# A REEXAMINATION OF LORD'S WALD TEST FOR DIFFERENTIAL ITEM FUNCTIONING USING ITEM RESPONSE THEORY AND MODERN ERROR ESTIMATION

Michelle M. Langer

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the Department of Psychology (Quantitative).

Chapel Hill
2008

Approved by

Advisor: David Thissen, Ph.D.

Reader: Patrick Curran, Ph.D.

Reader: Melanie Green, Ph.D.

Reader: Robert MacCallum, Ph.D.

Reader: Abigail Panter, Ph.D.

## ABSTRACT

Michelle M. Langer: A Reexamination of Lord's Wald Test for Differential Item Functioning using Item Response Theory and Modern Error Estimation
(Under the direction of David Thissen, Ph.D.)

The detection of differential item functioning (DIF) is an essential step in increasing the validity of a test for all groups. The item response theory (IRT) model comparison approach has been shown to be the most flexible and powerful method for DIF detection; however, it is computationally-intensive, requiring many model-refittings. The Wald test, originally employed by Lord for DIF detection, is asymptotically equivalent to this approach and requires only one model fitting. In this research, the Wald test for DIF detection was improved from Lord's original conception through modern error estimation, concurrent calibration, maximum marginal likelihood item parameter estimation, conditional DIF tests, and extensions to commonly used IRT models as well as multiple groups.

This research examined the Type I error and power of the Wald test by varying the magnitude of DIF, the mean difference between groups, test length, and the sample size per group. Data were simulated under the graded response model and the three-parameter logistic (3PL) model. An additional simulation study compared the IRT model comparison approach to the Wald test under the two-parameter logistic model. The results indicated that the Wald test performs well detecting DIF. The performance improves with larger sample sizes, greater magnitudes of DIF, greater test lengths, and the random assignment estimation procedure. The use of larger sample sizes and greater test lengths is most critical for situations employing the 3PL model. The Wald test also performs well compared to the IRT model comparison approach, although the results of the two methods should converge asymptotically.

This research also demonstrated the flexibility of the Wald test through its straightforward extension to multiple groups. An example was used to demonstrate the effectiveness of the Wald test and compare it to the IRT model comparison approach. The Wald test was able to accurately identify the source of DIF. However, the IRT model comparison approach appeared more powerful but confounded the results of the DIF tests, due to combining groups. Several considerations for designing a DIF detection framework given multiple groups were outlined, particularly the superiority of the Wald test when given unequal sample sizes.

# ACKNOWLEDGMENTS

I could not have completed this dissertation without the contributions of many. First and foremost, I would like to thank Dr. David Thissen for his consistent guidance and support for this project, and for my graduate school career in general. His endless creativity, insight, and superhero-like abilities to provide immediate and thorough feedback were invaluable. I greatly appreciate his generous support, including research assistantships, travel funding, and introductions to others in our field. I am forever grateful to have had such mentoring and the opportunity to learn from one of the great minds in psychometrics.

Thank you to my committee members: Dr. Patrick Curran, Dr. Melanie Green, Dr. Robert MacCallum, and Dr. Abigail Panter. I am indebted to them for their time, the unique perspective each brought to our discussions, and their thoughtful comments.

Thank you to Li Cai, whose statistical expertise and ingenuity inspired my work. I am fortunate to have shared my graduate school experience with such a creative fellow student who was always willing to take the time to explain any advanced statistical technique with words I could understand.

Thank you to the loves of my life, my friends: Kelly, Mo, Mary, Kira, Kacy, Cheryl, Erin, Anne, Mary Thomas, Jessica, Margaret, Alex, Amanda, Lauren, Jaime, Dorothy, Heather, and Melissa. They kept me sane, indulged my complaints, and were there for me every step of the way. I not only would not have made it through graduate school without them, but I would not be who I am today. I am truly lucky to have such support in my life.

Finally, deep appreciation goes to my family. They have always believed in me, and I could not have made it to this point without their generous support, love, and encouragement. I could never have hoped or expected to be so blessed. Thank you from the bottom of my heart.

## TABLE OF CONTENTS

Page

# LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

# Introduction

Differential item functioning (DIF) refers to a psychometric difference in how an item functions across groups. An item that performs differently must necessarily be less valid, in some senses, for at least one of the groups. As a result, an effort to detect and eliminate DIF from tests seeks to increase the validity of the test for all groups.

A variety of methods for detecting DIF have been developed. These methods include the Mantel-Haenszel procedure (MH; Holland & Thayer, 1988; Mantel, 1963; Mantel & Haenszel, 1959), logistic regression (Swaminathan & Rogers, 1990), proportion difference measures (Dorans & Kulick, 1983, 1986; Dorans & Schmitt, 1991), SIBTEST (Shealy & Stout, 1993), the test of *b* difference (Lord, 1977, 1980), an item drift method (Bock, Muraki, & Pfeiffengerger, 1988; Muraki & Englehard, 1989), Lord's Wald test[1] (Lord, 1980), empirical sampling distributions for DIF indices (Shepard, Camilli, & Williams, 1984), model comparison measures (Judd & McClelland, 1989; Thissen, Steinberg, & Wainer, 1993), DFIT (Raju, van der Linden, & Fleer, 1995), and simple area indices (Hambleton & Rogers, 1989; Rudner, 1977; Rudner, Getson, & Knight, 1980). Of these methods, the IRT model comparison approach to DIF detection, using the likelihood ratio test, has been shown to be the most flexible and powerful (Thissen, Steinberg, & Wainer, 1993; Teresi, Kleinman, & Ocepek-Welikson, 2000; Wainer, 1995). However, this approach can be computationally-intensive, involving many model re-fittings.

[1] Lord referred to his test as the chi-square test; however, the general procedure is more commonly called the Wald test.

Lord’s Wald test (1977, 1980) for DIF detection is asymptotically equivalent to the likelihood ratio test (Thissen, Steinberg, & Wainer, 1993) and less computationally-intensive. However, the Wald test has not been frequently employed as a DIF detection method. Lord’s original implementation of the Wald test involved a different conception of the IRT model with less than desirable standard error estimation. However, given approximate normality and good estimates of the error covariance matrix, the Wald test could be improved upon as a more practical alternative to the likelihood ratio test. Good estimates of the item parameter error covariance matrix are an additional requirement for the Wald test in comparison to the likelihood ratio test, but modern error estimation could be implemented to ease computation.

The goal of this project is to reexamine Lord’s Wald test (1977, 1980) for DIF detection using modern error estimation. First, DIF detection methods and procedures will be discussed. Second, a simulation study to detect DIF using the Wald test with modern error estimation in two groups will be outlined. Third, simulation results will be presented and discussed. Fourth, the extension of the Wald test for DIF detection with modern error estimation to more than two groups will be examined with an example.

Given good error estimates and an effective implementation in the two-group case, the Wald test for DIF detection can be straightforwardly extended to multiple groups. Current DIF detection procedures have been developed exclusively for the two-group case. All of these approaches involve comparisons between a reference group and a focal group. However, it is often desirable to assess DIF for several focal groups. Numerous focal groups have been identified as important candidates for DIF analysis for measures of educational achievement: Asian Americans, African Americans, Hispanic groups, Native Americans, women, and examinees with disabilities (Zieky, 1993). In discriminating among groups, Linn

(1993) further suggested that Hispanic groups be subdivided to distinguish among Puerto Ricans, Mexican Americans, and Cubans. The practical need for considering multiple focal groups is demonstrated by numerous studies examining DIF among multiple ethnic groups (Schmitt, 1988; Schmitt & Dorans, 1990; Zwick & Ercikan, 1989) and multiple languages of administration (Angoff & Sharon, 1974; Ellis & Kimmel, 1992). Given the prevalence of multiple group DIF assessments, DIF studies would benefit from the availability of statistical procedures that test for DIF simultaneously across multiple groups.

## Differential Item Functioning Detection Methods

Methods of DIF detection can generally be categorized within two broad approaches: observed-score approaches and latent variable/item response theory (IRT) approaches. Observed-score approaches include the Mantel-Haenszel procedure (MH; Holland & Thayer, 1988; Mantel, 1963; Mantel & Haenszel, 1959), logistic regression (Swaminathan & Rogers, 1990), proportion difference measures (Dorans & Kulick, 1983, 1986; Dorans & Schmitt, 1991), and SIBTEST (Shealy & Stout, 1993). Although applications of these observed-score approaches require few assumptions and are relatively easy to implement, their results may be sample-specific and, thus, inadequate for ensuring measurement invariance (Budgell, Raju, Quartetti, 1995; Hulin, Drasgow, Parsons, 1983). Given that the IRT assumptions hold, results from an IRT approach theoretically generalize beyond the sample being studied to the intended population.

Latent variable/item response theory (IRT) approaches propose a latent trait, usually denoted $\theta$, that underlies the item responses. These approaches offer an advantage over the observed score approaches due to their ability to unconfound group mean differences on the latent trait and DIF. In practical applications using a latent variable/IRT approach, one of the most commonly used models is the three-parameter logistic (3PL) model, frequently used for responses to multiple-choice items in educational research, which assumes that the probability that an examinee with trait value $\theta$ will respond correctly to item $i$ is

$$P_i(\theta) = g_i + \frac{1 - g_i}{1 + \exp(-1.7a_i\theta + c_i)},$$

where $a_i$ is the item discrimination, or slope, $c_i$ is the item intercept, and $g_i$ is the lower asymptote. This slope-intercept form of the 3PL model is used for computational ease; however, the literature often refers to a location parameter, the *b* parameter, which is *–c/a*. The item parameters and trait values are estimated from examinees' responses to a set of items. Predictions based on the estimated parameters and traits are then compared to observed data to examine whether the model actually fits the data (Yen, 1986).

An alternative model often used in psychological research is Samejima’s (1969, 1997) graded response model (GRM), a formulation that permits estimation of multiple $c_{ij}$ parameters[2] per item (*j* from 1 to *m*-1) associated with *m* response categories (e.g., items with the response scale “Strongly Disagree”, “Disagree”, “Neutral”, “Agree”, and “Strongly Agree”). The formula, also in slope-intercept form, for a GRM trace line is

$$P(x_i = j|\theta) = \frac{1}{1 + \exp(-a_i\theta + c_{ij})} - \frac{1}{1 + \exp(-a_i\theta + c_{ij+1})]}, \qquad (2)$$

which states that the probability of responding in category *j* is the difference between the probability of responding in category *j* or higher and the probability of responding in category *j*+1 or higher.

DIF detection methods based on IRT include the test of *b* difference (Lord, 1977, 1980), an item drift method (Bock, Muraki, & Pfeiffengerger, 1988; Muraki & Englehard, 1989),

[2] Analogous to the 3PL model, the literature often refers to location parameters, $b_{ij}$s, which are - $c_{ij}/a$.

Lord’s Wald test (1977, 1980), empirical sampling distributions for DIF indices (Shepard, Camilli, & Williams, 1984), model comparison measures (Judd & McClelland, 1989; Thissen, Steinberg, & Wainer, 1993), DFIT (Raju, van der Linden, & Fleer, 1995), and simple area indices (Hambleton & Rogers, 1989; Rudner, 1977; Rudner, Getson, & Knight, 1980). Of these methods, two are good candidates for situations involving more than two groups: model comparison measures and Lord’s Wald test.

### *IRT Model Comparison Approach*

The IRT model comparison approach to DIF detection has been shown to be more powerful than other methods (Thissen, Steinberg, & Wainer, 1993; Teresi, Kleinman, & Ocepek-Welikson, 2000; Wainer, 1995) and is implemented in available software (IRTLRDIF; Thissen, 2001). Under reasonable conditions, model-based likelihood ratio tests are closely related to the most powerful test given by the Neyman-Pearson (1928) lemma. This optimality of power, decreasing the chances of accepting the null hypothesis of no DIF, lends credibility to this type of test as one of the most powerful DIF detection tools.

Tests of statistical significance using the IRT model comparison approach (Thissen, Steinberg, & Wainer, 1993) always involve the comparison of two models, a compact model and an augmented model (Judd & McClelland, 1989). The compact model is hierarchically nested within the augmented model, which includes all of the parameters of the compact model as well as additional parameters. In DIF detection, the compact model involves the likelihood of the parameter estimates for a given item *i*, constrained to be DIF-free, compared to the likelihood of the augmented model that allows for additional parameters representing differences between the item *i* parameters for the reference and focal groups. The goal of this approach is to test whether these additional parameters in the augmented model are significantly different from zero. The form of these likelihood ratio tests is always

$$G^2(d.f.) = 2\log\left[\frac{Likelihood[A]}{Likelihood[C]}\right],$$

where Likelihood [·] represents likelihood of the data given maximum likelihood estimates of the parameters of the model, and *d.f.* is the difference between the number of parameters in the augmented model and the number of parameters in the compact model. Under very general assumptions, the value of $G^2(d.f.)$ is distributed as $\chi^2(d.f.)$ under the null hypothesis (Rao, 1973). Significant $G^2(d.f.)$ values result in a rejection of the null hypothesis of no DIF, and thus the compact model. The test of the significance of DIF is on *k* degrees of freedom, where *k* is the number of item parameters differing between the reference and focal groups.

The main disadvantage of the IRT model comparison approach is that the number of models fitted increases with each additional set of hypotheses under consideration. The procedure involves fitting the model twice per hypothesis, once for the compact model and once for the augmented model. With more than two groups, the number of models required to be fitted results in an extravagant number of fittings and computational time. Lord’s Wald test (1977, 1980) is asymptotically equivalent and can be performed with just one fitting, making it more easily extended to multiple groups.

### *Lord’s Wald Test*

Lord’s Wald test (1977, 1980) for DIF detection compares vectors of IRT item parameters between groups. If, for a given item, the vectors of its parameters differ significantly between groups, then the trace lines differ across groups, and thus the item functions differentially for the groups studied. For two groups, Lord first proposed a test to evaluate the significance of the DIF for the location parameters only, using:

$$d_i = \frac{\hat{b}_{F_i} - \hat{b}_{R_i}}{\sqrt{Var(\hat{b}_{F_i}) + Var(\hat{b}_{R_i})}},$$

in which $\hat{b}_{g_i}$ is the maximum likelihood estimate of the parameter $b_i$ in group $g$ and Var ($\hat{b}_{g_i}$) is the corresponding estimate of the sampling variance of $\hat{b}_{g_i}$.

Lord observed that probability statements could be made by referring $d_i$ to tables of the standard normal distribution. Lord also extended this test for differences between the discrimination parameters, developing the more general test of the joint difference between [$a_i$, $b_i$] for the two groups,

$$\chi_i^2 = \mathbf{v}_i{}'\mathbf{\Sigma}_i^{-1}\mathbf{v}_i,$$

where $\mathbf{v}_i{}'$ is [ $\hat{a}_{F_i}$ - $\hat{a}_{R_i}$ , $\hat{b}_{F_i}$ - $\hat{b}_{R_i}$ ], $\mathbf{\Sigma}_i$ is the estimate of the sampling variance-covariance matrix of the differences between the item parameter estimates, and $\chi_i^2$ is distributed on 2 *d.f.* for large samples. Alternatively, an equivalent procedure tests the difference between intercept parameters, $c_i$'s, instead of location parameters, $b_i$'s[3]. For this test, $\mathbf{v}_i{}'$ is [ $\hat{a}_{F_i}$ - $\hat{a}_{R_i}$ , $\hat{c}_{F_i}$ - $\hat{c}_{R_i}$ ].

For the 3PL model, Lord did not propose a test for the differences between the $g_i$ parameters; he specified that they should be constrained to be equal for all groups. However, the test of slopes and intercepts can be extended to include a test for the differences between the $g_i$ parameters. This test of the joint difference between [$a_i$, $c_i$, $g_i$] for the two groups

[3] Reparameterization to slope-intercept form improves the stability of parameter estimation by working in a less correlated space.

follows the same general equation as above, allowing $\mathbf{v}_i'$ to be [$\hat{a}_{F_i}$ - $\hat{a}_{R_i}$, $\hat{c}_{F_i}$ - $\hat{c}_{R_i}$, $\hat{g}_{F_i}$ - $\hat{g}_{R_i}$ ][4]. $\mathbf{\Sigma}_i$ is the estimate of the sampling variance-covariance matrix of the differences between the three item parameter estimates, and $\chi_i^2$ is distributed on 3 *d.f.* for large samples.

The Wald test can also be extended to include conditional DIF tests. For the graded model, the unconditional test of DIF in the *a* parameters is

$$Z_{a_i}^2 = \frac{(\hat{a}_{F_i} - \hat{a}_{R_i})^2}{\sigma_{\hat{a}_i}^2},$$

where $\sigma_{\hat{a}_i}^2$ is the variance of the difference between the *a* parameter estimates, and $Z_{a_i}^2$ is chi-square distributed on 1 *d.f.* for large samples. Conditioning on the equal *a* parameter estimates, the conditional test of DIF in the *c* parameters may be computed as the difference between the overall chi-square test and the unconditional test of DIF in the *a* parameters,

$$Z_{c_i|a_i}^2 = \chi_i^2 - Z_{a_i}^2,$$

where $Z_{c_i|a_i}^2$ is chi-square distributed on (*j* – 1) *d.f.* for *j* categories in large samples. Because the commonly-used IRT location parameter *b* equals -*c/a*, when conditioned on equal *a* parameters, this is also a test of *b*-DIF.

