Iowa Research Online
The University of Iowa's Institutional Repository

University of Iowa
Iowa Research Online

Theses and Dissertations

Spring 2011

Differential item functioning procedures for polytomous items when examinee sample sizes are small

Scott William Wood
University of Iowa

Copyright 2011 Scott William Wood

This dissertation is available at Iowa Research Online: http://ir.uiowa.edu/etd/1110

Recommended Citation
Wood, Scott William. "Differential item functioning procedures for polytomous items when examinee sample sizes are small." PhD (Doctor of Philosophy) thesis, University of Iowa, 2011.
http://ir.uiowa.edu/etd/1110.

Follow this and additional works at: http://ir.uiowa.edu/etd


Part of the Educational Psychology Commons

DIFFERENTIAL ITEM FUNCTIONING PROCEDURES FOR POLYTOMOUS ITEMS WHEN EXAMINEE SAMPLE SIZES ARE SMALL

by

Scott William Wood

An Abstract

Of a thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
in the Graduate College of
The University of Iowa

May 2011

Thesis Supervisor: Associate Professor Timothy Ansley

ABSTRACT

As part of test score validity, differential item functioning (DIF) is a quantitative characteristic used to evaluate potential item bias. In applications where a small number of examinees take a test, statistical power of DIF detection methods may be affected. Researchers have proposed modifications to DIF detection methods to account for small focal group examinee sizes for the case when items are dichotomously scored. These methods, however, have not been applied to polytomously scored items.

Simulated polytomous item response strings were used to study the Type I error rates and statistical power of three popular DIF detection methods (Mantel test/Cox's β, Liu-Agresti statistic, HW3) and three modifications proposed for contingency tables (empirical Bayesian, randomization, log-linear smoothing). The simulation considered two small sample size conditions, the case with 40 reference group and 40 focal group examinees and the case with 400 reference group and 40 focal group examinees.

In order to compare statistical power rates, it was necessary to calculate the Type I error rates for the DIF detection methods and their modifications. Under most simulation conditions, the unmodified, randomization-based, and log-linear smoothing-based Mantel and Liu-Agresti tests yielded Type I error rates around 5%. The HW3 statistic was found to yield higher Type I error rates than expected for the 40 reference group examinees case, rendering power calculations for these cases meaningless. Results from the simulation suggested that the unmodified Mantel and Liu-Agresti tests yielded the highest statistical power rates for the pervasive-constant and pervasive-convergent patterns of DIF, as compared to other DIF method alternatives. Power rates improved by several percentage points if log-linear smoothing methods were applied to the contingency tables prior to using the Mantel or Liu-Agresti tests. Power rates did not improve if Bayesian methods or randomization tests were applied to the contingency tables prior to using the Mantel or Liu-Agresti tests. ANOVA tests showed that

statistical power was higher when 400 reference examinees were used versus 40 reference examinees, when impact was present among examinees versus when impact was not present, and when the studied item was excluded from the anchor test versus when the studied item was included in the anchor test. Statistical power rates were generally too low to merit practical use of these methods in isolation, at least under the conditions of this study.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

DIFFERENTIAL ITEM FUNCTIONING PROCEDURES FOR POLYTOMOUS ITEMS WHEN EXAMINEE SAMPLE SIZES ARE SMALL

by

Scott William Wood

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Psychological and Quantitative Foundations
in the Graduate College of
The University of Iowa

May 2011

Thesis Supervisor: Associate Professor Timothy Ansley

Copyright by

SCOTT WILLIAM WOOD

2011

All Rights Reserved

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Scott William Wood

has been approved by the Examining Committee
for the thesis requirement for the Doctor of Philosophy
degree in Psychological and Quantitative Foundations at the May 2011
graduation.

Thesis Committee: _____
Timothy Ansley, Thesis Supervisor

_____
Kathleen Banks

_____
Michael J. Kolen

_____
Catherine Welch

_____
Peter Hlebowitsh

To April, Alyssa, Kurt, & Jennifer

I don’t care to belong to any club that will have me as a member.
Groucho Marx (1890-1977)

ACKNOWLEDGMENTS

Funding for this dissertation came from a variety of sources, including the University of Iowa's Department of Statistics and Actuarial Science, the University of Iowa's Department of Geography, the University of Arkansas for Medical Sciences, Iowa Testing Programs, the Iowa City Community School District, the University of Iowa's Department of Psychological and Quantitative Foundations, the University of Iowa's Science Education Program, the University of Iowa's Graduate College, Pacific Metrics Corporation, Kurt and Jennifer Wood, and April González. Thank you all for your support.

Special thanks to the members of my dissertation committee for devoting time to this project. Very special thanks to Dr. Timothy Ansley, for being a patient adviser and a fantastic editor. There is something to be said about a person who patiently circled every instance when I used the wrong verb tense or forgot that "data" is a plural noun and should be treated as such. Thank you, Dr. Ansley, for your editing prowess.

I would also like to acknowledge two professors who have been very important mentors during my time in graduate school. First, to Dr. Arkady Shemyakin of the University of St. Thomas' Department of Mathematics, thank you for your patience and support. For three summers, I had the chance to do simulation-based statistics research under the supervision of Dr. Shemyakin. Because of the lessons learned from that project, I am a better researcher and programmer today. Also, I would like to thank Dr. Blake Whitten of the University of Iowa's Department of Statistics and Actuarial Science and Department of Economics. His advice and optimism gave me the support and strength necessary to continue pursuing graduate education. Dr. Whitten, thank you for being such a great mentor.

In the Educational Measurement and Statistics program, I have had the chance to work with some very nice, and very intelligent, students, most of whom now have their

Doctorates. Thank you for your support during our time in the program. It is an honor to be a colleague amongst some very talented new psychometricians. In particular, thanks to Thomas Proctor, Jonathan Beard, and Tawnya Knupp. It was very educational to watch them complete their dissertations; I hope I was able to learn from their troubles.

To my friends and family, thank you all for your patience as I worked on this project. I promise to return all of your phone calls and e-mails soon. In particular, thank you to my friends from the University of St. Thomas and Hudson, Wisconsin. I never planned to be away from the Twin Cities for this long, but I look forward to coming back soon.

To April González, thank you for your support. April and I met in the College of Education in August 2005. I was her emotional support as she took her Master's comprehensive exams, and she was my emotional support when I took PhD exams and wrote this book. She has been there for all the high points and the low moments during the last six years. I would not have completed this project without her kindness, patience, and love. April, I love you, and I look forward to many years together as colleagues and best friends.

And finally, thanks to you, the reader, for at least reading the Acknowledgements section of this dissertation. For those of you willing to dive into the details of this project, I hope that you learn something new and interesting about differential item functioning.

ABSTRACT

As part of test score validity, differential item functioning (DIF) is a quantitative characteristic used to evaluate potential item bias. In applications where a small number of examinees take a test, statistical power of DIF detection methods may be affected. Researchers have proposed modifications to DIF detection methods to account for small focal group examinee sizes for the case when items are dichotomously scored. These methods, however, have not been applied to polytomously scored items.

Simulated polytomous item response strings were used to study the Type I error rates and statistical power of three popular DIF detection methods (Mantel test/Cox's β, Liu-Agresti statistic, HW3) and three modifications proposed for contingency tables (empirical Bayesian, randomization, log-linear smoothing). The simulation considered two small sample size conditions, the case with 40 reference group and 40 focal group examinees and the case with 400 reference group and 40 focal group examinees.

In order to compare statistical power rates, it was necessary to calculate the Type I error rates for the DIF detection methods and their modifications. Under most simulation conditions, the unmodified, randomization-based, and log-linear smoothing-based Mantel and Liu-Agresti tests yielded Type I error rates around 5%. The HW3 statistic was found to yield higher Type I error rates than expected for the 40 reference group examinees case, rendering power calculations for these cases meaningless. Results from the simulation suggested that the unmodified Mantel and Liu-Agresti tests yielded the highest statistical power rates for the pervasive-constant and pervasive-convergent patterns of DIF, as compared to other DIF method alternatives. Power rates improved by several percentage points if log-linear smoothing methods were applied to the contingency tables prior to using the Mantel or Liu-Agresti tests. Power rates did not improve if Bayesian methods or randomization tests were applied to the contingency tables prior to using the Mantel or Liu-Agresti tests. ANOVA tests showed that

statistical power was higher when 400 reference examinees were used versus 40 reference examinees, when impact was present among examinees versus when impact was not present, and when the studied item was excluded from the anchor test versus when the studied item was included in the anchor test. Statistical power rates were generally too low to merit practical use of these methods in isolation, at least under the conditions of this study.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

Test fairness is an important consideration in the development of educational and psychological assessments (Camilli, 2006). The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) devote four of fifteen chapters to fairness issues. The *Code of Professional Responsibilities in Educational Measurement* (Schmeiser, Geisinger, Johnson-Lewis, Roeber, & Schafer, 1995) also addresses test fairness, expecting measurement professionals to "perform all … responsibilities with honesty, integrity, due care, and fairness." Additional fairness standards have been published as the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) and the *ETS Standards for Quality and Fairness* (Educational Testing Service, 2002). Although there are many ways of considering test fairness, the focus here will be on the set of statistical procedures intended to detect differential item functioning (DIF).

Ideally, the probability of answering an item correctly should only depend on an examinee's ability level on the construct being measured by the item, along with any relevant item parameters such as difficulty or discrimination. If the probability also depends on a construct-irrelevant subgroup factor, such as race, ethnicity, gender, or socioeconomic status, then DIF may be present. Items with DIF may incorporate construct-irrelevant variance into test scores, thereby reducing score validity. Good testing practice suggests developers identify items with DIF and avoid using them in operational tests "unless they are judged to be fair and are required for valid measurements" (Zieky, 2006, p. 374).

Some DIF detection techniques are based on inferential statistics, using the item responses from the test under consideration. Many methods are available for

dichotomously scored items (Camilli, 2006; Clauser & Mazor, 1998; Hills, 1990; Millsap & Everson, 1993; Penfield & Camilli, 2007), including the Mantel-Haenszel test (Holland & Thayer, 1988), the standardization statistic (Dorans, 1989), the SIBTEST statistic (Shealy & Stout, 1993b), logistic regression (Swaminathan & Rogers, 1990), Lord's chi-square test (Lord, 1977; Lord, 1980), Raju's area measures (Raju, 1988; Raju, 1990), and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988; Thissen, Steinberg, & Wainer, 1993). Many of these methods have been adapted for use in polytomous items, including the standardization statistic (Zwick & Thayer, 1996), the SIBTEST statistic (Chang, Mazzeo, & Roussos, 1996), logistic regression (French & Miller, 1996), and the likelihood ratio test (Kim & Cohen, 1998). Additional DIF detection methods for polytomous items include the Mantel test (Zwick, Donoghue, & Grima, 1993), the HW1 and HW3 statistics (Welch & Hoover, 1993), logistic discriminant function analysis (Miller & Spray, 1993), Cox's β statistic (Camilli & Congdon, 1999), and the Liu-Agresti statistic (Penfield & Algina, 2003).

Because some DIF detecting methods are based on inferential statistics, sample size can affect the statistical power of detection. There are several reasons why examinee sample sizes could be smaller than desired, particularly when using polytomous items. For example, testing programs are increasingly incorporating performance assessments in their tests (Lane & Stone, 2006; Welch, 2006). Performance assessments, typically scored polytomously, require larger samples of examinees for item analysis and DIF detection. The costs required to administer and score these performance items, however, may prevent the test developer from acquiring such large samples.

A number of additional testing situations exist which could lead to small sample sizes. For example, the size of a minority group (commonly called the *focal* group in DIF research) could be much smaller than the size of the majority group (commonly called the *reference* group). This is possible for some ethnicities in some locations, such as Native Americans, whose number could be very small compared to a majority White examinee

group (Sinharay, Dorans, Grant, & Blew, 2008). It may be difficult, or too costly, to obtain enough focal group examinees for analysis. Small sample sizes are also common in studies that use translated tests (Fidalgo, Hashimoto, Bartram, & Muñiz, 2007; Muñiz, Hambleton, & Xing, 2001) and computerized adaptive tests (Fidalgo et al., 2007; Zwick, 2000; Zwick & Thayer, 2002). In addition, there are some testing contracts that require a thorough DIF analysis despite a small number of examinees taking the test (Parshall & Miller, 1995; Puhan, Moses, Yu, & Dorans, 2009).

In situations where small sample sizes are encountered, statistical power can be affected in such a way that the inferences made in DIF analyses could be unstable. Many solutions have been proposed to address the small sample size problem, but these solutions have usually been applied to the dichotomous case only (Camilli & Smith, 1990; Chaimongkol, 2005; Fidalgo, Ferreres, & Muñiz, 2004a; 2001; Parshall & Miller, 1995; Puhan et al., 2009; Sinharay et al., 2008; Swanson, Clauser, Case, Nungester, & Featherman, 2002; Zwick, Thayer, & Lewis, 1999; Zwick & Thayer, 2002; Zwick & Thayer, 2003). Little research has verified if these methods could be used effectively with polytomous items (Camilli & Congdon, 1999; Meyer, Huynh, & Seaman, 2004; Penfield & Algina, 2003; Vaughn, 2006).

For example, jackknife procedures and exact tests have been applied to dichotomous items to provide better estimates of the Mantel-Haenszel statistic (Camilli & Smith, 1990; Parshall & Miller, 1995). For small sample sizes, exact methods were slightly more conservative than the asymptotic Mantel-Haenszel test. Although research has considered using exact tests in the polytomous case (Meyer et al., 2004), only empirical data have been studied. Further, jackknife methods do not appear to have been extended to the polytomous case. Simulated data could help assess how DIF effect size, sample size, and additional factors affect the results of jackknife and exact methods.

Another modification to DIF detecting procedures that has been successful with dichotomous items is the use of Bayesian inferential statistics. First developed to assess

DIF in computerized adaptive tests, the empirical Bayes Mantel-Haenszel test uses information from the entire test to compute the prior distribution of the Mantel-Haenszel odds ratio (Zwick et al., 1999; Zwick & Thayer, 2002; Zwick & Thayer, 2003). A full Bayes Mantel-Haenszel test is available when groups of test items have similar characteristics (Sinharay et al., 2008). Bayesian logistic regression is available via hierarchical logistic regression, using normal priors on the regression coefficients (Chaimongkol, 2005; Swanson et al., 2002). By incorporating prior distributions, the sampling variance of the DIF statistic is typically reduced, leading to more precise inferences.

Little research exists regarding the use of Bayesian statistics in polytomous DIF detection. Although Bayesian versions of logistic regression exist for polytomous items (Vaughn, 2006), small sample sizes have not been considered. An empirical example of Bayesian inference using Cox's β statistic is available (Camilli & Congdon, 1999), though there is little knowledge of how well a Bayesian version of Cox's β would work under simulated conditions. Although research indicates that the Liu-Agresti statistic could be useful in a Bayesian DIF analysis (Penfield & Algina, 2003), the Bayesian version of the statistic has yet to be applied to simulated data.

A third adaptation used with small sample sizes in DIF detection is the application of log-linear smoothing to the total observed test score distribution (Puhan et al., 2009). In most DIF detection methods, the total score on a predetermined subset of items (often the complete test) is used as an estimate of ability. Log-linear smoothing introduces a small amount of bias to better estimate the population distribution of total scores. Log-linear smoothing has been effective on small samples when used with the dichotomous version of SIBTEST. Smoothing techniques in DIF analysis have yet to be applied to polytomous items or other DIF detection methods.

Additional methods for addressing the small sample size problem have been proposed. For example, increasing the significance level of the hypothesis test from α =

0.05 to $\alpha = 0.20$ can reduce Type II errors and improve power (Fidalgo et al., 2004a). Although this increases the number of false positives, this would decrease the number of false negatives. Arguably, it would be better for the test developer to remove items that did not truly exhibit DIF—the result of false positives—than to include items that did exhibit DIF—the result of false negatives (Camilli, 2006). Additional DIF detection methods include Angoff's transformed item difficulty, or delta, plot (Angoff & Ford, 1973; Angoff & Sharon, 1974; Angoff, 1982) and conditional *p*-value plots (Muñiz et al., 2001). Both Angoff's plot and the conditional *p*-value plots can be adapted for use in polytomous items, though the effect of using these methods on polytomous items is unknown.

Many techniques have been proposed for DIF analysis in the small sample case, including jackknife procedures, exact methods, Bayesian inference, log-linear smoothing, liberal alpha values, Angoff's transformed item difficulty plot, and conditional *p*-values. With the exception of some Bayesian methods, almost no research has applied these techniques to polytomous items. As developers continue to create tests with polytomously-scored items (Lane & Stone, 2006; Welch, 2006), particularly during a time where performance assessments are becoming increasingly popular, there will be a need to expand DIF methodology to ensure fairness and score validity. Psychometricians need to be aware of the statistical properties of polytomous DIF detection techniques for large and small sample sizes. Research should provide insight about which methods are preferred for small samples and which methods should be avoided.

Because there is a lack of research, a Monte Carlo simulation is proposed which allows for the study of the statistical properties of polytomous DIF techniques in the small sample case. The simulation will allow various techniques to be compared directly to each other under a variety of testing conditions. For example, the simulation will address the number of items used in the anchor test (defined here as the subset of test items used to produce an estimate of examinee ability) and the pattern of DIF. The

simulation will allow for the study of the Type I error rates for the various methods, to confirm that the "true" Type I error rate matches the level of significance defined by the analyst. The simulation will also allow for the study of statistical power, the probability that an item containing DIF is correctly flagged for DIF. The best methods will be those that control Type I error rates and have the highest statistical power.

Through this Monte Carlo simulation, the following research questions will be addressed:

1. What is the Type I error rate for the following small-sample polytomous DIF detection procedures: Mantel test, Liu-Agresti statistic, and HW3 statistic?
2. Of those methods listed in Question 1 that produce Type I error rates that are approximately equal to the a priori significance level ($\alpha = 0.05$), which of the methods yield the highest statistical power?
3. What is the Type I error rate for the following modifications to small-sample polytomous DIF detection procedures: empirical Bayesian, randomization methods, and log-linear smoothing?
4. Of those modifications listed in Question 3 where the simulated Type I error rates are approximately equal to the a priori significance level ($\alpha = 0.05$), which methods are the most statistically powerful?
5. Which of the following factors affect(s) the statistical power of the polytomous DIF detection methods listed in Questions 1 and 3: pattern of DIF in the studied item, reference group sample size, inclusion of the studied item in the anchor, and difference between the focal and reference group means?

In the next chapter, the literature relevant to this study will be reviewed. Chapter 3 will present the simulation methodology to be used to answer the above research questions. The results of the simulation will be provided in Chapter 4; discussion of these results will be provided in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

The purpose of this chapter is to summarize the terminology and methodology used in the study of polytomous differential item functioning (DIF). The first section establishes the definitions of common terms used in the DIF detection literature. Next, DIF detection methods for polytomous items will be described in detail. Variations proposed for small sample sizes will be discussed in the third section. Next, a review of Monte Carlo DIF studies from 1990 through the present will establish the factors most commonly addressed in published DIF simulations. Finally, a review of empirically-based DIF studies from 1990 through the present will establish how polytomous DIF detection methods and related modifications have been used in various settings.

Relevant Terms

Fairness, *bias*, and *differential item functioning* are three commonly used terms in the measurement literature, but the definitions used by the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) suggest that these terms are not synonymous. Fairness, as defined by the *Standards*, is "the principle that every test taker should be assessed in an equitable way" (American Educational Research Association et al., 1999, p. 175). It is a construct that applies not just to test items, but the development of test materials, administration of test items, and dissemination of test information. Fairness "must be considered in all aspects of the testing process" (Joint Committee on Testing Practices, 2004). It is not only an issue to be addressed by item writers, but also test users, administrators, educators, evaluators, and researchers (Schmeiser et al., 1995).

Whereas fairness applies to the entire test process, bias applies to the unintended consequences on an item or a test. The *Standards* define bias as "construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of

examinees" (American Educational Research Association et al., 1999, p. 172). In educational testing, these construct-irrelevant components may include gender, race, ethnicity, socioeconomic status, and religion. Other definitions of bias were established by Cleary (1968), Cleary and Hilton (1968), and Roussos and Stout (1996a).

There are several methods for evaluating item bias, including the use of sensitivity reviews, differential validity studies, and differential item functioning (DIF) detection methods. A sensitivity review, or fairness review, uses expert judgments to study the language and content of the items to determine any potential bias before the items are administered (Ramsey, 1993). Examples of potentially biased language or content include, but are not limited to, the use of stereotypes, the use of humor to insult, the use of specialized vocabulary that may put a group at a disadvantage, and the use of controversial topics (Zieky, 2006). Sensitivity reviews are common in large-scale testing programs, including the *Iowa Tests of Basic Skills* (Hoover, Dunbar, & Frisbie, 2003), the *SAT* (Educational Testing Service, 2009), and the *ACT* (ACT, 2008).

When items have been administered to examinees, statistical procedures can be applied to the item response strings or total scores to identify potential bias. Differential validity studies use an external criterion to study the relationship between total observed test score and subgroup membership (Anastasi & Urbina, 1997; Cole & Moss, 1989). Studies for educational tests employ external criteria such as high school grade point average (GPA), college GPA, or a related test. The analyst compares the correlation between the criterion and the total test score (e.g. Wood & Ansley, 2008) or the prediction equations of total test score regressed on the criterion (e.g. Sireci & Talento-Miller, 2006) for the two subgroups. Differences in the Pearson correlations, regression slopes, or regression intercepts suggest that the test may function differently for different subgroups.

Predictive validity studies may be difficult to produce if an appropriate criterion score is not available to the analyst (Scheuneman, 1976). It is also possible that a

criterion variable is available, but only for a convenience sample. To get around this problem, researchers developed ways of assessing item bias using information internal to the test being analyzed. During the 1970s, methods were proposed using differences in item difficulties, differences in item discriminations, differences in item response theory (IRT) item characteristic curves, and various chi-square tests (Shepard, Camilli, & Averill, 1981). These item bias detection methods soon became known as *differential item functioning* (*DIF*) detecting methods.

The *Standards* define *differential item functioning* as "a statistical property of a test item in which different groups of test takers who have the same total test score have different average item scores" (American Educational Research Association et al., 1999, p. 175). Let Y be the score obtained on the studied item, which may be dichotomous or polytomous. Assume that individual examinees can be allocated to the focal (minority) group or the reference (majority) group. Define X as the observed total score over a predetermined subset of test items (known as an *anchor*). Then, DIF is present if, for the two groups of examinees ($i = 1$ and $i = 2$), there exists at least one value of X such that

$$E[Y \mid X, i = 1] \neq E[Y \mid X, i = 2] \tag{1a}$$

(Chang et al., 1996). An alternative definition of DIF uses latent ability θ in lieu of observed total test score; DIF is present if there exists at least one value of θ such that:

$$E[Y \mid \theta, i = 1] \neq E[Y \mid \theta, i = 2] \tag{1b}$$

(Chang et al., 1996). These equations are the foundation for current DIF detecting methods.

