show an effect due to the presence of impact. A significant implementation x impact x sample size interaction was found in which the presence of impact substantially increased the error rate of the constant conditions and to a much lesser extent the all-other. This deleterious effect was magnified as sample size increased.

In summary, IRT-LR clearly performed worse than either nonparametric method, with inflated error rates observed in every condition. Additionally, it was found in the all-other conditions that increased DIF led to increased error rates - an effect caused by the greater contamination within the baseline model. Performance improved somewhat in the constant anchor conditions, with the highest observed error occurring in the large-sample and long-test conditions, suggesting that *p*-values smaller than .05 and effect size measures should be strongly considered in conjunction with IRT-LR in practice. In contrast, the LOGREG method provided excellent results with the all-other implementation in almost every condition. However, Type I error rates increased concomitantly with sample size for the constant anchor implementation in conditions involving impact. CSIBTEST provided good Type I error control in most of the all-other

conditions but error rates worsened when DTF was present and as DIF increased. This suggests that CSIBTEST, like IRT-LR, is susceptible to contamination within the anchor set; yet, even with contamination, the all-other CSIBTEST conditions demonstrated lower error rates than the constant. Also, error rates for both nonparametric procedures increased substantially when short tests were examined using the constant implementation, this indicates that a longer matching subtest is needed for that approach to be effective with these procedures.

Table 6.

ANOVA Results for Type I Error Rate by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	161.52*	.05	15.73	.00	5.78	.00	I * N * DIF	2	18.34*	.01	3.38	.00	0.32	.00
DTF	1	21.76*	.01	5.10	.00	14.14	.01	I * N * DTF	2	19.60*	.01	0.07	.00	1.26	.00
Impact (M)	1	3.19	.00	2.45	.00	239.80*	.13	I * N * M	2	1.45	.00	0.15	.00	103.62*	.11
Implementation (I)	1	119.04*	.04	372.64*	.09	20.78*	.01	L * DIF * DTF	1	0.01	.00	22.45*	.01	0.04	.00
Length (L)	1	279.70*	.09	326.27*	.08	19.82*	.01	L * M * DIF	1	0.01	.00	17.06*	.00	0.05	.00
Sample size (N)	2	26.16*	.02	58.73*	.03	120.84*	.13	L * M * DTF	1	1.91	.00	11.06	.00	0.00	.00
DIF * DTF	1	5.54	.00	5.25	.00	3.35	.00	L * N * DIF	2	4.77	.00	2.31	.00	0.54	.00
I * DIF	1	118.76*	.04	19.23	.00	2.94	.00	L * N * DTF	2	0.05	.00	1.56	.00	0.62	.00
I * DTF	1	80.75*	.03	5.83	.00	8.11	.00	L * N * M	2	0.61	.00	0.30	.00	17.87*	.02
I * L	1	1812.91*	.57	3074.86*	.72	40.65*	.02	M * DIF * DTF	1	2.42	.00	2.02	.00	2.44	.00
I * M	1	40.46*	.01	0.03	.00	276.06*	.15	N * DIF * DTF	2	12.85*	.01	0.56	.00	1.14	.00
I * N	2	17.08*	.01	22.22*	.01	58.11*	.06	N * M * DIF	2	3.89	.00	1.74	.00	0.19	.00
L * DIF	1	2.88	.00	12.06	.00	1.64	.00	N * M * DTF	2	2.71	.00	0.82	.00	6.43	.01
L * DTF	1	0.86	.00	11.28	.00	6.23	.00	I * L * DIF * DTF	1	0.16	.00	18.51*	.00	1.61	.00
L * M	1	37.02*	.01	0.32	.00	46.12*	.02	I * L * M * DIF	1	1.18	.00	15.79	.00	0.27	.00
L * N	2	1.50	.00	3.11	.00	18.10*	.02	I * L * M * DTF	1	1.58	.00	10.68	.00	3.66	.00
M * DIF	1	0.57	.00	19.44*	.00	0.89	.00	I * L * N * DIF	2	0.01	.00	0.52	.00	2.31	.00
M * DTF	1	5.95	.00	9.68	.00	18.67*	.01	I * L * N * DTF	2	3.66	.00	2.81	.00	4.22	.00
N * DIF	2	3.81	.00	2.71	.00	2.69	.00	I * L * N * M	2	0.90	.00	0.14	.00	11.34	.01
N * DTF	2	33.95*	.02	1.52	.00	1.00	.00	I * M * DIF * DTF	1	1.80	.00	4.92	.00	0.35	.00
N * M	2	1.24	.00	0.07	.00	161.95*	.17	I * N * DIF * DTF	2	19.41*	.01	0.86	.00	0.58	.00
I * DIF * DTF	1	24.64*	.01	7.79	.00	1.05	.00	I * N * M * DIF	2	0.53	.00	0.59	.00	2.17	.00
I * L * DIF	1	0.21	.00	10.47	.00	8.04	.00	I * N * M * DTF	2	4.90	.00	0.51	.00	1.33	.00
I * L * DTF	1	0.07	.00	22.84*	.01	8.93	.00	L * M * DIF * DTF	1	0.62	.00	2.46	.00	1.00	.00
I * L * M	1	46.21*	.01	0.11	.00	39.45*	.02	L * N * DIF * DTF	2	0.36	.00	1.40	.00	0.22	.00
I * L * N	2	1.28	.00	4.68	.00	19.11*	.02	L * N * M * DIF	2	1.23	.00	0.67	.00	0.23	.00
I * M * DIF	1	6.02	.00	11.89	.00	5.89	.00	L * N * M * DTF	2	1.09	.00	2.18	.00	0.74	.00
I * M * DTF	1	0.55	.00	15.15	.00	0.18	.00	N * M * DIF * DTF	2	1.78	.00	0.95	.00	0.92	.00

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 7.

Type I Error Rate by Study Procedures and Variables

				All-other implementation					Constant implementation				
					.4 DIF		.8 DIF			.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	No DIF	DTF	No DTF	DTF	No DTF	No DIF	DTF	No DTF	DTF	No DTF
CSIBTEST													
	15	250	-.5	.04	.04	.02	.08	.08	.38	.31	.34	.32	.37
	15	250	.0	.06	.02	.01	.04	.06	.26	.29	.30	.31	.29
	15	500	-.5	.03	.04	.03	.12	.08	.35	.34	.34	.35	.40
	15	500	.0	.05	.06	.03	.15	.06	.30	.29	.27	.26	.31
	15	1000	-.5	.04	.06	.05	.17	.06	.39	.38	.34	.31	.34
	15	1000	.0	.05	.08	.02	.25	.07	.29	.30	.31	.31	.28
	30	250	-.5	.01	.02	.01	.07	.07	.04	.03	.05	.04	.08
	30	250	.0	.06	.00	.01	.03	.08	.08	.06	.07	.06	.07
	30	500	-.5	.02	.05	.03	.10	.06	.05	.05	.06	.06	.08
	30	500	.0	.07	.03	.02	.09	.09	.08	.06	.07	.06	.07
	30	1000	-.5	.03	.05	.03	.22	.07	.06	.05	.07	.05	.10
	30	1000	.0	.05	.06	.03	.30	.09	.07	.08	.07	.07	.08
IRT-LR													
	15	250	-.5	.86	.82	.85	.78	.81	.11	.12	.10	.11	.10
	15	250	.0	.87	.64	.86	.81	.82	.11	.11	.09	.09	.08
	15	500	-.5	.97	.92	.97	.88	.90	.14	.11	.14	.10	.10
	15	500	.0	.97	.69	.96	.91	.94	.11	.11	.10	.10	.10
	15	1000	-.5	1.00	.99	1.00	.95	.78	.13	.16	.16	.18	.15
	15	1000	.0	.99	.76	1.00	.97	1.00	.14	.15	.13	.16	.16
	30	250	-.5	.23	.29	.22	.39	.35	.13	.14	.17	.15	.14
	30	250	.0	.21	.25	.22	.35	.34	.14	.12	.13	.12	.14
	30	500	-.5	.25	.36	.32	.42	.44	.16	.16	.17	.16	.20
	30	500	.0	.21	.36	.29	.40	.42	.17	.16	.17	.16	.17
	30	1000	-.5	.31	.46	.46	.46	.49	.21	.26	.25	.22	.24
	30	1000	.0	.28	.45	.43	.46	.48	.26	.23	.20	.22	.22
LOGREG													
	15	250	-.5	.01	.00	.02	.02	.01	.02	.03	.05	.02	.02
	15	250	.0	.00	.01	.00	.00	.01	.00	.00	.00	.01	.00
	15	500	-.5	.03	.02	.04	.01	.01	.12	.12	.14	.12	.19
	15	500	.0	.01	.01	.01	.05	.01	.01	.01	.01	.01	.01
	15	1000	-.5	.11	.02	.11	.01	.06	.37	.38	.36	.41	.41
	15	1000	.0	.01	.02	.02	.08	.02	.02	.02	.01	.02	.01
	30	250	-.5	.01	.00	.01	.01	.01	.01	.01	.01	.00	.02
	30	250	.0	.00	.00	.01	.01	.01	.00	.00	.04	.00	.01
	30	500	-.5	.03	.03	.02	.02	.06	.06	.04	.11	.04	.12
	30	500	.0	.00	.01	.03	.07	.05	.01	.01	.04	.02	.03
	30	1000	-.5	.04	.04	.05	.07	.10	.15	.14	.28	.17	.25
	30	1000	.0	.00	.01	.03	.13	.06	.01	.01	.03	.01	.02

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Overall Power

A significant difference was not found among the procedures for overall power, $F(2,285) = 3.19$, $p > .05$, and likewise for the post hoc tests. Nevertheless, though not statically significant, it was observed that LOGREG (.71) had higher power than CSIBTEST (.64). The average power rate for IRT-LR was .70 but is difficult to interpret due its inflated Type I error rate.

Table 8 presents the ANOVA results by procedure and Table 9 mean power for each procedure by manipulation. Main effects for sample size and DIF magnitude were found, with greater observed power as either increased. For sample size, this finding supports Hypothesis 4, which stated that greater power would be observed as sample size increased. It was also predicted that observed power would increase with test length (H3). To test this hypothesis, the main effect of test length was examined. Despite a significant finding, the observed cell differences demonstrated minimal mean differences. For IRT-LR and LOGREG the observed differences were negligible. For CSIBTEST, greater power was observed in the 15-item conditions but this was due to an inflated Type I error rate.

Many significant interactions were found that included the implementation manipulation. An interaction of implementation x length was found for all procedures; but, for CSIBTEST and IRT-LR, it was attributable to elevated Type I error rates and for LOGREG the cell differences were minor. A significant implementation x DIF interaction was also found across procedures. For IRT-LR, an examination of the cell means revealed that this finding was attributable to the main effect of DIF and the inflated Type I error rates found in the all-other conditions. For CSIBTEST and

LOGREG, it was observed that, although greater power was achieved in the constant conditions when DIF was 0.4 (due to elevated Type I error rates), the all-other conditions exhibited greater power than the constant when DIF was 0.8. Moreover, an effect for implementation x DTF was detected for all procedures. Better power was observed for the constant conditions with maximum DTF; on the contrary, the all-other showed better performance in the minimal DTF conditions.

Relatedly, Hypothesis 2 posited that, due to the greater contamination present in the anchor set as DIF and DTF increased, greater power would be observed for the constant anchor implementation relative to the all-other. To test this, the interaction of implementation x DIF x DTF was investigated; it was found to be significant for the nonparametric procedures, although a review of cell means showed that the hypothesis was only partially supported. For CSIBTEST and LOGREG, it was found that the all-other conditions showed better power in the minimal DTF condition with 0.4 DIF and a minor benefit in 0.8, whereas DTF had a beneficial effect in the constant conditions. However, contrary to expectation, it was found for both that the all-other approach outperformed the constant.

When observed error rates are considered, it appears that LOGREG exhibited the greatest overall power regardless of implementation. CSIBTEST and IRT-LR performed similarly though both demonstrated Type I error rates above .05. Again all procedures performed better with the approach that is suggested in the literature (all-other for the nonparametric procedures and constant for IRT-LR). Also, it is noteworthy that the power of the nonparametric methods was not affected by impact. The presence of DTF had mixed effects. When the all-other approach was used, the minimal DTF conditions

yielded better detection, which is consistent with notions concerning the effects of contamination in the anchor item subset; but the reverse was true in the constant implementation conditions. For the constant conditions, this is attributable to the fact that in order to minimize DTF, both groups were allowed to benefit from the presence of DIF whereas, to maximize DTF, only the reference group benefitted. These results are consistent with past research; for example, Wang and Yeh (2003) found greater power for the constant approach to IRT-LR when DIF favored one group (relative to equivalent conditions in which DIF favored both groups). Regardless, results indicated that despite contamination within the anchor subtest the all-other approaches generally outperformed the constant in terms of power and Type I error.

Table 8.

ANOVA Results for Overall Power by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	2862.86*	.49	3615.46*	.42	4256.47*	.47	I * N * DIF	2	7.55	.00	4.40	.00	7.34	.00
DTF	1	25.75*	.00	69.23*	.01	194.34*	.02	I * N * DTF	2	0.67	.00	2.12	.00	5.76	.00
Impact (M)	1	9.90	.00	9.03	.00	61.03*	.01	I * N * M	2	0.22	.00	2.79	.00	0.38	.00
Implementation (I)	1	39.69*	.01	205.06*	.02	34.31*	.00	L * DIF * DTF	1	0.28	.00	6.68	.00	162.13*	.02
Length (L)	1	156.79*	.03	37.96*	.00	32.60*	.00	L * M * DIF	1	0.21	.00	0.24	.00	8.17	.00
Sample size (N)	2	749.25*	.26	1057.43*	.25	1165.50*	.26	L * M * DTF	1	0.11	.00	0.45	.00	6.95	.00
DIF * DTF	1	11.48	.00	1.65	.00	55.88*	.01	L * N * DIF	2	0.57	.00	3.47	.00	1.74	.00
I * DIF	1	183.87*	.03	1167.31*	.14	202.90*	.02	L * N * DTF	2	0.76	.00	1.11	.00	5.36	.00
I * DTF	1	76.04*	.01	70.37*	.01	406.06*	.04	L * N * M	2	2.12	.00	3.42	.00	0.12	.00
I * L	1	413.55*	.07	562.19*	.07	57.74*	.01	M * DIF * DTF	1	2.09	.00	0.04	.00	69.11*	.01
I * M	1	118.08*	.02	0.42	.00	21.54*	.00	N * DIF * DTF	2	0.86	.00	5.93	.00	5.89	.00
I * N	2	57.72*	.02	17.88*	.00	1.13	.00	N * M * DIF	2	0.01	.00	16.03*	.00	3.02	.00
L * DIF	1	90.11*	.02	53.64*	.01	92.38*	.01	N * M * DTF	2	0.06	.00	0.56	.00	3.17	.00
L * DTF	1	0.66	.00	14.94	.00	39.49*	.00	I * L * DIF * DTF	1	0.34	.00	3.11	.00	85.86*	.01
L * M	1	3.54	.00	2.84	.00	14.84	.00	I * L * M * DIF	1	0.71	.00	2.49	.00	0.05	.00
L * N	2	6.33	.00	24.02*	.01	4.08	.00	I * L * M * DTF	1	3.87	.00	0.89	.00	2.84	.00
M * DIF	1	0.04	.00	1.60	.00	90.18*	.01	I * L * N * DIF	2	2.15	.00	5.05	.00	0.01	.00
M * DTF	1	1.50	.00	0.25	.00	168.25*	.02	I * L * N * DTF	2	0.03	.00	1.50	.00	6.59	.00
N * DIF	2	9.44	.00	63.89*	.01	121.17*	.03	I * L * N * M	2	1.81	.00	1.90	.00	3.57	.00
N * DTF	2	3.89	.00	0.17	.00	3.34	.00	I * M * DIF * DTF	1	7.38	.00	0.34	.00	8.46	.00
N * M	2	2.81	.00	1.07	.00	1.89	.00	I * N * DIF * DTF	2	0.50	.00	1.59	.00	16.28*	.00
I * DIF * DTF	1	25.99*	.00	7.67	.00	20.56*	.00	I * N * M * DIF	2	1.87	.00	2.03	.00	1.72	.00
I * L * DIF	1	19.18*	.00	237.00*	.03	71.59*	.01	I * N * M * DTF	2	0.28	.00	0.60	.00	2.09	.00
I * L * DTF	1	0.31	.00	0.47	.00	76.54*	.01	L * M * DIF * DTF	1	0.26	.00	0.03	.00	3.51	.00
I * L * M	1	13.46	.00	0.71	.00	1.71	.00	L * N * DIF * DTF	2	0.39	.00	1.56	.00	2.88	.00
I * L * N	2	4.94	.00	13.13*	.00	3.53	.00	L * N * M * DIF	2	0.54	.00	1.14	.00	1.26	.00
I * M * DIF	1	27.36*	.00	2.65	.00	5.75	.00	L * N * M * DTF	2	0.50	.00	0.26	.00	1.31	.00
I * M * DTF	1	7.64	.00	0.80	.00	17.62	.00	N * M * DIF * DTF	2	1.29	.00	0.86	.00	1.86	.00

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 9.