For the 3PL model, the unconditional test of the DIF in the *g* parameters is

$$Z_{g_i}^2 = \frac{(\hat{g}_{F_i} - \hat{g}_{R_i})^2}{\sigma_{\hat{g}_i}^2},$$

where $\sigma_{\hat{g}_i}^2$ is the variance of the difference between the *g* parameter estimates, and $Z_{g_i}^2$ is chi-square distributed on 1 *d.f.* for large samples. Conditioning on the *g* parameter estimates, the conditional test of DIF in the *a* parameters is

[4] Computationally, the logit of $g_i$ is used. This transformation improves the normality of the likelihood and removes problematic boundary conditions.

$$Z^2_{a_i|g_i} = \frac{u_i^2}{\sigma^2_{\hat{a}_i|\hat{g}_i}},$$

where

$$u_i = [\hat{a}_{F_i} - \hat{a}_{R_i}] - \frac{\sigma_{\hat{a}\hat{g}_i}}{\sigma^2_{\hat{g}_i}}[\hat{g}_{F_i} - \hat{g}_{R_i}],$$

$$\sigma^2_{\hat{a}_i|\hat{g}_i} = \sigma^2_{\hat{a}_i} - \frac{\sigma^2_{\hat{a}\hat{g}_i}}{\sigma^2_{\hat{g}_i}},$$

$\sigma_{\hat{a}\hat{g}_i}$ is the covariance matrix of the difference between the *a* parameter estimates and the difference between the *g* parameter estimates, $\sigma^2_{\hat{a}_i}$ is the variance of the difference between the *a* parameter estimates, $\sigma^2_{\hat{g}_i}$ is the variance of the difference between the *g* parameter estimates, and $Z^2_{a_i|g_i}$ is chi-square distributed on 1 *d.f.* for large samples (Bock, 1985). This conditional test of the *a* parameters is similar to univariate regression, in which the regression weight is the covariance of the difference between the *a* parameter estimates and the difference between the *g* parameter estimates divided by the variance of the difference between the *a* parameter estimates.

Conditioning on both the *a* and *g* parameter estimates, the conditional test of DIF in the *c* parameters is simply the difference between the overall chi-square test and the sum of the unconditional test of DIF in the *g* parameters and the conditional test of DIF in the *a* parameters,

$$Z^2_{c_i|a_ig_i} = \chi^2_i - [Z^2_{g_i} + Z^2_{a_i|g_i}],$$

where $Z^2_{c_i|a_ig_i}$ is chi-square distributed on 1 *d.f.* for large samples. (This is also a test for location differences, commonly called *b*-DIF.)

### *Linking Procedures*

A necessary preliminary step of DIF analyses is to place groups of examinees on a common scale of measurement. To link groups together, Lord used the Stocking-Lord approach (1983), an ad hoc method, to put estimates on the same scale. However, DIF detection using the Wald test could be improved by using an IRT-based linking procedure, such as concurrent calibration (Kolen & Brennan, 2004).

The main benefit of using an IRT-based linking procedure is that it is then possible to develop an estimate of an examinee's ability that is independent of the set of items to which the examinee responds. The observed-score methods cannot accomplish this compensation for intended and unintended differences in item "difficulty" and sample ability (Skaggs & Lissitz, 1986). Furthermore, an IRT-based linking approach provides conversions that are independent of the group or groups used to obtain them. Another benefit of using an IRT-based linking approach is the accuracy of the linking along the entire score scale.

### *Anchoring Situations*

Several different situations, in various contexts, exist for which DIF detection methods can be implemented. DIF methods have been extended beyond their original use of simply separating group and item differences, separating so-called "impact" from DIF (Angoff, 1993). The first situation involves the case of two randomized groups with the same population mean. This circumstance does not involve the issue of determining an anchor with which to link the two groups; the statistical theory is straightforward. This case is exemplified by Steinberg's (1994, 2001) research on context and serial-order effects in personality measurement, using a randomized-assignment experiment. The second, more traditional, situation allows for nonrandom groups and, thus, differing population means.

This circumstance necessarily involves selecting an anchor set of items to link the two groups.  Once anchor items are chosen, the statistical theory follows to use the anchor in effectively linking the groups together and testing group differences on the candidate items; this method is employed in the current version of IRTLRDIF (Thissen, 2001). This case is most often present in the educational context, when DIF detection methods are used to test differences between nonrandom groups such as gender or ethnicity.

Given nonrandom groups as in the second situation, the third situation arises in the absence of a pre-specified linking anchor test. The IRTLRDIF (2001) method uses an *all-other* procedure in which each item, in turn, is treated as a candidate item while the other items are treated as anchor items linking the groups together. Logistic regression (Swaminathan & Rogers, 1990) and Mantel-Haenszel (Holland & Thayer, 1988; Mantel, 1963; Mantel & Haenszel, 1959) methods use similar tactics, including candidate items in the score that is used as the conditioning variable.  A proposed Wald test-based alternative will employ a two-stage estimation procedure. The first stage constrains the item parameters to be the same in both groups to estimate the population mean and standard deviation of the focal group relative to the reference group, assuming no DIF in any items. The second stage then treats that estimated population mean and standard deviation as fixed, and allows the item parameters to differ for the detection of DIF.

### *Estimation Methods*

In Lord's (1977, 1980) original implementation of the Wald test for DIF detection, standard error estimates obtained using joint maximum likelihood (JML; as implemented in LOGIST [Wood, Wingersky, & Lord, 1976]) were computed with θ considered a fixed variable (in the statistical sense). These standard error estimates are not accurate for the modern conception of θ as a latent random variable. In a simulation study, McLaughlin and Drasgow (1987) observed that this inaccuracy of the standard errors resulted in overall proportions of significant DIF (in the null case) as high as 11 times the nominal α level. However, item parameter estimation procedures have greatly improved since Lord’s time, and with those improvements have come better estimates of the item parameter error covariance matrix.

For tests with data from an adequate number of examinees, maximum marginal likelihood (MML) item parameter estimation performs quite well. As an example of the superiority of MML estimation to JML estimation, Drasgow (1989) examined item parameter bias for the two-parameter logistic (2PL) model with tests ranging in length from 5 to 25 items and samples ranging in size from 200 to 1000. Average biases in both *a* and *b* parameter estimates were consistently much larger using JML estimation than for MML estimation. Thus, another way the Wald test for DIF detection can be improved is to use MML estimation rather than JML estimation.

Arguably the most fundamental problem with Lord’s Wald test (1977, 1980) for DIF detection was the less than desirable standard error estimates. This problem hopefully can be circumvented using the Supplemented EM (SEM) algorithm (Meng & Rubin, 1991) to obtain the item parameter error covariance matrix. This algorithm provides a convenient computational procedure for estimating the information matrix for item parameters (Cai, in press) and subsequently can provide more accurate standard errors for the estimated item parameters than the methods implemented in packages such as Multilog (Thissen, Chen, & Bock, 2003).

# Method

### *Empirical Parameter Distributions*

To answer the questions posed in the introduction, a simulation study was conducted. Data were simulated using the 3PL model and the GRM. For the 3PL model, empirical inspiration for the parameter distributions was based on the 1998 National Assessment of Educational Progress (NAEP) Reading assessment, representing an educational context. Tables of item parameters are reported in *The NAEP 1998 Technical Report* (Allen, Donoghue, & Schoeps, 2001).

The 1998 NAEP Reading assessment 3PL item parameters were compiled, and empirical parameter distributions were examined. The distribution of *a* parameters appears to be symmetric, suggesting that it could be approximated by a normal distribution. The mean is 1.03, and the standard deviation is 0.37. Thus, similarly distributed *a* parameters could be drawn from a normal (1, $0.3^2$) distribution. The distribution of the *b* parameters also appears roughly symmetric, suggesting that it could be approximated by a normal distribution. The mean is -0.13, and the standard deviation is 0.85. Thus, similarly distributed *b* parameters could be drawn from a normal (0.0, $0.8^2$) distribution, representing items centered at an average difficulty level. The *g* parameters are also distributed symmetrically and appear amenable to a normal approximation. The mean is 0.26, and the standard deviation is 0.06. Thus, similarly distributed *g* parameters could be drawn from a normal (0.25, $0.05^2$) distribution, corresponding to items with four response alternatives and subsequently a ¼ chance of getting an item correct due to guessing.

For the GRM, empirical inspiration for the parameter distributions was based on psychological scales. A list of item parameters from 15 tests has been compiled (see Hill, 2004). Empirical histograms indicate that the distribution of the *a* parameters appears to be reasonably symmetric and suitable to be approximated by a normal (1.7, $.3^2$) distribution. Similarly, the *b* parameters are reasonably symmetric and suitable to be approximated by a normal distribution.

To generate item parameters for the GRM, $d_{\mathrm{j}}$ values, the difference between the $b_j$ parameters, are employed, in addition to $b_1$, the leftmost threshold. The $d_{\mathrm{j}}$ value is the distance between $b_j$ and $b_{j+1}$. For each item, there are one fewer $d_j$ values than there are $b_j$ parameters. The empirical histograms of the $d_j$ parameters for the 15 tests indicate that the $d_j$ parameters are roughly symmetric, suggesting that they could be approximated by a normal (1.0, $.2^2$) distribution for scales with 5, 7, or 9 response categories. Because the *b* parameters are suitable to be approximated by a normal distribution, drawing $b_1$ from a normal (-1.5, $.5^2$) should result in a $b_j$ distribution centered around 0 for 5 response categories[5]. If the $d_j$s are at their expected values, then the subsequent $b_j$ parameters would be -.5, .5, and 1.5.

*Design*

To reexamine Lord's Wald test (1977, 1980), a simulation study was conducted. All data were simulated for two groups using the 3PL model or the GRM. Item parameter estimation used MML, with the SEM algorithm to compute the parameter error covariance matrix and concurrent estimation of the population parameters for the focal group. DIF was detected using the extensions of the Wald test. Data were simulated under the random assignment and two-stage estimation anchoring situations. The number of replications was 100.

The number of items was $n$ = 5, 20, and 40, with $N$ = 250 and 1000 simulees per group. The number of response categories for items following the GRM was 5[6]. Mean differences

[5] Only 5 response categories are being considered in this simulation.

[6] Additional numbers of response categories may be considered in future investigations.

between groups were 0 and .6 standard deviations (with both populations having the same standard deviation)[7]. For the mean difference of .6 standard deviations, the focal group's population mean was -.6, and the reference group's population mean was 0.

In manipulating the amount of DIF simulated, several factors were varied. First, the magnitude of DIF was varied in both *a* and *b* item parameters. Items simulated with DIF had DIF present in both *a* and *b* parameters. For the GRM, the focal group's *a* parameters for the DIF items were simulated to be the magnitude of the *a* parameters of those items in the reference group multiplied by a factor of either 1.25 or 0.875. For the 3PL model, the focal group's *a* parameters for the DIF items were simulated to be either twice or one half the magnitude of the *a* parameters of those items in the reference group. Additionally, for the GRM, the focal group's *b* parameters for the DIF items were simulated to be either .1 or .2 greater than those items in the reference group. For the 3PL, the focal group's *b* parameters for the DIF items were simulated to be either .4 or .8 greater than those items in the reference group[8]. The proportion of items with simulated DIF was 0 and .2. As a result of all varied factors, the total number of cells in the proposed simulation was 180; refer to Table 1, a tabular description of the simulation design.

*Evaluation Criteria*

For each cell, empirical Type I error and the power of DIF detection were calculated. For each set of replications, the proportion of times a DIF-free item was mistakenly identified as a DIF item provides an estimate of Type I error. On the other hand, the proportion of times a DIF item was correctly identified provides an estimate of power. These

[7] This does not apply to the random assignment condition.

[8] The factors for simulating DIF were smaller for data simulated using the GRM versus the 3PL model due to the greater number of parameters being estimated. These values led to moderate amounts of DIF to effectively reexamine Lord's Wald test.

proportions were further averaged across DIF-free and DIF items, respectively, to form the average Type I error and the average power.

*IRTLRDIF Comparison*

The power of the Wald test was also compared, in a small separate study, to that of the IRT model comparison approach, using IRTLRDIF. Little comparable research on IRTLRDIF currently exists in the literature, with the exception of an unpublished simulation study by Brian Habing, described in a presentation at the annual meeting of the National Council on Measurement in Education in 2001. Habing used the IRTLRDIF approach (and several other methods) to evaluate DIF detection for conditions with *a*-DIF or *b*-DIF, and to estimate Type I error. An additional 36 simulation cells using the Wald test were run to mimic the simulation conditions used by Habing. This simulation study also allows us to examine the performance of the Wald test in detecting *a*-DIF and *b*-DIF separately when only one occurs.

The simulation conditions used by Habing and replicated in this simulation study involved 20-item tests, for which only one item was the candidate item. This was the only item simulated to have DIF or not, and evaluated for DIF. As a result, the remaining 19 items functioned as a DIF-free anchor set. Items were simulated under the 2PL model, which is the 3PL model with the *g* parameter set to zero or the GRM with only two categories. The *a* parameters for the anchor items were drawn from a lognormal distribution[9], such that

$$a = \exp(z),$$

and $z$ is drawn from a normal (0, $0.35^2$) distribution. The *b* parameters for the anchor items were drawn from a normal (0, 1) distribution. The number of replications was 400.

[9] This distribution is based on parameters published by Lord and Novick (1968).

For the 12 cells used to estimate Type I error, the candidate item's *a* parameters were fixed at 1 in both groups. The candidate item's *b* parameters were fixed at 0 or 1 in both groups. The sample size for each group was 250 or 1000 simulees. The mean difference between groups was 0, .5, and 1 standard deviation (with both populations having the same standard deviation). The means of both groups were centered around 0 so that the pair of means for each cell was effectively (0, 0), (-0.25, 0.25) or (-0.5, 0.5).

For the 12 cells with simulated *a*-DIF in the candidate item, the reference group's *a* parameter for the candidate item was fixed at 1.20072 or 1.35251, and the focal group's *a* parameter for the candidate item was fixed at 0.83283 or 0.73937, respectively. These fixed *a* parameters reflect differences in parameter magnitude of approximately .4 and .6, respectively, centered around an *a* parameter value of 1. The candidate item's *b* parameter was fixed at either 0 or 1 in both groups. The sample size was 250 simulees per group. Similar to the cells without simulated DIF, the mean difference between groups was 0, .5, and 1 standard deviation (with both populations having the same standard deviation), with the two group means centered around 0.

For the 12 cells with simulated *b*-DIF in the candidate item, the reference group's *b* parameter for the candidate item was fixed at 0.15, 0.25, 1.15 or 1.25, and the focal group's *b* parameter for the candidate item was fixed at -0.15, -0.25, 0.85, or 0.75, respectively. These fixed *b* parameters reflect differences in parameter magnitude of approximately .3 or .5, centered around a *b* parameter value of 0 or 1. The candidate item's *a* parameter was fixed at 1 in both groups. The sample size was 250 simulees per group. Analogous to all other cells for this comparison simulation study, the mean difference between groups was 0, .5, and 1 standard deviation (with both populations having the same standard deviation), with the two group means centered around 0.

## Simulation Results

For each cell, empirical Type I error and the power of DIF detection were calculated. For the cells without simulated DIF, the proportion of times an item was mistakenly identified as a DIF item provides an estimate of Type I error, alpha. For the cells with simulated DIF, the proportion of times a DIF item was correctly identified provides an estimate of power. On the other hand, the proportion of times a DIF-free item, in the presence of DIF items, was mistakenly identified as a DIF item provides an estimate of the false alarm rate.

### *Graded Response Model*

For the cells simulating data using the GRM, the estimated alpha rates are presented in Table 2[10]. Under random assignment, across all sample size and test length conditions, the estimated alpha rate falls within a 95% confidence interval of .05 as expected. The random assignment cells simulating larger sample sizes per group (1000 versus 250) exhibit an alpha rate of exactly .05[11]. Under the two-stage estimation procedure, almost all of the estimated alpha rates fall below the 95% confidence interval around .05. This underestimate of alpha is likely a result of a portion of the random DIF being absorbed into the estimated mean difference between the groups in the first stage of the estimation procedure. Alpha rates appear not to differ based on whether the mean difference between groups was simulated to be 0 or -.6.

Table 3 provides the chi-square means and variances for the no-DIF cells using the GRM. Given five response categories, the GRM estimates five parameters: one *a* parameter and

[10] Overall DIF results are presented rather than separate *a*-DIF and *b*-DIF results. None of the individual IRT parameter alpha rates were large, and the overall alpha rate provides a complete summary.

[11] To two decimal places.

four *b* parameters. As a result, we would expect the chi-square means to be approximately 5. Similarly, a chi-square distribution should possess a variance twice its mean, so we would expect the chi-square variances to be approximately 10. The pattern of chi-square means and variances in Table 3 naturally follows the same pattern as the alpha rates in Table 2. Under random assignment, the estimated chi-square means and variances are relatively close to their expected values, with the larger sample size per group conditions providing a better approximation. Under the two-stage estimation procedure, the estimated chi-square means and variances fall shy of their expected values, reflected in the lower alpha rates of Table 2 and again explained by the absorption of some random DIF into the estimated mean difference between the groups.

For the conditions simulating DIF for 20% of the items, the estimated false alarm rates are presented in Table 4[12]. Under random assignment, with the exception of one cell out of the 24, the estimated false alarm rate falls within a 95% confidence interval of .05 as expected. Nearly all of the random assignment cells simulating larger sample sizes per group (1000 versus 250) exhibit false alarm rates of .05[13]. Under the two-stage estimation procedure, almost all of the estimated false alarm rates fall short of the 95% confidence interval around .05. Similar to the underestimate of the alpha rate, these false alarm rates can likely be attributed to some of the random DIF being absorbed into the estimated mean difference between the groups in the first stage of the estimation procedure. False alarm rates appear not to differ based on whether the mean difference between groups was simulated to be 0 or -.6. There is some improvement for the cells simulating larger sample sizes per group (1000

[12] Overall DIF results are presented rather than separate *a*-DIF and *b*-DIF results. None of the individual IRT parameter false alarm rates were large, and the overall false alarm rate provides a complete summary.
[13] To two decimal places.

versus 250). These results follow the same pattern of the alpha rates, suggesting that the presence of DIF simulated in 20% of the items is not affecting the alpha/false alarm rate.

Table 5 provides the chi-square means and variances for the statistics used to estimate the false alarm rates using the GRM. As with the alpha rate estimates, we would expect the chi-square means to be approximately 5 and the chi-square variances to be approximately 10. The pattern of chi-square means and variances in Table 5 follows the same pattern as the false alarm rates in Table 4. Under random assignment, the estimated chi-square means and variances are relatively close to their expected values, with the larger sample size per group providing a better approximation. Under the two-stage estimation procedure, the larger sample size per group (1000 versus 250) also provides a better approximation. However, for the two-stage estimation cells, the estimated chi-square means and variances fall shy of their expected values, reflected in the lower false alarm rates of Table 4 and likewise explained by the absorption of some random DIF into the estimated mean difference between the groups.