If Equation 1a or 1b holds, then DIF is present, and the item characteristic curves (or item-test regressions) will differ. These differences can be classified into one of several patterns. Most researchers define *uniform DIF* as the translation of one item characteristic curve (ICC) or item response function (IRF) to the left or right a number of

units (Mellenbergh, 1982). Using the mathematical definition of observed score DIF, uniform DIF occurs when:

$$E[Y \mid X = x, i = 1] = E[Y \mid X = x + c, i = 2] \quad (2)$$

for some non-zero constant *c* (Hanson, 1998; Mellenbergh, 1982). Figure A1a provides an example of an item that exhibits uniform DIF. Most dichotomous DIF detection methods are able to detect uniform DIF (Clauser & Mazor, 1998).

If Equation 1a or 1b holds, but Equation 2 does not hold for any value of *c*, then *non-uniform DIF* has occurred[1]. If the ICCs (IRFs) exhibit non-uniform DIF but do not intersect on the entire domain of X (or θ), then *unidirectional DIF* has occurred (Hanson, 1998; Mellenbergh, 1982). Mathematically, this occurs when

$$E[Y \mid X = x, i = 1] - E[Y \mid X = x, i = 2] > 0 \qquad \forall x$$

or

$$E[Y \mid X = x, i = 1] - E[Y \mid X = x, i = 2] < 0 \qquad \forall x$$

(Hanson, 1998). Figure A1b shows an example of unidirectional DIF. If the ICCs (IRFs) exhibit non-uniform DIF and intersect at some point in the domain, then *crossing DIF* has occurred (Hanson, 1998; Mellenbergh, 1982). In this case, assuming the ICCs (IRFs) are not identical, then there exists values of x such that

$$E[Y \mid X = x, i = 1] - E[Y \mid X = x, i = 2] = 0$$

(Hanson, 1998). Crossing DIF can be especially problematic, as DIF favors one group over a subset of the domain X (θ) but favors the other group over a different non-overlapping subset of the domain X (θ). Figure A1c shows an example of crossing DIF. Many DIF detection methods have difficulty detecting DIF when unidirectional or crossing DIF is present (Clauser & Mazor, 1998; Marañón, Garcia, & San Luis Costas,

[1] Hanson (1998) preferred the term *parallel DIF* when Equation 2 holds, while defining uniform DIF with respect to conditional odds ratios. This distinction will not be used in this research.

1997; Mazor, Clauser, & Hambleton, 1994; Narayanan & Swaminathan, 1996; Swaminathan & Rogers, 1990). Methods specifically designed to address non-uniform DIF include logistic regression methods (Swaminathan & Rogers, 1990) and their logit model counterparts (Mellenbergh, 1982).

DIF patterns become more complex when considering polytomous items. Of the *J* possible scores that an examinee can receive on a polytomous item, zero, one, several, or all scores could exhibit DIF. Furthermore, some of the *J* scores may be biased in favor of the focal group, while others may be biased in favor of the reference group. A two-dimensional taxonomy has been developed to distinguish among various patterns of DIF (Penfield, Alvarez, & Lee, 2009). The first dimension, pervasiveness, identifies the number of scores (out of *J*) that contain DIF. If all *J* categories contain DIF, then the item contains *pervasive DIF*. If some of the *J* categories contain DIF, then the item contains *non-pervasive DIF*. The second dimension, consistency, identifies the direction and magnitude of DIF. *Constant DIF* occurs when the contaminated scores differ by the same amount and always favor the same group. *Convergent DIF* occurs when the contaminated scores differ by different amounts, but always favor the same group. *Divergent DIF* occurs when the contaminated scores differ by different amounts; some differences favor the reference group while others favor the focal group. Figure A2 contains examples of the six types of polytomous DIF, according to the above taxonomy. The first column of graphs shows pervasive DIF, and the second column shows non-pervasive DIF. The first row shows constant DIF, the second row shows convergent DIF, and the third row shows divergent DIF.

Prior to the development of a polytomous DIF taxonomy, researchers used their own terminology to describe DIF patterns. The terms *balanced DIF*, *constant DIF*, *shift-low DIF*, and *shift-high DIF* were common to several researchers (Chang et al., 1996; Su & Wang, 2005; Wang & Su, 2004a). Balanced DIF is a special case of divergent, non-pervasive DIF, where the first score category contains DIF that favors one group and the

J^{th} score category contains the same magnitude DIF that favors the opposite group. Constant DIF, as defined by these authors, is more accurately described as constant, pervasive DIF. Shift-low and shift-high DIF describe the situation when only the lowest (or highest) score category contains DIF. Using the polytomous DIF taxonomy, this would be classified as constant, non-pervasive DIF.

If the item is defined as coming from Samejima's graded response model, Bock's nominal model, or Muraki's partial credit model, then all of the above patterns of DIF can be defined as the addition of a constant to the difficulty parameters of one or more score categories while keeping the discrimination parameter the same. Many simulation studies have defined DIF in this way. However, there is an alternative method for defining DIF in a model. Some researchers have defined DIF as an item that measures a "nuisance dimension" in addition to a true ability dimension (Ackerman, 1992; Roussos & Stout, 1996a; Shealy & Stout, 1993b). The inclusion of a nuisance dimension produces potential construct-irrelevant variance that may result in DIF. The statistical theory behind the SIBTEST procedure (Shealy & Stout, 1993b) defines DIF using multidimensional item response theory, auxiliary dimensions, and nuisance dimensions.

Polytomous DIF Detection Methods

Methodology reviews reveal that there are many ways of detecting DIF in polytomous items (Penfield & Lam, 2000; Penfield & Camilli, 2007; Potenza & Dorans, 1995). This section will describe techniques that can be used when test data are summarized in a three-dimensional contingency table, which include the Mantel test, Cox's β, the Liu-Agresti statistic, HW1, and HW3. Additional DIF detection techniques for polytomous items will be described briefly at the end of this section.

The notation used for these test statistics is summarized in Table A1 for a single studied item (Zwick et al., 1993). Rows represent the focal and reference group examinees ($i = F, R$). Row value is a nominal variable with fixed marginal totals.

Columns represent the possible scores possible on the studied item ($j = 1, \ldots, J$). Those examinees in the first column received a score of y_1, those in the second column received a score of y_2, and those in the last column received a score of y_J. It is assumed that column totals are random and $y_1 < y_2 < \ldots < y_J$, so j is ordered. Examinees are divided into categories based on the possible scores obtainable on the anchor test (x_1, x_2, …, or x_K). These categories form K subtables in the 2 x J x K contingency table. In the small sample case, it is usually necessary to combine adjacent subtables to create sample sizes adequate for analysis (Donoghue & Allen, 1993). Subtable marginal totals are assumed to be random, and k is ordered. As shown in Table A1, the value n_{ijk} is the number of examinees in group i that obtained a score of y_j on the studied item and obtained a score of x_k on the anchor test. Marginal sums are represented by "+" subscripts.

The Mantel Test

The Mantel-Haenszel statistic is arguably the most popular DIF detection statistic for dichotomous items, as evidenced by its use in large-scale testing programs (ACT, 2008; Forsyth, Ansley, Feldt, & Alnot, 2003; Hoover et al., 2003) and in research studies (Ackerman & Evans, 1992; Allen & Donoghue, 1996; Camilli & Penfield, 1992; Clauser, Mazor, & Hambleton, 1991; Clauser, 1993; Harvey, 1990; Mazor, Clauser, & Hambleton, 1992; Mazor et al., 1994; Raju, Bode, & Larsen, 1989; Ryan, 1991; Shin, 1992; Uttaro & Millsap, 1994). The Mantel-Haenszel test for dichotomous items and the Mantel test for polytomous items are easy to compute in most statistical software programs.

The original Mantel-Haenszel (MH) statistic is a non-parametric procedure used to study 2 x 2 x K contingency tables, where K is the number of 2 x 2 subtables (Conover, 1999). The statistic was first used in medical research to assess if table factors were independent (Mantel & Haenszel, 1959). An extension for I x J x K tables, where there are K tables with I rows and J columns, was created by Mantel (1963) and is called the

Mantel test. The MH statistic was first applied to DIF analysis in the late 1980's (Dorans & Holland, 1993; Holland, 1985; Holland & Thayer, 1988). Its popularity led to its use as the Mantel test in polytomous items several years later (Welch & Hoover, 1993; Welch & Miller, 1995; Zwick et al., 1993). The MH and Mantel statistics seem to have replaced earlier chi-square statistics developed by Scheuneman (1979) and Camilli (1979).

To test the null hypothesis of no DIF in the studied item versus the alternative that there is DIF in the studied item, the analyst calculates the test statistic

$$T_k = \sum_{j=1}^{J} y_j n_{Fjk} \ ,$$

the weighted sum of scores in the focal group, within table *k*. Under the hypothesis of no DIF, one can assume that the rows, columns, and tables are independent; it can also be assumed that the frequency counts are distributed as hypergeometric variables, given the fixed row totals and random column and table totals. Assuming independence and assuming the frequency counts are distributed as hypergeometric variables, where the expected value of T_k is

$$E(T_k) = \frac{n_{F+k}}{n_{++k}} \sum_{j=1}^{J} y_j n_{+jk} \ ,$$

and the variance of T_k is

$$V(T_k) = \frac{n_{R+k} n_{F+k}}{n_{++k}^2 (n_{++k} - 1)} \left[\left(n_{++k} \sum_{j=1}^{J} y_j^2 n_{+jk} \right) - \left(\sum_{j=1}^{J} y_j n_{+jk} \right)^2 \right].$$

To test the null hypothesis, the Wald statistic

$$\chi^2 = \frac{\left[\sum_{k=1}^{K} T_k - \sum_{k=1}^{K} E(T_k) \right]^2}{\sum_{k=1}^{K} V(T_k)}$$

is used, where χ^2 is distributed as a chi-square variable with 1 degree of freedom. Large values of χ^2 provide evidence that the studied item contains DIF. The Mantel test can be computed using the *R* function given in the Appendix (Table B1). To illustrate the Mantel test, consider the example data in Table A2. Based on these data, $\Sigma T_k = 74$, $\Sigma E(T_k) = 61.69$, $\Sigma V(T_k) = 19.81$, $\chi^2 = 7.65$, and the *p*-value = 0.0057. Therefore, the analyst rejects the null hypothesis and flags the item for DIF.

If the item under consideration is dichotomous, then $y_1 = 0$ and $y_2 = 1$. Using these values, the Mantel test reduces to the MH test statistic (see Conover, 1999). A more general test statistic is available for the case where the analyst wishes to study more than 2 subgroups simultaneously (Fidalgo & Madeira, 2008), though DIF analysts have not typically considered the case when $I > 2$.

The Mantel test and the MH test have many advantages. The computations are easy, and the method is conceptually easy to understand. The Mantel and MH test have been extensively studied and documented, and many research studies have used the MH test as a benchmark for comparison. One disadvantage, however, is the inability for the Mantel and MH tests to detect non-uniform DIF (Marañón et al., 1997). A variation of the MH test to address non-uniform DIF has been proposed (Mazor et al., 1994), though researchers seem to prefer methods based on logistic regression (Miller & Spray, 1993; Swaminathan & Rogers, 1990; Zumbo, 1999) over this MH adaptation.

Cox's β

A mathematically equivalent, but conceptually different, approach to the Mantel test is known as Cox's β statistic (Cox, 1958). This statistic is an estimate of the non-centrality parameter of a non-central multivariate hypergeometric distribution. Note that the null hypothesis of the Mantel test assumes that the frequency counts are derived from a (central) multivariate hypergeometric distribution, so it is reasonable that the alternative hypothesis assumes a non-central multivariate hypergeometric distribution with

parameter β. Values of β significantly different from zero suggest that the alternative hypothesis holds, requiring the analyst to flag the studied item for DIF. Cox's β was proposed as a DIF statistic that could be used in Bayesian inference (Camilli & Congdon, 1999) as well as an effect size for DIF detection. Cox's β is not as commonly applied as the Mantel test, but it has been cited in the works of Penfield (Penfield & Algina, 2003; Penfield & Algina, 2006; Penfield, 2007; Penfield & Camilli, 2007).

The derivation of Cox's β uses the notation found in Table A1. It is assumed that the row totals are fixed, the column and table totals are random, and the data come from a multivariate hypergeometric distribution. An estimate of the sample statistic $\hat{\beta}$, using the notation of the previous section, is (Camilli & Congdon, 1999)

$$\hat{\beta} = \frac{\sum_{k=1}^{K} T_k - \sum_{k=1}^{K} E(T_k)}{\sum_{k=1}^{K} V(T_k)}.$$

The variance of the estimate of the non-centrality parameter is

$$V(\hat{\beta}) = \frac{1}{\sum_{k=1}^{K} V(T_k)}.$$

To test the null hypothesis that $\beta = 0$ (the studied item contains no DIF) versus the alternative that $\beta \neq 0$ (the studied item contains DIF), the test statistic

$$Z = \frac{\hat{\beta}}{\sqrt{V(\hat{\beta})}}$$

can be used. Under the null hypothesis, Z is distributed as a standard normal distribution. Equivalently, when testing the two-sided hypothesis, Z^2 is distributed as chi-square with 1 degree of freedom. Larger values of Z^2, or larger values of $| Z |$, provide evidence that the item contains DIF.

Table B2 contains an *R* function that calculates Cox's β and its test statistic Z. To illustrate this method, Table A2 will be used again as an example. Given the 2 x 3 x 4

contingency table, $\hat{\beta}$ = 0.62, $V(\hat{\beta})$ = .05, Z = 2.77, and Z^2 = 7.65. Note that Z^2 is equivalent to the value of χ^2 from the Mantel test. From these results, the analyst would reject the null hypothesis and flag the studied item for DIF.

The hypothesis test for Cox's β can be shown to be equivalent to the test statistic χ^2 used by the Mantel test. Despite this equivalency, the true utility of Cox's β comes from its use as an effect size. Because the sampling distribution of $\hat{\beta}$ is symmetric around zero and approximately normally distributed (under the assumption of no DIF), it is an ideal statistic to use in the Bayesian methods to be described later in this chapter. A different approach to 2 x J x K contingency tables, to be presented next, uses the definition of odds ratio to develop a test statistic.

Liu-Agresti Statistic

The Liu-Agresti statistic was originally proposed as a way of calculating an aggregate odds ratio for I x J x K contingency tables (Liu & Agresti, 1996). The statistic was eventually recommended as a way of evaluating a polytomous item for DIF (Penfield & Algina, 2003; Penfield, 2007). Though it has not gained traction in the literature, it has the potential of supplementing the dichotomous DIF detection techniques based on odds ratios (Penfield & Camilli, 2007).

To develop the Liu-Agresti statistic, a single 2 x 2 table (say, table k') is assumed. Using the notation of Table A1, the sample odds ratio (Agresti, 2003) for table k' is defined as

$$\hat{\psi}_{k'} = \frac{n_{F1k'} n_{R2k'}}{n_{F2k'} n_{R1k'}}.$$

An odds ratio of 1 suggests that the odds of a correct response by the focal group are equivalent to the odds of a correct response by the reference group. Values of an odds ratio can range from 0 to positive infinity.

The analyst needs a way of combining the *K* odds ratios into a single omnibus statistic. A common approach used in dichotomous DIF is the Mantel-Haenszel odds ratio (Penfield & Camilli, 2007):

$$\hat{\psi}_{MH} = \frac{\sum_{k=1}^{K} \frac{n_{F1k} n_{R2k}}{n_{++k}}}{\sum_{k=1}^{K} \frac{n_{F2k} n_{R1k}}{n_{++k}}}.$$

Because the range of the odds ratio is not symmetric around 1, many researchers take the natural logarithm of $\hat{\psi}_{MH}$. This produces a statistic which ranges from negative to positive infinity and is symmetric around zero. The variance of log $\hat{\psi}_{MH}$ was derived by Hauck (Agresti, 2003; 1979) as

$$\begin{aligned} V(\log \hat{\psi}_{MH}) = & \frac{1}{2\left(\sum_{k=1}^{K} \frac{n_{11k} n_{22k}}{n_{++k}}\right)^2} \sum_{k=1}^{K} \frac{(n_{11k} + n_{22k}) n_{11k} n_{22k}}{n_{++k}^2} \\ & + \frac{1}{2\left(\sum_{k=1}^{K} \frac{n_{12k} n_{21k}}{n_{++k}}\right)^2} \sum_{k=1}^{K} \frac{(n_{12k} + n_{21k}) n_{12k} n_{21k}}{n_{++k}^2} \\ & + \frac{1}{2\left(\sum_{k=1}^{K} \frac{n_{11k} n_{22k}}{n_{++k}}\right)\left(\sum_{k=1}^{K} \frac{n_{12k} n_{21k}}{n_{++k}}\right)} \sum_{k=1}^{K} \left[\frac{(n_{12k} + n_{21k}) n_{12k} n_{21k}}{n_{++k}^2} + \frac{(n_{11k} + n_{22k}) n_{11k} n_{22k}}{n_{++k}^2} \right] \end{aligned}$$

though a simplified formula for the variance is given by Penfield and Camilli (2007):

$$V(\log \hat{\psi}_{MH}) = \frac{\sum_{k=1}^{K} \frac{1}{n_{++k}^2} \left(n_{11k} n_{22k} + n_{12k} n_{21k} \log \hat{\theta}_{MH}\right)\left(n_{11k} + n_{22k} + n_{12k} \log \hat{\theta}_{MH} + n_{21k} \log \hat{\theta}_{MH}\right)}{2\left(\sum_{k=1}^{K} \frac{n_{11k} n_{22k}}{n_{++k}}\right)^2}$$

To test the null hypothesis that log $\psi_{MH} = 0$ (no DIF is present) versus the alternative hypothesis that log $\psi_{MH} \neq 0$ (DIF is present in the studied item), the analyst can produce the Wald statistic

$$Z^2 = \frac{(\log \hat{\psi}_{MH})^2}{V(\log \hat{\psi}_{MH})}.$$

Z^2 is distributed as a chi-square variable with 1 degree of freedom. Large values of Z^2 provide evidence that the item contains DIF and should be flagged.

Liu and Agresti (1996) proposed a way to generalize the above test to the I x J x K case. For the purpose of DIF analysis, I is set to 2. Liu and Agresti defined the quantity

$$n^*_{ij'k} = \sum_{j=1}^{j'} n_{ijk}$$

as the number of examinees in group i and anchor score category k that achieved a score of $y_{j'}$ or less. When partnered with the quantity $n_{i+k} - n^*_{ij'k}$, the number of examinees in group i and anchor score category k that achieved a score higher than $y_{j'}$, an odds ratio can be created. For ordered column variables, Liu and Agresti proposed using

$$\hat{\psi}_{LA} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J-1} \frac{n^*_{Fjk}(n_{R+k} - n^*_{Rjk})}{n_{++k}}}{\sum_{k=1}^{K}\sum_{j=1}^{J-1} \frac{n^*_{Rjk}(n_{F+k} - n^*_{Rjk})}{n_{++k}}}$$

as a common odds ratio. For the dichotomous case where $J = 2$, $\hat{\psi}_{LA}$ simplifies to the Mantel-Haenszel common odds ratio $\hat{\psi}_{MH}$.

Like the Mantel-Haenszel common odds ratio, analysts prefer to work with the natural logarithm of the Liu-Agresti statistic. This transforms the range of $\hat{\psi}_{LA}$ so values are symmetric around zero. The variance of the natural logarithm of the Liu-Agresti statistic, using their original notation (Liu & Agresti, 1996), is

$$V(\log \hat{\psi}_{LA}) = \frac{\sum_{k=1}^{K} \hat{\xi}_k(\hat{\psi}_{LA})}{\hat{\psi}^2_{LA}\left(\sum_{k=1}^{K}\sum_{j=1}^{J-1} \frac{n^*_{Rjk}(n_{F+k} - n^*_{Fjk})}{n_{++k}}\right)^2},$$

where

$$\hat{\xi}_k(\hat{\psi}_{LA}) = \sum_{j'=1}^{J-1}\sum_{j=1}^{J-1}\hat{\phi}_{jj'k}(\hat{\psi}_{LA})$$

and

$$\hat{\phi}_{jj'k}(\hat{\psi}_{LA}) =$$
$$\frac{n_{R+k}n_{F+k}}{n_{++k}^2}\left\{\frac{\hat{\psi}_{LA}n^*_{Fjk}(n_{R+k}-n^*_{Rj'k})}{n_{R+k}}\left[1+(\hat{\psi}_{LA}-1)\frac{n^*_{Fj'k}}{n_{F+k}}\right]+\frac{n^*_{Rjk}(n_{F+k}-n^*_{Fj'k})}{n_{F+k}}\left[\hat{\psi}_{LA}-(\hat{\psi}_{LA}-1)\frac{n^*_{Rj'k}}{n_{R+k}}\right]\right\}$$

A Wald statistic can be used to test the null hypothesis that the logarithm of the Liu-Agresti common odds ratio equals 0 (DIF is not present) versus the alternative hypothesis that log $\hat{\psi}_{LA} \neq 0$ (DIF is present). The test statistic

$$Z^2 = \frac{(\log\hat{\psi}_{LA})^2}{V(\log\hat{\psi}_{LA})}$$

is distributed as chi-square with one degree of freedom. Larger values of Z^2 support the alternative hypothesis. A directional hypothesis test is also available by using

$$Z = \frac{\log\hat{\psi}_{LA}}{\sqrt{V(\log\hat{\psi}_{LA})}}.$$

The Liu-Agresti statistic is not readily available in *SAS* or *R*, although an *R* function is presented in Table B3. To illustrate $\hat{\psi}_{LA}$, Table A2 will again be used as an example. Based on the information presented, $\hat{\psi}_{LA}$ = 0.33 and log $\hat{\psi}_{LA}$ = -1.11. The variance of log $\hat{\psi}_{LA}$ = .16, so the test statistic Z^2 = 7.59. The *p*-value of Z^2 is 0.0059, so the analyst would reject the null hypothesis and flag the item for DIF.

HW1 and HW3

The HW1 and HW3 statistics were likely among the first attempts to produce a DIF detection statistic for polytomous items (Welch & Hoover, 1993; Welch & Miller, 1995). Both statistics are based on the two sample *t*-test. The HW1 and HW3 statistics

were studied throughout the 1990s, often as a comparison to other DIF detection techniques (Zwick et al., 1993; Zwick & Thayer, 1996).