Overall Power by Study Variables

				All-other implementation				Constant implementation			
				.4 DIF		.8 DIF		.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	DTF	No DTF	DTF	No DTF	DTF	No DTF	DTF	No DTF
CSIBTEST											
	15	250	-.5	.19	.31	.54	.64	.60	.61	.76	.71
	15	250	.0	.30	.45	.69	.67	.51	.48	.68	.70
	15	500	-.5	.33	.47	.77	.85	.74	.71	.83	.80
	15	500	.0	.45	.57	.86	.83	.57	.62	.77	.77
	15	1000	-.5	.53	.65	.89	.96	.80	.71	.85	.83
	15	1000	.0	.67	.73	.97	.94	.67	.67	.81	.84
	30	250	-.5	.18	.31	.63	.74	.27	.24	.56	.54
	30	250	.0	.26	.45	.74	.75	.28	.27	.58	.58
	30	500	-.5	.38	.50	.84	.92	.43	.40	.68	.69
	30	500	.0	.46	.56	.94	.92	.42	.39	.69	.68
	30	1000	-.5	.59	.68	.99	.99	.56	.54	.80	.76
	30	1000	.0	.68	.78	1.00	.98	.52	.38	.79	.79
IRT-LR											
	15	250	-.5	.54	.65	.67	.68	.32	.26	.71	.64
	15	250	.0	.56	.68	.76	.68	.26	.20	.79	.73
	15	500	-.5	.76	.83	.82	.75	.45	.40	.88	.83
	15	500	.0	.82	.86	.84	.83	.44	.39	.92	.88
	15	1000	-.5	.94	.94	.91	.95	.58	.50	.99	.94
	15	1000	.0	.97	.92	.89	.87	.63	.57	.98	.94
	30	250	-.5	.50	.47	.72	.71	.42	.34	.84	.73
	30	250	.0	.44	.45	.74	.72	.37	.29	.86	.75
	30	500	-.5	.58	.56	.78	.79	.57	.43	.95	.86
	30	500	.0	.57	.57	.80	.78	.60	.46	.96	.88
	30	1000	-.5	.66	.64	.82	.81	.72	.64	1.00	.94
	30	1000	.0	.70	.66	.82	.80	.78	.68	1.00	.96
LOGREG											
	15	250	-.5	.41	.28	.79	.80	.57	.35	.88	.72
	15	250	.0	.24	.38	.78	.87	.29	.34	.74	.75
	15	500	-.5	.56	.53	.94	.97	.83	.50	1.00	.89
	15	500	.0	.44	.47	.98	.98	.44	.44	.96	.95
	15	1000	-.5	.80	.74	1.00	.99	.99	.59	1.00	.99
	15	1000	.0	.59	.74	1.00	1.00	.65	.67	1.00	1.00
	30	250	-.5	.35	.35	.82	.79	.51	.29	.89	.31
	30	250	.0	.24	.38	.80	.89	.35	.38	.88	.36
	30	500	-.5	.52	.56	.96	.95	.71	.52	.99	.53
	30	500	.0	.43	.58	.96	.98	.47	.54	.98	.54
	30	1000	-.5	.73	.75	1.00	1.00	.93	.76	1.00	.74
	30	1000	.0	.70	.80	1.00	1.00	.73	.75	1.00	.74

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Power to Detect Uniform DIF

It was found that the power of the procedures to detect uniform DIF significantly differed, $F(2,285) = 7.37$, $p < .05$. Specifically, post hoc analysis revealed that LOGREG (.87) and CSIBTEST (.87) performed the same. In relation to IRT-LR (.77), both were found to be significantly more effective.

Table 10 presents the ANOVA results by procedure and Table 11 mean power for each procedure by manipulation. Consistent with overall power, main effects for sample size and DIF magnitude were found although the effects were weaker for LOGREG than CSIBTEST and IRT-LR. A main effect for DTF was found for IRT-LR and LOGREG in which the maximum DTF conditions showed better power. Relatedly, a significant implementation x DTF interaction was found for all procedures, the effect of which differed by procedure. For CSIBTEST, the all-other conditions benefitted from minimal DTF whereas the constant benefitted from maximum DTF. IRT-LR and LOGREG demonstrated little difference in the all-other conditions but the constant showed an increase in power when DTF was present.

For length, a significant main effect was found for LOGREG and IRT-LR but the associated cell differences were minor. However, a significant implementation x length interaction was observed for all procedures. For CSIBTEST and IRT-LR it was caused by the inflated Type I error rates observed in the 15-item conditions, and for LOGREG the observed cell differences were trivial.

LOGREG again showed a significant effect related to impact. Specifically, an effect for impact x DIF x DTF was found in which power to detect uniform DIF dropped when either impact or DTF occurred (but not when they co-occurred). Further

examination revealed a relation to implementation. That is, in the minimal DTF conditions, the constant approach showed lower power when impact was present. For the all-other implementation, this was the case in the 0.4 DIF conditions.

The nonparametric procedures exhibited moderate to high power for the detection of uniform DIF. In regard to implementation, the all-other approach performed better than the constant for LOGREG, though it should be noted that it was the only procedure whose effectiveness was affected by impact. Both implementations of CSIBTEST performed comparably; however, results for the constant conditions were less consistent and had greater associated Type I error rates. For IRT-LR, despite its high error rates, power for both implementations was lower than what was found for the other procedures. As was seen for overall power, the effect of DTF varied, with it producing a negative effect upon power in the all-other conditions but a positive one in the constant conditions. As mentioned previously, this is attributable to the change in the direction of DIF that occurred in the minimal DTF conditions.

Table 10.

ANOVA Results for Power to Detect Uniform DIF by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	702.71*	.32	996.75*	.35	53.95*	.09	I * N * DIF	2	3.81	.00	3.55	.00	4.49	.02
DTF	1	12.96	.01	250.35*	.09	50.56*	.09	I * N * DTF	2	13.62*	.01	15.58*	.01	2.29	.01
Impact (M)	1	28.17*	.01	6.28	.00	38.47*	.06	I * N * M	2	10.32	.01	1.54	.00	0.69	.00
Implementation (I)	1	2.32	.00	0.47	.00	17.56	.03	L * DIF * DTF	1	0.11	.00	10.52	.00	10.01	.02
Length (L)	1	18.10	.01	47.66*	.02	0.03	.00	L * M * DIF	1	0.80	.00	0.54	.00	9.71	.02
Sample size (N)	2	296.89*	.27	326.04*	.23	30.30*	.10	L * M * DTF	1	0.53	.00	4.78	.00	0.83	.00
DIF * DTF	1	15.09	.01	0.15	.00	7.16	.01	L * N * DIF	2	5.14	.00	5.34	.00	0.37	.00
I * DIF	1	0.25	.00	6.78	.00	0.06	.00	L * N * DTF	2	1.94	.00	4.14	.00	2.22	.01
I * DTF	1	137.14*	.06	144.34*	.05	41.18*	.07	L * N * M	2	3.07	.00	2.57	.00	0.45	.00
I * L	1	35.55*	.02	87.17*	.03	1.95	.00	M * DIF * DTF	1	0.86	.00	0.74	.00	25.30*	.04
I * M	1	53.70*	.02	1.96	.00	15.11	.03	N * DIF * DTF	2	7.79	.01	6.03	.00	0.13	.00
I * N	2	3.20	.00	10.89	.01	0.33	.00	N * M * DIF	2	2.26	.00	7.69	.01	0.06	.00
L * DIF	1	26.09*	.01	34.48*	.01	14.51	.02	N * M * DTF	2	0.61	.00	0.48	.00	7.01	.02
L * DTF	1	0.17	.00	17.04	.01	0.64	.00	I * L * DIF * DTF	1	0.31	.00	22.95*	.01	5.36	.01
L * M	1	0.53	.00	0.70	.00	0.05	.00	I * L * M * DIF	1	0.15	.00	0.22	.00	2.38	.00
L * N	2	3.05	.00	19.38*	.01	0.75	.00	I * L * M * DTF	1	2.87	.00	6.28	.00	1.15	.00
M * DIF	1	8.84	.00	0.70	.00	8.08	.01	I * L * N * DIF	2	2.09	.00	2.63	.00	0.01	.00
M * DTF	1	0.03	.00	1.33	.00	70.72*	.12	I * L * N * DTF	2	0.51	.00	5.63	.00	2.65	.01
N * DIF	2	113.02*	.10	81.26*	.06	13.53*	.05	I * L * N * M	2	0.81	.00	2.02	.00	2.19	.01
N * DTF	2	8.58	.01	1.32	.00	2.92	.01	I * M * DIF * DTF	1	0.49	.00	0.09	.00	0.49	.00
N * M	2	13.36*	.01	3.90	.00	1.25	.00	I * N * DIF * DTF	2	1.00	.00	3.28	.00	0.13	.00
I * DIF * DTF	1	63.27*	.03	46.32*	.02	3.85	.01	I * N * M * DIF	2	3.86	.00	0.73	.00	2.39	.01
I * L * DIF	1	30.32*	.01	50.77*	.02	6.60	.01	I * N * M * DTF	2	0.22	.00	0.89	.00	0.22	.00
I * L * DTF	1	0.11	.00	39.87*	.01	2.79	.00	L * M * DIF * DTF	1	0.11	.00	0.27	.00	14.58	.02
I * L * M	1	0.02	.00	0.51	.00	1.87	.00	L * N * DIF * DTF	2	1.23	.00	6.20	.00	0.17	.00
I * L * N	2	2.06	.00	5.64	.00	2.20	.01	L * N * M * DIF	2	2.84	.00	0.69	.00	0.20	.00
I * M * DIF	1	34.20*	.02	5.68	.00	0.26	.00	L * N * M * DTF	2	0.17	.00	0.39	.00	0.20	.00
I * M * DTF	1	4.74	.00	10.84	.00	16.51	.03	N * M * DIF * DTF	2	1.59	.00	4.68	.00	2.71	.01

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 11.

Power to Detect Uniform DIF by Study Variables

				All-other implementation				Constant implementation			
				.4 DIF		.8 DIF		.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	DTF	No DTF	DTF	No DTF	DTF	No DTF	DTF	No DTF
CSIBTEST											
	15	250	-.5	.33	.58	.82	.92	.82	.67	.96	.86
	15	250	.0	.57	.87	.94	.96	.71	.63	.94	.91
	15	500	-.5	.57	.84	.99	1.00	.96	.84	1.00	.96
	15	500	.0	.79	.95	1.00	1.00	.87	.88	1.00	.96
	15	1000	-.5	.85	.99	1.00	1.00	1.00	.84	1.00	1.00
	15	1000	.0	.97	1.00	1.00	1.00	.96	.96	1.00	.99
	30	250	-.5	.23	.54	.83	.97	.55	.45	.94	.88
	30	250	.0	.50	.94	.96	1.00	.52	.47	.96	.91
	30	500	-.5	.70	.93	1.00	1.00	.78	.71	.99	.98
	30	500	.0	.83	.98	1.00	1.00	.72	.69	1.00	.98
	30	1000	-.5	.92	1.00	1.00	1.00	.98	.87	1.00	1.00
	30	1000	.0	.97	1.00	1.00	1.00	.91	.70	1.00	1.00
IRT-LR											
	15	250	-.5	.23	.46	.73	.73	.52	.17	.94	.72
	15	250	.0	.19	.48	.92	.58	.40	.19	.99	.83
	15	500	-.5	.40	.58	.91	.68	.65	.51	1.00	.95
	15	500	.0	.66	.63	1.00	.78	.72	.51	1.00	.99
	15	1000	-.5	.75	.73	1.00	1.00	.97	.72	1.00	1.00
	15	1000	.0	.92	.70	1.00	.75	.98	.81	1.00	1.00
	30	250	-.5	.63	.61	.95	.95	.76	.31	.90	.58
	30	250	.0	.59	.59	.99	.99	.62	.29	.90	.55
	30	500	-.5	.63	.60	1.00	1.00	.92	.43	.98	.67
	30	500	.0	.71	.67	1.00	1.00	.94	.44	.98	.69
	30	1000	-.5	.78	.74	1.00	1.00	1.00	.76	1.00	.78
	30	1000	.0	.89	.85	1.00	1.00	1.00	.78	1.00	.85
LOGREG											
	15	250	-.5	.71	.05	1.00	1.00	1.00	.02	1.00	.88
	15	250	.0	.44	.99	1.00	1.00	.60	.75	1.00	1.00
	15	500	-.5	.98	.64	1.00	1.00	1.00	.02	1.00	1.00
	15	500	.0	.94	1.00	1.00	1.00	.97	.98	1.00	1.00
	15	1000	-.5	1.00	.99	1.00	1.00	1.00	.09	1.00	1.00
	15	1000	.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	30	250	-.5	.71	.48	1.00	1.00	.99	.03	1.00	.04
	30	250	.0	.58	.80	1.00	1.00	.83	.71	1.00	.76
	30	500	-.5	.98	.93	1.00	1.00	1.00	.38	1.00	.32
	30	500	.0	.90	1.00	1.00	1.00	.98	.95	1.00	.97
	30	1000	-.5	1.00	1.00	1.00	1.00	1.00	.80	1.00	.77
	30	1000	.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Power to Detect Nonuniform DIF

Analysis revealed that the procedures significantly differed in their ability to detect nonuniform DIF, $F(2,285) = 22.39$, $p < .05$. It was found that IRT-LR (.69) performed better than the other procedures, though its elevated Type I error rate should be recalled. Additionally, LOGREG (.51) was found to have better detection than CSIBTEST (.39), which counters the expectation set forth in Hypothesis 5 that CSIBTEST would show greater power to detect nonuniform DIF than LOGREG. Also, it was predicted that LOGREG would exhibit greater power for uniform DIF detection than nonuniform (H6). Support was found, $F(4,475) = 48.06$, $p < .05$, and post hoc tests showed that LOGREG's power to detect uniform DIF (.69) was greater than its power to detect nonuniform (.61).

Table 12 presents the ANOVA results by procedure and Table 13 the mean power for each procedure by manipulation. Again, all procedures benefitted from greater sample size and DIF magnitude. A significant implementation x DIF interaction was found for all procedures in which the all-other conditions of CSIBTEST and LOGREG showed poor power in the 0.4 DIF conditions but good power in the 0.8 conditions. Relative to the all-other, the constant approach produced higher power in the 0.4 DIF conditions and lower in the 0.8. Results for the IRT-LR all-other conditions were due to its inflated Type I error rate.

A comparison with the results for uniform DIF confirms the findings of many previous studies indicating that nonuniform DIF is more difficult to detect. In general, power to detect nonuniform DIF was lower for CSIBTEST and LOGREG, and similar for IRT-LR. Greater power was observed for LOGREG than CSIBTEST; this is surprising

given that it was specifically designed to perform better than its predecessor SIBTEST when IRFs cross. Unlike findings for uniform DIF, neither impact nor DTF had any obvious effects. For the other manipulations, power was substantially higher in the large DIF conditions and similar across implementations when taking Type I errors into consideration. In sum, it appears that when nonuniform DIF is suspected any of the study procedures is a viable option, although the selected method should be implemented with its recommended approach (all-other for nonparametric methods and constant for IRT-LR) for a more controlled Type I error rate. Additionally, it appears that the efficacy of the procedures to detect nonuniform DIF was unaffected by many of the study manipulations.

Table 12.