The proportions of overall DIF, *a*-DIF, and *b*-DIF[14] detected at the .05 level for the cells simulating data using the GRM under random assignment are displayed in Figure 1 as an estimate of power. In general, power is most affected by the sample size per group. With 1000 simulees per group, the DIF items are detected at least 50% of the time. With 250 simulees per group, the DIF items are detected less than 50% of the time. Inflating the *a* parameters by a factor of 1.25 induces greater DIF detection as compared to reducing the *a* parameters by a multiple of .875. Shifting the *b* parameters by a factor of .2 also results in greater DIF detection as compared to shifting the *b* parameters by .1. For the sample size of

[14] *b*-DIF is actually a test of the significance of the difference between intercept (*c*) parameters conditional on equal slope (*a*) parameters. Because the DIF was generated with a shift in the *b* parameter in the slope-threshold form of the model, and because it is interpreted as *b*-DIF in practice, this test is referred to as *b*-DIF in the figures.

only 250 simulees per group, increasing the test length from 5 items to either 20 or 40 items did not greatly improve DIF detection. On the other hand, for the sample size of 1000 simulees per group, increasing the test length from 5 items to 20 items tended to increase DIF detection, although a test length of 40 items did not add much power over a test length of 20 items. Across simulation conditions for the GRM under random assignment, power was greatest, almost 100%, for the cells simulating sample sizes of 1000 per group, *a*-DIF of 1.25 and *b*-DIF of .2. On the other hand, power was minimally above a .05 chance level for the cells simulating sample sizes of 250 per group, *a*-DIF of .875 and *b*-DIF of .1.

The proportions of overall DIF, *a*-DIF, and *b*-DIF detected at the .05 level for the cells simulating data using the GRM with two-stage estimation and no simulated mean difference between the groups are displayed in Figure 2 as an estimate of power. The pattern of results is very similar to that of random assignment. However, power is generally lower for these two-stage estimation cells as compared to random assignment, particularly when the sample size is only 250 per group. This can likely be attributed to a portion of the simulated DIF being absorbed into the estimated mean difference between the groups in the first stage of the estimation procedure. This is consonant with the observation that the differences in the proportions of *b*-DIF detection between the two-stage estimation and random assignment cells tend to be greater than the differences in the proportions of *a*-DIF detection with 250 per group.

As an estimate of power, Figure 3 presents the proportions of overall DIF, *a*-DIF, and *b*-DIF detected at the .05 level for the cells simulating data using the GRM with two-stage estimation and a -.6 simulated mean difference between the groups. The pattern of power for detecting DIF is nearly identical to that for two-stage estimation with no simulated mean difference between the groups. The only noticeable difference is a slight general reduction in power for cells simulating a -.6 mean difference between the groups as compared to no mean difference. This can be explained by the additional difficulty in estimating a -.6 mean difference between the groups.

### *3PL Model*

For the simulation cells using the 3PL model, some items exhibited infinite slope estimates due to the error covariance matrix becoming singular. These items were not included in the computations for the estimated alpha rates, false alarm rates, and power. The percentages of items per cell displaying such extreme slopes are included in Appendix A. These instances occurred for cells with few items and small sample sizes. The 3PL model is difficult to estimate with a small sample size because the effective sample size for the *g* parameter includes only the subset of the sample who did not know the answer. The 3PL model is not used in practice when the sample size is only 250 or with only 5 items.  Although one would not necessarily compute power under these conditions, this simulation further demonstrates that the model should not be used in such cases. A possible option for such situations may be the use of informative priors to prevent extreme slopes.

For the cells simulating data using the 3PL model, the estimated alpha rates are presented in Table 6[15]. Across all simulation conditions for the 3PL model, the estimated alpha rate falls below a 95% confidence interval around .05. The cell simulating a sample size of 1000 per group for a test length of 40 items under random assignment should provide results closest to the nominal value; however, the estimated alpha rate is only .02, significantly less than the expected .05. These lower overall alpha rates can be attributed to near-zero alpha

[15] Overall DIF results are presented rather than separate *a*-DIF, *b*-DIF, and *g*-DIF results. None of the individual IRT parameter alpha rates were large, and the overall alpha rate provides a complete summary.

rates for detecting *g*-DIF due to the prior on the *g* parameter. The prior shrinks estimates to the same value, reducing the chance that any two *g* parameters will differ. As a result, the 3PL DIF test is very conservative because it is nominally a 3 *d.f.* test, however the *g* parameter does not exhibit a full range of variation and it is more like a 2 *d.f.* test with a 3 *d.f.* criterion.

Table 7 provides the chi-square means and variances for the no-DIF cells using the 3PL model. The 3PL model estimates three parameters: one *a* parameter, one *b* parameter, and one *g* parameter. As a result, we would expect the chi-square means to be approximately 3. Similarly, a chi-square distribution should possess a variance twice its mean, so we would expect the chi-square variances to be approximately 6. The pattern of chi-square means and variances in Table 7 follows the same pattern as the alpha rates in Table 6. Across all simulation conditions, the estimated chi-square means and variances fall shy of their expected values, reflected in the lower overall alpha rates in Table 6. Again, these underestimates are due to the prior on the *g* parameter producing near-zero alpha rates for detecting *g*-DIF. There is some improvement for the cells simulating larger sample sizes per group (1000 versus 250), as well as the cells with greater test length (40 items versus 5 items); however, such differences are small relative to the underestimation of alpha across all cells.

In hopes of better estimating alpha, a small study was conducted using the 3PL model without a prior on the *g* parameter. Without a prior on the *g* parameter, the sample sizes need to be large for effective estimation. Simulation runs were conducted for both the random assignment and the two-stage estimation procedures, with only the zero mean difference between groups simulated for the two-stage estimation procedure. For each estimation procedure, three simulation cells were run: a sample size of 4000 examinees per group paired with 40 items, a sample size of 8000 examinees per group paired with 40 items, and a sample size of 8000 per group paired with 80 items. The results of this small study are displayed in Table 8. In general, removing the prior on the *g* parameter greatly improved the estimates of alpha. Alpha rates appear to be approaching .05 with greater sample sizes per group. Also, the chi-square means and variances are nearing their expected values of 3 and 6. Again, the random assignment procedure outperforms the two-stage estimation procedure due to a portion of the simulated DIF being absorbed into the estimated mean difference between the groups in the first stage of the two-stage estimation procedure. These results support the hypothesis that the prior on the *g* parameter was producing near-zero alpha rates for *g*-DIF detection, rendering the Wald test conservative.

For the conditions simulating DIF for 20% of the items, the estimated false alarm rates are presented in Table 9[16]. Under random assignment, the estimated false alarm rate falls below the 95% confidence interval of .05 for all cells. Under the two-stage estimation procedure, almost all of the estimated false alarm rates fall outside of the 95% confidence interval around .05. Similar to the alpha rates in Table 6, the underestimated false alarm rates can likely be attributed to the prior on the *g* parameter producing near-zero alpha rates for detecting *g*-DIF. However, for the two-stage estimation procedure, the cells with sample sizes of 1000 per group and simulated *b*-DIF of .8 for the DIF items exhibit false alarm rates greater than the expected .05 value. It is likely that the .8 shift in the *b* parameters of the DIF items is shifting the estimated mean difference in stage 1 of the two-stage estimation procedure, which subsequently causes some false alarm in the DIF-free items. False alarm

[16] Overall DIF results are presented rather than separate *a*-DIF, *b*-DIF, and *g*-DIF results. None of the individual IRT parameter false alarm rates were large, and the overall false alarm rate provides a complete summary.

rates appear not to differ greatly based on whether the mean difference between groups was simulated to be 0 or -.6. Overall, these results do not follow the same pattern as the alpha rates, suggesting that the presence of DIF simulated in 20% of the items using the 3PL model results in a false alarm rate that is different from the pure alpha rate due to difficulties estimating the mean difference between groups.

Table 10 provides the chi-square means and variances for the statistics used to estimate the false alarm rates using the 3PL model. As with the alpha rate estimates, we would expect the chi-square means to be approximately 3 and the chi-square variances to be approximately 6. The pattern of chi-square means and variances in Table 10 follows the same pattern as the false alarm rates in Table 9. For the two-stage estimation procedure, the cells with sample sizes of 1000 per group and simulated *b*-DIF of .8 for the DIF items exhibit large chi-square values, corresponding to the inflated false alarm rates in Table 9 due to the effect of *b*-DIF on the estimated mean difference between groups. Under all other conditions, the estimated chi-square means and variances fall shy of their expected values, reflected in the lower false alarm rates of Table 9 and likewise explained by the prior on the *g* parameter producing near-zero alpha rates for detecting *g*-DIF.

The proportions of overall DIF, *a*-DIF, *b*-DIF, and *g*-DIF detected at the .05 level for the cells simulating data using the 3PL model under random assignment are displayed in Figure 4 as an estimate of power. In general, power is most affected by the sample size per group. With 1000 simulees per group, the DIF items are for the most part detected more than 50% of the time. With 250 simulees per group, the DIF items are generally detected less than 50% of the time. This reduction in power can be partially attributed to the near zero rates of *g*-DIF detection due to the prior on the *g* parameter. Halving the *a* parameters induces slightly greater DIF detection as compared to doubling the *a* parameters. However, shifting the *b* parameters has a greater effect on DIF detection than halving or doubling the *a* parameters: a shift of .8 greatly increases DIF detection as compared to a shift of .4. Additionally, for sample sizes of only 250 simulees per group, greater levels of *b*-DIF tend to be detected in comparison to overall DIF. This may be due to the near-zero rates of *g*-DIF detection, a greater effect of *b*-DIF as compared to *a*-DIF, and difficulties estimating the *g* asymptote complicating *a*-DIF and *b*-DIF separation. For sample sizes of only 250 simulees per group, increasing the test length from 5 items to either 20 or 40 items did not greatly improve DIF detection. On the other hand, for the sample size of 1000 simulees per group, increasing the test length from 5 items to 20 items tended to increase DIF detection, although a test length of 40 items did not add as much power. Across simulation conditions for the 3PL model under random assignment, power was greatest for the cells simulating sample sizes of 1000 per group, *a*-DIF of 2 and *b*-DIF of .8.

The proportions of overall DIF, *a*-DIF, *b*-DIF, and *g*-DIF detected at the .05 level for the cells simulating data using the 3PL model with two-stage estimation and no simulated mean difference between the groups are displayed in Figure 5. The pattern of results is very similar to that of random assignment. However, power is generally lower for these two-stage estimation cells as compared to random assignment, particularly when the sample size is only 250 per group. This can likely be attributed to the low alpha and false alarm rates due to the near-zero rates of *g*-DIF detection resulting from the prior on the *g* parameter. As with the GRM, a portion of the simulated DIF may also be absorbed into the estimated mean difference between the groups in the first stage of the estimation procedure, reducing power. This is consonant with the observation that the differences in the proportions of *b*-DIF detection between the two-stage estimation and random assignment cells are often greater than the differences in the proportions of *a*-DIF detection.

As an estimate of power, Figure 6 presents the proportions of overall DIF, *a*-DIF, *b*-DIF, and *g*-DIF detected at the .05 level for the cells simulating data using the 3PL model with two-stage estimation and a -.6 simulated mean difference between the groups. The pattern of power for DIF detection is very similar to that for two-stage estimation with no simulated mean difference between the groups. The primary difference is a slight general reduction in power for cells simulating a -.6 mean difference between the groups as compared to no mean difference; this difference is most salient when the sample size is 250 examinees per group. This can be explained by the additional difficulty in estimating a -.6 mean difference between the groups.

*IRTLRDIF Comparison*

For the simulation cells mimicking Habing's IRTLRDIF study, empirical Type I error and the power of DIF detection were calculated. For the cells without simulated DIF, the proportion of times the candidate item was mistakenly identified as a DIF item provides an estimate of Type I error, alpha. For the cells with simulated DIF, the proportion of times the candidate item was correctly identified provides an estimate of power. False alarm rates are not reported.

For the cells without simulated DIF, the estimated alpha rates at the .05 level are presented in Table 11. Across all conditions, the estimate of the overall alpha rate obtained by Habing using IRTLRDIF is higher than the estimate using the Wald test. The overall alpha rate for the Wald test improves, approaching the expected .05 rate, with sample sizes of 1000 simulees versus 250 simulees. There are not large differences in alpha rates between cells varying the mean difference between groups or the *b* parameter. The average overall alpha rate estimated for the Wald test is .025. This may be due to a platykurtic likelihood or an overcorrection of the parameter error variances by the SEM procedure. To quantify the latter possibility: if the distribution of the standardized difference between the item parameters approximated a normal (0, $1.125^2$) distribution instead of the assumed (0, 1) distribution, we would expect alpha rates of .025, which corresponds to our observed estimate. So if the standard errors of the parameter estimates are as little as 12.5% too large, the result would be the observed lower-than-nominal alpha rates.

The difference in alpha rate between IRTLRDIF and the Wald test is reflected more greatly in underestimates of alpha rates for *b*-DIF detection than *a*-DIF detection. This may be due to the conditional nature of the *b*-DIF detection using the Wald test; when *a*-DIF is present, even due to sampling error, *b*-DIF is evaluated based on equal *a* parameters.

For the cells simulating only *a*-DIF in the candidate item, the proportions of overall DIF, *a*-DIF, and *b*-DIF detected at the .05 level are presented in Table 12 as an estimate of power. Across the majority of conditions, the overall power obtained by Habing using IRTLRDIF tends to be higher than the estimate using the Wald test. However, the estimates obtained using the Wald test are not always substantially lower than those for the IRTLRDIF approach. Some of the underestimation may be due to the differences in estimated alpha rates, as the Wald test appears to be conservative with lower alpha rates. There are not large differences in power between cells varying the mean difference. However, both IRTLRDIF and the Wald test produce greater rates of DIF detection given greater differences simulated between the *a* parameters.

These results also suggest that the Wald test can effectively detect *a*-DIF in the absence of *b*-DIF. Some *b*-DIF is being detected above the .05 level, but this is due to the conditional nature of the test; the *b*-DIF detection test using the Wald test is conditional on the *a* parameters set equal. This “false-alarm” *b*-DIF is not an area of concern given that *b*-DIF is generally not interpreted in the presence of *a*-DIF.

For the cells simulating only *b*-DIF in the candidate item, the proportions of overall DIF, *a*-DIF, and *b*-DIF detected at the .05 level are presented in Table 13. Again, the overall power obtained by Habing using IRTLRDIF is a little higher than that estimated using the Wald test. However, greater power is evident for both IRTLRDIF and the Wald test for the cells with greater differences simulated between the *b* parameters. These results also suggest that the Wald test can effectively detect *b*-DIF in the absence of *a*-DIF. Unlike the detection of only *a*-DIF, there is no evidence of increased “false-alarm” *a*-DIF; this is because the *a*-DIF detection test is not conditional on another parameter test.

## Discussion of Simulation Results

The implications of the simulation results are optimistic. For the GRM, the Wald test performed well, producing nearly nominal alpha rates and false alarm rates, as well as the correspondingly expected chi-square means and variances, with improved performance given larger sample size. Similarly, the Wald test provided evidence of adequate power, with greater power attributed to greater magnitudes of simulated DIF, larger sample size, and in some cases, greater test length. The two-stage estimation procedure, in comparison to random assignment, is somewhat conservative due to a portion of the random DIF being absorbed into the estimated mean difference between the groups at stage 1. This also results in somewhat lower power, particularly when sample size is also small.

On the other hand, the Wald test under the 3PL model yielded lower alpha and false alarm rates, as well as lower chi-square means and variances, and lower power than expected. However, this performance was primarily due to a near-zero alpha rate for detecting *g*-DIF as a result of the prior on the *g* parameter. The results of a small study removing this prior confirm this assertion and suggest that performance is further improved with larger sample size and the use of the random assignment estimation procedure versus the two-stage estimation procedure. This is reassuring, given that the 3PL is not used in practice for small sample sizes or short test lengths. For the two-stage estimation procedure, the false alarm rate and power were affected by difficulties estimating the mean difference between groups. In general, under the 3PL model, greater power was evident with greater sample size, greater magnitude of simulated DIF, and in some cases, greater test length.

For the simulation cells mimicking Habing's IRTLRDIF study, the Wald test performed with a lower than nominal alpha rate. This may be due to a platykurtic likelihood or an overcorrection of the parameter error variances by the SEM procedure. However, Habing's study only used small sample sizes; alpha rates improve with larger sample sizes and greater test lengths, suggesting that the two methods will converge asymptotically. Power is somewhat greater for IRTLRDIF, for the conditions simulated by Habing, as compared to the Wald test due to the lower alpha rates. Although IRTLRDIF appears to have more power than the Wald test at small sample sizes, this is a relatively minor issue due to the small magnitude of the simulated DIF; in practice, only a small proportion of DIF items fall in that gray area separating the performance of these two methods. This comparison also demonstrates that the Wald test can effectively separate *a*-DIF and *b*-DIF when they occur in isolation of one another, with the caveat that some *b*-DIF is detected when *a*-DIF is present due to the conditional nature of the test. Fortunately, *b*-DIF is generally not interpreted in the presence of *a*-DIF.

In summary, the Wald test performs well, detecting DIF as expected. The performance of the test improves with large sample sizes and the random assignment estimation procedure. Furthermore, in general, the power of DIF detection increases with greater magnitudes of simulated DIF and greater test length. The use of large sample sizes and greater test lengths is especially important when the 3PL model is employed. The Wald test holds up well against IRTLRDIF, although the results of the two methods should converge, again, with larger sample sizes. The comparison study with IRTLRDIF also demonstrates the ability of the Wald test to effectively isolate *a*-DIF and *b*-DIF when they do not occur jointly.