The HW1 statistic assumes that the number of examinees in each stratum *k* is approximately equal (Welch & Hoover, 1993). The average score on the studied item for the focal group examinees in anchor stratum *k* is:

$$\sum_{j=1}^{J} \frac{y_j n_{Fjk}}{n_{F+k}};$$

a similar statistic for the reference group simply replaces the subscript *F* with *R*. A *t*-test statistic used to compare the focal and reference groups in stratum *k* is

$$t_k = \frac{\left(\sum_{j=1}^{J} \frac{y_j n_{Fjk}}{n_{F+k}}\right) - \left(\sum_{j=1}^{J} \frac{y_j n_{Rjk}}{n_{R+k}}\right)}{\sqrt{\frac{S^2_{F+k} n_{F+k} + S^2_{R+k} n_{R+k}}{n_{F+k} + n_{R+k} - 2}\left(\frac{1}{n_{F+k}} + \frac{1}{n_{R+k}}\right)}},$$

where S^2_{F+k} and S^2_{R+k} are the variances of the studied item scores for the focal and reference groups on stratum *k*. All other variables are defined in Table A1.

Defining

$$\nu_k = n_{F+k} + n_{R+k} - 2$$

as the degrees of freedom in the *k*th *t*-test, t_k and ν_k can be combined to produce the HW1 statistic:

$$HW1 = \frac{\sum_{k=1}^{K} t_k}{\sqrt{\sum_{k=1}^{K} \frac{\nu_k}{\nu_k - 2}}}.$$

HW1 is used to test the null hypothesis that the sum of the *t*-test statistics equals zero (the studied item contains no DIF) versus the alternative hypothesis that the sum of the *t*-test statistics does not equal 0 (the studied item contains DIF). HW1 is distributed as a standard normal variable. Large values of |HW1| suggest that the alternative hypothesis

holds. Using Table A2 as an example, the sum of the *t*-test statistics equals 5.62, HW1 is 2.68, and the *p*-value is 0.0073. This provides evidence that the studied item should be flagged for DIF.

A modification to the HW1 statistic is preferred when the number of examinees in each stratum differs (Welch & Hoover, 1993). The HW3 statistic uses effect sizes for the *t*-test to assess an item for DIF. For stratum *k*, the effect size is

$$ES_k = \frac{\left(\sum_{j=1}^{J} \frac{y_j n_{Fjk}}{n_{F+k}}\right) - \left(\sum_{j=1}^{J} \frac{y_j n_{Rjk}}{n_{R+k}}\right)}{\sqrt{\frac{S_{F+k}^2 n_{F+k} + S_{R+k}^2 n_{R+k}}{n_{F+k} + n_{R+k} - 2}}}.$$

A correction factor necessary to make the effect size unbiased is

$$CF_k = 1 - \frac{3}{4\nu_k - 1}.$$

The unbiased effect size, denoted d_k, is

$$d_k = CF_k \times ES_k.$$

The variance of d_k is

$$S_k^2 = CF_k^2 \frac{\nu_k}{\nu_k - 2}\left(\frac{1}{n_{F+k}} + \frac{1}{n_{R+k}}\right) + d_k^2\left(\frac{CF_k^2 \nu_k}{\nu_k - 2} - 1\right).$$

The unbiased effect sizes d_k can be combined to produce a weighted sum of effect sizes *D*:

$$D = \frac{\sum_{k=1}^{K} \frac{d_k}{S_k^2}}{\sum_{k=1}^{K} \frac{1}{S_k^2}}.$$

The HW3 statistic

$$HW3 = \frac{D}{\sqrt{\sum_{k=1}^{K} \frac{1}{S_k^2}}}$$

is distributed as a standard normal variable. HW3 is used to test the null hypothesis that the sum of the unbiased effect sizes equals zero (the studied item contains no DIF) versus the alternative hypothesis that the sum of the unbiased effect sizes does not equal zero (the studied item contains DIF). Large values of | HW3 | provide evidence that the item should be flagged for DIF.

A function for calculating HW3 in *R* is available in Table B4. Table A2 will once again provide an example of the calculation of HW3. Using the contingency tables, it can be shown that $D = 0.54$, HW3 = 2.59, and the *p*-value for the test is 0.0097. Thus, the analyst rejects the null hypothesis and flags the item for DIF.

Additional Methods for Identifying Polytomous DIF

The simulation to be described in Chapter 3 will utilize the DIF detection methods described above. There are several other methods, however, that have been presented for assessing polytomous items for DIF. These methods include the standardized mean difference (SMD) statistic, SIBTEST, logistic regression, and logistic discriminant function analysis. This section will briefly describe these methods, although they will not be investigated in the simulation.

The dichotomous version of the standardized mean difference (SMD) statistic was proposed as a way to condense the large amounts of information in an item-test regression into a single value (Dorans & Kulick, 1983; Dorans & Kulick, 1986). The dichotomous SMD statistic is frequently associated with the dichotomous MH statistic since both statistics are reasonably easy to compute (Donoghue, Holland, & Thayer, 1993; Dorans & Holland, 1993; Dorans & Kulick, 2006; Lu, 1996; Wright, 1986; Zwick & Thayer, 1996).

Using the notation from Table A1, the polytomous SMD statistic is typically defined as

$$SMD = \left(\sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \frac{\sum_{j=1}^{J} y_j n_{Fjk}}{n_{F+k}} \right) - \left(\sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \frac{\sum_{j=1}^{J} y_j n_{Rjk}}{n_{R+k}} \right)$$

(Zwick & Thayer, 1996). Large values of | SMD | provide evidence that the studied item contains DIF. Two standard errors are available for SMD, depending on the distributional assumption made about the data (Zwick & Thayer, 1996). Zwick and Thayer (1996) recommended the hypergeometric version of the SMD variance over the independently distributed multinomial version of the SMD variance.

To test the null hypothesis that the population value of SMD is zero (DIF does not exist) versus the alternative that the population value of SMD differs from zero (DIF exists in the studied item), the analyst can use

$$Z = \frac{SMD}{\sqrt{V(SMD)}}$$

as a test statistic. Z is distributed as a standard normal variable. Large values of | Z | provide evidence that the studied item should be flagged for DIF. Table B5 provides a function for computing SMD in *R*. Note that the function computes the variance under the multivariate hypergeometric and the independent multinomial assumptions. Using Table A2 as a sample data set, SMD equals 0.50. The variance of SMD under the multivariate hypergeometric distribution is 0.03225, which implies that Z = 2.78 and *p*-value = 0.0054. If one were to assume independent multinomial distributions, then the variance of SMD is 0.03, which implies that Z = 2.97 and *p*-value 0.0030. In either case, the analyst would reject the null hypothesis and flag the item for DIF.

The SIBTEST statistic was developed as a way of addressing the multi-dimensionality definition of DIF (Chang et al., 1996; Shealy & Stout, 1991; Shealy &

Stout, 1993b). SIBTEST is supported with a large research base (Banks & Walker, 2006; Bolt & Stout, 1996; Bolt, 2000; Douglas, Stout, & DiBello, 1996; Fidalgo, Ferreres, & Muñiz, 2004b; Finch, 2005; Gotzman & Boughton, 2004; Jiang & Stout, 1998; Klockars & Lee, 2008; Lee & Klockars, 2005; Nandakumar & Roussos, 1997; Nandakumar & Roussos, 2001; Narayanan & Swaminathan, 1994; Ross, 2007; Roussos & Stout, 1996b). The SIBTEST statistic relies on a theoretical approach that attempts to approximate item response theory bias detection with observed data and classical test theory.

SIBTEST attempts to calculate the following weighted bias index using observed scores. Let $E[Y \mid \theta, i]$ be the expected score on the studied item for examinees in group i with a given latent ability θ. One way of measuring the amount of DIF in an item is to compute the (weighted) area between $E[Y \mid \theta, R]$ and $E[Y \mid \theta, F]$ over the domain θ (Shealy & Stout, 1993b):

$$B_{\text{weighted}} = \int_{\theta} f(\theta \mid F)(E[Y \mid \theta, R] - E[Y \mid \theta, F])d\theta ,$$

where $f(\theta \mid F)$ is the distribution of latent abilities for the focal group. The area $B_{\text{unweighted}}$ should equal zero if DIF is not present in the studied item. For dichotomous items, closed-form equations exist to calculate this area and its standard error for the 1-parameter and 2-parameter logistic models (Raju, 1988; Raju, 1990).

Using the notation of Table A1, one can approximate $f(\theta \mid F)$ as n_{F+k} / n_{F++}, the proportion of focal group examinees in anchor score category k, and

$$E[Y \mid i, \theta] \approx E[Y \mid i, X = k] = \overline{Y}_{ik} = \frac{\sum_{j=1}^{J} y_j n_{ijk}}{n_{i+k}} .$$

Substituting these approximations into the formula for B_{weighted} yields the uncorrected SIBTEST statistic

$$B_{\text{uncorrected}} = \sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \left(E[Y \mid i = R, X = k] - E[Y \mid i = F, X = k] \right).$$

Note that $B_{\text{uncorrected}}$ is the same (ignoring sign) as the SMD statistic defined in the previous section. The original research on SIBTEST acknowledged that $B_{\text{uncorrected}}$ and *SMD* were derived independently (Shealy & Stout, 1993b).

Shealy and Stout (1993b) argued that $B_{\text{uncorrected}}$ can be an inflated estimate of the true amount of bias. To counteract this, the raw anchor scores are converted to estimated true scores using Kelley's regressed scores. Kelley's regressed estimators are then used to produce adjusted estimates of $E[Y \mid i, \theta]$, labeled $E^*[Y \mid i, X]$, using a linear approximation. These estimates are then substituted into $B_{\text{uncorrected}}$ to produce a corrected SIBTEST statistic (Shealy & Stout, 1993b):

$$B_{\text{corrected}} = \sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \left(E^*[Y \mid R, X = k] - E^*[Y \mid F, X = k] \right).$$

This is the polytomous SIBTEST statistic proposed by Chang, Mazzeo, and Roussos (1996).

To test the null hypothesis that there is no DIF present in the studied item ($B_{\text{weighted}} = 0$) versus the alternative that there is DIF present in the studied item ($B_{\text{weighted}} \neq 0$), the test statistic

$$Z = \frac{B_{\text{corrected}}}{\sqrt{\sum_{k=1}^{K} \left(\frac{n_{F+k}}{n_{F++}} \right)^2 \left(\frac{V(Y \mid k, R)}{n_{R+k}} + \frac{V(Y \mid k, F)}{n_{F+k}} \right)}}$$

is calculated, where $V(Y \mid k,i)$ is the variance of the studied item scores for group *i* and observed score category *k*. *Z* is distributed as a standard normal variable. Large values of $|Z|$ provide evidence that the null hypothesis should be rejected, and the item should be flagged for DIF. Table B6 contains an *R* function that computes the SIBTEST statistic. Using the data in Table A2, along with some additional information, an analyst can calculate the SIBTEST statistic. Assume that the midpoints of the anchor score intervals are used for X_{ki}, the average anchor score for the focal group is 14.5, the average anchor score for the reference group is 15.5, and the score reliability is 0.80 for both groups.

Then, it can be shown that $B_{\text{corrected}} = -0.50$, the variance of $B_{\text{corrected}}$ is 0.03, $Z = -2.82$, and the p-value is 0.0048.

When the item characteristic curves (or item-test regressions) for the focal group cross the item characteristic curves for the reference group, the above DIF detection methods may not be appropriate. Furthermore, the above tests can be considered omnibus tests; they are not designed to evaluate DIF at a specific score level. Polytomous logistic regression provides the analyst a way of evaluating DIF with these two considerations in mind.

Polytomous logistic regression (Agresti, 2003; French & Miller, 1996) is a DIF detection method that requires the analyst to dichotomize the responses to the polytomously scored studied item and uses this dichotomous response variable in a logistic regression model with anchor score X, focal/reference group indicator variable G, and interaction term $X*G$ as independent variables. The difference between model deviances, where deviance is defined as -2 times the natural logarithm of the likelihood function, is used as a test statistic for assessing DIF. Depending on the models used, either crossing or non-crossing DIF can be assessed. To assess a single item for DIF using polytomous logistic regression, $3(J - 1)$ regression models are necessary.

As a way to reduce the number of regression models necessary to evaluate DIF, logistic discriminant function analysis (LDFA) has been proposed as an alternative to polytomous logistic regression (Miller & Spray, 1993). In LDFA, the subgroup indicator variable G is used as the dependent variable, and anchor score X, studied item score Y and interaction term $X*Y$ are used as independent variables in the model. Because the dependent variable is defined as a dichotomous variable, only three regression models are necessary to test an item for crossing and non-crossing DIF. Like polytomous logistic regression, the difference between model deviances is used as a test statistic.

Summary

This section presented common approaches to DIF detection in the polytomous case: the Mantel test, Cox's β statistic, the Liu-Agresti common odds ratio, HW1 & HW3. Additional methods based on contingency tables, the standardized mean difference statistic and the SIBTEST statistic, were also presented. Logistic regression methods are available to study uniform and non-uniform DIF, including polytomous logistic regression and logistic discriminant function analysis (LDFA). Although methods using item response theory (IRT), including Raju's area measures (Raju, 1988; Raju, 1990) and the IRT likelihood ratio test (Thissen, Steinberg, & Gerrard, 1986; Thissen et al., 1988; Thissen et al., 1993) have been suggested, IRT methods usually require large sample sizes. Therefore, their use in testing situations with small examinee sizes is not recommended.

In the next section, a number of solutions to the small sample size problem will be described. These modifications, when used with the contingency table methods described above, could be useful to the analyst who must conduct a DIF analysis when examinees sample sizes are small.

Modifications for Small Sample Sizes

Bayesian Modifications

Bayesian statistics provide an alternative approach to the traditional frequentist view of parameter estimation, hypothesis testing, and confidence intervals. While frequentist methods assume that the parameter of interest, ξ, is a constant value, Bayesian methods assume that ξ is a random variable. Prior to analysis, Bayesian statisticians will describe their beliefs about ξ through a prior distribution $g(\xi)$. After data, **x**, are collected, the statistician will update the prior through the relationship

$$f(\xi \mid \mathbf{x}) \propto g(\xi) L(\mathbf{x} \mid \xi),$$

where $L(\mathbf{x} \mid \xi)$ is the likelihood function from the data. The updated distribution $f(\xi \mid \mathbf{x})$ is called the posterior distribution. Inferences on ξ are based on the posterior distribution (Gill, 2002).

Bayesian methods have slowly been adapted to psychometric applications. One such example is the use of Bayesian modal estimates in calculating an examinee's latent ability in item response theory (Wainer & Mislevy, 2000). The earliest use of Bayesian methods in DIF was Longford's model of item bias (Longford, Holland, & Thayer, 1993). Another Bayesian approach to DIF is the use of hierarchical logistic regression models (Chaimongkol, 2005; Swanson et al., 2002; Vaughn, 2006). Hierarchical logistic regression allows the analyst to incorporate information about the test items and the examinees as independent variables. Hierarchical logistic regression also allows the analyst to study the relationships between item-level, examinee-level, and school-level variables.

The Bayesian DIF detecting method known as *empirical Bayes DIF* or *true DIF* (Zwick et al., 1999) will be considered here. Originally developed for dichotomous items, it can also be used for polytomous items. Cox's β, the logarithm of the Liu-Agresti statistic, HW3, the SMD statistic, and SIBTEST can all be modified to incorporate empirical Bayes. Define ξ_w as the true DIF parameter for item w (w = 1, 2, …, W). Assume that the prior distribution of ξ_w (for all values of w) is normal with mean μ and variance τ^2, where μ is the mean and τ^2 is the variance of the observed values of $\hat{\xi}_w$ over all W items on the test:

$$f(\xi_w) \sim Normal(\mu, \tau^2).$$

It is assumed that the observed statistic $\hat{\xi}_w$ is normally distributed with mean ($\hat{\xi}_w$) and variance (σ^2_w) determined from the statistic's squared standard error

$$L(\hat{\xi}_w \mid \xi_w) \sim Normal(\hat{\xi}_w, \sigma_w^2).$$

Given this information, the updated posterior distribution is also normal:

$$g(\xi_w \mid \hat{\xi}_w) \sim Normal\left(\frac{\tau^2}{\sigma_w^2 + \tau^2} \hat{\xi}_w + \frac{\sigma_w^2}{\sigma_w^2 + \tau^2} \mu, \frac{\tau^2 \sigma_w^2}{\sigma_w^2 + \tau^2} \right).$$

To test for DIF, the posterior $g(\xi_w \mid \hat{\xi}_w)$ is used as the basis of a confidence interval or hypothesis test. If zero does not appear in the confidence interval, then the null hypothesis $\xi_w = 0$ is rejected, and the item should be flagged for DIF.

In some instances, the variance of the prior distribution can be improved. If the Mantel-Haenszel statistic is assumed to contain measurement error, then the variance of the prior distribution can be improved by removing the measurement error (Fidalgo et al., 2007; Sinharay et al., 2008; Zwick et al., 1999; Zwick & Thayer, 2002). To do this, Zwick, Thayer, and Lewis (1999) recommended calculating the prior variance as the variance of the DIF statistics minus the average DIF statistic's squared standard error over the anchor test items. Although this method may improve the estimation of the prior variance, it is possible to obtain a negative prior variance, especially when examinee sample sizes are small (Fidalgo et al., 2007). This would happen if the average DIF statistic's squared standard error is greater than the variance of the DIF statistics. Because of the risk of negative prior variances, the modified prior variance suggested by Zwick et al will not be used in the proposed simulation study of Chapter 3.

When considering test content, an analyst may decide that a single prior distribution is inadequate. If the test items can be divided into g mutually exclusive groups $G_1, G_2, \ldots, G_g$ based on context, then the analyst can use *full Bayes* to detect DIF (Sinharay et al., 2008). In full Bayes DIF, each of the g groups of items gets its own prior distribution

$$f(\xi_{wG}) \sim Normal(\mu_G, \tau_G^2),$$

where μ_G and τ^2_G are the mean and variance of $\hat{\xi}_w$ for only the items in group G. Analysis proceeds in the same way as empirical Bayes. Sinharay et al. (2008) provided an example of this approach on a reading test, dividing the items into four groups based on item type and classification and using four prior distributions.

To date, the empirical Bayes method has only been applied to dichotomous items through the Mantel-Haenszel common odds ratio statistic (Fidalgo et al., 2007; Sinharay et al., 2008; Zwick et al., 1999; Zwick & Thayer, 2002; Zwick & Thayer, 2003). A brief illustrative example using Bayesian techniques with Cox's β is available (Camilli & Congdon, 1999), though the researchers did not consider any factors that could effect Type I error rate and statistical power. Research on the Liu-Agresti statistic promotes its use in a Bayesian context, though no studies or examples appear to have been published (Penfield & Algina, 2003) regarding this use.

Exact and Randomization Modifications

One approach to determine the statistical significance of a statistic, useful when sample sizes are small, is to use exact methods. For contingency tables, the statistician enumerates all possible tables with the same marginal totals as the original table. For each table, a relevant sample statistic is computed. The histogram of the sample statistics is the sampling distribution for the population statistic, given the marginal row and column totals. This sampling distribution can be used to determine the *p*-value of the observed data (Conover, 1999; Meyer et al., 2004; Parshall & Miller, 1995).

To illustrate this process, consider the hypothetical 2 x 3 x 4 contingency table of Table A2. To run an exact test, the statistician would have to determine all possible 2 x 3 x 4 contingency tables where the first subtable contains 13 focal group examinees, 12 reference group examinees, 5 examinees that scored 0 on the studied item, 7 examinees that scored 1, and 13 examinees that scored 2; the second subtable contains 11 focal group examinees, 15 reference group examinees, 10 examinees scored 0 on the studied item, 3 examinees that scored 1, and 13 examinees that scored 2; and so on. For each of these tables, the statistician computes a relevant test statistic $\hat{\xi}$, which could include the numerator of the Mantel test statistic, Cox's β, the logarithm of the Liu-Agresti statistic, the HW3 statistic, the SMD statistic, or the SIBTEST statistic. If the observed test

statistic for the original data is greater than the $1 - (\alpha/2)$ percentile or less than the $(\alpha/2)$ percentile, then the item should be flagged for DIF.

In practice, there are too many 2 x J x K contingency tables to consider for the exact test. For the 2 x 3 x 4 contingency table in Table A2, the analyst would have to determine $({}_{25}C_{13})\,({}_{26}C_{11})\,({}_{26}C_{13})\,({}_{23}C_{13}) \approx 4.78 \times 10^{26}$ possible tables to obtain the exact sampling distribution. To solve this problem, a randomization test can be used to approximate the sampling distribution (Camilli & Smith, 1990). Using random multivariate hypergeometric variates, a percentage of the total tables can be sampled. This abridged information can then be used to produce a sampling distribution from which a *p*-value can be approximated.

Research using dichotomous items suggests that the exact and randomization tests with the Mantel-Haenszel test are as effective at detecting DIF as the asymptotic Mantel-Haenszel test (Camilli & Smith, 1990), though exact and randomization tests may be slightly more conservative (Parshall & Miller, 1995). Research using Likert items revealed similar results, although only empirical data were considered (Meyer et al., 2004). Sample sizes of 299 and 76 were used for the focal and reference groups; it is unclear if smaller sample sizes would cause the randomization test to be more accurate than asymptotic methods.

Log-Linear Smoothing Modification

A recent modification proposed for the case of small sample sizes is the use of log-linear smoothing to remove noise from the score frequencies (Puhan et al., 2009). Smoothing methods are common in psychometrics, especially in test equating (Kolen & Brennan, 2004) via log-linear, four-parameter beta, cubic spine, or kernel smoothing methods. By incorporating a small amount of systematic error, smoothing aids in reducing the random error in a sample of examinees.

Smoothing methods were proposed for use in dichotomous DIF detection for sample sizes as small as 300 reference group and 50 focal group examinees (Puhan et al., 2009). For almost all of the studied items, the researchers found that population estimates of the true amount of SIBTEST bias were improved by smoothing the frequency counts with log-linear smoothing. Furthermore, the root mean square error of the population estimates was smaller for the smoothed estimates than the unsmoothed estimates. Although smoothed and unsmoothed estimates were similar for very large sample sizes, estimation was improved for smaller samples.

Smoothing can modify any of the polytomous DIF techniques that use 2 x J x K contingency tables. The examinees are divided into 2J subgroups, based on focal or reference group membership and studied item score. For each subgroup, the ordered pairs (X_k, N_k) can be produced, where N_k is the number of people in the reference (or focal) group that scored y_j on the studied item and X_k on the anchor test. These ordered pairs are smoothed using the regression

$$\log(N_k) = \beta_0 + \beta_1 X_k + \beta_2 X_k^2 + \beta_3 X_k^3 .$$

Puhan et al. (2009) recommend using a polynomial of degree 3, which causes the smoothed frequencies to maintain the same mean, variance, and skewness as the unsmoothed frequencies. Log-linear smoothing is repeated for the remaining 2J – 1 subgroups. An algorithm using Newton's Method to estimate the β_is is available in Holland and Thayer (2000).