ANOVA Results for Power to Detect Nonuniform DIF by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	1517.68*	.43	1027.79*	.29	989.89*	.54	I * N * DIF	2	5.60	.00	40.35*	.02	1.84	.00
DTF	1	0.08	.00	7.21	.00	16.12	.01	I * N * DTF	2	1.99	.00	0.82	.00	0.17	.00
Impact (M)	1	10.43	.00	34.86*	.01	20.69*	.01	I * N * M	2	0.54	.00	3.97	.00	2.50	.00
Implementation (I)	1	380.79*	.11	54.76*	.02	22.37*	.01	L * DIF * DTF	1	0.11	.00	0.12	.00	28.07*	.02
Length (L)	1	29.21*	.01	54.76*	.02	0.17	.00	L * M * DIF	1	0.22	.00	0.97	.00	5.34	.00
Sample size (N)	2	227.00*	.13	324.43*	.19	150.19*	.16	L * M * DTF	1	1.54	.00	2.20	.00	0.02	.00
DIF * DTF	1	1.60	.00	0.42	.00	28.55*	.02	L * N * DIF	2	3.59	.00	9.25	.01	4.80	.01
I * DIF	1	547.46*	.15	1020.54*	.29	65.65*	.04	L * N * DTF	2	0.59	.00	0.72	.00	0.42	.00
I * DTF	1	3.77	.00	7.46	.00	19.77*	.01	L * N * M	2	1.81	.00	2.49	.00	0.64	.00
I * L	1	37.58*	.01	22.55*	.01	19.47*	.01	M * DIF * DTF	1	0.40	.00	2.41	.00	0.38	.00
I * M	1	7.24	.00	10.00	.00	55.24*	.03	N * DIF * DTF	2	0.68	.00	2.32	.00	0.61	.00
I * N	2	62.23*	.03	2.07	.00	6.40	.01	N * M * DIF	2	0.79	.00	8.45	.00	3.69	.00
L * DIF	1	177.25*	.05	75.16*	.02	5.24	.00	N * M * DTF	2	0.28	.00	2.50	.00	0.72	.00
L * DTF	1	1.48	.00	0.61	.00	16.85	.01	I * L * DIF * DTF	1	0.34	.00	2.20	.00	25.02*	.01
L * M	1	2.38	.00	3.58	.00	14.52	.01	I * L * M * DIF	1	0.12	.00	2.41	.00	0.01	.00
L * N	2	17.60*	.01	10.26	.01	3.41	.00	I * L * M * DTF	1	1.48	.00	3.25	.00	1.60	.00
M * DIF	1	4.24	.00	1.82	.00	7.46	.00	I * L * N * DIF	2	15.68*	.01	10.49	.01	0.05	.00
M * DTF	1	5.81	.00	3.50	.00	0.74	.00	I * L * N * DTF	2	0.11	.00	1.12	.00	0.86	.00
N * DIF	2	20.34*	.01	8.04	.00	11.94	.01	I * L * N * M	2	0.24	.00	0.25	.00	1.55	.00
N * DTF	2	0.23	.00	0.76	.00	0.58	.00	I * M * DIF * DTF	1	4.43	.00	1.31	.00	1.46	.00
N * M	2	1.35	.00	0.14	.00	5.95	.01	I * N * DIF * DTF	2	0.95	.00	1.26	.00	1.07	.00
I * DIF * DTF	1	0.67	.00	1.76	.00	16.76	.01	I * N * M * DIF	2	5.16	.00	8.42	.00	14.29*	.02
I * L * DIF	1	35.88*	.01	204.21*	.06	8.03	.00	I * N * M * DTF	2	1.47	.00	1.27	.00	0.51	.00
I * L * DTF	1	2.61	.00	1.42	.00	12.19	.01	L * M * DIF * DTF	1	0.00	.00	3.25	.00	0.01	.00
I * L * M	1	6.50	.00	1.36	.00	3.53	.00	L * N * DIF * DTF	2	1.47	.00	1.87	.00	0.93	.00
I * L * N	2	9.46	.01	17.44*	.01	0.52	.00	L * N * M * DIF	2	0.36	.00	0.68	.00	8.43	.01
I * M * DIF	1	1.91	.00	2.41	.00	2.77	.00	L * N * M * DTF	2	1.89	.00	0.14	.00	0.86	.00
I * M * DTF	1	5.05	.00	3.67	.00	0.75	.00	N * M * DIF * DTF	2	2.73	.00	0.16	.00	0.43	.00

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 13.

Power to Detect Nonuniform DIF by Study Variables

				All-other implementation				Constant implementation			
				.4 DIF		.8 DIF		.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	DTF	No DTF	DTF	No DTF	DTF	No DTF	DTF	No DTF
CSIBTEST											
	15	250	-.5	.10	.06	.40	.42	.23	.42	.32	.32
	15	250	.0	.08	.04	.56	.38	.33	.25	.25	.31
	15	500	-.5	.12	.14	.63	.64	.33	.39	.29	.30
	15	500	.0	.15	.18	.77	.67	.18	.35	.25	.26
	15	1000	-.5	.19	.27	.76	.94	.31	.39	.28	.22
	15	1000	.0	.46	.31	.93	.88	.31	.32	.26	.35
	30	250	-.5	.06	.06	.57	.52	.09	.09	.20	.25
	30	250	.0	.07	.07	.69	.61	.09	.13	.32	.32
	30	500	-.5	.15	.17	.86	.92	.14	.14	.44	.43
	30	500	.0	.22	.15	.96	.93	.15	.15	.49	.45
	30	1000	-.5	.35	.38	1.00	1.00	.17	.22	.63	.52
	30	1000	.0	.45	.44	.99	.99	.21	.14	.67	.70
IRT-LR											
	15	250	-.5	.60	.53	.34	.29	.18	.10	.65	.73
	15	250	.0	.74	.62	.47	.34	.16	.15	.89	.89
	15	500	-.5	.77	.88	.62	.52	.25	.24	.94	.94
	15	500	.0	.84	.90	.76	.59	.28	.31	1.00	1.00
	15	1000	-.5	.96	.99	.72	1.00	.33	.30	1.00	1.00
	15	1000	.0	.98	.99	.93	.70	.54	.52	1.00	1.00
	30	250	-.5	.61	.54	.76	.66	.21	.21	.91	.90
	30	250	.0	.57	.57	.75	.65	.22	.21	.97	.99
	30	500	-.5	.66	.58	.87	.82	.32	.34	1.00	1.00
	30	500	.0	.63	.60	.91	.82	.44	.47	1.00	1.00
	30	1000	-.5	.74	.69	.98	.96	.58	.63	1.00	1.00
	30	1000	.0	.78	.61	1.00	.95	.87	.84	1.00	1.00
LOGREG											
	15	250	-.5	.02	.01	.52	.63	.03	.06	.68	.61
	15	250	.0	.01	.02	.85	.68	.00	.01	.18	.28
	15	500	-.5	.08	.08	.87	.99	.49	.43	.98	.97
	15	500	.0	.10	.10	1.00	.96	.04	.03	.81	.78
	15	1000	-.5	.34	.56	1.00	.99	.98	.84	1.00	1.00
	15	1000	.0	.29	.29	1.00	1.00	.05	.03	.98	.99
	30	250	-.5	.05	.04	.88	.71	.02	.04	.75	.03
	30	250	.0	.01	.06	.97	.97	.01	.09	.81	.03
	30	500	-.5	.11	.10	1.00	.98	.21	.29	.98	.32
	30	500	.0	.13	.23	1.00	1.00	.07	.20	.97	.14
	30	1000	-.5	.22	.47	1.00	1.00	.74	.71	1.00	.69
	30	1000	.0	.74	.51	1.00	1.00	.29	.42	1.00	.37

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Power to Detect Unidirectional Mixed DIF

Analysis revealed that the procedures significantly differed in their ability to detect unidirectional mixed DIF, $F(5,285) = 33.57$, $p < .05$. It was found that IRT-LR (.70) demonstrated lower average power than the other procedures. No significant difference was found between LOGREG (.98) and CSIBTEST (.91).

Table 14 presents the ANOVA results by procedure and Table 15 the mean power for each procedure by manipulation. Like previous findings, main effects for magnitude of DIF and sample size were found. A significant implementation x test length interaction was found for CSIBTEST and IRT-LR. For CSIBTEST, it was found that the all-other conditions produced similar power across test lengths; however, in the constant conditions, power declined in the 30-item conditions (this was attributable to the elevated Type I error rate observed in the 15-item, constant conditions). For IRT-LR, a sharp drop in power was seen in the 30-item conditions of the all-other approach relative to the 15-item conditions and to the constant implementation conditions. Significant effects related to impact were observed for LOGREG but these could be attributed to the elevated Type I error rates found when impact was present.

It was again found that CSIBTEST and LOGREG performed comparably to each other but better than IRT-LR. In fact, the nonparametric procedures were very effective for the detection of unidirectional mixed DIF with average detection rates near 1.00. Findings suggest that the all-other implementations of LOGREG and CSIBTEST provide good to excellent power while maintaining average Type I error rates near or below the nominal level of .05. The constant IRT-LR implementation also provided good power, but the results must be interpreted cautiously in light of its elevated error rates. Despite

concerns about contamination of the anchor subtest, the DTF manipulation did not have an effect upon detection for the all-other implementations. Interestingly, all of the procedures demonstrated excellent detection of unidirectional mixed DIF, though the amount of DIF present in these items was equivalent to the amount in the items known to possess uniform and nonuniform DIF (0.4 and 0.8).

Table 14.

ANOVA Results for Power to Detect Unidirectional Mixed DIF by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	1173.71*	.28	446.66*	.03	37.11*	.07	I * N * DIF	2	1.54	.00	15.28*	.00	5.61	.02
DTF	1	69.36*	.02	32.40*	.00	4.34	.01	I * N * DTF	2	11.57	.01	6.40	.00	10.59	.04
Impact (M)	1	17.82	.00	10.56	.00	29.42*	.05	I * N * M	2	21.14*	.01	4.86	.00	0.02	.00
Implementation (I)	1	34.07*	.01	5100.35*	.31	1.79	.00	L * DIF * DTF	1	2.53	.00	27.15*	.00	1.68	.00
Length (L)	1	89.91*	.02	3993.84*	.24	2.35	.00	L * M * DIF	1	0.73	.00	2.34	.00	1.21	.00
Sample size (N)	2	423.81*	.20	247.52*	.03	46.70*	.17	L * M * DTF	1	0.12	.00	0.89	.00	1.58	.00
DIF * DTF	1	46.32*	.01	28.97*	.00	10.64	.02	L * N * DIF	2	13.61*	.01	46.67*	.01	0.26	.00
I * DIF	1	12.79	.00	22.69*	.00	6.33	.01	L * N * DTF	2	0.14	.00	7.42	.00	0.05	.00
I * DTF	1	117.07*	.03	24.70*	.00	11.96	.02	L * N * M	2	5.31	.00	0.10	.00	0.12	.00
I * L	1	218.69*	.05	5041.57*	.31	0.22	.00	M * DIF * DTF	1	5.65	.00	0.65	.00	2.23	.00
I * M	1	124.17*	.03	4.92	.00	0.00	.00	N * DIF * DTF	2	26.78*	.01	5.89	.00	10.45	.04
I * N	2	0.74	.00	6.80	.00	2.03	.01	N * M * DIF	2	7.89	.00	6.97	.00	20.17*	.07
L * DIF	1	75.65*	.02	147.81*	.01	0.13	.00	N * M * DTF	2	2.75	.00	0.14	.00	0.97	.00
L * DTF	1	3.67	.00	29.72*	.00	0.01	.00	I * L * DIF * DTF	1	0.09	.00	17.18	.00	2.00	.00
L * M	1	0.98	.00	0.01	.00	0.10	.00	I * L * M * DIF	1	0.39	.00	0.00	.00	0.01	.00
L * N	2	17.65*	.01	39.75*	.00	2.09	.01	I * L * M * DTF	1	1.85	.00	1.10	.00	0.82	.00
M * DIF	1	5.01	.00	7.48	.00	21.51*	.04	I * L * N * DIF	2	8.19	.00	11.02	.00	1.94	.01
M * DTF	1	9.41	.00	0.01	.00	0.50	.00	I * L * N * DTF	2	3.49	.00	4.78	.00	7.52	.03
N * DIF	2	254.28*	.12	200.61*	.02	34.65*	.13	I * L * N * M	2	7.54	.00	0.66	.00	0.77	.00
N * DTF	2	38.72*	.02	8.32	.00	5.26	.02	I * M * DIF * DTF	1	0.28	.00	0.09	.00	0.97	.00
N * M	2	16.20*	.01	3.73	.00	25.46*	.09	I * N * DIF * DTF	2	6.96	.00	3.82	.00	5.33	.02
I * DIF * DTF	1	98.87*	.02	21.72*	.00	5.20	.01	I * N * M * DIF	2	10.83	.01	3.07	.00	0.44	.00
I * L * DIF	1	138.98*	.03	21.40*	.00	2.72	.00	I * N * M * DTF	2	2.97	.00	0.01	.00	2.64	.01
I * L * DTF	1	2.68	.00	19.23*	.00	6.73	.01	L * M * DIF * DTF	1	0.65	.00	0.32	.00	0.22	.00
I * L * M	1	9.99	.00	1.85	.00	0.82	.00	L * N * DIF * DTF	2	0.38	.00	7.22	.00	1.33	.00
I * L * N	2	27.19*	.01	6.70	.00	0.20	.00	L * N * M * DIF	2	5.90	.00	0.01	.00	0.81	.00
I * M * DIF	1	99.79*	.02	9.06	.00	0.56	.00	L * N * M * DTF	2	0.92	.00	0.01	.00	1.24	.00
I * M * DTF	1	4.22	.00	0.40	.00	3.13	.01	N * M * DIF * DTF	2	2.71	.00	0.38	.00	2.32	.01

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 15.

Power to Detect Unidirectional Mixed DIF by Study Variables

				All-other implementation				Constant implementation			
				.4 DIF		.8 DIF		.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	DTF	No DTF	DTF	No DTF	DTF	No DTF	DTF	No DTF
CSIBTEST											
	15	250	-.5	.38	.75	.88	.98	.92	.89	1.00	.98
	15	250	.0	.64	.95	1.00	1.00	.73	.76	.96	.96
	15	500	-.5	.65	.90	.99	1.00	.98	.98	.99	1.00
	15	500	.0	.88	.98	1.00	1.00	.91	.86	1.00	1.00
	15	1000	-.5	.93	1.00	1.00	1.00	1.00	.98	1.00	1.00
	15	1000	.0	.97	.99	1.00	1.00	.99	.95	1.00	1.00
	30	250	-.5	.45	.72	.99	1.00	.47	.50	.92	.91
	30	250	.0	.71	.97	1.00	1.00	.47	.53	.93	.97
	30	500	-.5	.78	.99	1.00	1.00	.81	.77	.98	1.00
	30	500	.0	.93	.99	1.00	1.00	.78	.71	1.00	1.00
	30	1000	-.5	.97	1.00	1.00	1.00	.94	.92	1.00	1.00
	30	1000	.0	1.00	1.00	1.00	1.00	.92	.75	1.00	1.00
IRT-LR											
	15	250	-.5	.34	.76	.95	.99	.62	.66	.99	.99
	15	250	.0	.32	.78	1.00	1.00	.42	.48	1.00	1.00
	15	500	-.5	.74	.96	1.00	1.00	.88	.88	1.00	1.00
	15	500	.0	.71	.99	1.00	1.00	.81	.82	1.00	1.00
	15	1000	-.5	.98	1.00	1.00	1.00	.99	1.00	1.00	1.00
	15	1000	.0	.95	1.00	1.00	1.00	.99	1.00	1.00	1.00
	30	250	-.5	.06	.08	.07	.07	.79	.79	1.00	1.00
	30	250	.0	.06	.04	.06	.04	.65	.64	1.00	1.00
	30	500	-.5	.09	.11	.07	.14	.98	.94	1.00	1.00
	30	500	.0	.07	.08	.07	.08	.94	.96	1.00	1.00
	30	1000	-.5	.08	.11	.13	.11	1.00	1.00	1.00	1.00
	30	1000	.0	.14	.11	.10	.07	1.00	1.00	1.00	1.00
LOGREG											
	15	250	-.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	15	250	.0	.71	.88	1.00	1.00	.81	.90	1.00	1.00
	15	500	-.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	15	500	.0	.99	1.00	1.00	1.00	1.00	.98	1.00	1.00
	15	1000	-.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	15	1000	.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	30	250	-.5	.81	1.00	1.00	1.00	1.00	1.00	1.00	.98
	30	250	.0	.60	.95	1.00	1.00	.91	.81	1.00	.90
	30	500	-.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	30	500	.0	1.00	.99	1.00	1.00	1.00	1.00	1.00	.98
	30	1000	-.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	30	1000	.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Power to Detect Crossing Mixed DIF

ANOVA revealed that the procedures significantly differed in their ability to detect crossing mixed DIF, $F(2,285) = 11.66$, $p < .05$. For significant differences, it was found that LOGREG (.72) and IRT-LR (.68) did better than CSIBTEST (.51). The difference between LOGREG and IRT-LR was not significant.