## Extension to Multiple Groups

Most of the literature on DIF involves methods and procedures for the comparison of the performance of items for two groups. However, a practical need for considering more than two groups has been demonstrated by numerous studies; DIF methodology would benefit from statistical procedures that assess DIF simultaneously across multiple groups. Current DIF methods and procedures have been discussed with respect to their applicability to multiple group situations; two approaches, the Mantel-Haenszel (MH; 1959) method and Lord’s Wald test (1977, 1980) have been implemented to detect DIF in multiple group situations. When more than two groups are involved, the Wald test offers several advantages: it allows a single test of significance that may be more powerful than individual tests for each pair of groups, and it avoids the increase in Type I error associated with an individual test for each focal group.

The use of the MH procedure was investigated for DIF detection among multiple groups by Penfield (2001). The choice of this DIF detection procedure to extend to multiple groups is natural given its popularity. Assessing DIF across multiple groups using the MH procedure essentially involves performing individual tests for each pair of groups to be compared, leading to the problem of an increased probability of committing a Type I error. To control for a potentially spiraling Type I error, Penfield's study compared the MH chi-square statistic with no adjustment to the alpha level, the MH chi-square statistic with a Bonferroni adjusted alpha-level (BMH), and the Generalized Mantel-Haenszel (GMH; Somes, 1986) statistic that offers a single test of significance across all groups. The use of the GMH procedure for a polytomous group variable and dichotomous response variable is analogous to a previous application by Zwick, Donoghue, and Grima (1993) that involved a dichotomous group variable and a polytomous response variable.

Penfield's (2001) investigation of MH, BMH, and GMH for detecting DIF among multiple groups varied several factors in a simulation study, including the number of focal groups with DIF, sample size per group, the differences among the ability distributions of the reference and focal groups, and magnitude of matching criteria contamination. His results suggested that within the MH framework for detecting DIF in more than two groups, GMH is the most useful procedure because its Type I error rate remained at the nominal level of 0.05, and its power was consistently among the highest. If a significant value of GMH is obtained, the null hypothesis that there is no DIF among any of the groups is rejected. As a result, post hoc paired comparisons may be performed between each focal group and reference group using the BMH procedure to determine which groups exhibit DIF, while ensuring the Type I error rate across all comparisons does not exceed the intended nominal familywise error rate.

Although Penfield's (2001) study provided evidence that a MH method, specifically GMH, can effectively detect DIF among more than two groups, there are several limitations. First, Penfield's investigation considered only a consistent magnitude of DIF set by increasing the difficulty parameter and did not consider manipulating item discrimination between groups (probably because the MH procedure does not detect DIF in discrimination). Furthermore, a problem inherent to all MH methods is an inability to unconfound DIF with mean differences on the latent ability measured across groups (Stark, Cherynshenko, & Drasgow, 2004). Second, large sample sizes (e.g., N > 1000) were not included in the simulation conditions. This is particularly relevant to MH statistics, which employ chi-square tests, because statistical significance is often observed even though no substantially meaningful level of DIF is found in the data. Currently, a measure of effect size associated with GMH has not been proposed, and thus no method exists for assessing DIF in more than two groups that incorporates both statistical significance and effect size (Penfield, 2001).

Lord's Wald test (1977, 1980) for DIF detection was explored for the multiple group case by Kim, Cohen, and Park (1995). Kim et al. developed the $Q_j$ statistic to compare vectors of IRT item parameters between three or more groups. Kim, Cohen, and Park (1995) presented one example to illustrate the use of their proposed $Q_j$ statistic. The data contained responses to 14 items from three groups of 200 students each. The 2PL model was used to fit the data sets. MML estimation, as implemented in Bilog 3 (Mislevy & Bock, 1990), was used to obtain the item parameters, and an iterative linking using the test characteristic curve method developed by Stocking and Lord (1983) was used to place the groups on the same scale. For comparison, three pairwise multiple group DIF statistics were also obtained and tested with a Bonferroni-corrected Type I error rate. These two approaches yielded different equating coefficients due to the iterative linking procedure and, consequently, different sets of DIF items. Their results are discouraging but may be due to several constraints of the method.

The extension of Lord's Wald test (1977, 1980) to the multiple group case by Kim, Cohen, and Park (1994) suffered from several limitations. First, it was only examined in one case study and would benefit from a simulation study to fully explore the effects of various factors. Second, the $Q_j$ statistic does not consider the density of examinees in the sample along the ability continuum, and thus may signal DIF in regions of the ability scale with sparse data (Penfield, 2001). This constraint is known to negatively impact the performance of Lord's chi-square method (Camilli & Shepard, 1994), and likely has a similar effect on the performance of the $Q_j$ statistic. However, this statistic could be weighted to account for the density of examinees. Finally, the MML estimation of the item parameter error covariance matrix is not fully documented in the literature. A more complete attempt to obtain the item parameter error covariance matrix, and thus more accurate standard errors, should lead to better results.

### *Comparison of IRTLRDIF and the Wald Test Approaches to Multiple Group DIF Detection*

Both the Wald test and IRTLRDIF could be extended to detect DIF across multiple groups. However, each method has important pros and cons to weigh before developing a new DIF detecting system for multiple groups. IRTLRDIF has been shown to be powerful with alpha levels near nominal both in the small simulation study conducted by Habing and in the existing literature (Thissen, Steinberg, & Wainer, 1993; Teresi, Kleinman, & Ocepek-Welikson, 2000; Wainer, 1995). Given the current simulation study, the Wald test also produces alpha levels near nominal, with a conservative tendency. The conservative tendency renders the Wald test somewhat less powerful than IRTLRDIF, but this difference may not be practically meaningful given the rarity of items that would be differentially flagged by the two methods.

However, for $I$ items and ($G$ – 1) contrasts among $G$ groups, IRTLRDIF requires $2[I(G – 1)] + 1$ model fittings for the all-other anchor procedure, whereas the Wald test requires $I(G – 1) + 1$ model fittings. IRTLRDIF requires twice as many fittings because the procedure involves fitting the model twice per hypothesis, once for the compact model and once for the augmented model. Furthermore, the two-stage estimation procedure, which is not available for IRTLRDIF, can be used by the Wald test. The two-stage estimation procedure effectively uses all the items as the anchor in the second stage of the procedure, thus requiring only two model fittings, a substantially smaller number of fittings than IRTLRDIF using the all-other anchor procedure.

Given more than one contrast between unequally sized groups, IRTLRDIF is limited to all pair-wise comparisons or contrasts *ignoring* all others, e.g. combining groups for each contrast without consideration of the other contrasts/dependencies between the groups. In addition, these contrasts must be simple contrasts, e.g. the “plus and minus” type. In contrast, the Wald test can evaluate both contrasts ignoring all others or contrasts that *eliminate* dependence on other contrasts in a given order. The Wald test can also handle more complicated contrasts, such as testing for a linear trend. These additional contrasts provided by the Wald test could be produced by IRTLRDIF, but would require implementation of a linear model for the parameters in the IRT estimation program. On the other hand, such contrasts are only a trivial additional computation with the use of the Wald test.

To detect DIF across multiple groups, the IRTLRDIF procedure would first fit the compact model with all item parameters equal across groups to obtain population differences and the baseline -2LL. Next, two augmented models are fit for each contrast for each item. For each contrast, a model is fit with all parameters constrained equal for each “sign” (side) of the contrast. Another model is also fit with only the slope, *a*, parameters constrained equal for each side in the contrast. Finally, three likelihood ratio tests are computed: one testing all parameters to be different, one testing the slope, *a*, parameters to be different, and one testing the location, *b*, parameters to be different. These likelihood ratio tests yield a test of each contrast ignoring the dependence of that contrast on other contrasts. In order to include tests that eliminate dependence on other contrasts, a linear model for each item parameter would need to be implemented; then, the model would need to be refit with appropriate coefficients of the linear model set to zero.

Detecting DIF across multiple groups is more straightforward employing the Wald test. First, a model is fit with all item parameters equal across groups to obtain population differences. Next, the population differences are fixed at the values obtained in the previous step, and a model is fit with all item parameters freed, allowing different item parameters for each group. Subsequently, all relevant Wald tests can be computed, as well as the slope-intercept decomposition for each item. For unequal group sizes, contrasts comparable to those available using IRTLRDIF, e.g. ignoring between-contrast dependence, can be achieved by using non-orthogonal contrasts weighted by the sample size of each group. Additionally, contrasts that eliminate between-contrast dependence in a given order can be tested by orthogonalizing the contrasts in the metric of the group sample sizes.

The utility of the Wald test in detecting DIF with multiple groups can be illustrated using an example.

*Detecting DIF with Multiple Groups Using the Wald Test: An Example*

To explore the flexibility of the Wald test to detect DIF with multiple groups, items from the Everyday Discrimination Scale (EDS; Williams et al., 1997) were analyzed. This scale measures “chronic, routine, and relatively minor experiences of unfair treatment” (pp. 340). Nine items were originally included by Williams et al.; however, we only considered the following five items, a more unidimensional set according to item factor analyses conducted by Stucky and Gottfredson (2008):

1. You are treated with less courtesy than other people.
2. People act as if they think you are not smart.
3. People act as if they are afraid of you.
4. People act as if they’re better than you.
5. You are threatened or harassed.

The response options are: “never,” “less than once a year,” “a few times a year,” “a few times a month,” “at least once a week,” and “almost everyday.”

Data were collected in 2004 by a multi-site, multidisciplinary team of researchers working on the Educational Diversity Project (EDP; www.unc.edu/edp)[17]. Participants included 6,100 incoming law students from a nationally representative sample of 50 American Bar Association (ABA) approved law schools in the United States. Schools with very high minority populations were over-sampled. Of the 6,100 students in the sample, 4,079 (66.9%) are white, 589 (9.7%) are African-American, 508 (8.3%) are Asian-American, 493 (8.1%) are multi-racial, 327 (5.3%) are Latino, and 104 (1.7%) are unknown. The sample also includes 3,177 (52.1%) females and 2,921 (47.9%) males. DIF analyses were limited to females/males crossed by blacks/whites because these are the primary groups of interest for

[17] These data were generously provided by Abigail Panter for use in this illustration.

applied researchers. Sample sizes for these groups are: 2,085 (44.7%) white males, 1994 (42.7%) white females, 402 (8.6%) black females, and 187 (4.0%) black males. In the future, other groups could be considered as well.

Employing the first stage of the two-stage estimation procedure in which all item parameters are constrained to be equal across groups, Table 14 displays the estimated means and variances for the four groups defined by race crossed with gender[18]. For comparison, the all-other anchor procedure was also used to test DIF for item 1, constraining the item parameters for items 2-5 to be equal across the four groups. The resulting estimated group means and variances are also provided in Table 14. The estimated group means and variances do not substantially differ between these two methods. However, the all-other anchor procedure must be repeated for each item under consideration for DIF; results for the other items are not tabulated here.

IRTLRDIF, like other DIF methods currently limited to pair-wise comparisons, can only use combinations of relevant groups when conducting multiple group DIF analyses. For illustrative purposes, several variations of the all-other anchor procedure were used to estimate group means and variances that would mimic these combinations. Focusing only on the groups defined by race, the black females and black males were constrained to have the same estimated group mean, as well as the same estimated item parameters. Correspondingly, the white females and white males were given analogous constraints. Combining the groups across race, the resulting estimated group means and variances are shown in the line labeled “Combined Groups-Race” in Table 14. Because the black females

[18] Confirming the accuracy of the estimation procedure, estimated means and variances as well as estimated item parameters matched, within 2 decimal places, results from Multilog (Thissen, Chen, & Bock, 2003).

are a much larger group than the black males, this combination essentially compares black females versus whites.

In the same manner, groups were combined within males and females. The resulting estimated group means and variances are shown in the line labeled “Combined Groups-Gender” in Table 14. The estimated group mean for the two female groups (0.18) is nearly identical to the previously estimated group mean for white females (0.17) when all groups were considered separately due to the much larger sample size of the white females. As a result of these inequalities in sample size, this gender comparison more closely resembles a comparison limited to white females and white males.

To mimic the group combinations necessary to test an interaction between race and gender, the black females and white males were combined, with the same estimated group mean, as well as the same estimated item parameters. Accordingly, the white females and black males were given analogous constraints. The resulting estimated group means and variances using these constraints are shown in the line labeled “Combined Groups-Interaction” in Table 14. Again, unequal sample sizes result in a group comparison that essentially functions as white females versus white males.

The results of using IRTLRDIF to test each respective combination of groups for DIF in item 1 are reported in the first column of Table 15. This set of results required three model fittings, and evaluation of DIF in the other four items would require 12 more model fittings. This set of tests detecting DIF separately by race, gender, and their interaction using IRTLRDIF is essentially a set of comparisons, each of which ignores the others, given that each test is confounded with the other two tests. Disregarding this confounding, the results of these tests indicate significant race-DIF, gender-DIF, and DIF due to the race by gender interaction.

For a direct comparison with these results, the Wald test was employed for combined groups and the all-other anchor procedure, to produce analogous estimated item parameter estimates and estimated group means. Non-orthogonal contrasts (including weighting by sample size) were used for the Wald test. The matrix of contrasts used is:

$$\begin{bmatrix} .6825 & .3175 & -.4888 & -.5112 \\ .1678 & -.0823 & .8322 & -.9177 \\ .1616 & -.0857 & -.9143 & .8384 \end{bmatrix},$$

where the first row represents the race test, the second row represents the gender test, and the third row represents the interaction test. The order of the columns from left to right is black females, black males, white females, and white males, and the entries are proportional to sample size for each sign of the contrast. The results of these comparisons using the Wald test are provided in the second column of Table 15. Although results for IRTLRDIF and the Wald test are similar for the gender and interaction DIF tests, the Wald test yields a higher chi-square value for race DIF. However, it is difficult to interpret any of these statistics given the degree of confounding resulting from the group combinations.

To reduce error correlation between the comparisons, a set of orthogonal contrasts was also employed with the Wald test. The matrices of contrasts used are:

$$\begin{bmatrix} 15.484 & 7.203 & -11.090 & -11.596 \\ 3.769 & -3.769 & 30.097 & -30.097 \\ 10.650 & -10.650 & -10.650 & 10.650 \end{bmatrix} \text{ and } \begin{bmatrix} 5.730 & -2.811 & 28.420 & -31.339 \\ 14.870 & 7.628 & -14.870 & -7.628 \\ 10.650 & -10.650 & -10.650 & 10.650 \end{bmatrix},$$

where the order of the columns from left to right is black females, black males, white females, and white males. These contrasts are orthogonalized in the metric of the sample sizes (from Table 14), and then weighted by those sample sizes.

Repeated model fittings using combined groups and the all-other procedure were used to match item parameter and group mean equality constraints to those of IRTLRDIF. Due to their orthogonality, these contrasts resulted in a set of tests that followed an ignoring-eliminating order. When race is the first contrast, depicted in the matrix on the left above, it ignores dependencies on gender and the race by gender interaction. However, the second contrast, gender, is independent of race, and the third contrast, the race by gender interaction, is independent of both the race and gender tests. This order can also be reversed so that race is the first contrast and gender follows second; this is depicted in the matrix above on the right. The results of both ignoring-eliminating orders using orthogonal contrasts are reported in columns 3 and 4 of Table 15. Compared to the Wald tests using non-orthogonal contrasts, these results do not differ much for race or gender, but the chi-square values are much higher for the interaction. This could be due to a large interaction effect, or it could be due to the fact that these comparisons do not fully unconfound item parameter differences from differences among the means and standard deviations for all four groups.

To eliminate the effects of combining groups, all four groups' means and item parameters were estimated separately, again using items 2-5 as the anchor. The Wald test procedure was employed for each ignoring-eliminating order of the contrasts orthogonalized in the metric of the sample sizes. The results of both orders of elimination are provided in the fifth and sixth columns of Table 15. Using uncombined groups in the DIF analysis has a large effect compared to the previous models fit using combined groups; the chi-square values are much lower for all DIF comparisons. DIF due to gender[19] or the race by gender interaction is no longer significant. This suggests that combining pairs of the four groups before estimating the group means and standard deviations may have inflated the chi-square values, indicating greater levels of DIF than actually exist.

Basing the analyses on all four distinct means and standard deviations from Table 14, and eliminating the error correlation between comparisons by using orthogonalized contrasts, the DIF analyses should be more valid. The all-other anchor procedure is also a more valid procedure than the two-stage estimation procedure because the candidate item for DIF analysis is not included in the anchor set used to estimate the group means. The group means corresponding to these more valid applications of the Wald test are provided in the second row of Table 14. The expected score on item 1 for each group of interest is provided in Figure 7. The curves indicate some race DIF. Weak gender DIF appears to be present, but most of the DIF seems to be accounted for by the race DIF. Given the nearly parallel curves, there does not appear to be an interaction. These observations are reflected in the chi-square values in Table 15; not surprisingly, the significance of the gender DIF changes based on the ignoring-eliminating order of the contrast employed.

Although the use of all the other items as the anchor is more valid than the two-stage estimation procedure, it requires a separate item parameter estimation for every candidate item tested for DIF. Using the Wald test for each set of contrasts, with the two-stage estimation procedure and the two sets of orthogonalized contrasts in the ignoring-eliminating order, produces the results displayed in the seventh and eighth columns of Table 15. Including the candidate item in the estimation of the group means, given only five items,

[19] According to the race-gender ignoring-eliminating order, which tests for gender-DIF independent of race-DIF.

results in a somewhat biased anchor that seems to decrease the magnitude of DIF detected in this particular example. However, the results of both the two-stage estimation and all-other anchor procedures, holding all else constant, are not substantially different, and in this case, would not lead to different conclusions regarding the nature of DIF present in item 1.

In summary, the IRTLRDIF procedure confounded race-DIF with gender-DIF and the race by gender interaction-DIF. This confounding was due to combining groups, which ignores dependencies among the DIF tests, and fails to correct for all of the population distribution differences among the four groups. Essentially, group mean differences are estimated using “hybrid” populations. As a result, IRTLRDIF did not accurately identify the source of DIF. On the other hand, the Wald test produces more interpretable results, using estimated group means and standard deviations for all four groups, as well as orthogonal contrasts, which do not ignore dependencies between DIF tests. Using either the all-other anchor procedure or the two-stage estimation procedure, the Wald test was able to identify race as the primary source of DIF for item 1, without unnecessary and inaccurate confounding.