Smoothing provides a way of better estimating the population distribution of the frequencies for a given group and studied item score. Although research using smoothing on dichotomous items has shown promise, no research is currently available applying smoothing to polytomous items. Furthermore, Puhan et al. (2009) encourage expanding their research using simulated data, as opposed to the empirical test data available to them.

Additional Modifications to Address Small Sample Sizes

Another approach to determining the standard error of a statistic is to use the jackknife method (Quenouille, 1956). The jackknife method has been used in many applications, including applications in sample surveys (Levy & Lemeshow, 1999) and generalizability theory (Brennan, 2001). It has been studied in the detection of dichotomous DIF (Camilli & Smith, 1990), but not polytomous DIF.

The jackknife procedure can be applied to any of the contingency table-based methods, including the Mantel test's sample statistic, Cox's β, the logarithm of the Liu-Agresti statistic, HW3, the SMD statistic, and the SIBTEST statistic. Define ξ as the parameter of interest. The sample statistic is calculated using all of the available data; call this statistic $\hat{\xi}_{all}$. Define $\hat{\xi}_{-k}$ as the estimate of the population parameter using all of the sample data except the examinees who received a score of k on the anchor test. The K "pseudovalues" $\hat{\xi}_k^*$ are defined as

$$\hat{\xi}_k^* = K\hat{\xi}_{all} - (K-1)\hat{\xi}_{-k}$$

(Camilli & Smith, 1990). The mean of the pseudovalues is the jackknife estimator for ξ; the variance of the pseudovalues divided by K is the estimate for the squared standard error of $\hat{\xi}$.

A related procedure used to estimate the standard error is the bootstrap method (Efron & Tibshirani, 1986). A bootstrap estimate is computed by randomly sampling with replacement n_{+++} examinees from the original sample of examinees and calculating the estimate $\hat{\xi}$ for the bootstrap sample. By repeating this procedure a pre-determined number of times, a sampling distribution begins to emerge. The standard deviation of this sampling distribution is an estimate for the true standard error of $\hat{\xi}$.

Little research has applied the jackknife or the bootstrap method to DIF detection, though empirical and simulated data using dichotomous data (Camilli & Smith, 1990) suggest that a jackknife version of the Mantel-Haenszel test yields similar results to the

asymptotic Mantel-Haenszel test. This research considered larger sample sizes, though. It is believed that small sample size conditions have not been studied for either dichotomous or polytomous items.

Increasing the Significance Level

A final modification for improving DIF detection with small sample sizes is to increase the level of significance α (Cohen, Kim, & Wollack, 1996; Fidalgo et al., 2004a; Kim, Cohen, & Kim, 1994). By increasing α, the analyst is more likely to make a Type I error. Making a Type I error implies that the analyst flagged the studied item for DIF, though no true DIF was present in the item. A consequence to increasing α is a decrease in making a Type II error. Making a Type II error implies that the analyst did not flag the studied item for DIF, even though true DIF is present. Decreasing Type II error rates also implies that statistical power will increase.

In terms of test validity, it may be better to commit a Type I error than a Type II error (Camilli, 2006). An item mistakenly flagged for DIF would be studied further by an analyst or a sensitivity review panel. Most likely, the item would pass this additional screening and eventually be used on a future test administration. However, more damage occurs if a biased item escapes DIF detection. A biased item might find its way onto an administered test, and test validity would be jeopardized. An ounce of prevention (reviewing the item again) would be worth a pound of cure (justifying the use of a biased item on a working test).

Summary

A variety of recommendations have been proposed in the DIF detection research for the case when sample sizes are small. Some methods, such as Bayesian techniques, allow the analyst to incorporate prior knowledge in estimating bias. Other methods, such as exact and randomization tests and jackknife and bootstrap estimators, provide the analyst with a more precise estimate of a statistic's standard error or sampling

distribution. Smoothing methods remove some noise from the score frequencies to obtain better estimates of the examinees' ability distributions. Liberal alpha values allow the analyst to improve statistical power in a DIF detection hypothesis test. All of these methods have been applied, with success, to dichotomous test items and small examinee sample sizes. It is unclear how effective any of these modifications are to polytomous test items.

Tables B7-B9 contain *R* functions that allow the analyst to apply each of these modifications to three of the contingency table-based DIF detection methods: Mantel test/Cox's β, the Liu-Agresti statistic, and HW3. These functions work in tandem with the DIF detection methods defined by the *R* functions used in Tables B1-B6.

Factors Considered in DIF Detection Simulations

Some attention will now be devoted to simulation studies conducted to study DIF detection rates. An analysis of these studies will help to guide the present study by understanding which factors have been shown to affect DIF detection in large-sample simulation studies. Here, 114 simulation studies that were published between 1990 and 2009 are considered. The majority of simulations were published in measurement journals (such as the *Journal of Education Measurement* and *Educational and Psychological Measurement*), though additional simulations were found in research papers, conference presentations, and doctoral dissertations. These simulation studies considered DIF detection in both dichotomous and polytomous items. The ratio of the number of simulations considering dichotomous item DIF detection to the number of simulations considering polytomous item DIF detection was approximately 5:2.

In hypothesis testing, statistical power is a function of many variables, including sample size, effect size, and significance level (α). In the case of DIF detection, power is dependent on additional factors, including examinee characteristics, test characteristics, and DIF analysis decisions. Examinee factors include information about the sample sizes

and ability distributions for the focal and reference groups. Test factors include information about test length, DIF contamination, dimensionality, test score reliability, and missing data. DIF analysis factors include decisions the analyst must make in running a DIF simulation and analysis, including the size of *K*, effect size, item parameters affected, pattern of DIF, size of anchor test, inclusion of the studied item in the anchor, use of anchor purification, and significance level.

Examinee Factors

Statistical power in DIF analysis is dependent on several examinee characteristics: sample size, ratio of examinees between reference and focal groups, and distributional assumptions of the focal and reference abilities. It seems reasonable that more examinees lead to more precise inferences about bias. Of the simulation studies that reported examinee sample sizes, over 75% considered multiple sample sizes. Many studies considered large samples, 3,000 focal group examinees and 3,000 reference group examinees (Bolt & Gierl, 2006; Jiang & Stout, 1998; Shealy & Stout, 1991; Shealy & Stout, 1993a; Stout, Li, Nandakumar, & Bolt, 1997; Zwick & Thayer, 2002); larger samples have also been considered (Allen & Donoghue, 1996; Kim, 2000; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Muraki, 1999; Robitzsch & Rupp, 2009; Roussos, Schnipke, & Pashley, 1999). The smallest samples to be considered were focal and reference groups of 50 per group (Fidalgo et al., 2004a; Fidalgo et al., 2007; Muñiz et al., 2001). Generally, DIF detection methods made more correct decisions when examinee sample size was large.

Some researchers have argued that DIF detection inferences are more powerful when the size of the sampled focal group equals the size of the sampled reference group (Herrera & Gómez, 2008). Most simulation studies consider a 1:1 sample size ratio, which best replicates the situation when males and females are the reference and focal groups. Some studies have considered ratios of 2:1 (Williams & Beretvas, 2006) and 3:1

(Donoghue & Allen, 1993; Monahan & Ankenmann, 2005). Larger ratios, including 9:1 (Lei, Chen, & Yu, 2006; Zwick, Thayer, & Wingersky, 1995) and 10:1 (Luppescu, 2002; Narayanan & Swaminathan, 1994), have been used to simulate the case when a large majority and a small minority make up the examinees (for example, Caucasians and African-Americans in the rural Midwest). Typically, power to detect uniform DIF is reduced when the sample size ratio increases (Herrera & Gómez, 2008; Narayanan & Swaminathan, 1994), enough that it was a statistically significant factor in the study by Monahan and Ankenmann (2005). In one study (Lei et al., 2006), Type I error rates were inflated for the case of a 9:1 sample size ratio; in some cases, Type I error rates were three times larger for the 1:1 sample size ratio case than the 9:1 sample size ratio case.

Information about the theta (ability) distributions for the focal and reference groups can have an effect on DIF detection power. Often, θ is distributed as a standard normal variable. In practice, however, the mean of θ for the focal group (μ_F) differs from the mean of θ for the reference group (μ_R). This difference in means does not provide evidence of DIF (American Educational Research Association et al., 1999), but it may have an effect on DIF detection (Cole & Moss, 1989; Shepard et al., 1981). Approximately half of the DIF simulations studied from the last two decades have considered differences between μ_F and μ_R. Of those that considered differences in group means, half simulated the cases where group means are equal or differ by 1 standard deviation. Six studies used differences of 0 and 0.5 when constructing their simulations, while nine studies considered differences of 0, 0.5, and 1. Nine other studies defined a single, non-zero difference between the group means, but did not treat this difference as a controllable factor. The largest difference defined in a simulation was 1.5 (Aguerri, Galibert, Attorresi, & Marañón, 2009; Su & Wang, 2005; Wang & Su, 2004b). Depending on the model being used, large differences between group means can inflate Type I error rates and affect statistical power (Wang & Su, 2004a).

Though not as common, several researchers have considered the shape of the focal and reference θ-distributions deviating from a normal distribution with a variance of one. For example, the variances of the θ-distributions do not have to equal one (Bolt & Gierl, 2006; Monahan & Ankenmann, 2005). Monahan and Ankenmann (2005) showed that Type I error increased from approximately .05 to .08 when the focal group variance decreased from 1.0 to 0.25, all other conditions remaining constant. Bolt and Gierl (2006) found that Type I error increased under certain conditions when the difference between the standard deviations of the ability distributions equaled one. In addition, the variances do not have to be equal for the focal and reference groups (Roussos et al., 1999). Roussos et al. used variances of 0.7 and 0.8 for the focal and reference groups, values derived from real testing data. Other researchers have not limited θ-distributions to be normally distributed. Research incorporating skewness into the θ-distributions concluded that skewness had little effect on Type I error and statistical power (Kristjansson et al., 2005). Other research defined the θ-distributions using cubic splines; defining the distributions in this way affected statistical power when using DIF detection techniques that assume normality of the latent ability (Woods, 2008a; Woods, 2008b). The percentage of "false alarms" for skewed ability distributions was two to four times larger than the percentage of false alarms for normally-distributed abilities (Woods, 2008a).

Test Factors

Often, the entire test is used as an anchor to estimate an examinee's ability θ. Therefore, the relationship between test characteristics and DIF detection power has been studied in past simulation studies. Test characteristics that could affect power include the length of the test, anchor score reliability, proportion of anchor items contaminated with DIF, item response theory model, dimensionality, and treatment of missing values.

If an observed score is used as the anchor criterion, a common assumption with the methods described so far, the analyst should use a score with high reliability. An anchor with high reliability will have a small standard error, thus providing a fairly precise estimate of the true anchor score. If the entire test is to be used as an anchor, then the length of the test will influence the anchor and the ability to make correct decisions regarding item DIF. Of the simulations produced since 1990, over 80% of the studies explicitly stated the number of items used for the entire test. Approximately 20% of the studies used test length as a factor in their simulations (including Clauser, Mazor, & Hambleton, 1993; DeMars, 2009; Donoghue & Allen, 1993; Fidalgo, Mellenbergh, & Muñiz, 2000; Klockars & Lee, 2008; Wang & Su, 2004a; Whitmore & Schumacker, 1999; Woods, 2009). Test lengths of 10, 20, 25, 30, 40, 50, and 80 dichotomous items are common. Simulations using dichotomous items tend to use longer tests than simulations using polytomous items, which is consistent with typical testing practice. Klockars and Lee (2008) showed that as the test length increased for large examinee sample sizes (1000 or greater), statistical power increased when using SIBTEST. They further showed that test length had less of an effect for smaller examinees sample sizes, namely those at and below 500. Wang and Su (2005) suggested that test length did not effect Type I error rates or statistical powerof Mantel tests, using focal and reference group sample sizes of 500 each. Though not observing test length, research by Ackermann and Evans (1992) considered the effect of test score reliability on the power of the dichotomous Mantel-Haenszel and SIBTEST statistics. As reliability increased from .70 to .95, under most conditions, statistical power increased by three to six points.

In practice, test developers are unaware if an item truly contains DIF. As a result, the test may contain items contaminated with DIF. If the entire test is used as an anchor, these contaminated items may lead to incorrect estimates of true ability. Although a DIF-free test may be ideal, some researchers have modeled a percentage of test items with DIF in their simulations. Approximately 35% of the DIF simulation studies incorporated

some percentage of contaminated items. Common percentages of contamination in use include 10% or 20% (Cohen & Kim, 1993; Finch & French, 2007; Flowers, Oshima, & Raju, 1999; French & Maller, 2007; Hidalgo-Montesinos & López-Pina, 2002; Kim & Cohen, 1992; Narayanan & Swaminathan, 1994; Narayanan & Swaminathan, 1996; Oshima, Raju, & Nanda, 2006; Raju, van der Linden, & Fleer, 1995), which reflect common testing situations. Some researchers, however, have considered higher contamination levels, including 66% (Lautenschlager, Flaherty, & Park, 1994; Park & Lautenschlager, 1990), 80% (Woods, 2009), and 100% DIF contamination (Su & Wang, 2005). Research suggests that an increase in DIF contamination leads to a higher number of false positives and false negatives (Cohen & Kim, 1993; Hidalgo-Montesinos & López-Pina, 2002). Using a sample of 100 examinees, Cohen and Kim (1993) showed that, on average, up to 4.0 false negatives for a 20-item test and up to 11.4 false negatives for a 60-item test could occur when using Mantel's test or Raju's area measures on a DIF contaminated test. Hidalgo-Montensinos and Lopez-Pina (2002) showed that as the DIF contamination level increased from 5% to 20%, the percentage of false positives increased significantly, particularly when large effect sizes were present in the DIF items.

Most DIF detection simulations assume that the trait being measured by the test is unidimensional. However, some researchers have considered the case when the test measures two dimensions simultaneously (Klockars & Lee, 2008; Lautenschlager et al., 1994; Lim & Drasgow, 1990; Mazor, Hambleton, & Clauser, 1998; Park & Lautenschlager, 1990; Ross, 2007; Spray & Miller, 1992). Failure to address multidimensionality could lead to poor estimation of ability and, consequently, affect DIF detection. In one study, false positive rates of up to 50% occurred when a multidimensional anchor was treated as being unidimensional (Mazor et al., 1998).

Finally, in almost all DIF detection research, examinees are assumed to answer every item on the test. If an item is omitted or not reached, many researchers would score the item incorrect (i.e. Bolt, 2000). By marking the item incorrect, the analyst may

inadvertently alter the anchor test score and estimate ability incorrectly. Robitzsch and Rupp (2009) have recently considered the case where dichotomously scored items contain missing values. When using the Mantel-Haenszel test or logistic regression, they found that missing data mechanisms (ANOVA effect size $\eta^2 = .40$) and imputation methods ($\eta^2 = .14$) were statistically significant factors in describing item bias rates.

DIF Analysis Factors

In addition to examinee and test factors, there are factors that the analyst must consider when conducting a DIF analysis. The decisions made regarding the analysis can affect statistical power. Some of these decisions include assumptions about the item's effect size, its IRT parameters, the type of DIF being considered (uniform versus non-uniform), the size of K in the construction of $2 \times J \times K$ contingency tables, the number of items in the anchor test, the inclusion of the studied item in the anchor, the type of purification method used to remove contaminated items from the anchor, and the choice of significance level for the hypothesis tests.

In inferential statistical procedures, the ability to detect a particular alternative hypothesis depends on the effect size being considered. Effect sizes for DIF are typically defined in one of two ways. The more common approach is to translate the item characteristic curve (ICC) to the left or right by a constant amount (e.g., Hidalgo-Montesinos & López-Pina, 2004; Penfield, 2007; Penfield, 2008; Robitzsch & Rupp, 2009; Su & Wang, 2005). The other approach, used when simulating non-uniform DIF, is to transform the ICC of one group in such a way that the area between the ICCs becomes an effect size (e.g., Finch & French, 2008; Hidalgo & Gómez, 2006; Lei et al., 2006; Narayanan & Swaminathan, 1994; Swaminathan & Rogers, 1990). As expected in power analysis research, larger effect sizes lead to higher statistical power rates when conducting DIF analyses.

Many researchers have considered the relationship between the IRT parameters for the studied item and DIF detection power (Allen & Donoghue, 1996; Ankenmann, Witt, & Dunbar, 1999; Jiang & Stout, 1998; Marañón et al., 1997; Monahan & Ankenmann, 2005; Roussos & Stout, 1996b; Uttaro & Millsap, 1994). This has been done by defining the studied item to take a fixed set of *a*- and *b*-parameters. Type I error rates appeared to increase slightly as the *a*-parameter (discrimination) increased, while Type I error rates appeared to decrease slightly as the *b*-parameter (difficulty) increased (Jiang & Stout, 1998; Monahan & Ankenmann, 2005).

The type and pattern of DIF have been studied as factors for DIF detection rates. In the dichotomous case, patterns are usually limited to uniform and non-uniform DIF. Non-uniform DIF is often studied, since most DIF detection techniques have a more difficult time identifying this kind of pattern (Finch & French, 2007; Li & Stout, 1996; Marañón et al., 1997; Mazor et al., 1994; Narayanan & Swaminathan, 1996). Research suggests, for example, that the Mantel-Haenszel test yields high DIF detection rates for uniform DIF (between 0.96 and 1.00 when the reference group size is 1500), but low DIF detection rates for non-uniform DIF (between 0.38 and 0.74) (Herrera & Gómez, 2008). Additional research has studied the differences between uniform and non-uniform DIF for other DIF detection methods (Hidalgo-Montesinos & Gómez-Benito, 2003; Lopez Rivas, Stark, & Chernyshenko, 2008; Oshima, Raju, & Flowers, 1997; Whitmore & Schumacker, 1999; Woods, 2009). In the polytomous case, there are many more DIF patterns that researchers have studied (Ankenmann et al., 1999; Chang et al., 1996; Meade, Lautenschlager, & Johnson, 2007; Penfield et al., 2009; Su & Wang, 2005; Wang & Su, 2004a; Zwick et al., 1993). For example, Meade, Lautenschlager, and Johnson (2007) found that power rates were statistically different when the *a*-parameter differed between the focal and reference groups compared to the power rates when the *a*-parameter did not differ between groups, all other conditions remaining constant (ANOVA effect size $\eta^2 = .019$).

When carrying out a DIF detection analysis, the analyst likely needs to produce 2 x *J* x *K* contingency tables, where *K* is the number of ordered categories based on the anchor test. If examinee sample size is small, it may be necessary to condense the number of tables to meet computational assumptions for contingency tables. This issue, where the analyst must decide the value of *K*, is called the thick/thin matching problem (Donoghue & Allen, 1993). Smaller values of *K* are used for thick matching; larger values of *K* imply thin matching. As *K* decreases, some information may be lost through the pooling of data, and DIF detecting power could be impacted. Research suggests that the rate of false positives increases by between 5 and 25 percentage points when thick matching is used with unequal ability distributions (Clauser, Mazor, & Hambleton, 1994). Further research suggests that certain thick matching methods could lead to highly biased estimates of the Mantel-Haenszel statistic (Donoghue & Allen, 1993).

As explained earlier, longer tests—when the entire test is used as the anchor test—lead to better estimates of examinee ability and should lead to better DIF detection rates. Several studies used anchor test length as a factor in DIF detection; their findings confirm that increasing the number of anchor items will improve statistical power (Lopez Rivas et al., 2008; Wang & Yeh, 2003; Woods, 2009). For example, Lopez Rivas, Stark, and Chernyshenko (2008) found that the number of items in their anchor (called referents in their research) had a significant effect on DIF detection power (ANOVA effect size η^2 = .98 for dichotomous items, η^2 = .60 for polytomous items). Wang and Yeh (2003) found that, when looking at polytomous items using item response theory, statistical power rates increased by 5 to 12 percentage points by increasing the size of the anchor from one item to ten items.

It seems plausible that the item under investigation should not be included in the anchor test. However, there are statistical justifications for including the studied item in the anchor test for dichotomous items (Zwick, 1990). When considering Rasch items, the Mantel-Haenszel test yields inflated Type I error rates when the studied item is not

included in the anchor and the ability distributions differ between the focal and reference groups. Research with simulated data has shown inflation also occurs for polytomous items when the studied item is not included in the anchor test and ability distributions differ (Zwick et al., 1993).

The anchor test may be contaminated with items that contain DIF. Because of this, use of the total observed score on the anchor test may be a biased estimate of true examinee ability. Many researchers have proposed *purification methods*, iterative procedures that calculate a DIF effect size or statistic for each item on the proposed anchor test and remove items that are flagged for DIF. Often, DIF statistics are recalculated based on the score on the reduced anchor test and any additional flagged items are removed from the anchor test (Lord, 1980). Purification methods have been shown to increase statistical power as much as 50% for the dichotomous Mantel-Haenszel test when ability distributions are equal and 20% of the items on the test contain DIF contamination (Clauser et al., 1993). Variations of purification methods have been studied using logit models (Kok, Mellenbergh, & van der Flier, 1985; Navas-Ara & Gómez-Benito, 2002; Van der Flier, Mellenbergh, Adèr, & Wijn, 1984), item response theory (Hidalgo-Montesinos & López-Pina, 2002; Lautenschlager et al., 1994; Park & Lautenschlager, 1990), the Mantel-Haenszel test (Fidalgo et al., 2000; Fidalgo et al., 2007; Navas-Ara & Gómez-Benito, 2002; Su & Wang, 2005; Wang & Su, 2004a), and logistic regression (French & Maller, 2007; Hidalgo-Montesinos & Gómez-Benito, 2003; Navas-Ara & Gómez-Benito, 2002; Su & Wang, 2005).

Finally, the analyst must decide on a significance level to use in DIF detection hypothesis tests. As discussed earlier, when alpha increases, statistical power increases. This is justified, as the consequences of committing a Type I error are not as severe as the consequences of committing a Type II error. Although many analysts define alpha as .05, and some analysts consider using a Bonferroni correction to produce a more conservative significance level, Fidalgo et al. have suggested increasing alpha to .20 to

increase statistical power (Fidalgo et al., 2004a). Several researchers have reported results for more than one significance level (Ankenmann et al., 1999; Cohen et al., 1996; Kim et al., 1994; Narayanan & Swaminathan, 1996).

Summary

In the last twenty years, over 100 simulation studies have been conducted to analyze factors that could influence DIF detection rates. These factors include, but are not limited to, examinee, test, and DIF analysis factors. Several of the factors discussed here will be used in the simulation study to be described in Chapter 3.