Table 16 presents the ANOVA results by procedure and Table 17 the mean power for each procedure by manipulation. All procedures demonstrated a main effect for sample size and DIF magnitude. As was found for nonuniform DIF, a significant implementation x DIF interaction was found for all procedures. Specifically, the all-other conditions of CSIBTEST and LOGREG exhibited lower power in the small DIF conditions but good power in the large, but the constant approach produced higher power than the all-other in the 0.4 conditions due to error and less in the 0.8. Findings for IRT-LR could be attributed to its inflated error rate.

Also, a significant implementation x DTF interaction was found for CSIBTEST and LOGREG that varied. For CSIBTEST, the all-other conditions benefitted from minimal DTF; the constant showed no difference. The LOGREG conditions also demonstrated improved detection in the all-other conditions when DTF was minimal but the constant showed a decline. Additionally, a significant interaction of implementation x test length was observed for CSIBTEST and IRT-LR procedures. The all-other conditions of CSIBTEST were found to benefit from increased test length whereas the constant conditions were affected by the elevated Type I error rate of the 15-item conditions. IRT-LR demonstrated improved power in the 30-item conditions over the 15, a gain that was more marked in the all-other conditions. Lastly, a significant impact main

effect was found for LOGREG and CSIBTEST but this effect for the latter was caused by elevated error rates.

The study procedures demonstrated worse detection of crossing mixed DIF than unidirectional mixed DIF though both exhibited equivalent magnitudes of DIF (0.4 or 0.8) *and* shifts in both the *a*- and *b*-parameters. This indicates that the presence of a crossing point reduces the likelihood that a DIF item will be detected by these procedures, which is consistent with past findings that nonuniform DIF is more difficult to detect than uniform. Again, the nonparametric procedures performed better when an all-other approach was employed and IRT-LR when a constant approach was employed. Considering Type I error rate, findings indicated that the all-other approach of LOGREG was best able to detect items exhibiting crossing mixed DIF**.**

Table 16.

ANOVA Results for Power to Detect Crossing Mixed DIF by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	2541.53*	.37	1309.98*	.27	447.00*	.34	I * N * DIF	2	15.54*	.00	36.72*	.02	1.46	.00
DTF	1	97.83*	.01	0.43	.00	0.00	.00	I * N * DTF	2	0.63	.00	0.05	.00	1.42	.00
Impact (M)	1	42.91*	.01	1.26	.00	151.62*	.11	I * N * M	2	4.36	.00	0.38	.00	3.74	.01
Implementation (I)	1	87.22*	.01	616.34*	.13	2.89	.00	L * DIF * DTF	1	0.23	.00	4.77	.00	9.18	.01
Length (L)	1	645.31*	.09	329.38*	.07	13.61	.01	L * M * DIF	1	2.42	.00	0.00	.00	10.89	.01
Sample size (N)	2	489.52*	.14	252.55*	.11	126.02*	.19	L * M * DTF	1	3.30	.00	4.37	.00	1.58	.00
DIF * DTF	1	5.88	.00	5.29	.00	18.17	.01	L * N * DIF	2	7.16	.00	26.10*	.01	0.17	.00
I * DIF	1	401.16*	.06	1454.24*	.30	33.38*	.03	L * N * DTF	2	0.88	.00	0.07	.00	0.38	.00
I * DTF	1	106.04*	.02	0.60	.00	33.26*	.03	L * N * M	2	3.35	.00	2.16	.00	1.27	.00
I * L	1	1359.92*	.20	64.84*	.01	3.50	.00	M * DIF * DTF	1	20.67*	.00	5.83	.00	4.82	.00
I * M	1	127.36*	.02	2.50	.00	12.29	.01	N * DIF * DTF	2	9.09	.00	0.50	.00	0.62	.00
I * N	2	106.01*	.03	10.45	.00	0.01	.00	N * M * DIF	2	0.66	.00	3.51	.00	4.82	.01
L * DIF	1	0.01	.00	94.39*	.02	2.70	.00	N * M * DTF	2	0.64	.00	0.25	.00	0.59	.00
L * DTF	1	0.23	.00	0.25	.00	3.67	.00	I * L * DIF * DTF	1	4.81	.00	9.51	.00	3.38	.00
L * M	1	20.14*	.00	2.58	.00	4.32	.00	I * L * M * DIF	1	10.30	.00	2.09	.00	2.67	.00
L * N	2	5.14	.00	11.79	.00	3.53	.01	I * L * M * DTF	1	2.07	.00	1.76	.00	0.75	.00
M * DIF	1	1.25	.00	2.58	.00	81.24*	.06	I * L * N * DIF	2	2.07	.00	0.44	.00	0.23	.00
M * DTF	1	0.03	.00	3.04	.00	0.21	.00	I * L * N * DTF	2	0.64	.00	0.67	.00	0.67	.00
N * DIF	2	7.46	.00	6.91	.00	49.37*	.07	I * L * N * M	2	0.19	.00	1.19	.00	0.24	.00
N * DTF	2	1.44	.00	0.40	.00	2.13	.00	I * M * DIF * DTF	1	12.66	.00	1.53	.00	3.59	.00
N * M	2	1.47	.00	0.39	.00	7.10	.01	I * N * DIF * DTF	2	9.42	.00	0.10	.00	7.38	.01
I * DIF * DTF	1	4.31	.00	9.95	.00	1.69	.00	I * N * M * DIF	2	6.49	.00	0.21	.00	1.46	.00
I * L * DIF	1	35.63*	.01	114.68*	.02	6.09	.00	I * N * M * DTF	2	0.25	.00	0.61	.00	0.72	.00
I * L * DTF	1	0.04	.00	1.06	.00	11.03	.01	L * M * DIF * DTF	1	0.47	.00	7.37	.00	2.74	.00
I * L * M	1	45.22*	.01	0.27	.00	1.22	.00	L * N * DIF * DTF	2	2.33	.00	1.08	.00	1.89	.00
I * L * N	2	5.15	.00	8.51	.00	1.23	.00	L * N * M * DIF	2	0.18	.00	0.60	.00	1.50	.00
I * M * DIF	1	1.75	.00	0.00	.00	0.35	.00	L * N * M * DTF	2	0.07	.00	0.58	.00	0.21	.00
I * M * DTF	1	0.39	.00	6.76	.00	1.12	.00	N * M * DIF * DTF	2	0.87	.00	2.09	.00	0.06	.00

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 17.

Power to Detect Crossing Mixed DIF by Study Variables

				All-other implementation				Constant implementation			
				.4 DIF		.8 DIF		.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	DTF	No DTF	DTF	No DTF	DTF	No DTF	DTF	No DTF
CSIBTEST											
	15	250	-.5	.08	.10	.36	.55	.52	.57	.80	.79
	15	250	.0	.06	.20	.47	.58	.37	.40	.65	.62
	15	500	-.5	.15	.25	.66	.87	.75	.75	.97	.93
	15	500	.0	.12	.37	.72	.83	.41	.53	.77	.75
	15	1000	-.5	.35	.51	.87	.98	.86	.75	1.00	.96
	15	1000	.0	.39	.68	.96	.95	.46	.47	.87	.85
	30	250	-.5	.08	.17	.51	.82	.10	.10	.31	.35
	30	250	.0	.02	.17	.61	.73	.12	.11	.27	.30
	30	500	-.5	.20	.31	.81	.96	.18	.14	.30	.41
	30	500	.0	.14	.41	.94	.97	.20	.15	.29	.30
	30	1000	-.5	.45	.65	.99	1.00	.26	.33	.46	.43
	30	1000	.0	.41	.75	1.00	1.00	.20	.13	.38	.33
IRT-LR											
	15	250	-.5	.54	.49	.35	.38	.16	.17	.67	.59
	15	250	.0	.57	.53	.39	.50	.14	.16	.73	.69
	15	500	-.5	.88	.75	.57	.54	.16	.19	.89	.86
	15	500	.0	.91	.79	.42	.76	.13	.17	.97	.99
	15	1000	-.5	.99	.98	.82	.75	.19	.23	1.00	1.00
	15	1000	.0	.99	.92	.54	.90	.31	.23	1.00	1.00
	30	250	-.5	.70	.63	.98	.99	.20	.20	.84	.86
	30	250	.0	.55	.57	1.00	1.00	.16	.16	.89	.87
	30	500	-.5	.91	.91	1.00	1.00	.23	.25	.97	.98
	30	500	.0	.85	.89	1.00	1.00	.31	.25	.99	.99
	30	1000	-.5	.99	1.00	1.00	1.00	.38	.43	1.00	1.00
	30	1000	.0	1.00	.99	1.00	1.00	.44	.45	1.00	1.00
LOGREG											
	15	250	-.5	.34	.34	.95	1.00	.78	.65	1.00	1.00
	15	250	.0	.03	.02	.74	1.00	.02	.04	.95	.80
	15	500	-.5	.59	.89	1.00	1.00	1.00	1.00	1.00	1.00
	15	500	.0	.11	.14	1.00	.99	.11	.17	.99	1.00
	15	1000	-.5	.95	.99	1.00	1.00	1.00	1.00	1.00	1.00
	15	1000	.0	.45	.80	1.00	1.00	.70	.77	1.00	1.00
	30	250	-.5	.12	.24	.72	.95	.47	.38	1.00	.47
	30	250	.0	.02	.06	.86	.95	.01	.15	1.00	.07
	30	500	-.5	.36	.70	.99	1.00	.93	.82	1.00	.88
	30	500	.0	.07	.41	.99	1.00	.24	.34	1.00	.34
	30	1000	-.5	.86	.97	1.00	1.00	1.00	1.00	1.00	.99
	30	1000	.0	.44	.81	1.00	1.00	.86	.74	1.00	.77

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Power to Detect Functionally Uniform DIF

A significant difference was found in the ability of the procedures to detect functionally uniform DIF, $F(2,285) = 9.59$, $p < .05$. Post hoc tests showed that IRT-LR (.67) exhibited better power than both nonparametric procedures. For the remaining conditions, comparable power was observed: CSIBTEST (.52) and IRT-LR (.47).

Hypothesis 7 stated that all three procedures would produce lower power for functionally uniform DIF detection relative to their detection rate for the other DIF prototypes. To test this, ANOVA results were examined. Although significant differences in the detection of the various DIF type were found for CSIBTEST, $F(4,475) = 91.55$, $p < .05$, and LOGREG, support was not found for the hypothesis. Neither of the procedures displayed significantly less power to detect functionally uniform DIF than the other DIF types: CSIBTEST detected nonuniform (.38) at a lower rate than functionally uniform (.52) and LOGREG showed a comparable rate for nonuniform (.51) and functionally uniform (.47). Note that a significant difference was not found for IRT-LR, $F(4,475) = 1.45$, $p > .05$, and that comparable power was observed across all of the DIF prototypes (.67 to .70) except uniform (.77).

Table 18 presents the ANOVA results by procedure and Table 19 the mean power for each procedure by manipulation. Like results for the other DIF prototypes, significant main effects were seen for sample size and DIF magnitude, with cell means showing better detection as either increased. Though many other significant factors were identified, these were attributable to increased Type I error rates or the effects of the sample or DIF magnitude manipulations.

When considered in conjunction with observed Type I error rates, it was found that the procedures when implemented in their recommended fashion demonstrated comparable detection rates for functionally uniform DIF. Of the manipulations, only sample size and DIF magnitude yielded meaningful effects, and both improved detection. Contrary to expectation, functionally uniform DIF was not detected at a significantly lower rate than the other DIF prototypes. Moreover, for CSIBTEST, it was found that it was detected at a better rate than nonuniform DIF.

Table 18.

ANOVA Results for Power to Detect Functionally Uniform DIF by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	1207.24*	.43	843.82*	.18	608.35*	.41	I * N * DIF	2	3.18	.00	28.99*	.01	3.73	.01
DTF	1	0.03	.00	41.70*	.01	51.08*	.03	I * N * DTF	2	0.94	.00	0.99	.00	1.02	.00
Impact (M)	1	17.22	.01	0.04	.00	0.93	.00	I * N * M	2	1.78	.00	0.28	.00	0.76	.00
Implementation (I)	1	106.49*	.04	2324.84*	.48	0.39	.00	L * DIF * DTF	1	0.03	.00	1.87	.00	22.72*	.02
Length (L)	1	139.77*	.05	18.01	.00	7.73	.01	L * M * DIF	1	0.57	.00	0.03	.00	1.94	.00
Sample size (N)	2	433.99*	.31	229.25*	.10	202.57*	.27	L * M * DTF	1	0.09	.00	0.53	.00	0.53	.00
DIF * DTF	1	0.06	.00	2.00	.00	6.37	.00	L * N * DIF	2	2.50	.00	5.56	.00	0.15	.00
I * DIF	1	7.74	.00	84.99*	.02	18.01	.01	L * N * DTF	2	0.28	.00	0.23	.00	0.15	.00
I * DTF	1	15.17	.01	50.69*	.01	69.96*	.05	L * N * M	2	0.52	.00	0.01	.00	0.94	.00
I * L	1	112.62*	.04	404.15*	.08	6.16	.00	M * DIF * DTF	1	2.12	.00	0.00	.00	4.63	.00
I * M	1	57.84*	.02	0.21	.00	0.03	.00	N * DIF * DTF	2	0.26	.00	1.23	.00	2.42	.00
I * N	2	21.15*	.02	94.92*	.04	0.34	.00	N * M * DIF	2	3.04	.00	0.22	.00	2.04	.00
L * DIF	1	28.94*	.01	151.28*	.03	21.61*	.01	N * M * DTF	2	0.14	.00	0.51	.00	2.56	.00
L * DTF	1	0.66	.00	0.09	.00	2.48	.00	I * L * DIF * DTF	1	0.57	.00	2.34	.00	11.42	.01
L * M	1	0.23	.00	0.15	.00	0.22	.00	I * L * M * DIF	1	0.06	.00	0.06	.00	0.27	.00
L * N	2	2.52	.00	10.42	.00	0.14	.00	I * L * M * DTF	1	1.44	.00	0.00	.00	1.15	.00
M * DIF	1	0.00	.00	2.85	.00	7.44	.01	I * L * N * DIF	2	1.29	.00	0.62	.00	0.75	.00
M * DTF	1	0.00	.00	0.02	.00	76.16*	.05	I * L * N * DTF	2	0.16	.00	0.24	.00	0.43	.00
N * DIF	2	24.58*	.02	14.19*	.01	13.28*	.02	I * L * N * M	2	3.19	.00	0.03	.00	0.02	.00
N * DTF	2	0.23	.00	1.40	.00	1.86	.00	I * M * DIF * DTF	1	11.21	.00	0.92	.00	0.07	.00
N * M	2	0.16	.00	0.19	.00	0.09	.00	I * N * DIF * DTF	2	4.71	.00	1.91	.00	4.97	.01
I * DIF * DTF	1	5.63	.00	1.35	.00	12.45	.01	I * N * M * DIF	2	5.73	.00	0.09	.00	0.73	.00
I * L * DIF	1	1.44	.00	55.34*	.01	6.98	.00	I * N * M * DTF	2	0.52	.00	0.96	.00	1.17	.00
I * L * DTF	1	0.14	.00	0.75	.00	5.19	.00	L * M * DIF * DTF	1	0.03	.00	0.15	.00	0.02	.00
I * L * M	1	1.69	.00	0.00	.00	0.05	.00	L * N * DIF * DTF	2	0.12	.00	2.54	.00	0.72	.00
I * L * N	2	0.88	.00	5.10	.00	0.63	.00	L * N * M * DIF	2	0.22	.00	0.18	.00	1.37	.00
I * M * DIF	1	7.85	.00	1.35	.00	2.97	.00	L * N * M * DTF	2	0.26	.00	0.16	.00	0.02	.00
I * M * DTF	1	9.58	.00	0.57	.00	2.97	.00	N * M * DIF * DTF	2	0.71	.00	0.89	.00	17.04*	.02

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 19.