### *Considerations When Performing Multiple Group DIF Analyses*

Based on the application of the Wald test in the preceding example, this method may be an effective tool for detecting DIF across multiple groups. However, there are several considerations to take into account when constructing a framework for DIF detection given multiple groups. Unequal sample sizes are frequently encountered in real data, particularly when groups are defined by demographic variables. Given their prevalence, the estimation of uncombined group means and the use of orthogonal contrasts weighted by sample size are of a high priority. These methods are easily implemented in the procedures outlined using the Wald test and should be given great weight in decisions involving the choice of the Wald test to detect DIF simultaneously across multiple groups or a set of paired comparisons using another DIF detection method, which  are limited to group means estimated using combined populations and non-orthogonal contrasts.

Once the Wald test has been chosen to detect DIF across multiple groups, it is necessary to choose between the use of all other items as the anchor or the two-stage estimation procedure. In this example, both methods provide comparable results. However, the all-other anchor method is more valid to a degree, as it does not include the candidate item being detected for DIF in the anchor. However, the remaining items used as an anchor are not necessarily free of DIF themselves, unless they have been previously subjected to rigorous DIF analyses. The use of the all-other anchor also requires the model to be fit repeatedly for each item under consideration. On the other hand, the two-stage estimation procedure requires only one model fitting. The computational ease and savings in time comes at the price of potentially biased results, given that the candidate item is included in the anchor during the first stage of the estimation procedure. In the end, given the similarity in results between the two procedures, it will ultimately be a decision based on resources available to the particular researcher conducting the DIF analyses.

A final consideration when conducting DIF analyses with multiple groups is the set of contrasts used. The preceding example highlighted the importance of contrasts that are orthogonal in the metric of the sample sizes. Such contrasts can be ordered with any of the given contrasts as a test that ignores the other effects, followed by a sequence of the other contrasts that eliminate those preceding. The possibilities of contrast order multiply given more contrasts. In the example, given only three contrasts of interest, two ignoring-eliminating orders were tested. Even with only two orders, the results varied slightly depending on which contrast was tested first. The researcher must decide whether to employ only one order, several select orders, or all possible orders of contrasts. The most practical method may be to use as many orders as necessary to obtain for each contrast of interest the results from only the eliminating tests, as these tests take into account the dependencies of the other tests. However, this approach requires more than one order of contrasts. Again, this decision is best made based on available resources, as determined by the researcher.

## Conclusion

In an effort to increase the validity of a test for all groups, effectively detecting and eliminating DIF is a necessary step. In this regard, the IRTLRDIF approach has been shown to be the most flexible and powerful method for DIF detection (Thissen, Steinberg, & Wainder, 1993; Teresi, Kleinman, & Ocepek-Welikson, 2000; Wainer, 1995). However, this method is computationally-intensive, requiring many model-refittings. The Wald test for DIF detection, asymptotically equivalent to IRTRLDIF and requiring only one model fitting, has been demonstrated in the current research as a practical alternative. In this research, the Wald test for DIF detection was improved from Lord's original conception (1977, 1980) through modern error estimation, concurrent calibration, MML item parameter estimation, conditional DIF tests, and extensions to commonly used IRT models as well as multiple groups.

The simulation research employing the improved Wald test suggests reasonable levels of DIF detection, in analogs to both educational and psychological contexts. The performance of the Wald test improves with larger sample sizes, greater magnitudes of DIF, greater test lengths, and the random assignment estimation procedure. The use of larger sample sizes and greater test lengths is most critical for situations employing the 3PL model. The Wald test performs reasonably well compared to IRTLRDIF, exhibiting somewhat less power, although the results of the two methods should converge asymptotically.

In addition to establishing the practicality of using the improved Wald test to detect DIF in the two-group case, a goal of this research has been to demonstrate the flexibility of the Wald test through its straightforward extension to multiple groups. Given the need for an accurate and efficient DIF detection method for multiple groups and the prevalence of multiple group DIF assessments, this extension is a natural step. A two by two example (with four groups) was used to demonstrate the effectiveness of the Wald test, as well as compare it to the IRTLRDIF procedure. As currently implemented, the IRTLRDIF approach confounds the results of the DIF tests with population distribution differences, due to combining groups. As a result, IRTLRDIF yields spuriously large test statistics and is unable to accurately identify the source of DIF. On the other hand, the Wald test using orthogonal contrasts, which do not ignore dependencies between DIF tests, is able to effectively estimate means and standard deviations for multiple groups and provide more interpretable results.

Several considerations for designing a DIF detection framework given multiple groups were outlined. The superiority of the procedures used with the Wald test is most salient with unequal sample sizes. Such situations are prevalent in the literature; focal groups identified as important candidates for DIF analysis include various ethnic groups, women, examinees with disabilities, and various modes and languages of administration. Other considerations include choosing between the all-other anchor procedure and the two-stage estimation procedure, as well as determining the set of contrasts used.

This research demonstrates the efficacy, accuracy, and flexibility of the improved Wald test for DIF detection. As this was only an initial investigation, further exploration of the

Wald test and its applications remain. Future directions include considering other IRT models, other manipulations of DIF, and other practical applications. The multiple group extension allows for much future examination: exploring various numbers of groups, combinations of sample sizes, and contrasts, including more complicated contrasts such as testing for linear trends. The Wald test has been shown to be a practical alternative to the IRTLRDIF approach for DIF detection, particularly in the multiple group case, and it is hoped that future explorations will further demonstrate the applicability and flexibility of the method.

Table 1

*Simulation design*

| Anchoring Situation | Randomized Groups | | Two-Stage Estimation | |
|---|---|---|---|---|
| Test Length | 5, 20, 40 (x3) | | 5, 20, 40 (x3) | |
| Sample Size per Group | 250, 1000 (x2) | | 250, 1000 (x2) | |
| Mean Difference Between Groups | 0 (x1) | | 0, -.6 (x2) | |
| Models | 3PL, GRM (x2) | | 3PL, GRM (x2) | |
| Amount/Proportion of DIF | 0 | 20% | 0 | 20% |
| Magnitude of DIF in *a*'s: Multiply by | - | .5, 2 (3PL)<br>.875, 1.25 (GRM)<br>(x2) | - | .5, 2 (3PL)<br>.875, 1.25 (GRM)<br>(x2) |
| Magnitude of DIF in *b*'s: Shift by | - | .4, .8 (3PL)<br>.1, .2 (GRM)<br>(x2) | - | .4, .8 (3PL)<br>.1, .2 (GRM)<br>(x2) |
| Number of Cells | 3 x 2 x 1 x 2<br>= 12 | 3 x 2 x 1 x 2 x 2 x 2<br>= 48 | 3 x 2 x 2 x 2<br>= 24 | 3 x 2 x 2 x 2 x 2 x 2<br>= 96 |
| Total Number of Cells | 12 + 48 + 24 = 96 = 180 | | | |

Table 2

*Estimated alpha rates at the .05 level for overall DIF using the GRM*

| N | Test length | Randomized Groups α | Two-Stage: 0 Mean Diff α | Two-Stage: -.6 Mean Diff α |
|---|---|---|---|---|
| 250 | 5 | .03 | .01* | .02* |
| 250 | 20 | .03 | .02* | .03* |
| 250 | 40 | .05 | .02* | .02* |
| 1000 | 5 | .05 | .03 | .03 |
| 1000 | 20 | .05 | .02* | .03* |
| 1000 | 40 | .05 | .03* | .02* |

*Estimated proportion falls outside of a 95% confidence interval around .05.

Table 3

*Chi-square means and variances for no-DIF cells using the GRM*

| | | Randomized Groups | | Two-Stage: 0 Mean Diff | | Two-Stage: -.6 Mean Diff | |
|---|---|---|---|---|---|---|---|
| N | Test length | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance |
| 250 | 5 | 4.67 | 7.82 | 3.72 | 4.89 | 3.57 | 6.29 |
| 250 | 20 | 4.74 | 8.13 | 4.18 | 7.18 | 4.25 | 7.45 |
| 250 | 40 | 5.03 | 9.50 | 4.30 | 7.18 | 4.29 | 7.24 |
| 1000 | 5 | 5.07 | 10.13 | 4.21 | 7.32 | 4.17 | 7.31 |
| 1000 | 20 | 5.10 | 9.93 | 4.28 | 7.45 | 4.30 | 7.89 |
| 1000 | 40 | 4.95 | 9.37 | 4.34 | 7.78 | 4.28 | 7.63 |

Table 4

*Estimated false alarm rates at the .05 level for overall DIF using the GRM*

| | | DIF Items | | | | |
|---|---|---|---|---|---|---|
| | Test | | | Randomized | Two-Stage: | Two-Stage: |
| N | length | *a*-DIF | *b*-DIF | Groups | 0 Mean Diff | -.6 Mean Diff |
| 250 | 5 | 0.875 | .1 | .03 | .01* | .02* |
| 250 | 5 | 0.875 | .2 | .03 | .01* | .01* |
| 250 | 5 | 1.250 | .1 | .03 | .01* | .01* |
| 250 | 5 | 1.250 | .2 | .03 | .01* | .02* |
| 250 | 20 | 0.875 | .1 | .03* | .02* | .03* |
| 250 | 20 | 0.875 | .2 | .04 | .02* | .03* |
| 250 | 20 | 1.250 | .1 | .04 | .02* | .03* |
| 250 | 20 | 1.250 | .2 | .04 | .02* | .03* |
| 250 | 40 | 0.875 | .1 | .04 | .02* | .02* |
| 250 | 40 | 0.875 | .2 | .04 | .02* | .02* |
| 250 | 40 | 1.250 | .1 | .05 | .02* | .02* |
| 250 | 40 | 1.250 | .2 | .04 | .02* | .02* |
| 1000 | 5 | 0.875 | .1 | .05 | .03 | .03 |
| 1000 | 5 | 0.875 | .2 | .05 | .03 | .03 |
| 1000 | 5 | 1.250 | .1 | .05 | .03 | .03 |
| 1000 | 5 | 1.250 | .2 | .05 | .04 | .03 |
| 1000 | 20 | 0.875 | .1 | .05 | .02* | .02* |
| 1000 | 20 | 0.875 | .2 | .05 | .03* | .03* |
| 1000 | 20 | 1.250 | .1 | .05 | .02* | .03* |

*Table 4: Estimated false alarm rates at the .05 level for overall DIF using the GRM (continued)*

| | | DIF Items | | | | |
|---|---|---|---|---|---|---|
| | Test | | | Randomized | Two-Stage: | Two-Stage: |
| N | length | *a*-DIF | *b*-DIF | Groups | 0 Mean Diff | -.6 Mean Diff |
| 1000 | 20 | 1.250 | .2 | .05 | .03* | .03* |
| 1000 | 40 | 0.875 | .1 | .05 | .03* | .03* |
| 1000 | 40 | 0.875 | .2 | .06 | .03* | .03* |
| 1000 | 40 | 1.250 | .1 | .05 | .03* | .03* |
| 1000 | 40 | 1.250 | .2 | .05 | .04 | .04 |

* Estimated proportion falls outside of a 95% confidence interval around .05.

Table 5

*Chi-square means and variances for no-DIF items in the presence of DIF items using the GRM*

| | | DIF Items | | Randomized Groups | | Two-Stage: 0 Mean Diff | | Two-Stage: -.6 Mean Diff | |
|---|---|---|---|---|---|---|---|---|---|
| N | Test length | *a*-DIF | *b*-DIF | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance |
| 250 | 5 | 0.875 | .1 | 4.76 | 7.53 | 3.84 | 5.16 | 3.71 | 6.34 |
| 250 | 5 | 0.875 | .2 | 4.76 | 7.53 | 3.90 | 5.22 | 3.90 | 5.22 |
| 250 | 5 | 1.250 | .1 | 4.75 | 7.47 | 3.85 | 5.12 | 3.85 | 5.12 |
| 250 | 5 | 1.250 | .2 | 4.75 | 7.49 | 3.94 | 5.28 | 3.83 | 6.48 |
| 250 | 20 | 0.875 | .1 | 4.77 | 8.26 | 4.25 | 6.79 | 4.33 | 7.48 |
| 250 | 20 | 0.875 | .2 | 4.76 | 8.21 | 4.30 | 6.90 | 4.38 | 7.58 |
| 250 | 20 | 1.250 | .1 | 4.79 | 8.46 | 4.23 | 6.62 | 4.35 | 7.53 |
| 250 | 20 | 1.250 | .2 | 4.77 | 8.23 | 4.36 | 7.05 | 4.44 | 7.69 |
| 250 | 40 | 0.875 | .1 | 5.00 | 9.45 | 4.28 | 7.14 | 4.29 | 7.09 |
| 250 | 40 | 0.875 | .2 | 5.00 | 9.42 | 4.34 | 7.23 | 4.33 | 7.20 |
| 250 | 40 | 1.250 | .1 | 5.04 | 9.55 | 4.31 | 7.23 | 4.27 | 7.12 |
| 250 | 40 | 1.250 | .2 | 5.04 | 9.53 | 4.39 | 7.40 | 4.36 | 7.18 |
| 1000 | 5 | 0.875 | .1 | 5.15 | 10.12 | 4.32 | 7.24 | 4.30 | 7.59 |
| 1000 | 5 | 0.875 | .2 | 5.15 | 10.05 | 4.54 | 7.57 | 4.50 | 7.90 |
| 1000 | 5 | 1.250 | .1 | 5.16 | 10.13 | 4.40 | 7.40 | 4.35 | 7.60 |
| 1000 | 5 | 1.250 | .2 | 5.16 | 10.09 | 4.76 | 8.07 | 4.72 | 8.23 |
| 1000 | 20 | 0.875 | .1 | 5.09 | 9.64 | 4.34 | 7.31 | 4.29 | 7.38 |
| 1000 | 20 | 0.875 | .2 | 5.09 | 9.62 | 4.56 | 7.62 | 4.49 | 7.71 |

*Table 5: Chi-square means and variances for no-DIF items in the presence of DIF items using the GRM (continued)*

| | | DIF Items | | Randomized Groups | | Two-Stage: 0 Mean Diff | | Two-Stage: -.6 Mean Diff | |
|---|---|---|---|---|---|---|---|---|---|
| N | Test length | *a*-DIF | *b*-DIF | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance |
| 1000 | 20 | 1.250 | .1 | 5.09 | 9.63 | 4.40 | 7.35 | 4.39 | 7.51 |
| 1000 | 20 | 1.250 | .2 | 5.10 | 9.66 | 4.75 | 7.94 | 4.74 | 7.96 |
| 1000 | 40 | 0.875 | .1 | 5.10 | 10.15 | 4.42 | 7.84 | 4.38 | 7.86 |
| 1000 | 40 | 0.875 | .2 | 5.10 | 10.17 | 4.62 | 8.23 | 4.57 | 8.28 |
| 1000 | 40 | 1.250 | .1 | 5.11 | 10.19 | 4.49 | 8.00 | 4.48 | 8.07 |
| 1000 | 40 | 1.250 | .2 | 5.10 | 10.16 | 4.85 | 8.75 | 4.85 | 8.88 |

Table 6

*Estimated alpha rates at the .05 level for overall DIF using the 3PL model*

| N | Test length | Randomized Groups α | Two-Stage: 0 Mean Diff α | Two-Stage: -.6 Mean Diff α |
|---|---|---|---|---|
| 250 | 5 | .00* | .00* | .00* |
| 250 | 20 | .00* | .00* | .00* |
| 250 | 40 | .01* | .00* | .00* |
| 1000 | 5 | .01* | .00* | .00* |
| 1000 | 20 | .01* | .00* | .01* |
| 1000 | 40 | .02* | .01* | .01* |

* Estimated proportion falls outside of a 95% confidence interval around .05.

Table 7

*Chi-square means and variances for no-DIF cells using the 3PL model*

| | | Randomized Groups | | Two-Stage: 0 Mean Diff | | Two-Stage: -.6 Mean Diff | |
|---|---|---|---|---|---|---|---|
| N | Test length | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance |
| 250 | 5 | 1.20 | 1.09 | 0.94 | 0.71 | 1.01 | 0.90 |
| 250 | 20 | 1.65 | 2.05 | 1.48 | 1.59 | 1.48 | 1.56 |
| 250 | 40 | 1.88 | 2.30 | 1.70 | 1.48 | 1.71 | 1.89 |
| 1000 | 5 | 1.53 | 2.01 | 1.23 | 1.20 | 1.29 | 1.40 |
| 1000 | 20 | 2.11 | 2.98 | 1.86 | 2.27 | 1.91 | 2.54 |
| 1000 | 40 | 2.31 | 3.67 | 2.08 | 2.97 | 2.11 | 2.90 |

Table 8

*Estimated alpha rates at the .05 level for overall DIF and corresponding chi-square means and variances using the 3PL model without a prior*

| | | Randomized Groups | | | Two-Stage: 0 Mean Diff | | |
|---|---|---|---|---|---|---|---|
| | | | $X^2$ | $X^2$ | | $X^2$ | $X^2$ |
| N | Test length | α | Mean | Variance | α | Mean | Variance |
| 4000 | 40 | .03* | 2.70 | 4.48 | .01* | 2.35 | 3.57 |
| 8000 | 40 | .04 | 2.92 | 5.51 | .02* | 2.46 | 4.04 |
| 8000 | 80 | .04* | 2.66 | 4.69 | .02* | 2.36 | 3.78 |

* Estimated proportion falls outside of a 95% confidence interval around .05.