DIF Detection Research Using Empirical Data

Research about polytomous DIF detection is not limited to simulation studies. Over the last twenty years, researchers have studied DIF using empirical data from polytomous items on national achievement tests (Hamilton, 1999; Miller & Spray, 1993; Penfield, Gattamorta, & Childs, 2009; Wainer, Sireci, & Thissen, 1991; Wainer, 1995; Zwick, Thayer, & Mazzeo, 1997), regional achievement tests (Kim, Cohen, Alagoz, & Kim, 2007; Penfield et al., 2009; Walker & Beretvas, 2001; Zenisky, Hambleton, & Robin, 2003), attitudinal scales (Collins, Raju, & Edwards, 2000; Edelen, McCaffrey, Marshall, & Jaycox, 2008; Meyer et al., 2004; Wang & Russell, 2005), and psychological inventories (Dorans & Kulick, 2006; Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Everson, Millsap, & Rodriguez, 1991; Gelin & Zumbo, 2003; Maller, 2001; Morales, Flowers, Gutiérrez, Kleinman, & Teresi, 2006; Santor, Ramsay, & Zuroff, 1994; Teresi et al., 2007). This section will summarize the characteristics presented in empirical DIF detection studies for polytomous items, including information about the tests, item characteristics, focal and reference groups, sample size, and DIF detection methods used by the researchers. In doing so, the reader can see how the proposed methods of Chapter 3 compare to DIF detection practice for polytomous items.

Several national achievement testing programs have published research about polytomous DIF detection using data from active and experimental sources, including an *SAT*-based prototype (Wainer et al., 1991), an experimental *ACT* mathematics test (Miller & Spray, 1993), the *Law School Admissions Test* (LSAT, Wainer, 1995), the *Advanced Placement Physics B Exam* (Zwick et al., 1997), a nationally administered writing assessment (Welch & Miller, 1995), and the mathematics section of the *School Achievement Indicators Program*, a Canadian testing program (Penfield et al., 2009). The research by Wainer (1991; 1995) used testlets—sets of dichotomous items which are based on the same reading comprehension prompt—to create polytomously scored "items", which were then evaluated for DIF using item response theory methodology. Each of four studied testlets contained up to nine score categories. Other research considered open ended or constructed response items based on mathematics (Miller & Spray, 1993; Penfield et al., 2009) or science (Hamilton, 1999; Zwick et al., 1997). The research by Welch and Miller (1995) used a single writing prompt with an anchor based on 40 multiple-choice writing items. Constructed response items had as few as four score categories per item (Miller & Spray, 1993) and as many as eleven (Welch & Miller, 1995) and sixteen (Zwick et al., 1997) score categories. Most tests considered between four and six polytomous test items, though the research by Penfield (2009) considered a test with thirty constructed response items.

Additional research has considered polytomous DIF detection on educational achievement tests that are not nationally based. Some examples utilized state assessments (Kim et al., 2007; Walker & Beretvas, 2001; Zenisky et al., 2003) and district assessments (Penfield et al., 2009), while others used instruction driven assessments (Lane, Wang, & Magone, 1996). Most of these cited studies considered mathematics and science items, with the exception of the research by Kim et al (Kim et al., 2007), which considered a state's kindergarten performance assessment. The number of polytomous items on most of the tests ranged from three (Walker & Beretvas, 2001) to

twelve (Zenisky et al., 2003), though one assessment contained thirty-six constructed response items (Lane et al., 1996). In most cases, the test contained both polytomous and multiple-choice items. Most polytomous items were scored according to three, four, or five categories.

Polytomous DIF detection research based on empirical data from attitudinal scales ranged in content from work satisfaction and description (Collins et al., 2000; Wang & Russell, 2005) to mathematics attitude (Meyer et al., 2004) to attitudes towards dating violence (Edelen et al., 2008). Wang and Russell (2005) evaluated DIF on five subscales, treating each subscale as a testlet. The research of Edelen et al. (2008) and Collins et al. (2000) considered scales of nine items and ten items, respectively. The mathematics attitude survey considered by Meyer et al. (2004) contained 30 items, each with four possible scores.

Finally, a number of psychological scales have been used to illustrate and evaluate DIF for polytomous items. DIF studies are available for depression scales (Gelin & Zumbo, 2003; Santor et al., 1994) such as the *Beck Depression Inventory*, cognitive scales (Crane, van Belle, & Larson, 2004) such as the *Mini-Mental State Exam* (Dorans & Kulick, 2006; Edelen et al., 2006; Morales et al., 2006), test anxiety scales (Everson et al., 1991), distress scales (Teresi et al., 2007), and intelligence scales (Maller, 2001). Psychological scales used in polytomous DIF research have contained as few as five (Dorans & Kulick, 2006; Edelen et al., 2006; Morales et al., 2006) and as many as 38 (Teresi et al., 2007) polytomously scored items. Most of the researched scales have used between four and six score categories per item.

Most of the studies cited above consider reasonably large examinee sample sizes. National and state-based educational testing programs typically have thousands of examinee response strings available for polytomous DIF detection. Total sample sizes for achievement tests in the above DIF detection studies were as high as 105,000 examinees (Kim et al., 2007), though some studies included total samples of 60,000

(Zenisky et al., 2003), 36,800 (Wainer, 1995), and 13,200 examinees (Zwick et al., 1997). Though examinee sample sizes were generally lower for attitudinal and psychological assessments, it was not uncommon to have a study use over 1,000 total participants. Some DIF analyses contained 1,500 (Dorans & Kulick, 2006; Edelen et al., 2006; Morales et al., 2006), 1,600 (Teresi et al., 2007), 2,000 (Collins et al., 2000), 2,500 (Edelen et al., 2008), or 4,300 (Wang & Russell, 2005) total participants.

Despite the number of studies that considered large examinee sample sizes, there are a few studies which evaluated polytomous DIF for small examinee sizes. The DIF detection study of Meyer et al. (2004), which considered a scale of mathematics attitude, used a total examinee sample size of 375. This was necessary, as their research studied the accuracy of exact and randomization methods on small sample sizes in polytomous DIF. Research on an instructional assessment for mathematics used a total examinee sample size of 460 (Lane et al., 1996). Additional polytomous DIF studies using small sample sizes included research on a test anxiety scale using 501 participants (Everson et al., 1991), a depression scale using 600 participants (Gelin & Zumbo, 2003), and a science assessment using 658 total examinees (Penfield et al., 2009).

The majority of the polytomous DIF detection studies using empirical data used males and females as the subgroups of interest (Collins et al., 2000; Crane et al., 2004; Edelen et al., 2008; Everson et al., 1991; Gelin & Zumbo, 2003; Hamilton, 1999; Kim et al., 2007; Lane et al., 1996; Maller, 2001; Meyer et al., 2004; Miller & Spray, 1993; Santor et al., 1994; Wainer et al., 1991; Welch & Miller, 1995; Zenisky et al., 2003; Zwick et al., 1997). In most of these cases, the ratio of male participants to female participants was approximately 1:1. There were two exceptions, the ratio of male to female examinees in the study of the *Advanced Placement Physics B Exam* was approximately 2:1 (Zwick et al., 1997), and the ratio of male to female participants in the study of the mathematics attitude scale was approximately 1:4 (Meyer et al., 2004).

Though not as common as gender, race was used as a subgroup factor in several of the polytomous DIF studies cited above. Of those studies that considered race, most evaluated DIF for White and African-American subgroups (Collins et al., 2000; Crane et al., 2004; Teresi et al., 2007; Wainer, 1995; Welch & Miller, 1995). The ratio of White participants to African-American participants ranged from 1:1 (Collins et al., 2000) to approximately 11:1 (Wainer, 1995). One study considered African-American and Hispanics as the subgroups of interest (Penfield et al., 2009). In addition to race, age was used to determine subgroups in two studies (Crane et al., 2004; Teresi et al., 2007).

One application of polytomous DIF detection is its use on translated tests or scales. Spanish and English versions of the *Mini-Mental State Exam* were administered to Spanish-speaking and English-speaking examinees to assess dichotomous and polytomous DIF (Dorans & Kulick, 2006; Edelen et al., 2006; Morales et al., 2006). Since the *School Achievement Indicators Program* is administered to Canadian students, researchers were able to investigate polytomous DIF for the English and French versions of the mathematics examination (Penfield et al., 2009). Polytomous DIF was also evaluated for a job description scale that was administered to over 2,600 Chinese participants and 1,600 American participants (Wang & Russell, 2005).

Many of the studies cited above used item response theory methodology to evaluate polytomous DIF (Crane et al., 2004; Edelen et al., 2006; Edelen et al., 2008; Kim et al., 2007; Maller, 2001; Teresi et al., 2007; Wainer et al., 1991; Wainer, 1995; Wang & Russell, 2005). Item response theory could be used in these studies, since large examinee sample sizes were available. This approach is not practical for the small sample case, though, as model estimation and convergence are jeopardized with smaller samples. Logistic regression methods, including polytomous logistic regression (French & Miller, 1996) and logistic discriminant function analysis (Miller & Spray, 1993), were also used in a number of studies (Crane et al., 2004; Gelin & Zumbo, 2003; Hamilton, 1999; Kim et al., 2007; Lane et al., 1996; Miller & Spray, 1993; Welch & Miller, 1995).

The Mantel test was frequently used as well (Dorans & Kulick, 2006; Kim et al., 2007; Maller, 2001; Meyer et al., 2004; Wainer, 1995). Additional methods used in polytomous DIF studies include HW3 (Welch & Miller, 1995), the Liu-Agresti statistic (Penfield et al., 2009; Penfield et al., 2009), the standardized mean difference statistic (Zenisky et al., 2003), confirmatory factor analysis (Everson et al., 1991), and nonparametric item response theory methods (Santor et al., 1994). Only one study appeared to utilize a small sample size modification to study polytomous DIF (Meyer et al., 2004).

Summary

Fairness is one of many characteristics that developers must consider when creating an educational or psychological test. Item bias evaluation is one practice used to establish evidence of test fairness. Differential item functioning is a statistically based method of evaluating potential item bias using the response strings of the examinees.

Several DIF detection methods are available for the case when test items are polytomously scored. Although methods exist using item response theory and logistic regression, the focus in this research will be methods based on 2 x J x K contingency tables. Contingency table-based methods include the Mantel test, Cox's β statistic, the Liu-Agresti common odds ratio, HW1, HW3, the standardized mean difference statistic, and the SIBTEST statistic.

When the number of examinees in the focal or reference group is small, the statistical power of a DIF detection method may be affected. Researchers have proposed ways to modify DIF detection methods in order to improve the statistical power, including the use of Bayesian methods, exact and randomization tests, jackknife methods, bootstrap methods, and log-linear smoothing methods. These modifications are meant to better understand the true sampling distribution of the DIF detection statistics when

sample sizes are small. Although these methods have been studied for dichotomous items, their utility in polytomous items is not as well known.

Simulation studies have been developed to better understand the Type I error rates and statistical power of many DIF detection methods. Researchers have altered such factors as examinee sample size, examinee latent ability distribution, test length, effect size, and DIF pattern to better understand how test factors, examinee factors, and DIF analysis factors affect statistical power. Further research about polytomous DIF detection has been conducted on empirical data sets from national and regional achievement tests, attitudinal scales, and psychological scales. Most empirical DIF studies used relatively large sample sizes and considered either gender or race to define the focal and reference groups.

In the next chapter, a simulation study will be described that will allow for the study of Type I error rates and statistical power for various polytomous DIF detection techniques and their small sample size modifications. The results of this research should help analysts to better understand which polytomous DIF detection methods and modifications are the most powerful when dealing with small numbers of examinees.

CHAPTER 3

METHODS

In this chapter, the simulation methods to be used to answer the research questions are described. This includes details about the simulation of examinees, anchor test items, the studied test item, and the DIF analyses to be considered. For convenience, the research questions addressed by this study are reprinted below.

1. What is the Type I error rate for the following small-sample polytomous DIF detection procedures: Mantel test, Liu-Agresti statistic, and HW3 statistic?
2. Of those methods listed in Question 1 that produce Type I error rates that are approximately equal to the a priori significance level ($\alpha = 0.05$), which of the methods yield the highest statistical power?
3. What is the Type I error rate for the following modifications to small-sample polytomous DIF detection procedures: empirical Bayesian, randomization methods, and log-linear smoothing?
4. Of those modifications listed in Question 3 where the simulated Type I error rates are approximately equal to the a priori significance level ($\alpha = 0.05$), which methods are the most statistically powerful?
5. Which of the following factors affect(s) the statistical power of the polytomous DIF detection methods listed in Questions 1 and 3: pattern of DIF in the studied item, reference group sample size, inclusion of the studied item in the anchor, and difference between the focal and reference group means?

Simulation Procedures

To simulate focal and reference group examinees, it is sufficient to simulate θ-values. Values of θ for the simulated reference group examinees were derived from a standard normal distribution with mean of zero and variance of one. Values of θ for the simulated focal group examinees were derived in one of two ways: (a) from a standard

normal distribution with mean of zero and variance of one or (b) from a normal distribution with mean of -0.5 and variance of one. When the mean θ for the focal group differs from the mean θ for the reference group, *impact* is said to exist (Dorans, 1989). As shown in the previous chapter, many DIF detection simulations have considered the case when the mean of the focal group's θ-distribution differs from the mean of the reference group's θ-distribution (including Penfield, 2008; Raju et al., 2009; Su & Wang, 2005; Woods, 2009). In addition, many recent DIF detection studies have simulated θ-values using the conditions listed above (Finch & French, 2008; Kristjansson et al., 2005; Lopez Rivas et al., 2008; Oshima et al., 2006; Raju et al., 2009; Robitzsch & Rupp, 2009; Su & Wang, 2005).

Two sample sizes were considered in this simulation: (a) 40 focal group and 40 reference group examinees and (b) 40 focal group and 400 reference group examinees. Condition (a) replicates the case when a small number of male and female examinees take the test; condition (b) replicates the case when a small minority and large majority take the test. Sample sizes as small as 50 focal group and 50 reference group examinees have been studied in past DIF detection simulations (Fidalgo et al., 2004a; Fidalgo et al., 2007; Muñiz et al., 2001), though these studies have considered dichotomous, but not polytomous, items. For studies of polytomous items, sample sizes as small as 100 focal group and 100 reference group examinees (Kim, 2000) and 250 focal group and 250 reference group examinees (Hidalgo-Montesinos & López-Pina, 2002; Penfield & Algina, 2003) have been considered.

Simulated examinees took an exam of 15 core items and 1 studied item. Most simulation studies using polytomous items considered test lengths of less than 50 items. Test lengths of 50 polytomous items were considered by Su and Wang (2005). Recent studies have considered tests made up of 10, 20, 30 or 40 polytomous items (Flowers et al., 1999; Wang & Su, 2004b; Williams & Beretvas, 2006; Woods, 2009). Given these studies, it is reasonable to simulate exams containing 16 polytomous items.

Item responses were simulated using Samejima's graded response model (GRM, Kolen & Brennan, 2004, p. 209-211; Samejima, 1997). The use of Samejima's GRM is common in DIF detection simulation studies using polytomous items (Bolt, 2002; Collins et al., 2000; DeMars, 2008; Hidalgo & Gómez, 2006; Hidalgo-Montesinos & Gómez-Benito, 2003; Kim, 2000; Meade et al., 2007; Raju et al., 2009; Woods, 2009). Define Y as a random variable that can take on one of J possible ordered scores $y_1, y_2, \ldots, y_J$ for a studied item. Let θ represent an examinee's latent ability being measured by the studied item. Then, Samejima's GRM is defined[2] as

$$P(Y \geq y_1 \mid \theta) = 1$$
$$P(Y \geq y_j \mid \theta) = \frac{\exp(a(\theta - b_j))}{1 + \exp(a(\theta - b_j))} \qquad j = 2,..., J$$

where a is the discrimination parameter common to all categories and b_j is the location parameter for category $j = 2, 3, \ldots, J$. Note that there is no b_1 parameter under this model. Although the a-parameter could differ for each score category, polytomous DIF research typically defines the model with a common a-parameter (Fidalgo & Madeira, 2008; Kim et al., 2007; Penfield, 2007; Penfield et al., 2009; Su & Wang, 2005) or simulates data using a common a-parameter (DeMars, 2008; Kristjansson et al., 2005; Meade et al., 2007; Raju et al., 2009). The category response functions are found by taking differences of cumulative response functions:

$$P(Y = y_j \mid \theta) = P(Y \geq y_j \mid \theta) - P(Y \geq y_{j+1} \mid \theta) \qquad j = 1,..., J-1$$
$$P(Y = y_J \mid \theta) = P(Y \geq y_J \mid \theta)$$

For reasons to be discussed shortly, J will equal 3.

There were no missing values, no omitted items, and no instance where multiple responses are keyed. Missing values would require additional analysis, as described by

[2] Some versions of Samejima's GRM replace a with $1.7a$ (see, for example, Kolen & Brennan, 2004, p. 209).

Robitzsch and Rupp (2009). Item parameters for the "core" items appear in Table A3; core items will be used as part of the anchor test to estimate an examinee's latent ability. These parameters are loosely based on the item parameters[3] found in Kim and Cohen (1998) which came from examinee responses on a mathematics assessment from the Wisconsin Student Assessment System, a performance assessment administered to 4th, 8th, and 10th graders.

The 16th item was designated as the "studied item". The studied item was the only item on the simulated test to be evaluated for DIF. Item parameters for the studied item appear in Table A4. The item parameters (b_2, b_3) = (0.01, 1.86) were constant for the reference group throughout the simulation. Depending on the condition of interest, the item parameters for the focal group were defined as: (a) (0.01, 1.86), in which case DIF is not present; (b) (0.46, 2.31), where pervasive constant DIF is present; (c) (0.46, 2.61), where pervasive convergent DIF is present; and (d) (0.46, 1.41), where pervasive divergent DIF is present. Pervasive, constant, convergent, and divergent DIF are defined in the polytomous DIF taxonomy created by Penfield, Alvarez, and Lee (2009). Figure A3 shows the cumulative category response functions for the studied item.

In order to run DIF analyses, an estimate of the examinee's latent ability θ is needed. The observed anchor score X was used as that estimate. X was calculated under one of two conditions: (a) the observed scores on the core items will be summed, or (b) the observed scores on the core items and the studied item will be summed. As discussed in the last chapter, statistical theory suggests that the studied item should always be included as part of the anchor test (Zwick, 1990; Zwick, Thayer, & Wingersky, 1994). Empirical evidence of the effects of including or excluding the studied item from the

[3] In Table A3, item parameter b_2 was computed by taking the average of β_{j1} and β_{j2} from Table 1 of Kim and Cohen (1998). Item parameter b_3 was computed by taking the average of β_{j3} and β_{j4} from Table 1 of Kim and Cohen (1998). Item parameter a is based on α_j from Table 1 of Kim and Cohen (1998).

anchor is minimal, though it is believed that Type I error rates should increase if the studied item is excluded from the anchor test (Zwick, 1990; Zwick et al., 1994).

In order to produce contingency tables based on the simulated responses, it was decided to split the range of the anchor test scores *X* into four categories (*K* = 4) based on the first quartile, median, and third quartile. Although the use of a small value of *K* is less than ideal (Donoghue & Allen, 1993), it is not possible to use too many anchor score categories *K* or too many polytomous item scores *J*. For example, using a focal group size of 40 and setting *K* equal to four allows for approximately 10 focal examinees to appear in each table. With *J* equal to three, approximately three or four focal group examinees would appear in each focal group cell of the contingency table. Increasing *J* or *K* would lead to an increase in the number of zero cells in the 2 x *J* x *K* contingency table. This could lead to a violation of assumptions that could affect inferential testing in much the same way that many zero cells could affect the chi-square test for contingency tables (Conover, 1999). Using *J* = 3 and *K* = 4 allows for the study of polytomous DIF techniques without risking the possibility of violating potential assumptions caused by sparse contingency tables.

Once the data have been simulated and the 2 x 3 x 4 contingency table has been produced, DIF procedures may begin. Three contingency table methods were considered: (a) Mantel test/Cox's β, (b) the natural logarithm of the Liu-Agresti polytomous common odds ratio, and (c) the HW3 statistic. Each DIF detection method was addressed in the following ways: (a) the empirical Bayes method, (b) the randomization test, (c) the log-linear smoothing method, and (d) unmodified. In all, 12 DIF detection methods (three contingency table methods times four versions of each method) were calculated for each set of examinees and response strings.

Variations of the simulated conditions allowed for the study of how the following factors affect Type I error rates and statistical power: (a) a priori impact (2 conditions: impact = 0 or -0.5), (b) number of examinees (2 conditions: 40 reference/40 focal, 400

reference/40 focal), (c) pattern of DIF in the studied item (4 conditions: no DIF, pervasive constant, pervasive convergent, pervasive divergent), and (d) inclusion or exclusion of the studied item from the anchor test (2 conditions: studied item is included, studied item is excluded). In all, 32 possible conditions (2 x 2 x 4 x 2) were simulated.

Initial attempts to simulate item response strings using the *R* statistical computing environment were found to be too slow given the number of simulation conditions and replications required for this research. Therefore, theta and item response string simulation were performed using the Java programming language on a personal computer running Windows Vista (see Table B12 for the code). For each set of simulation conditions, 1,000 replications were simulated and saved into a text file. It will be shown in the next section that 1,000 replications allowed for descriptive statistics to be computed with a reasonable margin of error.

Simulated examinees were created by randomly selecting normal deviates using the "nextGaussian" method from the "Random" class found in the "java.util.*" library. Although the seed used to produce these pseudorandom numbers could be set by the analyst, the default constructor uses the time when the method is called (in milliseconds) as the seed (Lewis & Loftus, 2001).

To simulate an item response, the probabilities of obtaining a score of 0, 1, or 2 on the item was calculated, based on the item parameters[4] a, b_2, and b_3 and the examinee latent ability parameter θ. Let the probability of scoring a 0 be p_0, the probability of scoring a 1 be p_1, and the probability of scoring a 2 be p_2. A uniform random variable U with range 0 to 1 is computed using the "random" method from the "Math" class in the base Java library. The simulated examinee scores a 0 if $U < p_0$. The simulated examinee scores a 1 if $p_0 \leq U < p_0 + p_1$; the examinee scores a 2 if $p_0 + p_1 \leq U$.

[4] Recall that the location parameter b_1 is not defined in Samejima's model.

Each text file of response strings was read into *R*, one replication at a time (see Table B13 for the relevant code). The anchor scores were calculated using the appropriate conditions, either by summing the scores of the core items or by summing the scores of the core items and the studied item. Once the anchor scores are calculated, the range of scores was divided into four categories using the "quantile" function. This allows for each of the (K =) 4 tables to have approximately the same number of examinees. The 2 x 3 x 4 contingency table for that replication can be produced using the "table" function. Each of the unmodified DIF detection techniques and its subsequent modifications were computed. The *R* program tallies the number of times an item was flagged out of the 1,000 replications for a given set of factors.