Power to Detect Functionally Uniform DIF by Study Variables

				All-other implementation				Constant implementation			
				.4 DIF		.8 DIF		.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	DTF	No DTF	DTF	No DTF	DTF	No DTF	DTF	No DTF
CSIBTEST											
	15	250	-.5	.06	.06	.24	.33	.50	.52	.73	.62
	15	250	.0	.15	.18	.48	.45	.43	.37	.62	.68
	15	500	-.5	.16	.20	.57	.72	.66	.58	.89	.82
	15	500	.0	.29	.38	.81	.66	.50	.49	.82	.90
	15	1000	-.5	.32	.49	.82	.89	.81	.58	.99	.98
	15	1000	.0	.55	.69	.95	.88	.65	.63	.93	.99
	30	250	-.5	.10	.05	.28	.42	.18	.09	.46	.34
	30	250	.0	.04	.10	.46	.43	.19	.14	.43	.43
	30	500	-.5	.08	.11	.57	.71	.26	.24	.72	.65
	30	500	.0	.21	.29	.82	.72	.26	.28	.67	.66
	30	1000	-.5	.27	.38	.95	.93	.46	.39	.94	.87
	30	1000	.0	.57	.69	.99	.93	.37	.17	.90	.90
IRT-LR											
	15	250	-.5	1.00	1.00	1.00	1.00	.10	.19	.29	.18
	15	250	.0	.99	1.00	1.00	1.00	.17	.04	.32	.24
	15	500	-.5	1.00	1.00	1.00	1.00	.29	.20	.58	.42
	15	500	.0	1.00	1.00	1.00	1.00	.24	.14	.63	.43
	15	1000	-.5	1.00	1.00	1.00	1.00	.40	.27	.93	.72
	15	1000	.0	1.00	1.00	1.00	1.00	.33	.31	.91	.69
	30	250	-.5	.52	.47	.86	.88	.17	.21	.54	.32
	30	250	.0	.44	.50	.89	.92	.23	.16	.56	.36
	30	500	-.5	.60	.63	.99	.99	.39	.22	.81	.68
	30	500	.0	.58	.63	1.00	.99	.35	.21	.82	.71
	30	1000	-.5	.71	.68	1.00	1.00	.65	.40	.99	.91
	30	1000	.0	.70	.73	1.00	1.00	.58	.35	1.00	.94
LOGREG											
	15	250	-.5	.00	.02	.48	.37	.04	.00	.71	.11
	15	250	.0	.02	.00	.33	.68	.03	.00	.58	.66
	15	500	-.5	.15	.03	.82	.87	.67	.03	1.00	.46
	15	500	.0	.04	.09	.89	.95	.09	.05	.98	.97
	15	1000	-.5	.73	.18	.99	.98	.96	.03	1.00	.95
	15	1000	.0	.23	.63	1.00	1.00	.48	.55	1.00	1.00
	30	250	-.5	.07	.02	.53	.31	.06	.01	.71	.03
	30	250	.0	.00	.03	.19	.53	.01	.15	.60	.05
	30	500	-.5	.18	.09	.80	.77	.42	.12	.97	.12
	30	500	.0	.03	.26	.82	.92	.08	.24	.94	.26
	30	1000	-.5	.56	.33	1.00	.98	.94	.29	1.00	.25
	30	1000	.0	.30	.67	1.00	1.00	.51	.61	1.00	.58

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Type III Error Rates

A significant difference was found among the procedures for Type III error rate, $F(2,285) = 2773.63$, $p < .05$. Post hoc tests revealed that CSIBTEST (.24) demonstrated lower Type III error rates than both IRT-LR (.62) and LOGREG (.96). Additionally, it was found that IRT-LR produced significantly lower error rates than LOGREG.

It was expected that the procedures would exhibit higher Type III error rates for functionally uniform DIF than for the other types of DIF (H12). To test this, a one-way ANOVA was run for each procedure. For CSIBTEST, a significant difference in error rate by DIF prototype was found, $F(4,475) = 55.21$, $p < .05$. Post hoc tests revealed that functionally uniform DIF (.34) demonstrated a significantly higher error rate than uniform (.11), nonuniform (.18), and unidirectional mixed DIF (.07) - but not crossing mixed DIF (.51). A significant difference was also found for IRT-LR, $F(4,475) = 108.57$, $p < .05$, but again the hypothesis was not supported. Specifically, the rate for functionally uniform (.78) was significantly greater than that for unidirectional mixed (.34) and crossing mixed DIF (.24), but not than the rate for uniform (.78) and nonuniform DIF (.98). For LOGREG a significant difference was found, $F(4,475) = 18.19$, $p < .05$, but again cell means did not confirm the hypothesis. Namely, it was found that the error rate for functionally uniform DIF (.88) was significantly lower than that for uniform (.96), nonuniform (.99), unidirectional mixed (.98), and crossing mixed (.98).

Table 22 presents the ANOVA results by procedure and Table 23 the mean error rates for each procedure by manipulation. A number of factors were found to affect the procedures, including many significant interactions; however, these could be accounted

for by the main effects that are outlined. Consistent with Hypothesis 9 and the findings for power, an effect for sample size was seen for all procedures in which error rates decreased as sample size grew. Also, it was specified that lower Type III error rates would be observed for the procedures in the larger DIF magnitude conditions (H8). A significant DIF magnitude effect was found for CSIBTEST and LOGREG. For LOGREG, a small drop in error rate was found in the 0.8 DIF condition relative to the 0.4. Conversely, for CSIBTEST, a slight increase was seen in the 0.8 conditions relative to the 0.4. This provides only partial support for Hypothesis 8. It was also stated that lower Type III error would be observed in the 30-item test conditions (H11). A significant main effect for test was found across procedures. Findings for CSIBTEST and LOGREG support the hypothesis. No cell differences were seen for IRT-LR. Additionally, a main effect was found for impact such that the procedures demonstrated lower error rates when it was absent.

A significant implementation main effect was seen for all procedures. For CSIBTEST, a lower Type III error rate was shown in the all-other conditions whereas LOGREG was lower in the constant; IRT-LR exhibited minor differences. A significant DTF effect was found for IRT-LR, in which error rate was lower in the maximum DTF conditions, and LOGREG, although it produced no mean difference in these conditions. Hypothesis 10 postulated that lower Type III error would be observed for the constant implementation relative to the all-other in conditions with increased contamination potential. To investigate this, the significance of the implementation x DIF x DTF interaction and its corresponding cell means were examined. The interaction was found to be significant for all three procedures however its effect on the cell means was not

consistent with the hypothesis. For CSIBTEST, the all-other conditions showed lower error rates than the constant though error rates in the all-other conditions were lowest in the 0.4 DIF with No DTF conditions. Results for IRT-LR showed comparable error rates across implementation in the 0.4 DIF conditions but, in the 0.8, lower error rates were found in the maximum DTF conditions. Error rates for LOGREG were found to be equally high across conditions when the all-other approach was used. When the constant approach was used, lower error rates were seen when DTF was present and in 0.8 DIF conditions.

In sum, none of the methods exhibited an average Type III error rate near .05. Of the procedures, CSIBTEST yielded the best rate followed by IRT-LR. In contrast to findings so far, LOGREG provided the worst performance with error rates near or equal to 1.0 in the majority of conditions. For the manipulations, it was found that increased sample size and the absence of impact reduced error rates for all procedures. Also, increased test length improved error rates for the nonparametric procedures. In line with findings thus far, CSIBTEST's performance was worsened by contamination when the all-other approach used. Additionally, it was found that error rates varied by DIF prototype across the procedures, although - differing from expectation - functionally uniform DIF was not consistently misclassified at a greater rate than the other prototypes.

Table 20.

ANOVA Results for Type III Error Rate by Study Procedures

		CSIBTEST		IRTLR		LOGREG				CSIBTEST		IRTLR		LOGREG	
Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2	Source	*df*	*F*	η^2	*F*	η^2	*F*	η^2
DIF	1	170.82*	.01	1.39	.00	149.91*	.02	I * N * DIF	2	99.94*	.01	75.62*	.02	157.82*	.05
DTF	1	10.91	.00	1912.80*	.21	159.29*	.02	I * N * DTF	2	3.66	.00	173.59*	.04	10.46*	.00
Impact (M)	1	7450.95*	.30	44.70*	.00	51.64*	.01	I * N * M	2	51.45*	.00	12.80*	.00	13.42*	.00
Implementation (I)	1	7396.11*	.30	81.49*	.01	2308.05*	.33	L * DIF * DTF	1	52.05*	.00	13.22*	.00	95.27*	.01
Length (L)	1	1179.41*	.05	11.93*	.00	65.20*	.01	L * M * DIF	1	64.73*	.00	6.20	.00	85.84*	.01
Sample size (N)	2	604.35*	.05	124.29*	.03	70.22*	.02	L * M * DTF	1	1.50	.00	6.83	.00	32.76*	.00
DIF * DTF	1	63.26*	.00	336.45*	.04	73.60*	.01	L * N * DIF	2	70.00*	.01	22.56*	.00	1.90	.00
I * DIF	1	484.81*	.02	6.08	.00	161.02*	.02	L * N * DTF	2	6.35	.00	101.58*	.02	20.08*	.01
I * DTF	1	5.67	.00	102.65*	.01	278.22*	.04	L * N * M	2	16.69*	.00	56.50*	.01	79.83*	.02
I * L	1	2229.03*	.09	656.56*	.07	14.17*	.00	M * DIF * DTF	1	31.62*	.00	126.55*	.01	0.28	.00
I * M	1	510.51*	.02	0.72	.00	25.69*	.00	N * DIF * DTF	2	13.49*	.00	5.91	.00	29.66*	.01
I * N	2	155.75*	.01	341.64*	.08	41.19*	.01	N * M * DIF	2	44.90*	.00	10.56*	.00	75.60*	.02
L * DIF	1	90.67*	.00	4.33	.00	0.97	.00	N * M * DTF	2	8.17*	.00	16.82*	.00	43.59*	.01
L * DTF	1	515.50*	.02	785.61*	.09	89.13*	.01	I * L * DIF * DTF	1	5.89	.00	0.49	.00	110.32*	.02
L * M	1	23.48*	.00	33.10*	.00	30.23*	.00	I * L * M * DIF	1	51.34*	.00	21.67*	.00	58.26*	.01
L * N	2	2.39	.00	166.32*	.04	14.26*	.00	I * L * M * DTF	1	0.37	.00	1.38	.00	38.49*	.01
M * DIF	1	401.58*	.02	196.79*	.02	133.32*	.02	I * L * N * DIF	2	0.16	.00	12.72*	.00	3.70	.00
M * DTF	1	12.96*	.00	61.82*	.01	14.91*	.00	I * L * N * DTF	2	54.63*	.00	33.85*	.01	25.98*	.01
N * DIF	2	27.56*	.00	47.30*	.01	115.72*	.03	I * L * N * M	2	1.91	.00	37.51*	.01	64.22*	.02
N * DTF	2	67.80*	.01	128.15*	.03	2.94	.00	I * M * DIF * DTF	1	9.90	.00	269.10*	.03	4.76	.00
N * M	2	6.13	.00	2.20	.00	13.50*	.00	I * N * DIF * DTF	2	5.73	.00	13.16*	.00	29.95*	.01
I * DIF * DTF	1	78.52*	.00	91.20*	.01	71.88*	.01	I * N * M * DIF	2	19.64*	.00	31.31*	.01	66.00*	.02
I * L * DIF	1	53.54*	.00	16.05*	.00	0.96	.00	I * N * M * DTF	2	2.27	.00	28.45*	.01	29.92*	.01
I * L * DTF	1	278.94*	.01	0.52	.00	112.32*	.02	L * M * DIF * DTF	1	5.21	.00	2.45	.00	43.32*	.01
I * L * M	1	41.28*	.00	67.79*	.01	28.53*	.00	L * N * DIF * DTF	2	3.83	.00	92.40*	.02	17.26*	.00
I * L * N	2	2.30	.00	19.34*	.00	23.69*	.01	L * N * M * DIF	2	20.15*	.00	121.13*	.03	25.34*	.01
I * M * DIF	1	120.00*	.00	60.95*	.01	223.66*	.03	L * N * M * DTF	2	13.83*	.00	36.89*	.01	20.85*	.01
I * M * DTF	1	60.30*	.00	129.20*	.01	20.96*	.00	N * M * DIF * DTF	2	41.46*	.00	89.26*	.02	16.84*	.00

Note. CSIBTEST = crossing simultaneous bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF indicates that DIF items may favor either group. * $p < .00089$.

Table 21.

Type III Error Rate by Study Variables

				All-other implementation				Constant implementation			
				.4 DIF		.8 DIF		.4 DIF		.8 DIF	
Procedure	Length	Sample	Impact	DTF	No DTF	DTF	No DTF	DTF	No DTF	DTF	No DTF
CSIBTEST											
	15	250	-.5	.25	.28	.29	.33	.38	.36	.35	.38
	15	250	.0	.17	.13	.12	.19	.30	.34	.33	.31
	15	500	-.5	.25	.26	.22	.31	.39	.36	.36	.38
	15	500	.0	.10	.08	.13	.18	.28	.34	.26	.28
	15	1000	-.5	.13	.21	.19	.31	.37	.36	.32	.37
	15	1000	.0	.06	.11	.07	.14	.33	.34	.25	.23
	30	250	-.5	.32	.27	.32	.27	.27	.23	.35	.29
	30	250	.0	.22	.14	.20	.16	.28	.25	.19	.20
	30	500	-.5	.27	.19	.26	.28	.23	.26	.30	.30
	30	500	.0	.13	.13	.20	.14	.23	.21	.20	.19
	30	1000	-.5	.17	.15	.28	.27	.24	.26	.30	.31
	30	1000	.0	.12	.07	.19	.11	.16	.24	.17	.18
IRT-LR											
	15	250	-.5	.61	.63	.53	.72	.63	.68	.64	.75
	15	250	.0	.65	.63	.56	.60	.64	.69	.59	.72
	15	500	-.5	.53	.61	.51	.74	.59	.71	.62	.72
	15	500	.0	.57	.68	.51	.65	.63	.69	.56	.70
	15	1000	-.5	.54	.64	.48	.72	.55	.60	.55	.60
	15	1000	.0	.55	.68	.59	.68	.54	.62	.54	.60
	30	250	-.5	.69	.65	.64	.63	.62	.63	.63	.59
	30	250	.0	.65	.66	.64	.65	.61	.61	.61	.61
	30	500	-.5	.64	.61	.62	.63	.60	.60	.60	.61
	30	500	.0	.63	.61	.62	.60	.60	.60	.57	.62
	30	1000	-.5	.62	.63	.61	1.00	.60	.59	.58	.60
	30	1000	.0	.61	.69	.60	.64	.59	.59	.57	.61
LOGREG											
	15	250	-.5	1.00	.99	1.00	1.00	.91	.90	.98	.95
	15	250	.0	1.00	1.00	1.00	.99	1.00	1.00	.98	.97
	15	500	-.5	1.00	1.00	.99	.99	.96	1.00	.89	.98
	15	500	.0	1.00	.96	.99	.99	.93	1.00	.92	.88
	15	1000	-.5	.99	.99	1.00	.99	.93	.91	.94	.97
	15	1000	.0	.98	1.00	.99	.99	.98	.99	.82	.83
	30	250	-.5	1.00	1.00	.99	.99	.97	.93	.95	.97
	30	250	.0	1.00	1.00	.99	.99	.80	.96	.89	.98
	30	500	-.5	.99	.99	.99	.99	.96	.96	.89	.96
	30	500	.0	.97	.96	.99	.98	.98	.94	.83	.97
	30	1000	-.5	.99	.99	.98	.97	.92	.96	.87	.96
	30	1000	.0	1.00	.98	.99	.99	.98	.94	.78	.95

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. DTF indicates that all DIF items favor the reference group. No DTF inidicates that DIF items may favor either group.

Classification of DIF Type

To identify whether any particular DIF type was systematically misidentified, Tables 22 through 25 present details about the classification of hits in the form of confusion matrices. Specifically, for all procedures, each test item (labeled by number in the tables) and the DIF type it is known to exhibit are listed. The total number of times a procedure flagged a given test item as a DIF item is shown under the column labeled "Total detections". The remaining columns indicate the number of times a detection (regardless of whether it was correct) met the criteria listed in Table 5 and was classified as the DIF type listed in the column label: "Uniform", "Nonuniform", or "Mixed". Note that it is possible for a significant omnibus test for DIF to not yield a significant follow-up test for a specific DIF prototype; the number of times this occurred is captured under the column labeled "No type indicated". For CSIBTEST, classification is based on the detection of a point at which the group IRFs cross, therefore it has no identifications in the "Mixed" column. Additionally, CSIBTEST always provides a classification whenever an item is flagged, thus it has no identifications in the "No type indicated" column.

For items that are known to exhibit "No DIF", all detections are false positives and reflect Type I errors. For test items known to exhibit DIF (uniform, nonuniform, functionally uniform, crossing mixed, and unidirectional mixed), all detections are true positives and represent power. Furthermore, for items know to possess DIF, bold entries highlight correct detections followed by accurate identifications; ideally, these values should be large relative to the other identification values for the same test item. Given

this, all identifications that are *not* bold represent misidentifications and therefore indicate Type III errors.