Table 9

*Estimated false alarm rates at the .05 level for overall DIF using the 3PL model*

| | | DIF Items | | | | |
|---|---|---|---|---|---|---|
| | Test | | | Randomized | Two-Stage: | Two-Stage: |
| N | length | *a*-DIF | *b*-DIF | Groups | 0 Mean Diff | -.6 Mean Diff |
| 250 | 5 | .5 | .4 | .00* | .00* | .00* |
| 250 | 5 | .5 | .8 | .00* | .00* | .00* |
| 250 | 5 | 2 | .4 | .00* | .00* | .00* |
| 250 | 5 | 2 | .8 | .01* | .01* | .00* |
| 250 | 20 | .5 | .4 | .00* | .00* | .00* |
| 250 | 20 | .5 | .8 | .00* | .00* | .00* |
| 250 | 20 | 2 | .4 | .00* | .00* | .00* |
| 250 | 20 | 2 | .8 | .00* | .01* | .00* |
| 250 | 40 | .5 | .4 | .00* | .00* | .00* |
| 250 | 40 | .5 | .8 | .01* | .00* | .00* |
| 250 | 40 | 2 | .4 | .01* | .00* | .00* |
| 250 | 40 | 2 | .8 | .00* | .01* | .01* |
| 1000 | 5 | .5 | .4 | .01* | .01* | .01* |
| 1000 | 5 | .5 | .8 | .01* | .05 | .03 |
| 1000 | 5 | 2 | .4 | .01* | .01* | .01* |
| 1000 | 5 | 2 | .8 | .01* | .14* | .07 |
| 1000 | 20 | .5 | .4 | .01* | .02* | .02* |
| 1000 | 20 | .5 | .8 | .01* | .04 | .03* |
| 1000 | 20 | 2 | .4 | .01* | .02* | .03* |

*Table 9: Estimated false alarm rates at the .05 level for overall DIF using the 3PL model (continued)*

|  |  | DIF Items |  |  |  |  |
|---|---|---|---|---|---|---|
| N | Test length | *a*-DIF | *b*-DIF | Randomized Groups | Two-Stage: 0 Mean Diff | Two-Stage: -.6 Mean Diff |
| 1000 | 20 | 2 | .8 | .01* | .10* | .11* |
| 1000 | 40 | .5 | .4 | .02* | .03* | .02* |
| 1000 | 40 | .5 | .8 | .02* | .06 | .05 |
| 1000 | 40 | 2 | .4 | .02* | .03* | .03* |
| 1000 | 40 | 2 | .8 | .02* | .10* | .11* |

* Estimated proportion falls outside of a 95% confidence interval around .05.

Table 10

*Chi-square means and variances for no-DIF items in the presence of DIF items using the 3PL model*

| | | DIF Items | | Randomized Groups | | Two-Stage: 0 Mean Diff | | Two-Stage: -.6 Mean Diff | |
|---|---|---|---|---|---|---|---|---|---|
| N | Test length | *a*-DIF | *b*-DIF | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance |
| 250 | 5 | .5 | .4 | 1.19 | 1.24 | 1.04 | 1.09 | 0.98 | 0.97 |
| 250 | 5 | .5 | .8 | 1.18 | 1.19 | 1.17 | 1.44 | 1.11 | 1.12 |
| 250 | 5 | 2 | .4 | 1.22 | 1.29 | 1.11 | 1.06 | 1.20 | 1.19 |
| 250 | 5 | 2 | .8 | 1.23 | 1.44 | 1.58 | 2.13 | 1.30 | 1.26 |
| 250 | 20 | .5 | .4 | 1.61 | 1.94 | 1.57 | 1.77 | 1.56 | 1.27 |
| 250 | 20 | .5 | .8 | 1.62 | 1.98 | 1.69 | 2.02 | 1.62 | 1.30 |
| 250 | 20 | 2 | .4 | 1.68 | 2.01 | 1.59 | 1.76 | 1.64 | 1.34 |
| 250 | 20 | 2 | .8 | 1.68 | 2.06 | 1.94 | 2.55 | 1.83 | 1.46 |
| 250 | 40 | .5 | .4 | 1.86 | 2.26 | 1.80 | 2.04 | 1.78 | 1.94 |
| 250 | 40 | .5 | .8 | 1.86 | 2.26 | 1.91 | 2.31 | 1.87 | 2.14 |
| 250 | 40 | 2 | .4 | 1.86 | 2.31 | 1.80 | 2.04 | 1.82 | 2.10 |
| 250 | 40 | 2 | .8 | 1.81 | 2.27 | 2.08 | 2.70 | 2.09 | 2.68 |
| 1000 | 5 | .5 | .4 | 1.54 | 2.17 | 1.67 | 2.61 | 1.54 | 1.44 |
| 1000 | 5 | .5 | .8 | 1.53 | 2.17 | 2.46 | 8.21 | 2.03 | 2.19 |
| 1000 | 5 | 2 | .4 | 1.56 | 2.11 | 1.88 | 2.73 | 1.86 | 1.72 |
| 1000 | 5 | 2 | .8 | 1.49 | 1.90 | 3.88 | 13.77 | 2.89 | 2.82 |
| 1000 | 20 | .5 | .4 | 2.08 | 2.91 | 2.35 | 3.62 | 2.28 | 1.85 |
| 1000 | 20 | .5 | .8 | 2.07 | 2.88 | 2.91 | 5.44 | 2.83 | 2.23 |

*Table 10: Chi-square means and variances for no-DIF items in the presence of DIF items using the 3PL model (continued)*

|  |  | DIF Items |  | Randomized Groups |  | Two-Stage: 0 Mean Diff |  | Two-Stage: -.6 Mean Diff |  |
|---|---|---|---|---|---|---|---|---|---|
| N | Test length | *a*-DIF | *b*-DIF | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance | $X^2$ Mean | $X^2$ Variance |
| 1000 | 20 | 2 | .4 | 2.10 | 2.98 | 2.37 | 3.64 | 2.52 | 2.05 |
| 1000 | 20 | 2 | .8 | 1.99 | 2.85 | 3.83 | 8.75 | 4.13 | 3.14 |
| 1000 | 40 | .5 | .4 | 2.29 | 3.64 | 2.64 | 4.70 | 2.54 | 4.01 |
| 1000 | 40 | .5 | .8 | 2.27 | 3.61 | 3.27 | 6.78 | 3.09 | 5.60 |
| 1000 | 40 | 2 | .4 | 2.28 | 3.61 | 2.55 | 4.36 | 2.69 | 4.44 |
| 1000 | 40 | 2 | .8 | 2.23 | 3.58 | 3.93 | 8.94 | 4.10 | 8.44 |

Table 11

*Comparison of Habing's IRTLRDIF results to the Wald test results: Estimated alpha rates at the .05 level*

|  | N | Mean Diff | Fixed *a*: Ref | Fixed *a*: Foc | Fixed *b*: Ref | Fixed *b*: Foc | Overall-DIF | *a*-DIF | *b*-DIF |
|---|---|---|---|---|---|---|---|---|---|
| IRTLRDIF | 250 | 0.00 | 1 | 1 | 0 | 0 | .06 | .06 | .05 |
| Wald Test | 250 | 0.00 | 1 | 1 | 0 | 0 | .03 | .05 | .02 |
| IRTLRDIF | 250 | 0.50 | 1 | 1 | 0 | 0 | .05 | .05 | .04 |
| Wald Test | 250 | 0.50 | 1 | 1 | 0 | 0 | .02 | .04 | .02 |
| IRTLRDIF | 250 | 1.00 | 1 | 1 | 0 | 0 | .07 | .06 | .05 |
| Wald Test | 250 | 1.00 | 1 | 1 | 0 | 0 | .03 | .03 | .02 |
| IRTLRDIF | 250 | 0.00 | 1 | 1 | 1 | 1 | .05 | .06 | .06 |
| Wald Test | 250 | 0.00 | 1 | 1 | 1 | 1 | .02 | .03 | .02 |
| IRTLRDIF | 250 | 0.50 | 1 | 1 | 1 | 1 | .03 | .04 | .04 |
| Wald Test | 250 | 0.50 | 1 | 1 | 1 | 1 | .01 | .02 | .03 |
| IRTLRDIF | 250 | 1.00 | 1 | 1 | 1 | 1 | .05 | .05 | .03 |
| Wald Test | 250 | 1.00 | 1 | 1 | 1 | 1 | .03 | .03 | .04 |
| IRTLRDIF | 1000 | 0.00 | 1 | 1 | 0 | 0 | .05 | .06 | .04 |
| Wald Test | 1000 | 0.00 | 1 | 1 | 0 | 0 | .02 | .02 | .02 |
| IRTLRDIF | 1000 | 0.50 | 1 | 1 | 0 | 0 | .05 | .06 | .05 |
| Wald Test | 1000 | 0.50 | 1 | 1 | 0 | 0 | .02 | .04 | .02 |
| IRTLRDIF | 1000 | 1.00 | 1 | 1 | 0 | 0 | .05 | .05 | .06 |
| Wald Test | 1000 | 1.00 | 1 | 1 | 0 | 0 | .04 | .05 | .03 |
| IRTLRDIF | 1000 | 0.00 | 1 | 1 | 1 | 1 | .06 | .06 | .04 |
| Wald Test | 1000 | 0.00 | 1 | 1 | 1 | 1 | .02 | .03 | .02 |
| IRTLRDIF | 1000 | 0.50 | 1 | 1 | 1 | 1 | .06 | .05 | .06 |
| Wald Test | 1000 | 0.50 | 1 | 1 | 1 | 1 | .03 | .03 | .02 |
| IRTLRDIF | 1000 | 1.00 | 1 | 1 | 1 | 1 | .06 | .03 | .06 |
| Wald Test | 1000 | 1.00 | 1 | 1 | 1 | 1 | .04 | .05 | .02 |

Table 12

*Comparison of Habing's IRTLRDIF results to the Wald test results: Proportion of only a-DIF detected*

|  | N | Mean Diff | Fixed *a*: Ref | Fixed *a*: Foc | Fixed *b*: Ref | Fixed *b*: Foc | Overall-DIF | *a*-DIF | *b*-DIF |
|---|---|---|---|---|---|---|---|---|---|
| IRTLRDIF | 250 | 0.00 | 1.20072 | 0.83283 | 0 | 0 | .27 | .36 | .04 |
| Wald Test | 250 | 0.00 | 1.20072 | 0.83283 | 0 | 0 | .22 | .35 | .02 |
| IRTLRDIF | 250 | 0.50 | 1.20072 | 0.83283 | 0 | 0 | .29 | .38 | .05 |
| Wald Test | 250 | 0.50 | 1.20072 | 0.83283 | 0 | 0 | .23 | .33 | .01 |
| IRTLRDIF | 250 | 1.00 | 1.20072 | 0.83283 | 0 | 0 | .32 | .37 | .08 |
| Wald Test | 250 | 1.00 | 1.20072 | 0.83283 | 0 | 0 | .22 | .33 | .03 |
| IRTLRDIF | 250 | 0.00 | 1.20072 | 0.83283 | 1 | 1 | .31 | .37 | .06 |
| Wald Test | 250 | 0.00 | 1.20072 | 0.83283 | 1 | 1 | .27 | .27 | .12 |
| IRTLRDIF | 250 | 0.50 | 1.20072 | 0.83283 | 1 | 1 | .41 | .36 | .22 |
| Wald Test | 250 | 0.50 | 1.20072 | 0.83283 | 1 | 1 | .25 | .25 | .11 |
| IRTLRDIF | 250 | 1.00 | 1.20072 | 0.83283 | 1 | 1 | .39 | .33 | .20 |
| Wald Test | 250 | 1.00 | 1.20072 | 0.83283 | 1 | 1 | .25 | .29 | .12 |
| IRTLRDIF | 250 | 0.00 | 1.35251 | 0.73937 | 0 | 0 | .72 | .79 | .06 |
| Wald Test | 250 | 0.00 | 1.35251 | 0.73937 | 0 | 0 | .58 | .73 | .01 |
| IRTLRDIF | 250 | 0.50 | 1.35251 | 0.73937 | 0 | 0 | .72 | .83 | .05 |
| Wald Test | 250 | 0.50 | 1.35251 | 0.73937 | 0 | 0 | .58 | .72 | .01 |
| IRTLRDIF | 250 | 1.00 | 1.35251 | 0.73937 | 0 | 0 | .70 | .77 | .06 |
| Wald Test | 250 | 1.00 | 1.35251 | 0.73937 | 0 | 0 | .53 | .70 | .03 |
| IRTLRDIF | 250 | 0.00 | 1.35251 | 0.73937 | 1 | 1 | .69 | .76 | .04 |
| Wald Test | 250 | 0.00 | 1.35251 | 0.73937 | 1 | 1 | .67 | .62 | .24 |
| IRTLRDIF | 250 | 0.50 | 1.35251 | 0.73937 | 1 | 1 | .78 | .71 | .45 |
| Wald Test | 250 | 0.50 | 1.35251 | 0.73937 | 1 | 1 | .70 | .66 | .23 |
| IRTLRDIF | 250 | 1.00 | 1.35251 | 0.73937 | 1 | 1 | .66 | .57 | .35 |
| Wald Test | 250 | 1.00 | 1.35251 | 0.73937 | 1 | 1 | .66 | .64 | .20 |

Table 13

*Comparison of Habing's IRTLRDIF results to the Wald test results: Proportion of only b-DIF detected*

| | N | Mean Diff | Fixed *a*: Ref | Fixed *a*: Foc | Fixed *b*: Ref | Fixed *b*: Foc | Overall-DIF | *a*-DIF | *b*-DIF |
|---|---|---|---|---|---|---|---|---|---|
| IRTLRDIF | 250 | 0.00 | 1 | 1 | 1.15 | 0.85 | .37 | .07 | .44 |
| Wald Test | 250 | 0.00 | 1 | 1 | 1.15 | 0.85 | .26 | .03 | .37 |
| IRTLRDIF | 250 | 0.50 | 1 | 1 | 1.15 | 0.85 | .32 | .03 | .46 |
| Wald Test | 250 | 0.50 | 1 | 1 | 1.15 | 0.85 | .22 | .02 | .34 |
| IRTLRDIF | 250 | 1.00 | 1 | 1 | 1.15 | 0.85 | .34 | .06 | .41 |
| Wald Test | 250 | 1.00 | 1 | 1 | 1.15 | 0.85 | .19 | .03 | .28 |
| IRTLRDIF | 250 | 0.00 | 1 | 1 | 0.15 | -0.15 | .49 | .05 | .58 |
| Wald Test | 250 | 0.00 | 1 | 1 | 0.15 | -0.15 | .32 | .04 | .45 |
| IRTLRDIF | 250 | 0.50 | 1 | 1 | 0.15 | -0.15 | .48 | .05 | .57 |
| Wald Test | 250 | 0.50 | 1 | 1 | 0.15 | -0.15 | .33 | .04 | .44 |
| IRTLRDIF | 250 | 1.00 | 1 | 1 | 0.15 | -0.15 | .47 | .05 | .57 |
| Wald Test | 250 | 1.00 | 1 | 1 | 0.15 | -0.15 | .27 | .03 | .37 |
| IRTLRDIF | 250 | 0.00 | 1 | 1 | 1.25 | 0.75 | .82 | .04 | .90 |
| Wald Test | 250 | 0.00 | 1 | 1 | 1.25 | 0.75 | .72 | .03 | .84 |
| IRTLRDIF | 250 | 0.50 | 1 | 1 | 1.25 | 0.75 | .83 | .06 | .91 |
| Wald Test | 250 | 0.50 | 1 | 1 | 1.25 | 0.75 | .69 | .03 | .82 |
| IRTLRDIF | 250 | 1.00 | 1 | 1 | 1.25 | 0.75 | .73 | .05 | .82 |
| Wald Test | 250 | 1.00 | 1 | 1 | 1.25 | 0.75 | .57 | .02 | .70 |
| IRTLRDIF | 250 | 0.00 | 1 | 1 | 0.25 | -0.25 | .92 | .07 | .96 |
| Wald Test | 250 | 0.00 | 1 | 1 | 0.25 | -0.25 | .83 | .03 | .92 |
| IRTLRDIF | 250 | 0.50 | 1 | 1 | 0.25 | -0.25 | .90 | .04 | .94 |
| Wald Test | 250 | 0.50 | 1 | 1 | 0.25 | -0.25 | .83 | .03 | .91 |
| IRTLRDIF | 250 | 1.00 | 1 | 1 | 0.25 | -0.25 | .88 | .05 | .92 |
| Wald Test | 250 | 1.00 | 1 | 1 | 0.25 | -0.25 | .76 | .04 | .84 |

Table 14

*Multiple groups DIF example: Estimated means and variances for relevant groups administered the EDS*

| | Black Female<br>(*N* = 402)<br>Mean (Var) | Black Male<br>(*N* = 187)<br>Mean (Var) | White Female<br>(*N* = 1994)<br>Mean (Var) | White Male<br>(*N* = 2085)<br>Mean (Var) |
|---|---|---|---|---|
| Two-Stage | 0.88 (0.89) | 1.21 (1.24) | 0.13 (0.60) | 0.00 (1.00)† |
| All-Other Anchor* | 0.85 (0.90) | 1.17 (1.22) | 0.17 (0.59) | 0.00 (1.00)† |
| Combined Groups-Race* | 0.98 (1.30) | 0.98 (1.30) | 0.00 (1.00)† | 0.00 (1.00)† |
| Combined Groups-Gender* | 0.18 (0.63) | 0.00 (1.00)† | 0.18 (0.63) | 0.00 (1.00)† |
| Combined Groups-Interaction* | 0.00 (1.00)† | 0.11 (0.67) | 0.11 (0.67) | 0.00 (1.00)† |

† Reference group.

* Means are estimated using items 2-5 as the anchor.