Addressing the Research Questions

Given the experimental conditions described above, the dependent variables are the Type I error rate and the statistical power of the various DIF detection procedures and their small sample size modifications. Type I error rate is the percentage of time that the DIF detection method flags an item for DIF when it has been defined to not contain any DIF. Statistical power is the percentage of time that the DIF detection method flags an item for DIF when the studied item has been defined to contain DIF using one of the four biased items defined in Table A4. The independent variables include a priori impact, the number of examinees, the pattern of DIF in the studied item, and inclusion or exclusion of the studied item from the anchor test.

The first research question investigates the empirical Type I error rates for the three main DIF detection methods: Cox's β, the natural logarithm of the Liu-Agresti statistic, and the HW3 statistic. Three factors were crossed to determine the simulation parameters: inclusion of the studied item in the anchor (2 conditions), impact (2 conditions), and examinee sample size (2 conditions). For each set of factors—8

combinations in all—1,000 replications were simulated. The studied item contained no DIF.

Ideally, all Type I error rates should be approximately 0.05 when a significance level of $\alpha = 0.05$ is used. If 1,000 replications are used to estimate the true Type I error rate for a given set of simulation conditions, then with 95% confidence, the margin of error for the Type I error rate[5] (as a proportion) is ±0.014 under the $\alpha = 0.05$ condition. Any Type I error rates that are less than 0.05 minus the margin of error will be identified in the tables with boldface. These rates will be more conservative than the true Type I error rate. Any Type I error rates that are larger than 0.05 plus the margin of error will be identified with boldface and italics. These rates suggest that the DIF detection method is more liberal than the true Type I error rate.

For those methods where the Type I error rates match the simulated significance level (within the margin of error), the second research question addresses which of the three contingency table methods yield the highest statistical power. Statistical power will be calculated as the proportion of replications where the item is (correctly) flagged for DIF. Statistical power will be evaluated using descriptive statistics. In addition to the factors considered in Research Question #1 (inclusion/exclusion of the studied item, impact, examinee sample size), the simulation also considered the three DIF patterns as defined in Table A4. With this additional factor, 24 combinations of factors were studied.

The power of DIF detection techniques for polytomous items has been shown to vary widely depending on effect size, alpha level, examinee sample size, and impact (Ankenmann et al., 1999). To determine a margin of error (with 95% confidence), it is assumed that the true power is 0.50. By estimating power at 0.50, the estimate of the

---

[5] The margin of error (with 95% confidence) for a proportion p based on sample size n is defined as $1.96\sqrt{p(1-p)/n}$.

margin of error will be maximized. If 1,000 replications are used to estimate statistical power for a given set of simulation conditions, then with 95% confidence, the margin of error for statistical power (as a proportion) is ±0.031. If the true value of power is higher than 0.50, then the margin of error will be less than ±0.031. Given these estimates of margin of error, it is believed that the number of replications will be sufficient to accurately estimate statistical power. The contingency table method (or methods) that yields the highest statistical power, after considering the margin of error, will be judged the preferred method.

To answer the third research question, contingency table modifications were used on the simulation replications to compute Type I error rates. For this study, modifications were used on Cox's β, the Liu-Agresti statistic, and the HW3 statistic. Modifications include the empirical Bayesian technique, the randomized test, and the use of log-linear smoothing on the score frequencies. As is the case with Research Question #1, 8 combinations of factors (inclusion/exclusion of the studied item, impact, examinee sample size) were considered.

Descriptive statistics of the Type I error rates will be reported. If 1,000 replications are used to estimate Type I error rates, then with 95% confidence, the margin of error for the Type I error rate as a proportion is ±0.014 under the $\alpha = 0.05$ condition. Any Type I error rates that fall outside of the estimated margin of error will be identified in relevant tables using boldface or boldface and italic print.

For the fourth research question, contingency table modifications were used to compute statistical power based on the assumption that the empirical Type I error is not statistically different from the a priori Type I error. The same three modifications were used on Cox's β statistic, the Liu-Agresti statistic, and the HW3 statistic. Descriptive statistics will be reported as the proportion of replications where the studied item was correctly flagged for DIF.

As was the case with Research Question #2, estimates of true statistical power vary across a range of potential values. Once again, it is assumed that statistical power is 0.50, so that the estimated margin of error is maximized. This provides a conservative estimate of the accuracy of the descriptive statistics. If 1,000 replications are used to estimate statistical power, then with 95% confidence, the margin of error for statistical power (as a proportion) is ±0.031. Assuming that true statistical power is higher than 0.50, the margin of error is likely going to be less than ±0.031.

The fifth research question addresses what factors affect the statistical power for the DIF detection methods and modifications considered in this simulation. To answer this question, analysis of variance (ANOVA) will be used to identify which main effects (examinee sample size, inclusion of the studied item, impact, DIF pattern) and two-way interactions explain differences in statistical power. The model will contain all main effects and all possible two-way interactions. Higher order interactions will be aggregated into the error term.

Chapter 4 contains the results of the simulation described in this chapter. Specifically, the descriptive statistics for Type I error rates and statistical power will be presented to help answer the first four research questions. ANOVA tables will be provided to help answer the final research question. The results presented in the next chapter will be discussed in Chapter 5.

CHAPTER 4

RESULTS

This chapter includes the results of the Monte Carlo simulation described in Chapter 3. To begin, there will be a discussion of instances when DIF statistics or their corresponding modifications could not be calculated. After this discussion, the results for each research question, as described at the end of Chapter 1 and the beginning of Chapter 3, will be addressed. Results for the first four research questions will utilize tables of descriptive statistics; results for the final research question will utilize ANOVA tables. All of these tables, starting with Table A5, can be found in Appendix A.

Instances When DIF Statistics Could Not Be Calculated

During the simulation, there were several occasions when a DIF statistic or a modification was unable to be computed. This section will describe the reasons why this may have occurred. Whenever a DIF statistic could not be estimated, the DIF statistic was labeled as missing, and the Type I error rates and statistical power rates were calculated with the remaining non-missing values.

Table A5 contains the numbers of replications (out of 1,000) where a DIF statistic could not be calculated. For many treatment conditions and DIF statistics, there were no issues, and analysis could be conducted using all 1,000 replications. However, as discussed below, there were three ways that a DIF detection statistic could not be computed under the methods described in Chapter 3. These situations can explain most of the non-zero entries in Table A5.

First, a DIF detection statistic may not have been produced because a 2 x 3 x 4 contingency table could not be created from the original simulated data. This happened in the rare case when none of the examinees scored "2" on the studied item. The simulation interpreted this situation as a 2 x 2 x 4 contingency table, which could not be handled using the *R* functions shown in Tables B1-B4. If this anomaly occurred, no DIF

statistics could be produced, and the replication was flagged across all 12 DIF detection methods in Table A5. This situation only occurred when 40 reference examinees were simulated with impact for the No DIF, Constant DIF, and Convergent DIF cases. For example, in the case of no DIF with simulated impact and the studied item was included in the anchor, 4 of the 1,000 replications failed. In the case of No DIF with simulated impact and the studied item was not included in the anchor test, 11 of the 1,000 replications failed. For the same conditions under constant DIF, 12 of the 1,000 and 9 of the 1,000 replications failed when the studied item was and was not in the anchor test, respectively. Under convergent DIF, 19 of the 1,000 and 21 of the 1,000 replications failed when the studied item was and was not in the anchor test, respectively.

Secondly, there were a number of times when a 2 x 3 x 4 contingency table could be produced, but the Liu-Agresti statistic could not be produced. Reviewing the formula for the variance of the Liu-Agresti statistic in Chapter 2 shows that the variance is not defined when n_{F+k} or n_{R+k} equals zero. There were several contingency tables where n_{F+4} equaled zero, specifically in cases where impact was present. This makes intuitive sense, since focal group examinees would be more commonly assigned low values of k as opposed to high values of k. Under the No DIF, Constant DIF, and Convergent DIF conditions, this situation affected less than 1% of the replications for a given treatment. Under the Divergent DIF condition, this situation occurred much more frequently, affecting up to 14.6% of the replications for a given treatment. This finding has not appeared in previous DIF detection research using the Liu-Agresti statistic.

Finally, there were a number of times when a 2 x 3 x 4 contingency table could be produced, but the HW3 statistic could not be produced. Reviewing the formula for the HW3 statistic, it can be shown that the effect size, ES_k, is undefined when the variances, S^2_{F+k} and S^2_{R+k}, equal zero. When the variances equal zero, the denominator of ES_k is zero, and ES_k is not defined. A variance can equal zero in one of two ways. First, there could have been no focal or reference group examinees for subgroup level k ($n_{F+k} = 0$ or

$n_{R+k} = 0$). Second, all n_{F+k} examinees or all n_{R+k} examinees, for a given subgroup k, obtained the same score on the studied item. In either case, the variance S^2_{F+k} or S^2_{R+k} will equal zero. This occurrence affected a number of replications, particularly in the case where 40 reference examinees were simulated with impact. In the worst case scenario, when divergent DIF with impact was simulated for the $n_R = 40$ case where the studied item was included in the anchor, 44% of the replications were affected by this situation (see Table A5).

Before addressing the research questions, a few additional notes about the analyses are presented here for completeness. When computing DIF statistics under the Bayesian modification, the above anomalies could appear for some of the items in the anchor test. As a result, prior distributions for Bayesian tests were based on only those anchor items where a DIF statistic could be computed.

In the case where the unmodified DIF statistic could not be calculated, the Bayesian version of the DIF statistic could not be produced, since the likelihood of the data $L(\hat{\xi}_w \mid \xi_w)$ could not be defined. Therefore, a missing value was recorded, and the replication was removed when calculating Type I error rates and statistical power rates. The number of undefined Bayesian DIF statistics can be found in Table A5.

When computing replications for the randomization modification, it was possible for the above anomalies to be present in one or more of the 200 replications. When this occurred, the randomization replication was recorded as a missing value, and randomization-based *p*-values were based on the remaining non-missing values. If no randomization replications could be produced, because the unmodified version of the test did not yield a calculable result, the overall *p*-value was recorded as a missing value and removed from the calculation of the Type I error rates and statistical power rates. The number of undefined randomization-based DIF statistics can be found in Table A5.

There were occasions when the log-linear smoothing algorithm was unable to converge to a solution. This can occur when there were few examinees used for

smoothing (in other words, n_{ij+} is small for a given group *i* and studied item score *j*). Occasionally, the smoothing algorithm could not converge, causing the convergence criterion ε to increase to infinity. Rarely, the linear system used in the algorithm was singular and did not have a unique solution. When any of the above situations occurred, the original data were used instead of a smoothed data vector. It is possible that log-linear smoothing could not be conducted on any of the six sets of examinees (2 groups by 3 studied item scores); in these cases, the conclusions made by the log-linear smoothed DIF statistics are identical to their unmodified DIF statistics. Unlike the previous anomalies, however, this situation was not flagged by the simulation program.

A situation that may have led to a non-calculable DIF statistic using log-linear smoothing methods occurred when trying to compute the *K* subtables of the 2 x 3 x 4 smoothed contingency table. Occasionally, when calculating the quartiles used to separate the examinees into the $K = 4$ subgroups, two adjacent quartiles may have been equal to each other (based on the way they were calculated in Table B11). When this happened, a smoothed 2 x 3 x 4 contingency table could not be produced; a smoothed 2 x 3 x 3 contingency table was produced instead. This situation was rare, but did occur several times out of the thousands of replications considered. From Table A5, this situation likely explains the non-zero entries for the situations when log-linear smoothing was used with reference group sample sizes of 400.

There were a number of simulation replications where a DIF statistic could not be produced, as suggested by the evidence in Table A5. A review of the DIF detection formulas suggests that this is possible when the 2 x 3 x 4 contingency table exhibits certain characteristics that are more likely to occur when extremely small sample sizes are used. When $n_{F+k} = 0$ or $n_{R+k} = 0$, the Liu-Agresti and HW3 statistics cannot be calculated without adjusting the original contingency tables. Furthermore, when $S^2_{F+k} = 0$ or $S^2_{R+k} = 0$, the HW3 statistic cannot be calculated without adjustment. Small sample

sizes also have an effect on the log-linear smoothing methodology; the smoothing algorithm may not converge due to the small amount of data used for smoothing.

When considering the rest of this chapter and the next, the reader should keep in mind that Type I error rates and statistical power rates may have been slightly affected by the replications where a DIF detection statistic could not be calculated. Care should be given, especially in the case where 40 reference examinees are simulated with impact, as up to 45% of the replications led to non-calculable DIF statistics (Table A5).

Research Question #1: Type I Error Rates for Unmodified DIF Statistics

The purpose of the first research question is to estimate the "true" Type I error rates for the Mantel test, the Liu-Agresti statistic, and the HW3 statistic for the simulated test at the $\alpha = 0.05$ significance level. If the Type I error rate obtained by the simulation is not close to α, then the statistical power calculations based on $\alpha = 0.05$ may not be accurate.

Type I error rates were calculated by assuming that the parameters of the studied item were identical for the focal and reference groups. The percentage of time that the studied item was (incorrectly) flagged for DIF is the estimated Type I error rate. These error rates appear in Table A6. Type I error rates are listed for each of eight treatment conditions based on the size of the reference group (40 or 400 examinees), whether impact was present, and whether the studied item was included in the calculation of the anchor score. Type I error rates that were smaller than 3.6% (5% minus the margin of error defined in Chapter 3) are highlighted in boldface print. Type I error rates that were larger than 6.4% (5% plus the margin of error) are highlighted in boldface and italic print. These conservative and liberal Type I error rates are beyond what is expected based on random sampling.

Most of the Type I error rates in Table A6 fall within the interval 5% ± 1.4%, the expected range of error rates given the a priori significance level and number of replications simulated. Six of the Type I error rates fell outside of the predicted interval. Two of the rates were larger than 6.4%. In the case where 40 reference group examinees were simulated with impact and the studied item was included in the anchor score, the Type I error rate for the Mantel test was 6.8%. In the case where 400 reference group examinees were simulated with no impact and the studied item was included in the anchor score, the Type I error rate for the Liu-Agresti test was 6.9%.

All four of the Type I error rates for the HW3 statistic with 40 reference and 40 focal examinees were smaller than expected. For example, in the case where 40 reference and 40 focal examinees were simulated with no impact and the studied item was included in the anchor score, only 2.2% of the replications were flagged for DIF. The HW3 statistic yielded Type I error rates that were within the margin of error when 400 reference examinees were simulated. Research by Welch (1993) suggests that small examinee sample sizes yield HW3 Type I error rates that are less than the a priori significance level, particularly for samples sizes of $n_R = n_F = 250$.

Most of the simulated Type I error rates matched the a priori significance levels. Therefore, the statistical power for eighteen of the twenty-four conditions described in Table A6 can be calculated without caution. As a reminder that caution needs to be considered for the six cases where Type I error rates differed significantly from 5%, these cells will be highlighted in boldface (or boldface italics) in the tables that contain statistical power calculations (Tables A7-A9).

Research Question #2: Statistical Power of Unmodified DIF Statistics

Based on the results of the first research question, the statistical power of the Mantel test, the Liu-Agresti statistic, and the HW3 statistic can be evaluated using the

simulation. In the simulation, the statistical power rates for three specific DIF patterns were considered. First, constant DIF was defined by adding a constant (0.35) to the b_2 and b_3 parameters of the reference group's studied item. Second, convergent DIF was defined by adding 0.35 to the reference group's b_2 parameter and adding 0.75 to the reference group's b_3 parameter. Finally, divergent DIF was defined by adding 0.35 to the reference group's b_2 parameter and subtracting 0.35 from the b_3 parameter. (Item characteristic functions for these three situations are provided in Figure A3.)

The statistical power for the Mantel test, the Liu-Agresti statistic, and the HW3 statistic under the constant DIF case can be found in Table A7. For the constant DIF case, it appears that the Mantel test and the Liu-Agresti statistic have approximately the same statistical power, ranging from 17.1% to 57.7%. Although the statistical power rates are similar, they are probably too low to be of practical use in an applied setting. For the case where 400 reference group examinees are simulated, the power of HW3 is slightly less than the power of the Mantel test and the Liu-Agresti statistic. Although the statistical power of HW3 is very low for the 40 reference examinee case, caution is necessary, as the empirical Type I error rate was significantly lower than the a priori significance level.

Table A8 contains power calculations for the simulated test containing convergent DIF. Most of the same trends that appeared for the constant DIF case also appear in the convergent DIF case. For example, the Mantel test and the Liu-Agresti statistic yielded approximately the same statistical power for all eight treatment conditions. Regardless, the statistical power rates, ranging from 32.4% to 66.6%, may be too low to be of use in an applied setting. All three unmodified tests produced similar statistical power for the 400 reference group treatment with no simulated impact. However, the HW3 statistic was slightly less powerful than either the Mantel test or the Liu-Agresti statistic for the 400 reference group treatment with simulated impact. Although the HW3 statistics were less powerful than the Mantel test and the Liu-Agresti statistic for all of the treatments

using 40 reference group examinees, caution must be used in making this conclusion, since the empirical Type I error rate was significantly less than the a priori significance level.

Table A9 presents the power results for the simulated test with divergent DIF. A surprising result from this set of simulations is that, with the exception of the 400 reference group examinees case with simulated impact and the studied item is not included in the anchor, most treatments yielded power less than 20.0%. In the divergent DIF case, the three DIF detection methods produced similar statistical power. However, the Liu-Agresti statistic was less powerful than the Mantel test and the HW3 statistic when impact was introduced in the simulation. In the case where 400 reference examinees were simulated, impact was included, and the studied item was not included in the anchor score, the Liu-Agresti statistic had 30.0% power, 12.4 percentage points lower than the Mantel test under the same treatment conditions. Regardless, the statistical power rates were too low to be of use in an applied setting.

For the constant DIF test and the convergent DIF test, the Mantel test and the Liu-Agresti statistic yielded similar levels of power. For the divergent DIF test, the Mantel test yielded slightly higher levels of power than the Liu-Agresti statistic. Across all three types of DIF patterns, an increase in reference group sample size led to an increase in statistical power. Likewise, including impact in the simulation improved statistical power. In most, but not all cases, removing the studied item from the anchor score increased the statistical power. As will be shown at the end of this chapter, ANOVA results suggests that there were significant differences in power when DIF pattern, reference group sample size, impact, and inclusion/exclusion of the studied item in the anchor score were used as main effect factors.

The magnitude of statistical power was typically too low to be of use in a practical DIF analysis. Statistical power rates of 70% or higher are usually thought to be useful in analysis. All of the power rates produced in the simulation, as shown in Tables

A7-A9, were less than 70%. Only 16 of the 72 power rates calculated were greater than 50%. The magnitudes of these rates would not provide acceptable statistical power for these DIF detection tests.

Research Question #3: Type I Error Rates for DIF Modifications

The third research question addresses the Type I error rates for the modified versions of the Mantel test, the Liu-Agresti statistic, and the HW3 statistic. These modifications include an empirical Bayesian version of each statistic, a randomization based test, and a log-linear smoothing based approach to each statistic. In order to properly study the statistical power of each statistic and modification, it is important that the Type I error rate computed from the simulated data is statistically similar to the a priori significance level of $\alpha = 0.05$. Tables A10, A11, and A12 present the Type I error rates for the Mantel test modifications, the Liu-Agresti statistic modifications, and the HW3 statistic modifications, respectively. As with the tables associated with Research Question #1, Type I error rates that were smaller than 5% minus the margin of error will be identified with boldface text. Type I error rates that were larger than 5% plus the margin of error will be identified with boldface italic text.

Table A10 contains the simulated Type I error rates for the original Mantel test and its modifications. Excluding the Bayesian modification for the moment, twenty-two of the twenty-four treatment/modification combinations yielded Type I error rates that were within 5% plus or minus the margin of error. The two exceptions, the 40 reference examinees case with impact and the studied item was included in the anchor test (unmodified Mantel test) and the 400 reference examinees case without impact and the studied item was included in the anchor test (randomization-based Mantel test), produced Type I error rates of 6.8% and 7.0%.

Recall from Chapter 2 that the Bayesian version of the test is not the same as the classically based hypothesis tests used for the unmodified, randomized, and log-linear smoothed tests. Because of this difference, it appears that the Type I error rates for the Bayesian tests were more conservative than the other tests. Type I error rates for the Bayesian tests ranged from 0.6% to 3.4%. Caution needs to be used when looking at the statistical power rates of Research Question #4 when considering the empirical Bayesian version of the Mantel test.

Table A11 contains the simulated Type I error rates for the Liu-Agresti statistic and its modifications. Like the Mantel test, twenty-two of the twenty-four treatment/non-Bayesian modifications produced Type I error rates that were within 5% plus or minus the margin of error. The two exceptions occurred for the simulated data set with 400 reference group examinees with no simulated impact and the studied item was used in the anchor test. Power for the unmodified Liu-Agresti statistic was 6.9%; power for the randomization test based Liu-Agresti statistic was 7.2%. Like the Bayesian Mantel test results, the Bayesian Liu-Agresti test produced Type I error rates that were significantly smaller than expected. The magnitudes of the Bayesian Liu-Agresti Type I error rates appear to be similar, but not necessarily statistically similar, to the Bayesian Mantel Type I error rates.

Table A12 contains the simulated Type I error rates for the HW3 statistic and its modifications. Unlike the Mantel and Liu-Agresti tests, many of the treatment-modification combinations yielded Type I error rates that were less than expected, especially when 40 reference group examinees were simulated. For the 40 reference examinees, all of the unmodified HW3 and Bayesian HW3 Type I error rates, as well as three of the four log-linear smoothing based HW3 Type I error rates, were less than expected. For the 400 reference examinees cases, all of the Bayesian HW3 Type I error rates were less than expected. In addition, the cases when the studied item was included in the anchor produced Type I error rates that were larger than expected for the

randomization-based HW3 statistic. When analyzing the statistical power results for the HW3 statistics in Research Question #4, caution is necessary, as the empirical Type I error rate is more conservative or more liberal than the a priori significance level.

Research Question #4: Statistical Power of DIF Modifications

The fourth research question addresses which small sample size modifications, Bayesian, randomization-based, or log-linear smoothing based, improved statistical power over an unmodified DIF test. Statistical power rates are presented in Tables A13-A21. Tables A13-A15 contain the statistical power rates for the simulated constant DIF test pattern. Tables A16-A18 contain the power calculations for the simulated convergent DIF test pattern. Finally, Tables A19-A21 contain the power calculations for the simulated divergent DIF test pattern. In all of these tables, the reader should exercise caution when trying to interpret the power calculations for the Bayesian tests. Not only does the Bayesian test differ conceptually from classical hypothesis testing, but the empirical Type I error rates were less than 5% minus the margin of error, suggesting that the true Type I error rate is more conservative than expected with an a priori significance level of 0.05.