Of the procedures, CSIBTEST provided the greatest number of accurate classifications. In the all-other conditions, items known to exhibit nonuniform DIF were accurately identified for 99% of detections (across both levels of test length). Items know to exhibit uniform DIF were correctly identified 83% of the time, which is similar to the observed rate for crossing mixed (81%) and unidirectional mixed DIF (85%). Functionally uniform DIF was correctly identified in 53% of detections. Of the false positives, 62% were classified as nonuniform. For the constant conditions, items known not to have DIF tended to be identified as nonuniform (64%) when incorrectly detected. 72% of nonuniform DIF items were correctly identified, which is lower than the rate for uniform (.94), unidirectional mixed (.97), and functionally uniform (.75). Additionally, correct detections of crossing mixed DIF items were properly identified only 16% of the time.

Both IRT-LR and LOGREG exhibited a low number of detections that were accurately identified in follow up tests. For IRT-LR, when an all-other approach was used, the procedure tended to classify false positives as uniform in the 15-item conditions (57%) and mixed in the 30-item (95%). In terms of identifying DIF type, classification accuracy was low for detections of the items known to possess nonuniform (3%) and functionally uniform DIF (0.06%). Rates improved somewhat for detections of uniform DIF (14%) and substantially for items known to exhibit unidirectional mixed (36%) and crossing mixed DIF (68%), it should be noted that 75% of detections related to items known to *not* exhibit mixed DIF received the classification of mixed, which draws into

question the veracity of these findings. Additionally, as can be seen in Table 24, Item 30 was never detected in the all-other conditions; this is likely attributable to poor estimation of the item's parameters, which demonstrated high discrimination and low difficulty. Findings for the constant conditions of IRT-LR presented a similar pattern. Namely, less than 1% of detections for items known to exhibit nonuniform and functionally uniform DIF were properly identified. The percentage of correct identifications for uniform was 19% of detections, and again improved for items known to exhibit unidirectional mixed (80%) and crossing mixed DIF items (97%). Although the latter two findings are again called into question as 88% of the non-mixed DIF items were also classified as mixed; similarly, 97% of false positives were labeled mixed.

For LOGREG in the all-other conditions, detections of known non-DIF items tended not to be identified as any particular DIF type (60%); this trend extended to the DIF items, resulting in a low rate of accurate identifications. That is, correct detections for items known to exhibit uniform, nonuniform, unidirectional mixed, and crossing mixed DIF were properly identified in follow up tests less than 1% of the time. For functionally uniform, the rate was about 2%. The constant conditions showed a minor improvement. Specifically, rates for uniform (5%), nonuniform (2%), and crossing mixed DIF (4%) marginally increased. Of the detections for items known to exhibit functionally uniform DIF, 31% were properly identified; on the other hand, for unidirectional mixed DIF the rate was 0.23%.

Table 22.

Confusion Matrix of Detections and Identifications by Procedure for 15-item All-other Implementation Conditions

Procedure	Item number	DIF type item is known to exhibit	Total detections	Identifications			
				Uniform	Nonuniform	Mixed	No type indicated
CSIB							
	6	No DIF	233	131	102	NA	NA
	7	No DIF	290	161	129	NA	NA
	8	No DIF	161	39	122	NA	NA
	9	No DIF	105	29	76	NA	NA
	10	No DIF	183	72	111	NA	NA
	11	Uniform	2133	**1785**	348	NA	NA
	12	Nonuniform	1041	21	**1020**	NA	NA
	13	Functionally Uniform	1169	**743**	426	NA	NA
	14	Crossing mixed	1234	423	**811**	NA	NA
	15	Unidirectional Mixed	2217	**2040**	177	NA	NA
IRT-LR							
	6	No DIF	2839	1907	252	217	463
	7	No DIF	2339	1443	184	106	606
	8	No DIF	3000	2418	0	581	1
	9	No DIF	2998	1246	211	766	775
	10	No DIF	2604	816	393	694	701
	11	Uniform	1728	**851**	62	713	102
	12	Nonuniform	2164	785	**161**	954	264
	13	Functionally Uniform	3000	2558	**1**	439	2
	14	Crossing mixed	2155	806	153	**696**	500
	15	Unidirectional Mixed	2212	1590	11	**502**	109
LOGREG							
	6	No DIF	92	6	12	14	60
	7	No DIF	125	8	19	2	96
	8	No DIF	53	0	2	5	46
	9	No DIF	30	3	0	2	25
	10	No DIF	53	2	0	4	47
	11	Uniform	2176	**32**	13	1	2130
	12	Nonuniform	1247	3	**1**	4	1239
	13	Functionally Uniform	1151	8	**16**	0	1127
	14	Crossing mixed	1674	10	12	**12**	1640
	15	Unidirectional Mixed	2363	23	134	**4**	2202

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. Bold font indicates an accurate detection and identification. NA = Not Applicable, CSIBTEST can only classify DIF as uniform or nonuniform.

Table 23.

Confusion Matrix of Detections and Identifications by Procedure for 15-item Constant Implementation Conditions

Procedure	Item number	DIF type item is known to exhibit	Total detections	Identifications			
				Uniform	Nonuniform	Mixed	No type indicated
CSIB							
	6	No DIF	999	738	261	NA	NA
	7	No DIF	1031	781	250	NA	NA
	8	No DIF	912	695	217	NA	NA
	9	No DIF	921	692	229	NA	NA
	10	No DIF	949	751	198	NA	NA
	11	Uniform	2390	**2333**	57	NA	NA
	12	Nonuniform	903	599	**304**	NA	NA
	13	Functionally Uniform	1855	**1729**	126	NA	NA
	14	Crossing mixed	1866	1770	**96**	NA	NA
	15	Unidirectional Mixed	2510	**2455**	55	NA	NA
IRT-LR							
	6	No DIF	411	3	6	400	2
	7	No DIF	453	19	13	415	6
	8	No DIF	466	0	4	462	0
	9	No DIF	223	0	0	223	0
	10	No DIF	268	6	1	259	2
	11	Uniform	1917	**412**	114	1346	45
	12	Nonuniform	1488	36	**18**	1432	2
	13	Functionally Uniform	982	0	**13**	969	0
	14	Crossing mixed	1321	81	1	**1239**	0
	15	Unidirectional Mixed	2234	612	16	**1562**	44
LOGREG							
	6	No DIF	466	35	23	5	403
	7	No DIF	618	52	12	5	549
	8	No DIF	78	30	2	0	46
	9	No DIF	58	17	2	1	38
	10	No DIF	222	20	48	2	152
	11	Uniform	2193	**119**	674	169	1231
	12	Nonuniform	1290	669	**29**	65	527
	13	Functionally Uniform	1244	80	**317**	17	830
	14	Crossing mixed	1962	349	393	**33**	1187
	15	Unidirectional Mixed	2542	433	67	**10**	2032

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. Bold font indicates an accurate detection and identification. NA = Not Applicable, CSIBTEST can only classify DIF as uniform or nonuniform.

Table 24.

Confusion Matrix of Detections and Identifications by Procedure for 30-item, All-other Implementation Conditions

				Identifications							Identifications			
Procedure	Item number	DIF type item is known to exhibit	Total detections	Uniform	Nonuniform	Mixed	No type indicated	Item number	DIF type item is known to exhibit	Total detections	Uniform	Nonuniform	Mixed	No type indicated
CSIB														
	6	No DIF	196	81	115	NA	NA	21	No DIF	238	45	193	NA	NA
	7	No DIF	231	104	127	NA	NA	22	No DIF	202	68	134	NA	NA
	8	No DIF	136	38	98	NA	NA	23	No DIF	209	72	137	NA	NA
	9	No DIF	110	25	85	NA	NA	24	No DIF	131	36	95	NA	NA
	10	No DIF	150	58	92	NA	NA	25	No DIF	231	100	131	NA	NA
	11	Uniform	2104	**1524**	580	NA	NA	26	Uniform	2195	**2018**	177	NA	NA
	12	Nonuniform	1212	8	**1204**	NA	NA	27	Nonuniform	1346	9	**1337**	NA	NA
	13	Functionally Uniform	1127	**570**	557	NA	NA	28	Functionally Uniform	1114	**478**	636	NA	NA
	14	Crossing mixed	1402	300	**1102**	NA	NA	29	Crossing mixed	1438	63	**1375**	NA	NA
	15	Unidirectional Mixed	2259	**1768**	491	NA	NA	30	Unidirectional Mixed	2264	**1926**	338	NA	NA
IRT-LR														
	6	No DIF	2790	0	0	2790	0	21	No DIF	960	0	20	940	0
	7	No DIF	609	0	0	609	0	22	No DIF	300	0	1	299	0
	8	No DIF	219	0	0	219	0	23	No DIF	593	0	3	590	0
	9	No DIF	279	7	0	271	1	24	No DIF	816	0	0	816	0
	10	No DIF	2024	402	78	1514	30	25	No DIF	1975	0	0	1975	0
	11	Uniform	1702	**42**	6	1653	1	26	Uniform	2852	**0**	0	2852	0
	12	Nonuniform	1272	0	**6**	1266	0	27	Nonuniform	2738	0	**0**	2738	0
	13	Functionally Uniform	1494	38	**1**	1455	0	28	Functionally Uniform	2504	0	**2**	2502	0
	14	Crossing mixed	2336	586	23	**1706**	21	29	Crossing mixed	2251	0	84	**2167**	0
	15	Unidirectional Mixed	498	10	1	**487**	0	30	Unidirectional Mixed	0	0	0	**0**	0
LOGREG														
	6	No DIF	100	8	12	34	46	21	No DIF	58	3	20	2	33
	7	No DIF	144	6	49	3	86	22	No DIF	130	4	65	37	24
	8	No DIF	57	4	1	5	47	23	No DIF	66	0	1	51	14
	9	No DIF	34	2	1	1	30	24	No DIF	89	61	1	9	18
	10	No DIF	48	1	0	2	45	25	No DIF	167	8	18	5	136
	11	Uniform	2205	**44**	32	2	2127	26	Uniform	2282	**18**	1337	128	799
	12	Nonuniform	1358	8	**2**	15	1333	27	Nonuniform	1483	500	**1**	737	245
	13	Functionally Uniform	1125	16	**34**	0	1075	28	Functionally Uniform	1156	570	**31**	60	495
	14	Crossing mixed	1624	11	8	**21**	1584	29	Crossing mixed	1691	94	805	**23**	769
	15	Unidirectional Mixed	2334	29	97	**7**	2201	30	Unidirectional Mixed	2351	904	18	**13**	1416

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. Bold font indicates an accurate detection and identification. NA = Not Applicable, CSIBTEST can only classify DIF as uniform or nonuniform.

Table 25.

Confusion Matrix of Detections and Identifications by Procedure for 30-item, Constant Implementation Conditions

				Identifications							Identifications			
Procedure	Item number	DIF type item is known to exhibit	Total detections	Uniform	Nonuniform	Mixed	No type indicated	Item number	DIF type item is known to exhibit	Total detections	Uniform	Nonuniform	Mixed	No type indicated
CSIB														
	6	No DIF	179	60	119	NA	NA	21	No DIF	198	90	108	NA	NA
	7	No DIF	211	71	140	NA	NA	22	No DIF	218	69	149	NA	NA
	8	No DIF	171	46	125	NA	NA	23	No DIF	189	75	114	NA	NA
	9	No DIF	181	62	119	NA	NA	24	No DIF	178	45	133	NA	NA
	10	No DIF	178	61	117	NA	NA	25	No DIF	194	57	137	NA	NA
	11	Uniform	2040	**1769**	271	NA	NA	26	Uniform	2048	**1974**	74	NA	NA
	12	Nonuniform	741	32	**709**	NA	NA	27	Nonuniform	741	27	**714**	NA	NA
	13	Functionally Uniform	1123	**722**	401	NA	NA	28	Functionally Uniform	1131	**626**	505	NA	NA
	14	Crossing mixed	740	592	**148**	NA	NA	29	Crossing mixed	537	279	**258**	NA	NA
	15	Unidirectional Mixed	2153	**2025**	128	NA	NA	30	Unidirectional Mixed	1971	**1937**	34	NA	NA
IRT-LR														
	6	No DIF	617	5	6	605	1	21	No DIF	266	0	4	262	0
	7	No DIF	678	32	10	630	6	22	No DIF	942	0	26	916	0
	8	No DIF	605	0	0	605	0	23	No DIF	265	0	2	263	0
	9	No DIF	212	0	0	212	0	24	No DIF	576	0	2	574	0
	10	No DIF	301	9	0	291	1	25	No DIF	886	2	19	865	0
	11	Uniform	2121	**607**	65	1304	145	26	Uniform	2110	**171**	65	1852	22
	12	Nonuniform	1715	37	**8**	1670	0	27	Nonuniform	1813	38	**20**	1755	0
	13	Functionally Uniform	1281	0	**4**	1277	0	28	Functionally Uniform	1444	0	**10**	1434	0
	14	Crossing mixed	1502	59	0	**1443**	0	29	Crossing mixed	1573	3	5	**1565**	0
	15	Unidirectional Mixed	2379	638	8	**1701**	32	30	Unidirectional Mixed	2345	1	25	**2319**	0
LOGREG														
	6	No DIF	188	47	9	5	127	21	No DIF	196	21	21	7	147
	7	No DIF	232	62	14	3	153	22	No DIF	213	78	2	5	128
	8	No DIF	74	34	3	1	36	23	No DIF	188	26	11	1	150
	9	No DIF	78	33	2	0	43	24	No DIF	125	31	12	7	75
	10	No DIF	130	28	12	3	87	25	No DIF	227	68	36	11	112
	11	Uniform	2019	**119**	554	270	1076	26	Uniform	2071	**70**	758	154	1089
	12	Nonuniform	967	641	**19**	73	234	27	Nonuniform	1101	675	**20**	110	296
	13	Functionally Uniform	1001	82	**314**	36	569	28	Functionally Uniform	1002	80	**366**	75	481
	14	Crossing mixed	1654	337	249	**110**	958	29	Crossing mixed	1682	541	71	**57**	1013
	15	Unidirectional Mixed	2433	379	29	**4**	2021	30	Unidirectional Mixed	2466	147	46	**3**	2270

Note. CSIBTEST = crossing simultaneous item bias test, IRT-LR = item response theory likelihood ratio test, and LOGREG = logistic regression. Bold font indicates an accurate detection and identification. NA = Not Applicable, CSIBTEST can only classify DIF as uniform or nonuniform.

CHAPTER 7

Discussion

The high stakes nature of testing in organizational settings has resulted in great concern over the potential for bias against a particular segment of the population (Sackett, Schmitt, Ellingson, & Kabin, 2001). Among non-psychometricians, tests on which the majority group on average achieves higher scores than a minority group are seen as biased, and this standard has informed most assessment related legislation in the United States (McAllister, 1993). However, mean test score differences between groups are not necessarily a symptom of biased tests; an important distinction must be made between impact - dissimilar performance due to a difference in the distribution of ability between the groups - and differential functioning, or the presence of performance differences for individuals of equal ability (Drasgow & Hulin, 1990; Shealy & Stout, 1993).

A test item that exhibits differential functioning (or DIF) is one for which examinees of equal ability, but from different portions of the population, have an unequal probability of endorsement (Hambleton & Swaminathan, 1985; Hulin et al., 1983; Lord, 1980). A test in which the cumulative effects of DIF influence total score would be said to exhibit DTF (Raju et al., 1995). Test developers and researchers are generally concerned with the detection of items that favor the majority or reference group within the population over a minority or focal group.

The presence of differential functioning signals problems within a test that should be addressed. Failure to do so can reduce the effectiveness of a selection system as factors other than those that an instrument was intended to assess may be affecting

observed scores (Ackerman, 1992; Hunter & Schmidt, 2000; Lord, 1980; Shealy & Stout, 1993). Furthermore, the use of assessments that exhibit substantial differential functioning favoring the reference group may create a situation in which selection rates violate the 4/5th rule used for determining the presence of adverse impact. This means that failure to detect differential functioning and accordingly act could result in missed opportunities for qualified minority applicants and in possible litigation against the test-user. Given the serious consequences of differential functioning for both test-taker and -user alike, the identification of procedures that can effectively detect DIF is important.

This simulation's objective was to examine the efficacy of three methods, CSIBTEST (Li & Stout, 1996), IRT-LR (Thissen et al., 1988), and LOGREG (Swaminathan & Rogers, 1990), to detect and classify various forms of DIF. Moreover, each procedure was implemented using both an all-other and a constant approach to the trait estimate (i.e., all test items but the one under study are used to estimate trait standing or a pre-selected subset is used, respectively). This was done to ascertain its effect upon the efficacy of the procedures and to ensure their comparability (typically, CSIBTEST and LOGREG utilize an all-other approach and for IRT-LR it is increasingly recommended that a constant approach be used). Within this section, a summary of the simulation's results are presented and related to past studies of these procedures. Additionally, the limitations of the present effort and ideas for future research are offered. Finally, the implications of the findings upon practice are discussed.