Table 15

*Comparison of IRTLRDIF results to the Wald test results for the multiple groups DIF example: EDS item 1*

| | | All-Other Anchor | | | | | | Two-Stage | |
|---|---|---|---|---|---|---|---|---|---|
| | | Combined Groups | | | | Uncombined Groups | | | |
| | | Ignoring Tests Only | | Race-Gender† | Gender-Race† | Race-Gender† | Gender-Race† | Race-Gender† | Gender-Race† |
| | | IRTLRDIF | Wald | Wald | Wald | Wald | Wald | Wald | Wald |
| Race | *d.f.* | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ |
| Overall-DIF | 6 | 36.2* | 48.2* | 48.2* | 48.7* | 24.7* | 24.6* | 19.2* | 18.3* |
| *a*-DIF | 1 | 2.8 | 4.5* | 4.5* | 4.5* | 3.1 | 2.7 | 2.5 | 2.3 |
| *b*-DIF | 5 | 33.4* | 43.7* | 43.7* | 44.2* | 21.6* | 21.9* | 16.8* | 16.0* |
| Gender | *d.f.* | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ |
| Overall-DIF | 6 | 23.1* | 21.9* | 21.3* | 21.9* | 12.3 | 13.6* | 9.6 | 11.7 |
| *a*-DIF | 1 | 0.9 | 1.0 | 1.0 | 1.0 | 0.1 | 0.3 | 0.0 | 0.1 |
| *b*-DIF | 5 | 22.3* | 20.9* | 20.4* | 20.9* | 12.2* | 13.3* | 9.6 | 11.6* |
| Interaction | *d.f.* | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ | $\chi^2$ |
| Overall-DIF | 6 | 23.6* | 23.5* | 37.3* | 37.3* | 6.4 | 6.4 | 5.0 | 5.0 |
| *a*-DIF | 1 | 3.1 | 3.9* | 6.2* | 6.2* | 5.0* | 5.0* | 3.6 | 3.6 |
| *b*-DIF | 5 | 20.5* | 19.6* | 31.1* | 31.1* | 1.4 | 1.4 | 1.4 | 1.4 |

† Ignoring-eliminating order.

* Significant at the .05 level.

Figure 1

Proportions of overall DIF, *a*-DIF, and *b*-DIF detected at the .05 level for each simulation condition under the GRM with randomized groups.

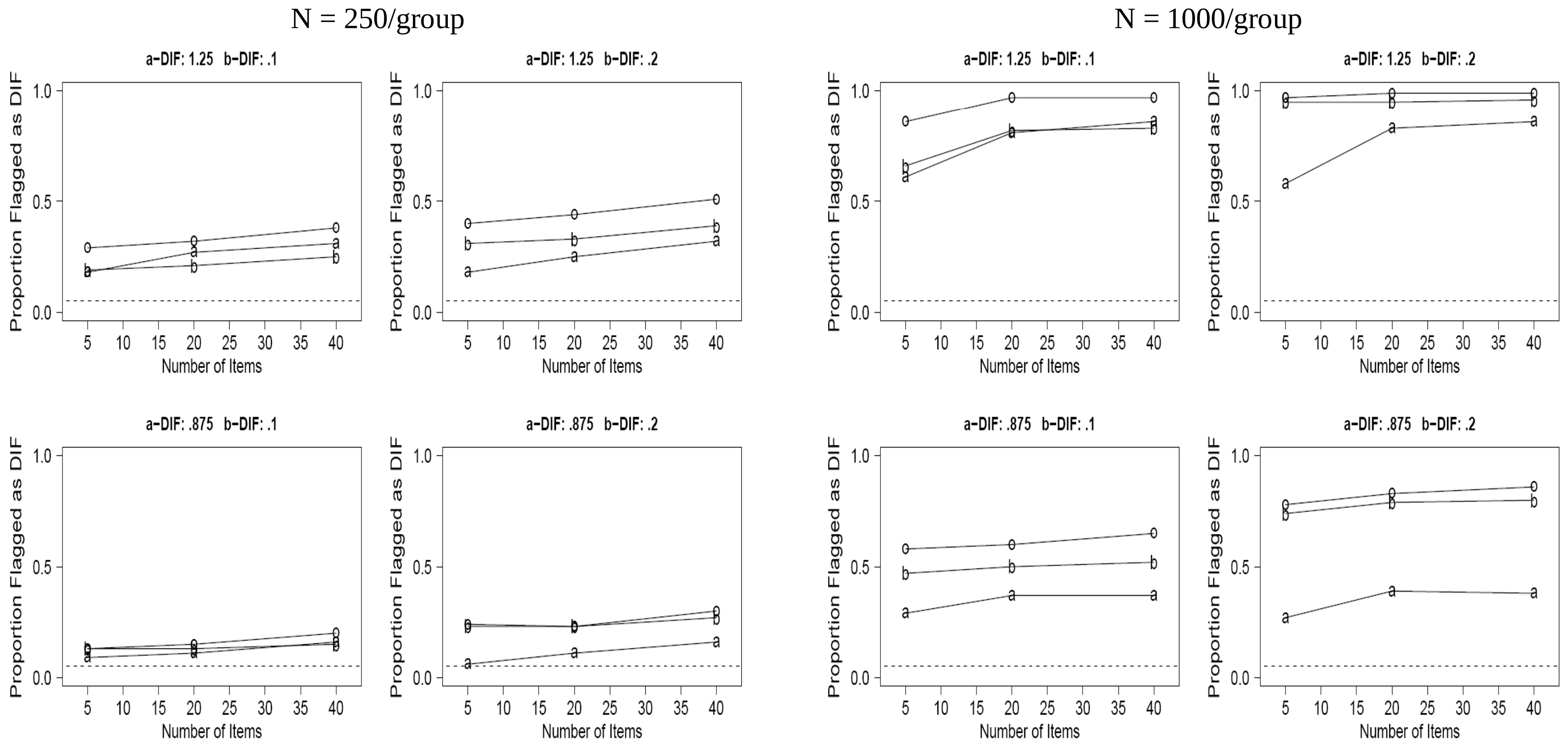


Figure 2

Proportions of overall DIF, *a*-DIF, and *b*-DIF detected at the .05 level for each simulation condition under the GRM with two-stage estimation and no simulated mean difference between the groups.

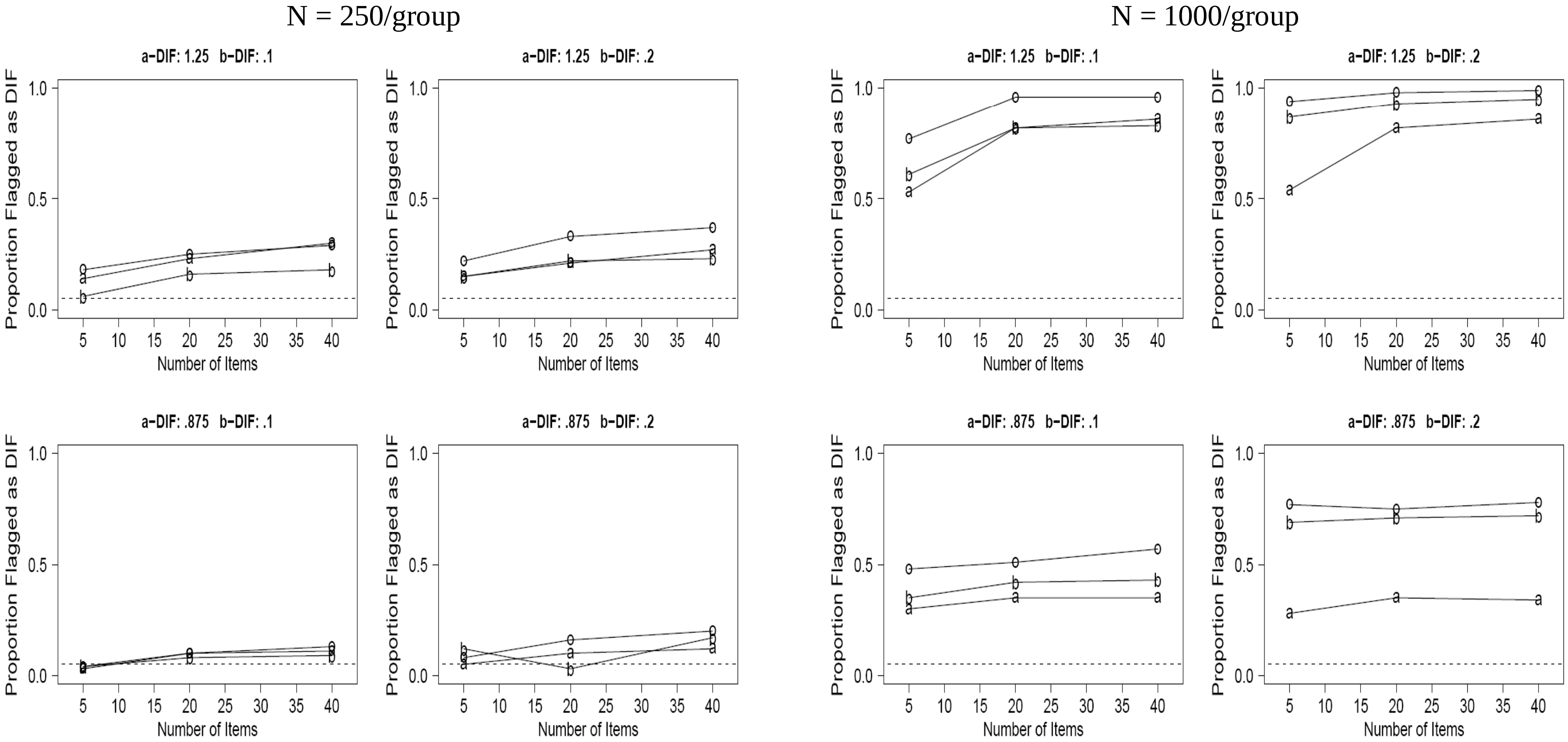


Figure 3

Proportions of overall DIF, *a*-DIF, and *b*-DIF detected at the .05 level for each simulation condition under the GRM with two-stage estimation and a -.6 simulated mean difference between the groups.

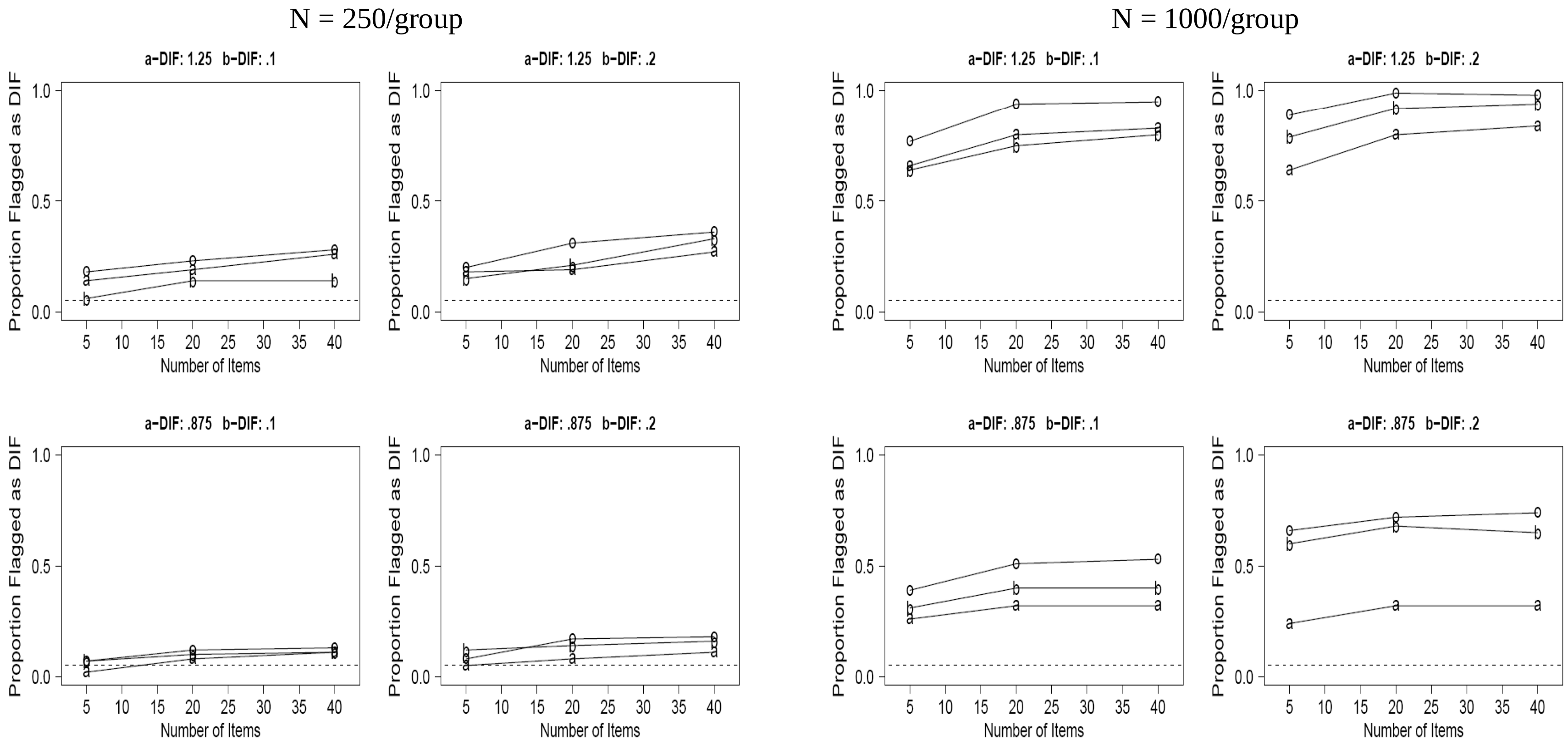


Figure 4

Proportions of overall DIF, *a*-DIF, *b*-DIF, and *g*-DIF detected at the .05 level for each simulation condition under the 3PL model with randomized groups.

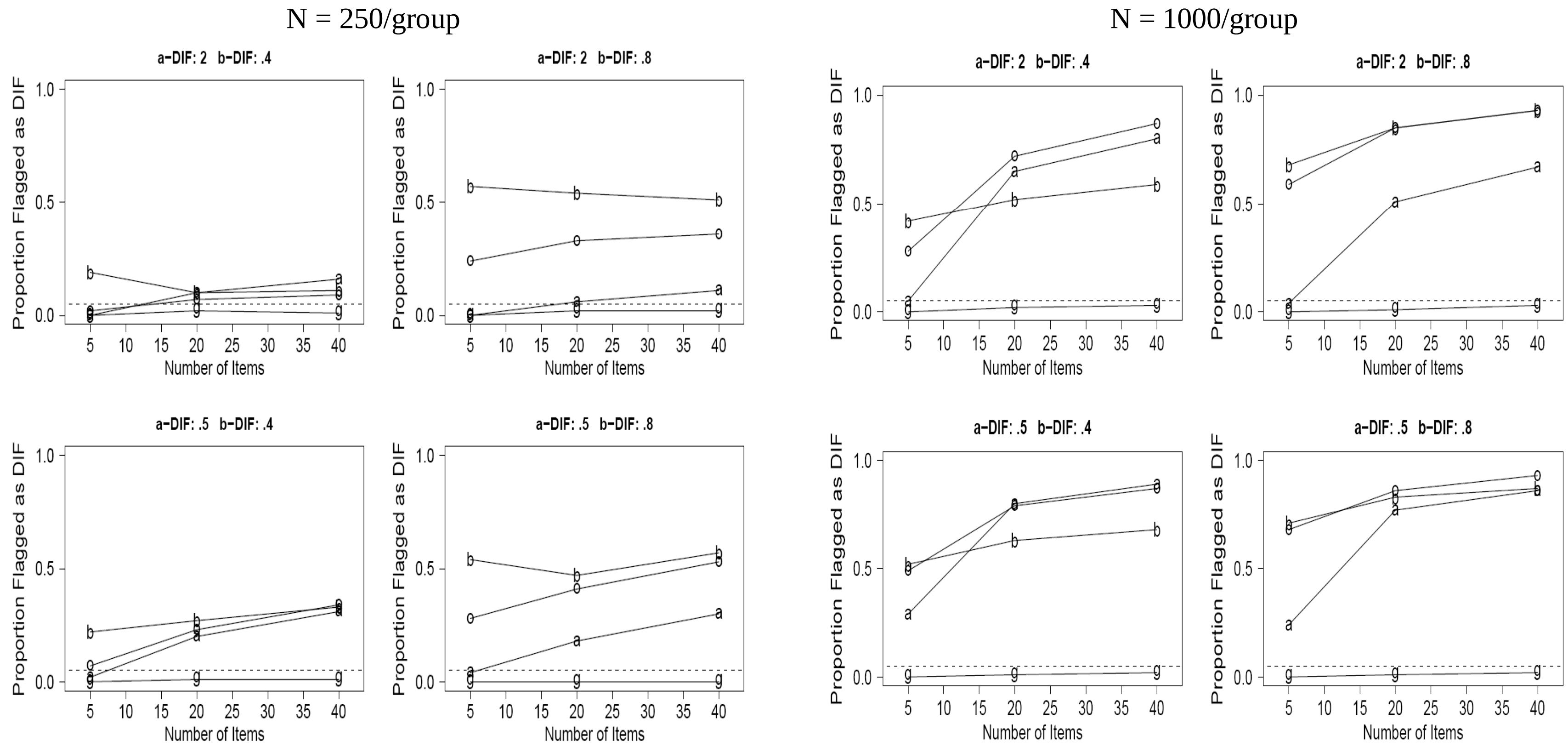


Figure 5

Proportions of overall DIF, *a*-DIF, *b*-DIF, and *g*-DIF detected at the .05 level for each simulation condition under the 3PL model with two-stage estimation and no simulated mean difference between the groups.