The statistical power rates for the various Mantel test modifications for the constant DIF test can be found in Table A13. Note that the first column, marked "Original", contains the statistical power for the Mantel test as reported in Research Question #2. The Bayesian modification seemed to reduce statistical power, though caution should be used in making this interpretation, as the Type I error rates from the Bayesian modification were more conservative than expected. The decrease in statistical power ranged from 11.1 percentage points (40 reference, no impact, studied item included) to 20.7 percentage points (400 reference, impact, studied item included). The statistical power for the randomization-based Mantel test was also smaller than power for

the original Mantel test, though the decrease in power was not as severe. The decrease in statistical power ranged from 4.1 percentage points (40 reference, no impact, studied item included) to 9.9 percentage points (40 reference, impact, studied item excluded). When log-linear smoothing was applied to the contingency tables, statistical power decreased a small amount for all four treatment conditions where impact was not simulated. However, statistical power increased a small amount for all four treatment conditions where impact was simulated. The change in statistical power was, at most, two percentage points in either direction. For the constant DIF case, it appears that log-linear smoothing may improve statistical power of the Mantel test, but only by a very small amount. Generally, the unmodified Mantel test was as powerful as, or more powerful than, any of the alternatives for the constant DIF simulated test. Despite this, all magnitudes of statistical power were lower than 70%, probably too low to be of use in an applied context.

The statistical power rates for the modified Liu-Agresti statistics, as applied to the constant DIF simulated test, appear in Table A14. Although caution must be used in analyzing the Bayesian statistical power rates, Table A14 shows that power was reduced when the Bayesian Liu-Agresti statistic was used as compared to the unmodified Liu-Agresti statistic. The decrease in power ranged from 8.9 percentage points (40 reference, no impact, studied item included) to 20.1 percentage points (400 reference, impact, studied item included). Power was slightly smaller for the randomization-based Liu-Agresti statistic compared to the unmodified Liu-Agresti statistic. The decrease in power ranged from 0.2 percentage points (40 reference, no impact, studied item included and 400 reference, impact, studied item excluded) to 2.1 percentage points (40 reference, impact, studied item excluded). For six of the eight treatments, power increased slightly when using the log-linear smoothing modification. Power increased from 0.4 percentage points (40 reference, no impact, studied item excluded) to 2.1 percentage points (40 reference, impact, studied item included). Power decreased slightly for the two

treatments where 400 reference group examinees were simulated without impact. The unmodified Liu-Agresti statistic was as powerful as, or more powerful, than the modified Liu-Agresti statistics for the constant DIF simulated test. The magnitudes of statistical power were all less than 70%, likely too low to be of use in an applied context.

Table A15 contains the power rates for the HW3 statistic and its modifications for the constant DIF simulated test. Many Type I error rates were significantly different from 5%. For this reason, conclusions about which modification is most powerful must be made with caution. For all eight treatment conditions, the power of the Bayesian HW3 statistic was less than the power of the unmodified HW3 statistic. For all eight treatment conditions, the power of the randomization-based HW3 statistic was slightly better than the unmodified HW3 statistic. The results of the log-linear smoothed HW3 statistic were mixed. The log-linear smoothed HW3 statistic had better power than the unmodified HW3 statistic for five of the eight treatment conditions. Although the significantly different Type I error rates may affect these conclusions, it appears that the randomization-based HW3 statistic may outperform the other HW3 statistics with regards to statistical power. In all cases, however, the magnitude of statistical power was less than 70%, most likely too low to be of use in an applied context.

Table A16 contains the power rates for the Mantel test and its modifications when applied to the convergent DIF simulated test. A comparison of the power rates for the original Mantel test versus the Bayesian Mantel test shows that the original Mantel test yielded higher statistical power. The difference in power ranged from 13.4 percentage points (40 reference, impact, studied item excluded) to 26.1 percentage points (400 reference, no impact, studied item included). Likewise, the original Mantel test yielded higher statistical power than the randomization-based Mantel test across all eight treatment conditions. The difference in power ranged from 5.8 percentage points (40 reference, no impact, studied item included) to 9.6 percentage points (40 reference, impact, studied item excluded). In seven of the eight treatment conditions, the log-linear

smoothed Mantel tests had higher statistical power rates than the unmodified Mantel test. The difference in power, however, was relatively small. The increase in power ranged from 0.2 percentage points (40 reference, no impact, studied item excluded) to 1.8 percentage points (400 reference, impact, studied item included). It seems that the original Mantel test or the log-linear smoothed Mantel test yielded the best statistical power rates for the convergent DIF simulated test. All statistical power rates were less than 70%, probably too low to be of practical use in a DIF analysis.

Table A17 provides the statistical power rates for the Liu-Agresti statistic and its modifications for the simulated test with convergent DIF. Like the Mantel test and its Bayesian counterpart, the unmodified Liu-Agresti statistic produced higher statistical power rates than the Bayesian Liu-Agresti statistic. The difference in power ranged from 15.6 percentage points (40 reference, impact, studied item excluded) to 24.0 percentage points (400 reference, no impact, studied item excluded or included). The unmodified Liu-Agresti statistic was also more powerful than the randomization-based Liu-Agresti statistic, but only by a few percentage points. Differences in power ranged from 1.0 percentage point (400 reference, impact, studied item included) to 4.0 percentage points (40 reference, impact, studied item excluded). In six of the eight treatment conditions, the log-linear smoothed Liu-Agresti statistic had slightly higher statistical power than the unmodified Liu-Agresti statistic. Differences ranged from 0.4 percentage points (40 reference, no impact, studied item included) to 2.6 percentage points (40 reference, impact, studied item included). Given the small improvement in power gained in using the log-linear smoothed Liu-Agresti statistic, it appears that the unmodified and log-linear smoothed Liu-Agresti statistics were the most powerful of the Liu-Agresti statistics for the convergent DIF simulated test. As before, all statistical power rates were less than 70%, perhaps too low to be of practical use.

Table A18 contains the statistical power rates for the HW3 statistics and its modifications for the convergent DIF simulated test. Because a number of the treatments

yielded Type I error rates that were significantly different from 5%, caution will be exercised by only reporting on general trends in Table A18. The unmodified HW3 statistic produced higher statistical power than the Bayesian HW3 statistic across all eight simulation conditions. The randomization-based HW3 statistic produced higher statistical power than the unmodified HW3 statistic in six of the eight simulation conditions (all except the 400 reference, no impact cases). The log-linear smoothed HW3 statistic was more powerful than the unmodified HW3 statistic in four of the eight treatment conditions; all four of the treatment conditions included simulated impact. Across the eight simulation conditions, the randomization-based HW3 appeared to outperform the other three versions of HW3. Like the Mantel and Liu-Agresti tests, the statistical power rates were lower than 70%, too low to be of practical use.

Table A19 contains the statistical power rates for the Mantel test and its modifications for the divergent DIF simulated test. Once again, the power rates for the unmodified Mantel test were greater than the power rates for the Bayesian Mantel test for all eight treatment conditions. The difference between the power rates ranged from 3.2 percentage points (40 reference, no impact, studied item included and excluded) to 7.9 percentage points (40 reference, impact, studied item excluded). The unmodified Mantel test yielded higher statistical power over the randomization based Mantel test in four of the eight simulation conditions, namely the four conditions where impact was simulated. The randomization test performed slightly better when impact was not simulated, but the difference in power was only between 0.4 and 1.3 percentage points. The performance of the log-linear smoothed Mantel test was slightly worse in six of the eight treatment conditions; the differences in power were no more than 1.2 percentage points in these six treatment conditions. It appears that the randomization-based Mantel test performed slightly better when impact was not present, while the log-linear smoothed and unmodified Mantel tests were slightly better when impact was present. In all cases, statistical power rates were too low to be of practical consideration.

The statistical power rates for the Liu-Agresti statistics and its modifications on the divergent DIF simulated test are presented in Table A20. The unmodified Liu-Agresti statistic produced higher statistical power than the Bayesian Liu-Agresti statistic across all eight treatment conditions. The differences between the power rates ranged from 2.8 percentage points (40 reference, no impact, studied item excluded) to 7.0 percentage points (40 reference, impact, studied item included). The randomization-based Liu-Agresti statistic had higher power than the unmodified Liu-Agresti statistic for six of the eight treatment conditions. Differences in power ranged from 0.3 percentage points in favor of the unmodified Liu-Agresti statistic (400 reference, no impact, studied item included) to 9.8 percentage points in favor of the randomization-based Liu-Agresti statistic (400 reference, impact, studied item excluded). Results for the log-linear smoothed Liu-Agresti statistic were mixed, as four of the treatment conditions slightly favored the unmodified Liu-Agresti statistic and the other four slightly favored the log-linear smoothed Liu-Agresti statistic. Differences in power rates ranged from 0.9 percentage points in favor of the unmodified Liu-Agresti statistic (400 reference, no impact, studied item included) to 2.9 percentage points in favor of the log-linear smoothed Liu-Agresti statistic (400 reference, impact, studied item excluded). For the divergent DIF case, the randomization-based and unmodified Liu-Agresti statistics appeared to yield similar statistical power rates. All power rates were lower than 70%, probably too low to be of practical use.

Table A21 contains the power rates for the HW3 statistics and its modifications for the divergent DIF simulated test. Because the majority of treatments yielded Type I error rates significantly different from 5%, caution must be used when analyzing these results. Therefore, only general trends will be reported here. The unmodified HW3 statistic yielded higher statistical power than the Bayesian HW3 statistic in seven of the eight treatment conditions (only the 400 reference, impact, studied item excluded treatment led to higher statistical power for the Bayesian HW3 statistic). The

randomization-based HW3 statistic gave slightly higher statistical power rates over the unmodified HW3 statistic for all eight treatment conditions. Four of the treatments led to slightly better statistical power for the unmodified HW3 statistic over the log-linear smoothed HW3 statistic; three led to better power for the log-linear smoothed HW3 statistic over the unmodified statistic (one treatment resulted in a tie). Based on the statistical power rates, the randomization-based HW3 seemed to yield slightly higher power under most circumstances when divergent DIF is present. All power rates were lower than 70%, too low to be of practical use.

Generally, the modifications to the Mantel, Liu-Agresti, and HW3 statistics did not help to improve statistical power. Under the assumption of constant DIF, the randomization-based and log-linear smoothing-based modifications to the Mantel and Liu-Agresti tests yielded power rates similar to those obtained from the unmodified Mantel and Liu-Agresti tests. Similarly, under the assumption of convergent DIF, the randomization-based and log-linear smoothing-based modifications to the Mantel and Liu-Agresti tests yielded power rates similar to those from the unmodified Mantel and Liu-Agresti tests. Under the assumption of divergent DIF, the results were mixed, depending on the set of treatment conditions simulated. Regardless of the DIF method and modification used, the magnitudes of statistical power were consistently below 70%. These low power rates would probably not be practical in psychometric applications. Although the Type I error rates and power rates were lower for the Bayesian-based DIF methods than their unmodified counterparts, caution is necessary when comparing Bayesian and frequentist hypothesis testing.

Research Question #5: Factors That Affect Power

The final research question used ANOVA methodology to explore which of four factors affected the percentage of time an item was flagged for DIF: DIF pattern, size of the reference group, inclusion or exclusion of impact, and the inclusion or exclusion of

the studied item in the anchor test score. The model included all main effects and two-way interactions. Higher order interaction terms were combined in the error term; this was done to increase the number of degrees of freedom in the error term. The dependent variable was defined as the percentage of time the studied item was flagged for DIF using method *M*, where method *M* was any of the twelve DIF detection methods presented in the first four research questions. The results of the twelve ANOVA tests are presented in the Appendix (Tables A22-A33).

Before analyzing the main effects, proper procedure requires looking at the interaction terms to see if any were statistically significant. The DIF pattern by reference group sample size interaction was statistically significant for all twelve DIF detection methods at the $\alpha = 0.05$ level and ten of the twelve DIF detection methods at the $\alpha = 0.01$ level (the exceptions were the Bayesian Mantel test and Bayesian HW3 statistic). The DIF pattern by impact interaction was statistically significant for eleven of the twelve DIF detection statistics at the $\alpha = 0.01$ level (the Bayesian HW3 statistic was the lone exception). All other two-way interactions were not statistically significant at the $\alpha = 0.01$ level, with the exception of the impact by inclusion of studied item interaction term when the Bayesian Liu-Agresti statistic was used (*p*-value = 0.007, Table A26).

Figures A4-A6 show the interaction plots for the DIF pattern by reference group sample size interaction effects for the twelve studied DIF detection methods. Figure A4 contains the interaction plots for the various Mantel tests. Figure A5 contains the interaction plots for the various Liu-Agresti tests; Figure A6 contains the interaction plots for the various HW3 tests. Note that in all cases, as one goes from "No DIF" to "Divergent" to "Constant" to "Convergent", the percentage of flagged items increased. In all cases, the increases from one DIF pattern to the next were greater for the "Large" sample size case of 400 reference group examinees than for the "Small" sample size case of 40 reference group examinees. Because the lines in the interaction plot are not parallel (or approximately parallel), this provides visual evidence that there was a DIF pattern by

sample size interaction effect. This evidence also appeared as statistically significant DIF pattern by sample size interaction effects in the ANOVA tables.

Figures A7-A9 show the interaction plots for the DIF pattern by impact interaction effects for the twelve studied DIF detection methods. Figure A7 contains the interaction plots for the Mantel tests. Figure A8 contains the interaction plots for the Liu-Agresti tests; Figure A9 contains the interaction plots for the HW3 tests. In all cases, as one goes from "No DIF" to "Divergent" to "Constant" to "Convergent", the percentage of flagged items increased, though the rate of increase for the cases where impact was simulated was larger than the rate of increase for the cases where impact was not simulated. This visual evidence supports the significant DIF pattern by impact interactions terms identified in the ANOVA tables.

Figure A10 presents the interaction plot for the Bayesian Liu-Agresti statistic's impact by inclusion of studied item interaction. With a *p*-value of 0.007, this was the only impact by inclusion of studied item interaction term to be significant at the 0.01 level. Given the large number of hypothesis tests considered in the ANOVA tables, coupled with the evidence that the other impact by inclusion of studied item interaction terms were not significant, this significant result may be a Type I error rather than a true interaction effect.

Having addressed the two-way interaction terms, attention can now be given to the main effects of the model. Descriptive statistics for the ANOVA main effects, including means and standard deviations, are available in Tables A34-A45. From the descriptive statistics (Tables A34-A45), it is clear that there were differences in the average percentage of flagged items depending on the DIF pattern. Although there was an interaction between DIF pattern and sample size, as well as an interaction between DIF pattern and impact, the average percentage of flagged items increased as one goes from "No DIF" to "Divergent" to "Constant" to "Convergent" regardless of the DIF detection method and modification used. For example, from Table A34, the average

percentage of flagged items for the unmodified Mantel test ranged from 5.6%, when no DIF was present, to 48.0%, when convergent DIF was present. From Table A35, the average percentage of flagged items for the Liu-Agresti statistic ranged from 5.7%, when no DIF was present, to 49.3%, when convergent DIF was present. Similar upward trends are evident from the remaining tables.

All twelve ANOVA tables show that reference group sample size was a significant main effect at the $\alpha = 0.01$ level. Tables A34-A45 provide evidence, based on descriptive statistics, that the percentage of flagged items was larger when the reference group sample size was 400 as opposed to when the reference group sample size was 40. For example, from Table A34, the percentage of flagged items was 31.9% when the reference group sample size was 400. The percentage of flagged items was only 20.1% when the reference group sample size was 40. From Table A35, the percentage of flagged items was 31.0% for the larger reference group sample size and 19.8% for the smaller reference group sample size. The largest difference occurred for the unmodified HW3 statistic (Table A36). The percentage of flagged items was 29.9% for the larger reference group and 14.8% for the smaller reference group.

All twelve ANOVA tables show that the inclusion or exclusion of impact was a significant main effect at the $\alpha = 0.01$ significance level. Tables A34-A45 provide evidence, based on descriptive statistics, that the percentage of flagged items was larger when impact was present as opposed to when impact was not present. For example, from Table A34, the percentage of flagged items was 31.2% when impact was present and 20.7% when impact was not present. From Table A35, the percentage of flagged items was 29.9% when impact was present and 20.9% when impact was not present. The largest differences between average percentage of flagged items occurred for the log-linear smoothed Mantel test (Table A43). The percentage of flagged items was 31.9% when impact was present but only 20.3% when impact was not present.

The final main effect, the inclusion or exclusion of the studied item in the anchor score, was significant in two of the twelve ANOVA tables at the $\alpha = 0.01$ significance level and eleven of the twelve ANOVA tables at the $\alpha = 0.05$ significance level. Tables A34-A45 provide evidence, based on descriptive statistics, that the percentage of flagged items was slightly larger when the studied item was excluded from the anchor as opposed to when the studied item was included from the anchor. For example, for the unmodified Mantel test (Table A34), the percentage of flagged items was 27.7% when the studied item was excluded from the anchor, but only 24.2% when the studied item was included in the anchor. For the unmodified Liu-Agresti statistic (Table A35), the percentage of flagged items was 27.1% when the studied item was excluded from the anchor, but only 23.8% when the studied item was included in the anchor. The largest difference occurred for the log-linear smoothed HW3 statistic. In this case (Table A45), the percentage of flagged items was 24.6% when the studied item was excluded in the anchor, but only 20.4% when the studied item was included in the anchor.

Summary

This chapter began with a description of circumstances when DIF detection statistics could not be computed in the simulation. Some contingency tables, namely those where n_{F+k} or $n_{R+k} = 0$ or S^2_{F+k} or $S^2_{R+k} = 0$ for some value k, led to undefined Liu-Agresti and HW3 statistics. The log-linear smoothing algorithm was also affected by extremely small sample sizes. For those situations where DIF detection statistics could be computed, the results were used in answering the five research questions defined in Chapters 1 and 3.

The first two research questions addressed the Type I error rates and the power rates for the unmodified Mantel test, the Liu-Agresti statistic, and the HW3 statistic. For the conditions where 40 reference and 40 focal group examinees were simulated, the Mantel and Liu-Agresti tests yielded Type I error rates similar to the a priori significance

level of 0.05. The HW3 statistic yielded Type I error rates lower than expected. For the conditions where 400 reference group examinees were simulated, all three methods produced Type I error rates around 0.05. Although statistical power rates were lower than 70% across all methods and treatments, the Mantel and Liu-Agresti tests yielded statistical power rates that were approximately equal.

The third and fourth research questions addressed the Type I error rates and power rates for the various small sample size modifications to the Mantel, Liu-Agresti, and HW3 statistics. Due to methodological differences between Bayesian and frequentist hypothesis testing, the Bayesian modifications produced Type I error rates lower than expected. Generally, the randomization-based and log-linear smoothing-based Mantel and Liu-Agresti tests produced Type I error rates that were approximately equal to the a priori significance level of 0.05. Under the constant DIF and convergent DIF conditions, the randomization-based Mantel and Liu-Agresti tests yielded statistical power rates similar to, or slightly less than, the unmodified Mantel and Liu-Agresti tests. Under these same conditions, the log-linear smoothing-based Mantel and Liu-Agresti tests yielded statistical power rates similar to, or slightly larger than, the unmodified tests. Under the divergent DIF conditions, the effects of modifying the Mantel and Liu-Agresti tests were mixed. The randomization-based and log-linear smoothing-based HW3 statistics appeared to improve statistical power rates (compared to the unmodified HW3 statistic) by several percentage points. For all DIF conditions, the magnitudes of the power rates were less than 70%, rates that are probably too low to be effective in practice.

The final research question analyzed which of four treatment conditions had an effect on the percentage of time the studied item was flagged for DIF. DIF pattern, inclusion of impact, and the reference group sample size main effects were significant for all DIF method/modification combinations at the $\alpha = 0.01$ level. As DIF pattern changed from no DIF to divergent DIF to constant DIF to convergent DIF, the percentage of items flagged for DIF increased. The percentage of items flagged for DIF was larger when

impact was included in the simulated data. The percentage of items flagged for DIF was also larger when the reference group sample size was larger. The inclusion/exclusion of the studied item in the anchor score main effect was significant for eleven of the twelve DIF method/modification combinations. Generally, excluding the studied item from the anchor score increased the percentage of items flagged for DIF. The DIF pattern by reference group sample size interaction term was significant for all twelve DIF method/modification combinations. The DIF pattern by impact interaction term was significant for eleven of the twelve DIF method/modification combinations. With one exception, all other two-way interaction terms were not significantly significant.

In the next and final chapter, the results will be analyzed and studied in the context of the DIF detection literature. Implications for statistical and psychometric practice will be provided. In addition, study limitations and future research directions will be discussed.

CHAPTER 5

DISCUSSION

The final chapter begins with a summary of the results presented in Chapter 4. Next, an explanation of the findings will be discussed. Findings will then be related to past literature, revealing that these findings both agree and disagree with prior research. Discussion will progress to show how these results can be applied in practical testing situations. Limitations, particularly limitations that might affect the generalizability of the results, will be discussed. Finally, further research questions that arose from this project will be provided.

Summary of Results

Before summarizing the results, the reader is reminded that there were some instances when a DIF detection method could not produce a valid test statistic (Table A5). Because these replications did not produce a valid test statistic, it is unclear if these replications would have led to an increase or decrease in the Type I error rates or statistical power rates. Special care should be taken when interpreting Type I error rates and power, particularly in cases where a large number of replications produced non-calculable test statistics, such as the case when the HW3 statistic was used on samples with 40 reference group examinees.

The first research question considered the Type I error rates for the Mantel test, the Liu-Agresti test statistic, and the HW3 test statistic under a variety of simulated treatments. The hypothesis was that the Type I error rates would be approximately 5% for all DIF detection methods and treatment conditions, within the margin of error. The results showed that most of the Type I error rates were approximately 5%. The primary exception occurred when the HW3 test statistic was used when evaluating DIF for groups of 40 reference group examinees and 40 focal group examinees. In this case, the Type I error rate was lower than the expected 5%.

The second research question asked what the statistical power rates were for the Mantel test, the Liu-Agresti statistic, and the HW3 statistic under a variety of simulation conditions. Under the simulated case where constant DIF was present in the studied item, the Mantel test and the Liu-Agresti statistic had similar statistical power. Both tests were slightly more powerful than the HW3 statistic, for those treatment conditions where statistical power could be directly compared (namely, those conditions for which reference group size was 400). Under the simulated case where convergent DIF was present for the studied item, the Liu-Agresti statistic had slightly more power than the Mantel test. Both tests had slightly more power than the HW3 statistic, for those treatment conditions where statistical power could be directly compared. Under the simulated case where divergent DIF was present in the studied item, the Mantel test had slightly more power than the Liu-Agresti statistic. Results were mixed for the HW3 statistic. Under some treatment conditions, HW3 actually had higher power rates than the Liu-Agresti statistic. Generally, however, power rates were typically too low for practical use.