Summary of Type I Error Results

Of the study procedures, it was found that LOGREG yielded the lowest average Type I error rate and that both CSIBTEST and IRT-LR demonstrated error rates that were greater the nominal .05 level. These findings were unanticipated as past research on the DIF detection methods included in this study has generally found that they all exhibit Type I errors near .05. For example, Li and Stout (1996) found that, when no DIF items were present, CSIBTEST demonstrated an overall error rate of .04 - a rate that was similar to the comparison procedures: SIBTEST and MH, two popular uniform DIF detection procedures. Moreover, in a second study, they found that CSIBTEST exhibited a more controlled Type I error rate than LOGREG, though this finding was at odds with past research by Rogers and Swaminathan (1993) that found similar rates for the two procedures. Additionally, many studies (e.g., Lopez Rivas et al., 2009; Stark et al., 2006) have found that IRT-LR produces well controlled error rates when a constant approach is used. Although dissimilar to previous findings, the pattern of results in this study can be explained by its inclusion of conditions that were not present in preceding investigations.

First, for CSIBTEST and LOGREG, how the matching subtest and trait estimate were constructed was manipulated. When an all-other approach was used, both CSIBTEST and LOGREG generally exhibited controlled Type I error rates; this finding accords with past studies (e.g., Finch & French, 2007). However, when a constant implementation was used, error rates for CSIBTEST in the 15-item test length condition increased markedly. Thus it appears that for shorter tests the constant approach to CSIBTEST may not be advisable because, as shown here, a 5-item anchor appears to be insufficient, but for longer tests it could be a viable alternative if enough anchor items can

be identified. For LOGREG, inflated error rates were found for the constant conditions when impact was present. This renders suspect its ability to reliably discern bias from differences in group ability, and therefore draws into question this implementation's usefulness. Additionally, constant LOGREG was also found to exhibit higher error in the short test length conditions, which again suggests that more anchor items may be needed for the constant implementation to be effective.

A second consideration is the inclusion of a DTF manipulation in the study. Namely, in the DTF conditions, where DIF items were designed to benefit only the reference group, CSIBEST demonstrated greater Type I error rates than LOGREG; an effect that again differed according to implementation. For CSIBTEST, when a constant implementation was used, error rates across DTF conditions were comparable but, in the all-other conditions, the DTF conditions had higher error rates and worsened as DIF magnitude increased. This confirmed that CSIBTEST is sensitive to the presence of contamination within the trait estimate. On the other hand, the LOGREG conditions showed no difference due to DTF.

Lastly, Type I error results for IRT-LR were unexpectedly high, especially for the all-other implementation where rates reached 1.0 in some conditions. The constant implementation yielded better results; nevertheless, even these rates were above .05 and generally greater than what was observed for the other study procedures under analogous conditions. This trend is consistent with past work, that is, that the constant approach yields lower error rates than the all-other (e.g., Stark et al., 2006); however, the magnitude of the error rates found in this study were greater than what is typically seen for either implementation. A possible explanation is that this study used a 3-PL model

and, as the *c*-parameter is not usually investigated for DIF, this parameter was not constrained in the comparison models. It is conceivable that constraining this parameter would have generated results consistent with past simulations (e.g., Lopez Rivas et al., 2009). Such a method to building comparison models, in which all item parameters are constrained, is recommended in the literature (Thissen, 2001; Stark, 2006); however, for this study, it would have abrogated the ability to investigate the accuracy of DIF classifications as defined in Table 5.

Summary of Power Results

This simulation investigated the capacity of the procedures to detect the five DIF prototypes illustrated in Figure 4, thus, providing a wide-ranging assessment of the power of these procedures that complements the available literature. For example, there was previously little information available on the ability of IRT-LR to detect nonuniform DIF and, when extant, it utilized a constant implementation (e.g., Finch & French, 2007). Also, past studies of CSIBTEST and LOGREG did not explore their efficacy to detect crossing mixed DIF or functionally uniform DIF (e.g., Hildago & López-Pina, 2004; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). Relatedly, the ability of CSIBTEST to detect uniform DIF has not been well investigated.

Consistent with past research and study hypotheses, it was found that the power of the procedures increased with larger sample size and greater DIF magnitude. Furthermore, as seen in past research (e.g., Wang & Yeh, 2003), it was found that the constant conditions demonstrated better power when DIF was designed such that only one group benefitted from it (i.e., conditions where DTF was present). At odds with expectations and concerns over the presence of contamination in the matching subtest, it

was found that - even when DTF was present - the all-other implementation of the CSIBTEST and LOGREG provided better results than the constant. When coupled with the findings of better Type I error rates, it is clear that the nonparametric procedures are more effective with an all-other approach.

Of the study procedures, it was found that LOGREG yielded the highest overall power. Additionally, its power was not significantly affected by the presence of impact. Likewise, it performed similarly across test length conditions, which was unexpected as longer tests provide more information for ability estimation, and this could have been expected to improve its reliability and accuracy. In relation to its efficacy to detect the DIF types relative to each other, power was greatest for items in which difficulty differed and IRFs did not cross (uniform and unidirectional mixed DIF). Whenever DIF involved only differences in item discrimination, performance worsened (nonuniform and functionally uniform DIF), which confirmed expectations. Crossing mixed DIF, in which IRFs crossed and both the *a*- and *b*-parameters were shifted, was detected better than aforementioned item types.

The all-other implementation of LOGREG consistently demonstrated excellent power and controlled Type I error rates. Additionally, it outperformed the constant approach and both implementations of CSIBTEST and IRT-LR. Its power to detect uniform and unidirectional mixed DIF was significantly greater than that for nonuniform, crossing mixed, and functionally uniform DIF. Furthermore, its effectiveness for detecting crossing mixed DIF was significantly greater than its ability to detect nonuniform and functionally uniform DIF. Although similar power was achieved by the constant implementation of LOGREG, its average Type I error rate exceeded .05 due to

its poor performance in the impact conditions. This implementation exhibited a similar pattern for detection of the DIF prototypes: uniform and unidirectional mixed DIF had the greatest rate of detection, followed by nonuniform, crossing mixed, and functionally uniform.

CSIBTEST demonstrated the lowest overall power. Contrary to expectation, it showed significantly better power for the detection of uniform and unidirectional mixed DIF than nonuniform, crossing mixed, and functionally uniform DIF. Also, its ability to detect crossing mixed and functionally uniform DIF was significantly better than its ability to detect nonuniform DIF. This pattern of results was surprising because CSIBTEST was specifically created to detect DIF in which IRFs cross. For the all-other implementation, unidirectional mixed DIF was most frequently detected followed by uniform. The remaining DIF types: nonuniform, crossing mixed, and functionally uniform, were flagged at a comparable rate. For the constant implementation, power was greatest for items in which location was shifted and a crossing point was not present (uniform and unidirectional mixed DIF). Among the other DIF prototypes, it was found that functionally uniform DIF was detected at a better rate than crossing mixed and nonuniform DIF.

Despite its elevated Type I error rate, it was found that IRT-LR produced overall power comparable to that of LOGREG. In terms of its ability to detect the various DIF types, no significant differences were found between them. In regard to implementation, the constant IRT-LR proved to have power akin to other the procedures but exhibited an inflated Type I error rate. For the DIF prototypes, it was found that unidirectional mixed generated the most hits, and similar rates were observed for uniform, nonuniform, and

crossing mixed DIF. Functionally uniform DIF showed the lowest power. The all-other implementation produced Type I errors rates that well exceeded the nominal level. It was also found that these rates were substantially worse in the short test conditions. For both implementations, the greatest error rates were observed in the large sample size and long test length conditions; this again suggests the need for *p*-values smaller than .05 when this procedure is used.

In terms of the procedures' effectiveness for detecting the DIF prototypes relative to one another, it was found for uniform DIF that LOGREG generated the greatest number of hits followed by CSIBTEST and then constant IRT-LR. For LOGREG, the all-other approach performed best, achieving near perfect detection in the small DIF conditions when groups were of equal ability and DIF items favored both groups. Power for the all-other conditions was lower when impact was present; the influence of impact upon the constant conditions was confounded by an inflated Type I error rate. The consequences of the impact and DTF manipulations upon CSIBTEST differed by implementation. For the constant implementation, neither the inclusion of impact nor DTF had a negative effect but, for the all-other conditions, both reduced power. Additionally, for the latter, augmented test length improved detection. For constant IRT-LR, none of the manipulations had a substantial effect upon observed uniform DIF detection.

Nonuniform power results indicated that LOGREG generated more hits than CSIBTEST. This does not agree with the findings of Naranyanan and Swaminathan (1996) and Finch and French (2007) though this is again attributable to the inclusion of different approaches to the trait estimates. That is, the nonuniform DIF detection rate for

the constant LOGREG conditions was substantially greater than that of the constant CSIBTEST conditions. Similarly, no advantage was gained for LOGREG by extending test length, which again counters past findings (e.g., Swaminathan & Rogers, 1990) but this is due to the decline in power observed in the 30-item, constant conditions. This suggests that in longer tests more anchor items are needed to effectively detect nonuniform DIF when a constant approach to LOGREG is used. Contrary to findings for uniform DIF, the impact and DTF were found to have no effect upon the nonparametric procedures for the detection of nonuniform DIF. The constant implementation of IRT-LR detected nonuniform DIF as well as the all-other implementations of the nonparametric procedures, but also exhibited a greater Type I error rate.

For unidirectional mixed DIF, the all-other implementation of LOGREG again generated the most hits and provided near perfect detection in many conditions. The impact, DTF, and test length manipulations had no real effect upon observed performance. Results for the constant conditions were comparable, with no relationships appearing between observed hits and the impact, DTF, and test length manipulations. For CSIBTEST, the all-other conditions provided better detection than the constant conditions. None of the manipulations were found to exert significant influence. As was found for the other procedures, IRT-LR was also not adversely affected by the manipulations. Additionally, it was found that its implementations had similar power in spite of the highly elevate Type I error rates seen in the all-other conditions.

For power to detect crossing mixed DIF, it was again found that LOGREG outperformed the other procedures. Both implementations of LOGREG demonstrated similar power though it was found that the constant conditions benefit from the

introduction of DTF. Both CSIBTEST implementations were comparably effective and the all-other produced better results in the absence of DTF. Again, it was found that the performance of IRT-LR was not tied to any of the study manipulations other than sample size and DIF magnitude. Interestingly, the procedures demonstrated lower detection of crossing mixed DIF than unidirectional mixed, which parallels findings for uniform and nonuniform DIF. This shows that the procedures are less effective when group IRFs cross near the middle of the ability range, as opposed to simply when differences in *a*-parameters are present.

The detection of functional uniform DIF did not vary substantially across procedures or due to the study manipulations (besides samples size and DIF magnitude). Additionally, it did *not* have the lowest associated power relative to the other DIF prototypes, which ran counter to expectations. IRT-LR exhibited the greatest sensitivity but also demonstrated an inflated Type I error rate; the power for the constant conditions was generally comparable to the other procedures. For CSIBTEST, both implementations were comparable when observed differences in Type I error rate are considered. The LOGREG implementations also generated similar results with no differences arising due to impact, DTF, or test length.

Summary of Type III Error and Classification Accuracy Results

Many procedures have been developed that are capable of detecting more than one DIF type; however, the issue of classification accuracy was only recently broached in the literature by Finch and French (2008). In their study, the rate at which a procedure detected the DIF type that was not present was examined. In other words, the frequency with which when testing for uniform DIF, nonuniform DIF items were detected and vice

versa; these were described as anomalous Type I errors. Their results indicated that CSIBTEST exhibited the greatest rate of anomalous Type I errors when uniform DIF was present, followed by LOGREG, and the constrained baseline IRT-LR. When nonuniform DIF was present, it was found that LOGREG exhibited a slightly higher rate than constrained baseline IRT-LR; CSIBTEST was not included in this condition.

This simulation extended this line of research by comparing the actual DIF classification assigned to an item by a procedure against the DIF type that is known to be present. This means that, for a detection of DIF to be considered a correct identification, it must be flagged as a DIF item *and* assigned the correct DIF type based on the criteria laid out in Table 5. The frequency with which DIF was found but misclassified was presented as the procedure's Type III error rate (Mosteller, 1948). It was anticipated that Type III error would be influenced by the same variables that affect power. Consistent with this expectation, it was found that increased sample size and the absence of impact reduced error rates for all procedures. Additionally, performance improved in the longer test conditions for the nonparametric procedures. Surprisingly, DIF magnitude did not affect observed error rates.

It was found that CSIBTEST generated the lowest Type III error rate, which supports the proposition made by Li and Stout (1996) that the k_c value can be used to classify DIF. For implementation, it was found that the all-other conditions outperformed the constant. Additionally, using a longer test improved performance. As evidenced by the confusion matrices, the all-other approach provided near perfect identification of nonuniform DIF. The remaining DIF types were accurately identified at a good rate except for functionally uniform. For false positives, it was found that they were usually

classified as nonuniform. Results for the constant implementation differed. Namely, near perfect detection was found for uniform and unidirectional mixed DIF. Nonuniform and functionally uniform were correctly identified at a good rate, and crossing mixed identification was poor. False positives also tended to be identified as nonuniform in these conditions.

IRT-LR demonstrated a substantially greater Type III error rate than CSIBTEST. This was not found to vary across implementations. In terms of classification accuracy, IRT-LR tended to classify detections (including false positives) as mixed DIF. For the all-other approach, classification accuracy was low for uniform, nonuniform, and functionally uniform DIF. Results for mixed DIF are questionable due to the procedure's strong tendency to label items as mixed. This same pattern of findings was observed in the constant conditions. In fact, 72% of the total detections (hits and false positives) generated by IRT-LR were labeled as mixed; this indicates that both 1 df follow up tests were found to be significant a majority of the time; this again suggests that the procedure would benefit from the use of a smaller *p*-value.

LOGREG exhibited highly elevated Type III error rates, with a slight improvement in the constant conditions. Likewise, its percentage of correct identifications was poor as its detections tended to not be identified as any particular DIF type, including true positives. This indicates that, although the addition of the group membership and group x trait estimate interaction terms produced a significant change in model fit, the individual terms rarely reached statistical significance in the follow up tests. This suggests that, contrary to what is often stated in the literature, the ability and interaction terms in the LOGREG model do not indicate the type of DIF detected. For

the all-other conditions, the percentage of correct identifications was near zero across DIF prototypes. The constant conditions yielded better results for the identification of functionally uniform DIF but the other DIF prototypes were again near zero.

Study Limitations and Future Research

Though many factors were investigated in this study, there are many considerations that warrant additional scrutiny. Namely, data were generated exclusively with a 3-PL model, which has many implications. First, this limits the applicability of results to dichotomous items. Second, and more importantly, is the *c*-parameter. For IRT-LR, the use of a 3-PLM introduced an additional parameter that could have been constrained during model construction. As previously mentioned, in the comparison models the lower asymptote was not fixed and it appears that this lead to the inflated Type I error findings for this procedure. Subsequent research should account for this by adopting the omnibus approach suggested by Stark et al. (2006) or using the IRTLRDIF program (Thissen, 2001), which implements a sequential approach to the analysis that conditions on the *c*-parameter before testing the equivalence of the *a*-parameter and conditions on both the *c*- and *a*-parameters before testing the *b*-parameter (again, in this study, MULTILOG was used so direct tests of parameters could be conducted as shown in Table 5).

The *c*-parameter also potentially influenced the findings for LOGREG. Specifically, unlike IRT-LR and CSIBTEST, LOGREG does not account for guessing. Furthermore, past findings have not shown a consistent effect attributable to the use of a 3-PLM, which should worsen performance because the assumption of a linear relationship between the logit and predictor is violated (Rogers & Swaminathan, 1993).

One possible reason is that the influence of the *c*-parameter may vary as a function of many variables - some of which were manipulated in this study. For example, if DTF against the focal was present, resulting in a test difficulty of 0.5 and the center of this group's ability distribution was centered at -0.5, a large proportion of the sample will exhibit test scores that fall in the region of the ability distribution that corresponds to the lower asymptote; this harms the reliability of the trait estimate and results in poor model-data fit. Conversely, for the reference group, if the TCC were centered at 0.0 and group ability was (0, 1), the majority of the sample would fall in the range of ability that is not impacted by the *c*-parameter, thus model-data fit should be acceptable.