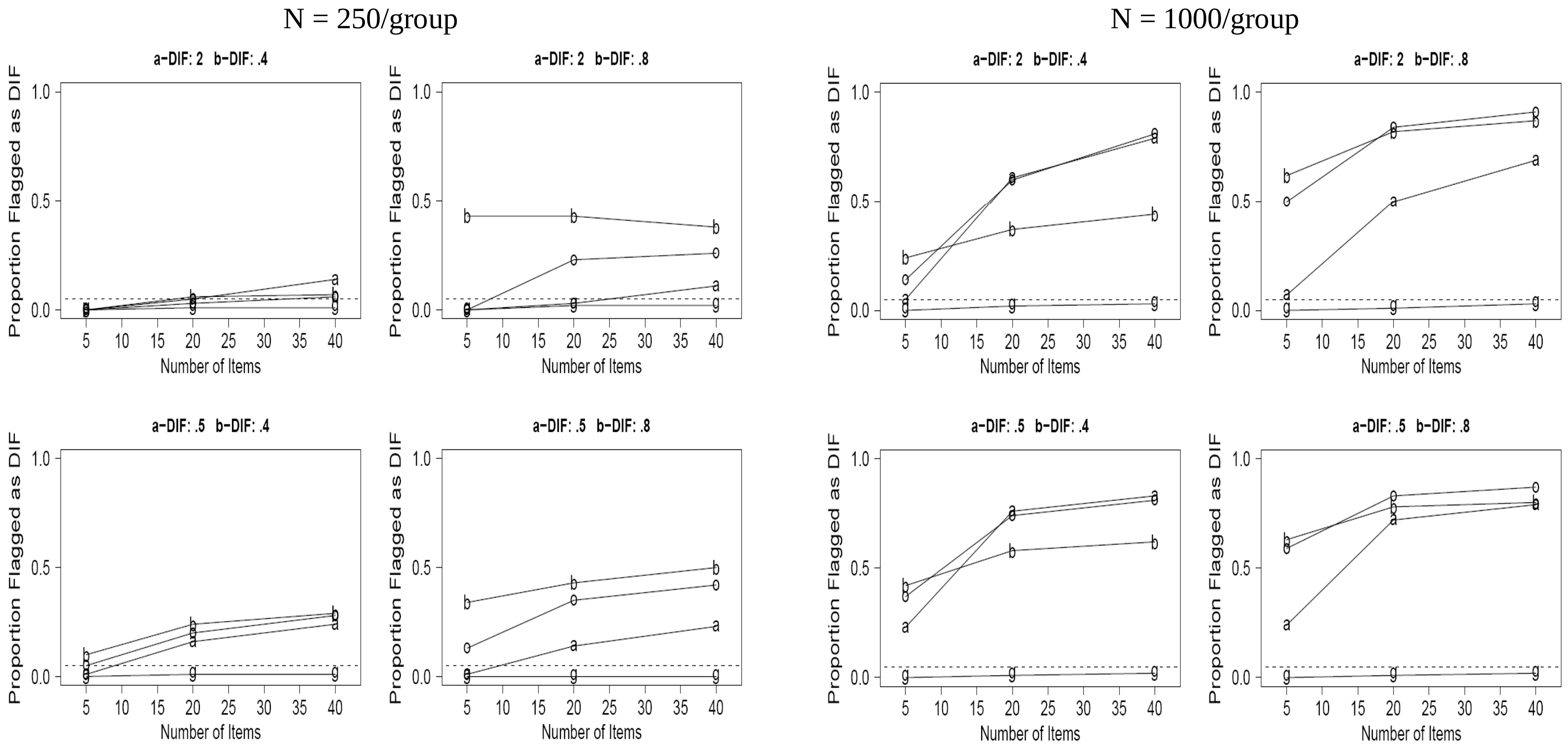


Figure 6

Proportions of overall DIF, *a*-DIF, *b*-DIF, and *g*-DIF detected at the .05 level for each simulation condition under the 3PL model with two-stage estimation and a -.6 simulated mean difference between the groups.

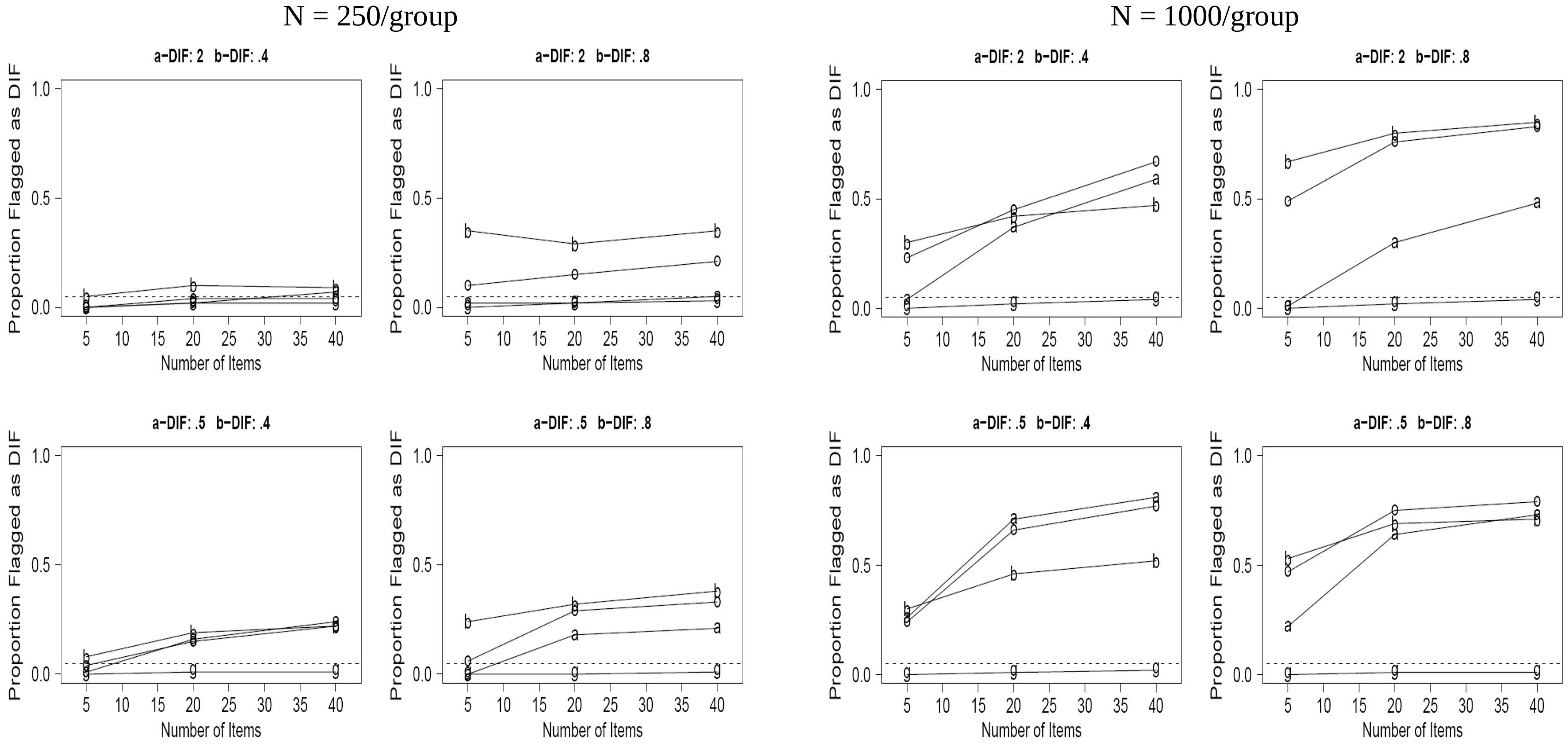
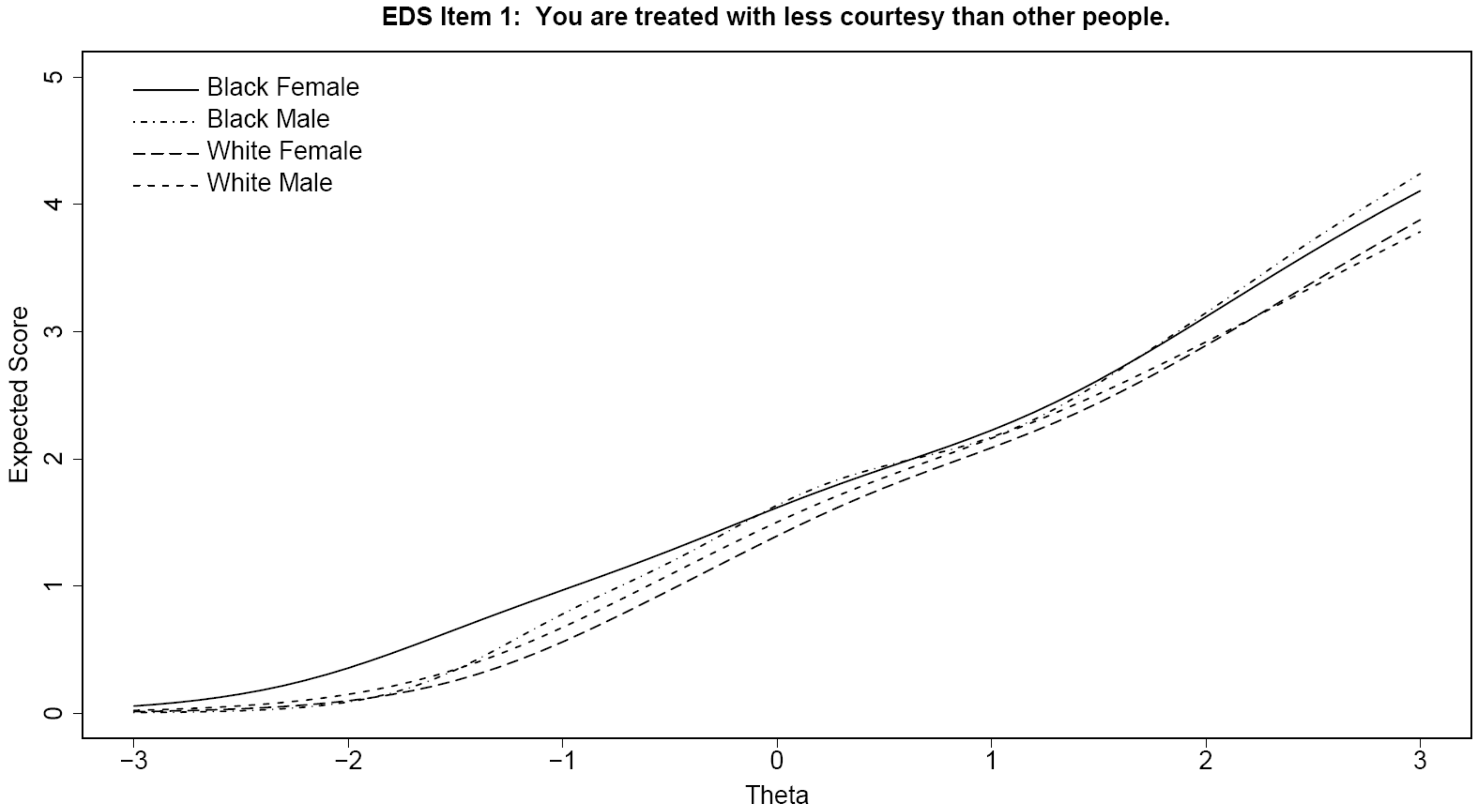


Figure 7

Multiple groups DIF example: Expected score on item 1 of the EDS, using the all-other anchor procedure.



Appendix A

## Tables of Percentages of Extreme Slopes using the 3PL Model

Table A1

*Percentages of items with extreme slopes for no-DIF simulation cells using the 3PL model*

| N | Test length | Randomized Groups | Two-Stage: 0 Mean Diff | Two-Stage: -.6 Mean Diff |
|---|---|---|---|---|
| 250 | 5 | 19.6 | 19.8 | 24.0 |
| 250 | 20 | 2.8 | 2.8 | 4.8 |
| 250 | 40 | 1.9 | 1.9 | 3.5 |
| 1000 | 5 | 7.0 | 7.0 | 9.4 |
| 1000 | 20 | 0.3 | 0.3 | 0.7 |
| 1000 | 40 | 0.2 | 0.2 | 0.4 |

Table A2

*Percentages of items with extreme slopes for DIF simulation cells using the 3PL model*

| | | DIF Items | | Randomized Groups | | Two-Stage: 0 Mean Diff | | Two-Stage: -.6 Mean Diff | |
|---|---|---|---|---|---|---|---|---|---|
| N | Test length | *a*-DIF | *b*-DIF | no-DIF items | DIF items | no-DIF items | DIF items | no-DIF items | DIF items |
| 250 | 5 | .5 | .4 | 20.3 | 14.0 | 20.3 | 14.0 | 26.3 | 23.0 |
| 250 | 5 | .5 | .8 | 19.3 | 15.0 | 19.3 | 15.0 | 25.5 | 30.0 |
| 250 | 5 | 2 | .4 | 19.8 | 38.0 | 19.3 | 38.0 | 25.0 | 35.0 |
| 250 | 5 | 2 | .8 | 19.3 | 42.0 | 19.5 | 42.0 | 23.8 | 51.0 |
| 250 | 20 | .5 | .4 | 3.0 | 3.0 | 3.0 | 3.0 | 6.0 | 3.3 |
| 250 | 20 | .5 | .8 | 3.3 | 2.8 | 3.4 | 2.5 | 5.8 | 3.8 |
| 250 | 20 | 2 | .4 | 2.5 | 11.3 | 2.6 | 11.3 | 4.5 | 18.3 |
| 250 | 20 | 2 | .8 | 2.8 | 17.5 | 3.9 | 18.3 | 4.6 | 24.5 |
| 250 | 40 | .5 | .4 | 2.0 | 2.9 | 2.0 | 2.9 | 3.8 | 2.8 |
| 250 | 40 | .5 | .8 | 2.1 | 3.3 | 2.1 | 3.1 | 3.5 | 4.4 |
| 250 | 40 | 2 | .4 | 1.8 | 6.4 | 1.8 | 6.6 | 3.0 | 11.0 |
| 250 | 40 | 2 | .8 | 1.6 | 10.0 | 1.6 | 10.3 | 3.0 | 17.9 |
| 1000 | 5 | .5 | .4 | 7.0 | 9.0 | 6.8 | 9.0 | 11.0 | 9.0 |
| 1000 | 5 | .5 | .8 | 7.5 | 8.0 | 7.3 | 8.0 | 12.3 | 11.0 |
| 1000 | 5 | 2 | .4 | 5.0 | 26.0 | 5.0 | 26.0 | 9.3 | 26.0 |
| 1000 | 5 | 2 | .8 | 6.8 | 31.0 | 6.8 | 32.0 | 9.8 | 31.0 |
| 1000 | 20 | .5 | .4 | 0.4 | 0.3 | 0.4 | 0.3 | 0.0 | 0.8 |
| 1000 | 20 | .5 | .8 | 0.3 | 0.3 | 0.3 | 0.3 | 0.8 | 0.5 |

*Table A2: Percentages of items with extreme slopes for DIF simulation cells using the 3PL model (continued)*

| | | DIF Items | | Randomized Groups | | Two-Stage: 0 Mean Diff | | Two-Stage: -.6 Mean Diff | |
|---|---|---|---|---|---|---|---|---|---|
| N | Test length | *a*-DIF | *b*-DIF | no-DIF items | DIF items | no-DIF items | DIF items | no-DIF items | DIF items |
| 1000 | 20 | 2 | .4 | 0.3 | 2.3 | 0.3 | 0.2 | 0.6 | 4.5 |
| 1000 | 20 | 2 | .8 | 0.3 | 5.5 | 0.3 | 5.5 | 0.6 | 7.0 |
| 1000 | 40 | .5 | .4 | 0.2 | 0.5 | 0.2 | 0.5 | 0.3 | 1.0 |
| 1000 | 40 | .5 | .8 | 0.2 | 0.5 | 0.2 | 0.5 | 0.2 | 1.1 |
| 1000 | 40 | 2 | .4 | 0.2 | 1.6 | 0.2 | 1.6 | 0.2 | 3.5 |
| 1000 | 40 | 2 | .8 | 0.2 | 3.1 | 0.2 | 3.0 | 0.2 | 5.9 |

# References

Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (Eds.). (2001). *The NAEP 1998 technical report.* Washington, DC: NCES. (Available online from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2001509).

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning,* (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.

Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement, 34,* 807–816.

Bock, R. D., Muraki, E., & Pfiffenberger, W. (1988) Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25,* 275--285.

Budgell, G. R., Raju, N. S., & Quartetti, D.A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309-321.

Cai, L. (in press). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December, 1977: An application of the standardization approach* (ETS Research Rep. No. RR-83-9). Princeton NJ: Educational Testing Service.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Dorans, N. J., & Schmitt, A. J. (1991). *Constructed response and differential item functioning: A pragmatic approach*. (Research Rep. No. 91-47). Princeton, NJ: Educational Testing Service.

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

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77,* 177–184.

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

Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model.* Unpublished masters thesis. University of North Carolina at Chapel Hill.

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, Inc.

Hulin, C., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Hillsdale NJ: Dow Jones-Irwin.

Judd, C. M., & McClelland G. H. (1989). *Data analysis: A model-comparison approach*. San Diego, CA: Harcourt Brace Jovanovich.

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32,* 261-276.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349–364). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga, *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690-700.

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

McLauglin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*, 161-173.

Meng, X., and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86,* 899-909.

Mislevy, R. J., & Bock, R. J. (1990). *BILOG3: Item analysis and test scoring with binary logistic model* (2nd ed.). Mooresville, IN: Scientific Software.

Muraki, E., & Englehard, G. (1989). *Examining differential item functioning with BIMAIN.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco CA.

Neyman, J. & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I and II. *Biometrika, 20,* 174-240, 263-294.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszal procedures. *Applied Measurement in Education, 14,* 235-259.

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.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: Wiley.

Rudner, L. M. (1977). *An evaluation of select approaches for biased item identification.* Unpublished doctoral dissertation, Catholic University of America, Washington DC.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5,* 213-233.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph,* No. 17, *34,* Part 2.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & Ronald K. Hambleton (Eds.), *Handbook of item response theory* (pp. 85-100). New York: Springer-Verlag.

Schmitt, A. P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement, 25,* 1–13.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27,* 67–81.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9,* 93-128.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician, 40,* 106–108.

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.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66,* 341-349.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, *81,* 332-342.

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

Stucky, B., & Gottfredson, N. (2008). *Using item response theory to revise the Everyday Discrimination Scale.* Unpublished manuscript, L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

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

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19,* 1651-1683.

Thissen, D. (2001). IRTLRDIF *v.2.0b: Software for the computation of statistics involved in item response theory likelihood-ratio tests for differential item functioning.* Unpublished manuscript, L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157-186.

Williams, D. R., Yu, Y., Jackson, J., & Anderson, N. B. (1997). Racial differences in physical and mental health: Socio-economic status, stress, and discrimination. *Journal of Health Psychology, 2(3),* 335-351.

Wood, R.L., Wingersky, M.S., & Lord, F.M. ( 1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM 76-6)* [Computer program]. Princeton NJ: Educational Testing Service.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299-325.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

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