As can be seen from this example, for 3-PLM, it is difficult to determine the unique influence of differential functioning, impact, and model-data fit as, by definition, DIF and impact require manipulations that will determine the influence of the *c*-parameter and thus fit. This may account for the inconsistency of past research regarding the effects of 3-PLM upon LOGREG. Future research should attempt to isolate the unique influence of model-data fit by manipulating the magnitude of the *c*-parameter (e.g., 0.0 for good fit, 0.2 for misfit, and 0.4 for severe misfit) as well as different levels of impact and DTF. Also, the introduction of contamination into the constant anchor set that is equal to the total DTF in the test would help clarify the findings related to implementation.

A second limitation of the study relates to the number of levels in the sample manipulation. Namely, sample sizes were equal across groups in all conditions. This could be varied to determine its effect upon the efficacy of the procedures as, in most instances where DIF is being investigated, the available sample for the focal group is

likely smaller than that of the reference. Also, the largest sample size included in the study was 1,000 per group. It is possible that for IRT-LR the inclusion of large sample size conditions (e.g., 3,000 or more) could have occasioned better outcomes as parameter estimates would be more accurate, especially given the use of the 3-PLM in this study.

Alpha level is another variable that influenced the observed results. In this study a *p*-value of .05 was used in all conditions, it is possible that the addition of a condition in which a smaller or corrected alpha level was employed would have assisted to control the Type I error rate observed in the IRT-LR conditions. On the other hand, this would likely have decreased the power of CSIBTEST and LOGREG, both of which exhibited generally good Type I error rates with the all-other approach even at a .05 alpha level.

Another avenue for future analysis is to examine the effects of the anchor items used to estimate trait level upon the effectiveness of the constant approach. That is, research (e.g., Lopez et al., 2009) has found that the number of items in the anchor set affected the efficacy of IRT-LR when a constant approach was employed, such could be the case for CSIBTEST and LOGREG. In this study, it was found that the constant implementation yielded worse results than the all-other for these procedures; however, whether this was due to the number of items in the anchor set versus test length cannot be determined because these were nested factors (anchor set was 5 in 15-item test and 10 in 30-item test). Succeeding simulations may wish to fully crossing these variables to investigate their effects (e.g., anchor set of 5 in 15- and 30-item tests and anchor of 10 in 15- and 30-item tests).

Finally, the effect of DIF direction upon the efficacy of the constant implementation should be further investigated. As stated before, to minimize DTF in this

study, the direction of DIF was changed such that three items favored the reference group (uniform, nonuniform, and crossing mixed) and two favored the focal group (unidirectional mixed and functionally uniform; see Appendix B for more details). Wang and Yeh (2003) found that the power of constant IRT-LR was improved when DIF favored only one group. This study found this same result but across all of the study procedures. Future simulations could better evaluate this finding by including conditions in which the type of DIF present is constant and only direction is varied (e.g., test includes only one-sided, uniform DIF). Also, the number of items that favor each group could be varied.

Conclusions and Recommendations

It was found that LOGREG, when implemented using an all-other approach, provided the best detection while maintaining controlled Type I error rates. Moreover, it had one of the highest detection rates for three of the five studied DIF prototypes: uniform, unidirectional mixed, and crossing mixed. This conclusion was surprising for two reasons. First, CSIBTEST was specifically designed to detect DIF in which IRFs cross, yet it consistently detected items with *a*-parameter shifts at a lower rate than items with *b*-parameter shifts, and both at a lower rate than LOGREG. Second, the use of the 3-PLM, could have reduced the observed power of LOGREG; thus it is plausible that had a 2-PLM been used the advantage of LOGREG over the other study procedures would have been even greater.

This finding is favorable news for practitioners. LOGREG is a commonly known analysis that provides a flexible, model-based approach to DIF detection that is easily conducted via widely used statistical software such as SPSS. Additionally, it requires

smaller sample sizes than procedures that estimate IRT parameters and demonstrated excellent power in 15-item tests. Findings also suggest that its performance is not harmed by contamination within the ability estimate; this means that purification is not a necessary step, further easing the procedure's implementation.

In relation to the DIF prototypes, it was found that the three procedures regardless of implementation detected uniform DIF at a greater rate than nonuniform. Additionally, the procedures achieved excellent unidirectional mixed DIF detection but not crossing mixed DIF. Also, contrary to expectation, it was *not* found that functionally uniform DIF had a significantly lower rate of detection than the other DIF types; only for LOGREG was it detected at the lowest rate. These findings are notable because, across all of the prototypes, the magnitude of DIF was the same and, in the case of mixed DIF, involved shifts of both the *a*- and *b*-parameters; therefore, it seems that the presence of a point at which the IRFs cross is the cause of the observed performance difference (as opposed to differences in the *a*-parameter).

The Type III error results show that the study procedures could not reliably identify the type of DIF detected. Specifically, all procedures exhibited Type III error rates greater than .05; CSIBTEST had the lowest overall rate at .24. This inability to properly classify the type of DIF detected may be problematic in contexts where that information is used for some end (e.g., the translation of a validated instrument into another language, Ellis, 1989). However, in the context of high-stake testing, where the objective is to identify and remove items that could affect group pass rates, this limitation may not be a concern.

Consider the scenario that an organization uses an assessment as an initial hurdle in a selection system. In this system, applicants who score below the cutoff point do not proceed to the next step and - for applicants who score above the cutoff point - scores are used for rank ordering with the highest scoring candidates moving forward (i.e., top-down selection). Under such a scenario, if a measure were to exhibit sufficient nonuniform DIF against one group that it accumulated to produce DTF, the group for which the test's discrimination - or *a*-parameter - was lower would be at a disadvantage. Such a situation is illustrated in Figure 10, it shows the TCCs for two groups and a hypothetical assessment's cut score represented by a vertical line that coincides with a trait standing of 0.0.

As can be seen, in this scenario, if one were to calculate the signed area across the entire ability range, neither group would have an advantage. However, in actuality, given that scores below the cut score are treated the same (i.e., these candidates do not move forward), only the group for whom the advantage occurs above the cut score receives any benefit. Therefore, the presence of nonuniform DIF results in the same outcome as if the test had uniformly favored one group across the entire ability range, such a scenario supports Rogers and Swaminathan's (1993) assertion that DIF analysis should focus on the detection of *all* forms of differential item functioning and not solely only on group differences in observed difficulty.

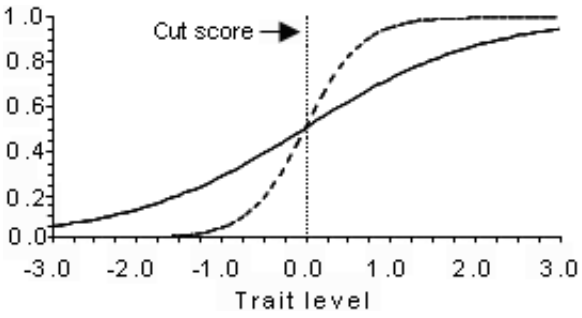

Figure 10. Two Test Characteristic Curves (TCCs) for hypothetical tests exhibiting nonuniform Differential Test Functioning (DTF) and a cut score that corresponds to a trait level of 0.0.

In selection contexts, it seems that the use of procedures that are sensitive to all forms of DIF is important, whereas the ability to accurately identify the type of DIF is of lesser importance as all forms have the potential to produce undesirable outcomes. Additionally, as was shown, factors that are not captured in a statistical test for differential functioning - such as the effect of cutoff scores and how test scores are used (e.g., top-down selection, banding, etc.), can result in a one-sided advantage for one group - even when TCCs cross.

With this in mind, future studies should look beyond hypothesis testing. That is, the literature has consistently shown that many DIF detection procedures are sensitive to small differences in item parameters that have no practical effect. This suggests that research should begin to investigate other considerations that could advise conclusions regarding the presence and nature of differential functioning. For example, effect size measures that translate test score differences directly into the outcome of interest, such as the one outlined by Stark et al. (2004), should be more widely used by researchers and practitioners to ascertain the consequences of observed differential functioning. The use

of such tools would help foster greater understanding of how differences in test functioning actually affect examinee outcomes. Furthermore, as suggested by Finch and French (2008), graphical representations such as the TCCs presented in Figure 10 could be used to identify the nature of differential functioning when it is detected. Such representations could depict considerations related to context, such as cutoff scores, which would promote awareness of how policies and testing practices also affect examinee outcomes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Education Research Association.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87-96.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick, *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrica*, 35, 179-197.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*, 129-147.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

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 the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.

Drasgow, F. & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 1, 2nd ed.* (pp. 755-636). Palo Alto, CA: Consulting Psychologists Press.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Drasgow, F. & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363-373.

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology, 34*, 437-442.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295.

Finch, W. H. & French, B. F. (2007). Detection of Crossing Differential Item Functioning. A comparison of four methods. *Educational and Psychological Measurement, 67*, 565-582.

Finch, W. H. & French, B. F. (2008). Anomalous Type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68*, 742-759.

Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10*, 544-565.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory*. Kluwer · Nijhoff Publishing: Boston, MA.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery. Washington, DC: National Academy Press.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. Applied Psychological Measurement, 20, 101-125.

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 performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Testing*. Homewood, IL: Dow Jones-Irwin.

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology, 3*, 131-150.

Jensen, A. R. (1980). Uses of sibling data in educational and psychological research. *American Educational Research Journal, 17*, 153-170.

Jiang, H. & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291-322.

Kirisci, L., Hsu, T.-C. & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162.

Li, H.-H. & Stout, W. (1996). A new procedure for detection of crossing Differential Item Functioning. *Psychometrika, 61*, 647-677.

Linn, R. L. (1982). Admissions testing on trial. *American Psychologist, 37*, 279-291.

Lord, F. M. (1980). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters upon DIF detection using the free-baseline likelihood ratio test. *Applied Psychological Measurement, 33*, 251-265.

Mazor, K. M., Kanjee, A. & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*, 131-144.

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. (2004b). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11*, 60-72.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-118.

Millsap. R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Mosteller, F. (1948). A k-sample slippage test for an extreme population. *The Annals of Mathematical Statistics, 19*, 58-65.

Narayanan, P. & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Raju, N. S. (1988). The area between two item response functions. *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. (1995). DFITD4: A program for the detection of differential test functioning [Computer program]. Charlotte, NC: University of North Carolina Charlotte.

Raju, N. S. & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis* (pp. 156-188). San Francisco, CA: Jossey-Bass.

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.

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

Reise, S. P., & Waller, N. G (2002). Item response theory for dichotomous assessment data. In F. Drasgow & N. Schmitt (Eds.), *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis* (pp. 88-122). San Francisco, CA: Jossey-Bass.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552-566.

Rogers, H. J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Journal of Applied Psychology, 17,* 105-116.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929-954.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56,* 302-318.

Shealy, R. & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects bias/DTF as well as item bias/DIF. *Psychometrika, 59*, 159-194.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Society for Industrial/Organizational Psychologists, (2003). *Principles for the Validation and Use of Personnel Selection Procedures*. Bowling Green, OH: Society for Industrial/Organizational Psychologists.

SPSS Inc. (2008). SPSS: Statistical Package for the Social Sciences [Computer program]. Chicago, IL: SPSS Inc.

Stark, S. (2000). 3PLGEN: A computer program for dichotomous data generation [Computer program]. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Stark, S., Chernyshenko, O.S., Chan, K.-Y., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*, 943-953.

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.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Stout, W. (1999). SIBTEST: Simultaneous Item Bias Test [Computer software]. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Stout, W. (1999b). CSIBTEST: Crossing Simultaneous Item Bias Test [Computer software]. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

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

Thissen, D. (2001). IRT-LRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Computer software]. Chapel Hill, NC: University of North Carolina.

Thissen, D. (2003). MULTILOG user's guide (Version 7) [Computer manual]. Mooresville, IN: Scientific Software.

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: Lawrence Erlbaum Associates.

Wang, W.-C. (2004). Effects of anchor item methods on differential item functioning detection within the family of Rasch models. *The Journal of Experimental Education, 72*, 221-261.

Wang, W.-C. & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498.

Zwick, R. & Erickan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*, 55-66.

Appendices

Appendix A: Glossary of Acronyms and Important Study Terms

ANOVA - analysis of variance.

CFA - confirmatory factor analysis.

Crossing DIF - DIF that results in neither group being favored consistently across the trait range.

CSIBTEST - crossing simultaneous item bias test.

df - degrees of freedom.

DIF - differential item functioning - when examinees from different groups have unequal item scores after conditioning on the primary trait a test is designed to measure.

DTF - differential test functioning - when examinees from different groups have unequal test scores after conditioning on the primary trait a test is designed to measure.

Functionally uniform DIF - nonuniform DIF that has the effect of unidirectional DIF due to the high trait level at which the group IRFs cross.

Impact - a difference in the trait distributions of two groups after the scores have been placed on a common metric.

IRF - item response function.

IRT - item response theory.

IRT-LR - item response theory likelihood ratio test.

LOGREG - logistic regression.

MH - Mantel-Haenszel.

Mixed DIF - DIF that results from group differences in both item discrimination and item difficulty, can result in either crossing or unidirectional DIF.

MML - marginal maximum likelihood estimation.

Nonuniform DIF - crossing DIF that results from group differences in item discrimination.

Power - the frequency with which DIF is identified in an item known to exhibit DIF.

SIBTEST - simultaneous item bias test.

TCC - test characteristic curve.

Type I error - the frequency with which DIF is identified in an item known to not exhibit DIF.

Type III error - the frequency with which the wrong DIF type is identified in an item known to exhibit DIF.

Unidirectional DIF - DIF that results in one group being favored consistently across the trait range.

Uniform DIF - unidirectional DIF that results from group differences in item difficulty.

Appendix B: Selecting DIF Item Types to Reduce DTF

In order to minimize DTF in this study, it was decided to design DIF items such that they do *not* all benefit the reference group. In addition, it was deemed necessary that any changes made to the DIF items in order to minimize DTF be the same across test length conditions in order to avoid introducing confounds related to test length and DIF item modifications. Thus, it was decided to alter individual DIF items (which are nested within the tests in groups of five) based on the extent to which they favored one group over another; to determine this, the signed area equation developed by Raju (1988) was used.

Unlike the unsigned area equation, the signed area equation: $(1 - c)(b_f - b_r)$, takes into account the fact that crossing DIF essentially "cancels", and provides an estimate of the area between the IRFs that is *not* negated by changes in group advantage across the trait continuum. Therefore, the resulting value provides an index of the degree to which an item exclusively favors one group.

Given the DIF prototypes to be included in this study, only the uniform, crossing mixed, and unidirectional mixed DIF items actually possessed a non-zero signed area value (Table B1). As can be seen, changing the advantage provided by any two of these items would have resulted in a total signed area for the DIF items that, in sum, favored the focal group; thus, it was decided to focus on the possible effects of altering either the uniform or unidirectional mixed DIF items, which had equal signed area values. In addition, although it possesses a zero signed area, the functionally uniform DIF item, given its unusual nature, could be considered to produce unidirectional DIF and was also investigated as a possible candidate for change. Therefore, the DTF generated by

changing either the uniform or unidirectional DIF item types to favor the focal group, with and without altering the functionally uniform DIF item type as well, was investigated using the computer program DFITD4 (Raju, 1995).

Table B1.

Signed Area Values for Study DIF Items by DIF Magnitude Conditions

Item	DIF type	.4 DIF	.8 DIF	Item	DIF type	.4 DIF	.8 DIF
11	Uniform	.4	.8	26	Uniform	.4	.8
12	Nonuniform	.0	.0	27	Nonuniform	.0	.0
13	Functionally uniform	.0	.0	28	Functionally uniform	.0	.0
14	Crossing mixed	.2	.4	29	Crossing mixed	.2	.4
15	Unidirectional mixed	.4	.8	30	Unidirectional mixed	.4	.8

Table B2 summaries the results of the DTF analyses. It was found that, generally, the least DTF was generated when the unidirectional mixed and functionally uniform DIF item types favored the focal group and the remaining DIF item types favored the reference group. Hence, the unidirectional mixed and functionally uniform DIF item types were designed to favor the focal group in the minimal DTF conditions.

Table B2.

Results of DTF Analyses by Test Length and DIF Magnitude

DIF item type(s) that favor focal group	Test length	DIF magnitude	DTF statistic
Unidirectional	15	.4	.02
	15	.8	.09
	30	.4	.15
	30	.8	.25
Unidirectional and functionally uniform	15	.4	.02
	15	.8	.04
	30	.4	.03
	30	.8	.28
Uniform	15	.4	.00
	15	.8	.13
	30	.4	.02
	30	.8	.45
Uniform and functionally uniform	15	.4	.01
	15	.8	.05
	30	.4	.07
	30	.8	.46