The third research question addressed the Type I error rates for the various modifications to the Mantel, Liu-Agresti, and HW3 tests. These modifications included the use of empirical Bayesian techniques, a randomization-based non-parametric test, and the use of log-linear smoothing methods to smooth the observed frequencies. Most of the Type I error rates were approximately 5% for the randomization-based Mantel test, the randomization-based Liu-Agresti test, the log-linear smoothing-based Mantel test, and the log-linear smoothing-based Liu-Agresti test. Type I error rates were significantly lower than 5% for all Bayesian-based versions of the Mantel, Liu-Agresti, and HW3 tests. Furthermore, the Type I error rates were significantly lower than 5% for many of the log-linear smoothing-based HW3 tests for treatment conditions using only 40 reference group examinees.

The fourth research question investigated the statistical power of the various modifications to the Mantel, Liu-Agresti, and HW3 tests. For the case where constant DIF was incorporated in the studied item, the log-linear smoothed Liu-Agresti test and the log-linear smoothed Mantel test yielded statistical power as high as or slightly higher than their unmodified counterparts. The randomization-based Liu-Agresti and randomization-based HW3 tests also yielded high rates of statistical power. Bayesian-based tests did not provide high power rates. For the cases where convergent DIF was incorporated in the studied item, the log-linear smoothing-based Liu-Agresti test and the log-linear smoothing-based Mantel test yielded the highest statistical power rates, slightly higher than the unmodified Liu-Agresti and Mantel tests, in some cases. The randomization-based Liu-Agresti and randomization-based HW3 tests were moderately powerful in detecting DIF, though typically not as powerful as the unmodified versions of these tests. The Bayesian-based tests did not provide high power rates under the convergent DIF condition. For the cases where divergent DIF was simulated in the studied item, statistical power rates were mixed. For the treatment conditions where impact was not simulated, power rates were very low for all DIF detection methods and their modifications. For treatment conditions where impact was simulated, the log-linear smoothed-based Mantel test and the randomization-based HW3 test, along with the unmodified Mantel test, yielded the highest power rates. The Bayesian-based methods provided low power across all treatment conditions. Generally, the power rates obtained using these modifications were too low to be of practical use.

The final research question considered factors that may affect statistical power, including DIF pattern, size of the reference group, the presence of impact, and the inclusion or exclusion of the studied item in the anchor test score. The percentage of flagged replications increased as the DIF pattern changed from no DIF to divergent DIF to constant DIF to convergent DIF. The percentage of flagged items was significantly larger for instances when the reference group sample size was 400 as opposed to 40. The

percentage of flagged items was significantly larger for simulations when impact was simulated compared to those simulations where impact was not included. For many of the DIF detection procedures and modifications, the percentage of flagged replications was significantly larger when the studied item was excluded from the anchor test than when the studied item was included in the anchor test. For all DIF detection procedures and modifications, there was an interaction effect between DIF pattern and sample size. The increases in power across DIF patterns were larger for the large reference group sample size than for the small reference group sample size. In most cases, the increases in power across DIF patterns were larger when impact was simulated than when impact was not simulated.

Synthesis of Findings

This section considers the results of this study in the context of earlier research. To begin, an explanation why the unmodified HW3 test produced Type I error rates that were smaller than expected for smaller sample sizes will be given. Next, consideration will be given to the finding that the unmodified Mantel test and the unmodified Liu-Agresti yielded similar statistical power rates. A discussion about the low Type I error rates for the Bayesian modified DIF statistics will be presented. This will be followed by a discussion of why the log-linear smoothing methods may have improved power for the Mantel and Liu-Agresti tests. Finally, this section will conclude with a discussion about some of the significant factors identified in Research Question #5. This discussion will include information about why the DIF patterns yielded very different power rates, why power increased by increasing the number of reference group examinees while keeping the number of focal group examinees constant, and why impact may have improved statistical power.

Research Question #1

Given the derivation of the Mantel, Liu-Agresti, and HW3 statistics, it is assumed that the Type I error rates would match the predefined significance level α. As shown by the first research question, the Type I error rate did match the a priori significance level under all conditions for the Mantel and Liu-Agresti statistics. The Type I error rates for the HW3 statistic, however, did not match the significance level when the reference group contained 40 examinees. The empirical Type I error rates were smaller than expected for this condition.

As noted, the results of this simulation suggest that the Type I error rate for the Mantel test (assuming $\alpha = 0.05$) is approximately 5%. Past simulation studies have also concluded that the Type I error rate for the Mantel test (assuming $\alpha = 0.05$) is approximately 5%. For example, Zwick, Donoghue, and Grima (1993) obtained a Type I error rate of 4.33% on a simulated polytomous item with $n_{F++} = n_{R++} = 500$ examinees. Given that the Type I error rate was based on 600 replications, 4.33% is within the margin of error assuming 5% is the true Type I error rate. Welch and Hoover found that the Type I error rate ranged from 4% to 6% for the Mantel test under a variety of conditions, including for sample sizes as small as $n_{F++} = n_{R++} = 250$ examinees (1993). Zwick and Thayer (1996) found that the Type I error rate for the Mantel test was 5% when impact was not present in the ability distributions and 3% when impact was present. These Type I error rates were computed using examinees sample sizes of $n_{F++} = n_{R++} = 500$. The results of the current study support the finding that the Type I error rate of the Mantel test is approximately 5%. Note, however, that the present research suggests that the Type I error rate was well-controlled, even for sample sizes as small as $n_{F++} = n_{R++} = 40$.

The results of this simulation suggest that the Type I error rate for the Liu-Agresti statistic (assuming $\alpha = 0.05$) was 5% for a variety of simulated conditions. Though less research is available using the Liu-Agresti statistic in simulations than the Mantel test, the

results of this research are supported by the literature. Research by Penfield and Algina (2003) used examinee sample sizes of $n_{F++} = n_{R++} = 500$ to simulate Type I error rates for the Liu-Agresti statistic. Based on 1,000 replications, Type I error rates ranged from 4.0% to 6.0%. New research by Carvajal and Skorupski (in press) considered the Type I error rates for the Liu-Agresti statistic under small sample size conditions. For samples as small as $n_{F++} = n_{R++} = 50$, Type I error rates ranged from 0% to 10%, although error rates were "generally similar" to 5%. High Type I error rates were believed to have been caused by the low discrimination parameter of the studied item and the inclusion of impact in the ability distributions. The results of this simulation, which included a studied item with high discrimination, support the research of Carvajal and Skorupski.

The results of this simulation suggest that the Type I error rate for the HW3 statistic (assuming $\alpha = 0.05$) was 5% when sample sizes were of moderate size and significantly less than 5% when sample sizes were very small. The simulation results of Welch and Hoover (Welch & Hoover, 1993) support this finding. Type I error rates were found to range from 3% to 8% using sample sizes as small as $n_{F++} = n_{R++} = 250$. Given that 100 replications were produced for each set of treatment conditions, Type I error rates as small as 3% and as large as 8% fall within the appropriate margin of error. Under the assumption of $n_{F++} = 40$ examinees and $n_{R++} = 400$ examinees, the results of this simulation support the findings in the literature.

Under the assumption of $n_{F++} = 40$ examinees and $n_{R++} = 40$ examinees, the Type I error rates were found to be lower than expected in this simulation study. This finding may have been caused by the lack of normality in the test data. Recall that the HW3 statistic is based on the effect size for the two-sample independent *t*-test. One of the assumptions required for the two-sample *t*-test to be valid is that the underlying populations are normally distributed. Failing normality, the *t*-test is still valid if the sample sizes are large enough (Kirk, 1999).

In this simulation, the test data (the score on the studied item for either the focal or reference group) can only take on discrete values of 0, 1, or 2. Because of the discrete nature of the data and the small number of potential values the data can take on, the data cannot be truly normally distributed. However, if the sample sizes are large enough, then the lack of normality can be overlooked (Kirk, 1999). Recall that the HW3 statistic was made up of a composite of $K = 4$ *t*-test effect sizes. As such, the 40 reference group examinees and 40 focal group examinees were divided approximately evenly among the four anchor test categories. Therefore, only approximately ten reference and ten focal group examinees were used in each of the 4 *t*-test effect sizes. Given ties in the anchor score obtained by simulated examinees, the sample sizes may have been less than ten for certain values of *K*. Because the sample data were not normally distributed and the sample sizes used in computing the *t*-test effect sizes were probably too small to overcome non-normality, this may have had an effect on the Type I error rates and the statistical power of the HW3 statistic for the case where 40 reference group examinees were sampled. It is unclear if the effects of non-normal data, especially severely non-normal data, severely impacted the Type I error rates of the HW3 test.

Research Question #2

In studying the second research question, it appears that the unmodified Mantel test and the unmodified Liu-Agresti statistic yielded similar power rates over the treatment conditions considered in the simulation. The statistical relationships between the Mantel test and the Liu-Agresti test may explain why their power rates were so similar. First, the dichotomous versions of these tests, the Mantel-Haenszel test and the Mantel-Haenszel odds ratio, are related; the Mantel-Haenszel test is used to test the null hypothesis that the Mantel-Haenszel odds ratio equals one (Penfield & Camilli, 2007). Second, under the assumption of constant DIF, the Mantel and Liu-Agresti tests are each proportional to the difference between the Samejima location parameters ($b_F - b_R$) under

the partial credit model and the graded response model, respectively (Penfield & Camilli, 2007). Third, research using the Liu-Agresti test statistic has shown it to produce statistical power rates similar to the Mantel test for large sample sizes. Penfield and Algina (2003) showed that the statistical power of the Liu-Agresti statistic was very similar to the statistical power of Cox's β, a mathematical equivalent to the Mantel test, for the case where $n_{F++} = n_{R++} = 500$. Note that the effect size in their simulation was similar in magnitude to the effect size in this simulation. As a result, both simulations obtained statistical power rates that were typically well below 50%. A new finding that can be inferred from the current simulation suggests that the power rates for the Liu-Agresti and Mantel tests were similar for sample sizes as low as $n_{F++} = n_{R++} = 40$.

The results of the current simulation suggest that the Mantel test had statistical power as large as or larger than the HW3 statistic for small sample sizes. Welch and Hoover (1993) studied statistical power rates for the Mantel test and HW3 under a variety of sample size conditions. For sample sizes $n_{F++} = n_{R++} = 1000$, $n_{F++} = n_{R++} = 500$, and $n_{F++} = 500$ and $n_{R++} = 1500$, it was shown that the HW3 statistic produced statistical power as high as or higher than the Mantel test. This may seem like a contradiction when compared to the current results, but for the case when $n_{F++} = n_{R++} = 250$, Welch and Hoover found that the HW3 statistic achieved lower statistical power rates than the Mantel test. The authors acknowledged that the smallest sample size case was "detrimental to the performance" of the HW3 and Mantel statistics (Welch & Hoover, 1993). The evidence of the current simulation suggests that the HW3 statistic may be of suspect utility for very small sample sizes.

It appears that there is no research where the HW3 statistic is compared to the Liu-Agresti statistic. Given the research cited above, it stands to reason that if the Liu-Agresti statistic yielded statistical power rates similar to the Mantel test, and if the Mantel test yielded statistical power rates better than the HW3 statistic for small sample sizes, then the Liu-Agresti statistic ought to yield better statistical power rates better than the

HW3 statistic for small sample sizes. The evidence presented in this current simulation supports this line of reasoning and may be a new finding in the literature.

Research Question #3

In studying the third research question, the results of the simulation showed that the randomization-based and log-linear smoothing-based DIF statistics yielded Type I error rates around the nominal 5% level at most treatment levels. The simulation showed, however, that the Bayesian-based DIF statistics yielded Type I error rates lower than expected. This decrease in statistical power for the Bayesian methods may have occurred because of the differences between Bayesian hypothesis testing and classical hypothesis testing.

Chapter 7 of Gill (2002) presents a thorough explanation of these differences. First, inference procedures in classical hypothesis testing and Bayesian hypothesis testing are based on different probability distributions. In classical testing, inference is based on $f(\mathbf{x} \mid \theta_0)$, where θ_0 is a fixed parameter based on the null hypothesis and $\mathbf{x}$ is a vector of the observed data. In Bayesian testing, inference is based on $f(\theta \mid \mathbf{x})$, where θ is now a random variable. This important distinction about θ affects how inference is conducted. Secondly, the concept of accepting or rejecting the null hypothesis differs between classical and Bayesian testing (Casella & Berger, 2002). Because θ is fixed in classical testing, the probability that the null hypothesis is true given the observed data is either zero or one. In Bayesian testing, however, θ is random, so the probability that the null hypothesis is true given the observed data can take on any value from zero to one, inclusive. Thirdly, the traditional forms of the hypothesis test differ between the classical and Bayesian paradigms. In classical testing, a test statistic is produced and compared to the distribution of the test statistic assuming the null hypothesis is true. In Bayesian testing, θ is a random variable, so the probability $P(\theta \in \Theta_0 \mid \mathbf{x})$ can be calculated, where Θ_0 is the set of parameter values represented by the null hypothesis (Casella & Berger,

2002). If $P(\theta \in \Theta_0 \mid \mathbf{x}) < ½$, then the analyst rejects the null hypothesis. Finally, in this simulation, two-sided hypothesis testing was used. This can easily be considered in the classical testing approach, but this cannot be automatically used in the Bayesian approach (Gill, 2002). Although this problem was circumvented by inverting 95% credible intervals to produce a "test statistic", this approach may have prevented the direct comparison between Bayesian and classical testing results. Although the above differences may make comparisons between Bayesian and classical testing results difficult, this has not stopped researchers from computing Type I error rates and statistical power rates under Bayesian testing (Fidalgo et al., 2007). Fidalgo et al. used a Bayesian loss function to calculate Type I error rates; in their study, the Bayesian Type I error rates were much larger than the classical frequentist counterparts. This discrepancy between the current simulation results and the results of Fidalgo et al. could be the result of using two different Bayesian approaches to evaluating DIF (distribution-based approaches versus loss functions).

Bayesian Type I error rates may have been affected by the factors described above. Bayesian results may also have been affected because of the model specifications used. Namely, the normality assumptions may not have been reasonable, or the prior distribution may not have been appropriate. Based on the statistical theory used to develop Cox's β, the logarithm of the Liu-Agresti statistic, and the HW3 statistic, it should be acceptable to assume that the statistics are normally distributed. The prior distribution may have had a larger effect on the results, though. The prior distributions for the Bayesian statistics were based on small sample sizes and a small number of items. While this simulation used 15 items (and sometimes fewer) to develop the prior distribution, past research using Bayesian DIF techniques incorporated many more items in the development of the prior. Zwick, Thayer, and Lewis (1999) provided an illustrative example with 76 dichotomous items and a simulated example with 36 dichotomous items. Zwick and Thayer (2002) used a test with 150 items to study the

empirical Bayesian Mantel-Haenszel test. A simulation using Empirical Bayesian methods with small sample sizes used a 34-item dichotomously-scored test to produce prior distributions (Fidalgo et al., 2007). Because a small number of polytomous items were used in this simulation, this may have had an effect on the prior distributions, which could have altered the final results.

The variance of the prior distribution was calculated slightly differently compared to past literature. In this simulation, the variance of the prior distribution was the variance of the DIF statistics over the items on the anchor test. Research discussed in Chapter 2, however, used the variance of the DIF statistics minus the average DIF statistic's squared standard error over the anchor test items (Fidalgo et al., 2007; Sinharay et al., 2008; Zwick et al., 1999; Zwick & Thayer, 2002). Preliminary computations using this modified prior variance revealed that many of the prior variances were negative. Because the majority of replications would have been thrown out of the simulation due to negative prior variances, it was decided to use simply the variance of the DIF statistics without subtracting the average DIF statistic's standard error squared. Using the variance of the DIF statistics (without subtracting) led to positive prior variances, but it also led to larger prior variances, which may have placed too much weight on the prior distribution.

To summarize, the Type I error rates for the Bayesian DIF statistics may have been lower than expected for a number of reasons. For example, it may not have been appropriate comparing the Bayesian hypothesis testing results to the classical hypothesis testing results. The assumption of using the normal distribution for the prior and the data may not have been appropriate. The small number of items used to produce the prior distribution may not have been adequate for stable estimation. Finally, the choice of prior distribution, namely the choice of the prior variance, may have been problematic, leading to posterior distributions weighted heavily towards the prior distribution.

The results of the simulation suggest that the randomization-based DIF statistics had Type I error rates that were approximately similar to the a priori significance level.

Of the research that considered randomization-based and exact methods of DIF detection, Parshall and Miller (Parshall & Miller, 1995) studied the Type I error rates of the exact and unmodified Mantel-Haenszel test for dichotomous items. Tables 6 and 7 of their research show that both methods yield approximately the same Type I error rates, as well as approximately the same average Mantel-Haenszel value and the same average standard error.

The results of the simulation suggest that the log-linear smoothing-based DIF statistics had Type I error rates that were approximately similar to the a priori significance level. The original article using log-linear smoothing in DIF analysis, using log-linear smoothing with SIBTEST on dichotomous items, did not study Type I error rates (Puhan et al., 2009). The results of this simulation may be the first study of Type I error rates when using log-linear smoothing on the Mantel, Liu-Agresti, and HW3 tests. The equivalence in Type I error rates may be attributed to the fact that the cells in the contingency tables did not change drastically, and in the cases where log-linear smoothing was not possible due to very small numbers of examinees, cells were not affected at all. These small changes probably did not affect the contingency tables in a major way, so Type I error rates were likely to be similar to the unmodified versions of the statistics.

Research Question #4

The results of this simulation suggest that the Bayesian modifications yielded slightly lower statistical power than the unmodified DIF statistics. In the last section, reasons why the Type I error rates were lower for the Bayesian-based modifications were discussed. Many of these reasons, including the differences between Bayesian and classical hypothesis testing and the influence of the prior distribution, may explain why the Bayesian-based modifications yielded lower statistical power rates.

There is another reason why statistical power was lower for the Bayesian methods. It was shown in Research Question #3 that Bayesian methods yielded Type I error rates that were smaller than expected. In short, the probability of making a Type I error was less than α. Statistical theory states that as a decreases, all other information remaining constant, the probability of making a Type II error (β) increases. Consequently, as the probability of making a Type II error increases, statistical power decreases. Since smaller Type I error rates were observed in the simulation, statistical power should naturally decrease.

Based on the simulation results, it seems that the randomization-based DIF tests did not improve statistical power. In some cases, the randomization-based DIF tests yielded slightly lower power rates. There are two considerations that may have led to this conclusion. One consideration is that the unmodified DIF detection statistics, with their asymptotically-based standard errors, may be robust when used with smaller sample sizes. Parshall and Miller's simulation results showed that the power rates for the exact Mantel-Haenszel test were slightly smaller than the power rates for the asymptotic Mantel-Haenszel test for dichotomously-scored items (Parshall & Miller, 1995, Tables 2, 4, and 7). Additional simulation research by Stephens-Bonty (2008) supported the finding that the exact Mantel-Haenszel test yielded statistical power similar to or slightly less than the statistical power of the asymptotic Mantel-Haenszel test. A second consideration is that the number of samples used in producing the randomization-based sampling distribution may have an effect on the *p*-value of the test, thereby having an effect on the empirically computed statistical power. Due to the size of this simulation, only 200 samples were used to produce the sampling distribution for each replication in the simulation. A larger number of samples would have produced a more precise picture of the sampling distribution with less sampling error. In the past, studies have used only 100 samples to produce the sampling distribution (Camilli & Smith, 1990) of the randomization-based Mantel-Haenszel statistic. Other studies were able to use exact tests

available in StatXact (Parshall & Miller, 1995; Stephens-Bonty, 2008) to compute *p*-values for the Mantel-Haenszel statistic. Unfortunately, exact versions of the Mantel, Liu-Agresti, and HW3 statistics were unavailable for this research. Because of the robustness of the asymptotic DIF statistics and the effects of the number of samples used to produce the *p*-value, statistical power of DIF statistics using randomization-based methods may not be an improvement over the unmodified DIF statistics.

Based on the simulation results, the statistical power of the log-linear smoothing-based DIF detection statistics was as high, or slightly higher than the unmodified versions of the DIF statistics. Puhan et al. (2009) did not discuss statistical power rates when using log-linear smoothing with SIBTEST; the current research may be the first to address statistical power using log-linear smoothing for any polytomous DIF detection method. As described in the previous section, the equivalence (or slight improvement) in statistical power over its unmodified counterpart may be attributed to the fact that the cells in the contingency tables did not change drastically, and in the cases where log-linear smoothing was not possible due to very small numbers of examinees, cells were not affected at all. The small changes likely did not affect the contingency tables in a major way, so statistical rates were similar.

Research Question #5

The results of the fifth research question showed that there were differences in power rates depending on the pattern of DIF appearing in the studied item, the size of the reference group, the inclusion or exclusion of the studied item in the anchor, and the inclusion of impact in the ability distributions. This section will review each of these findings and analyze the results with respect to previous research.

The results of the simulation showed that the percentage of times the studied item was flagged was affected by the pattern of DIF present in the studied item. The power rates for the studied item with convergent DIF were larger than the rates for the studied

item with constant DIF. This is reasonable for two reasons. First, although the difference between the *b*-parameters for the first score level was the same in the constant and convergent DIF cases, the difference between the *b*-parameters for the second score level was larger for the focal group than the reference group. Used as an effect size, this would imply that the convergent DIF case had a larger effect size than the constant DIF case. Second, given the *b*-parameters, the area between the focal and reference item characteristic functions was greater for the convergent DIF case than the constant DIF case. The area between focal and reference item characteristic functions has been used as an effect size for both dichotomous (Finch & French, 2008; Lei et al., 2006; Narayanan & Swaminathan, 1996; Raju, 1988; Raju, 1990) and polytomous items. The item with the larger area between the item characteristic functions has the larger effect size, and statistical tests of DIF for the item should have higher statistical power.

It is unclear, however, why the statistical power rates for the divergent DIF case were so low. Recent research has studied the statistical power of divergent DIF patterns (Penfield, 2007; Penfield, 2008). Like the results of this simulation, past research has revealed that the power rates for items with divergent DIF are typically much smaller than the constant and convergent cases. Under certain conditions, statistical power rates for items with divergent DIF were no better than the Type I error rates. Further research regarding divergent DIF may be necessary to better understand these unusual power rates.

It is expected that increasing the sample size will increase the power of the test (French & Maller, 2007; Herrera & Gómez, 2008; Ross, 2007). One conclusion that was not expected was that increasing only the size of the reference group, while keeping the size of the focal group constant, would be enough to increase statistical power. By studying the formulas for the test statistics, one can see how changing the size of the reference group will affect the statistics. It is believed that increasing n_{R++} will lead to a

decrease in the standard error of the Mantel, Liu-Agresti, and HW3 statistics, thereby increasing statistical power.

The role of impact in this simulation is a bit more difficult to explain. Impact did not seem to affect Type I error rates, but it did affect statistical power, especially in the divergent DIF case. Past research has generally shown that statistical power rates are smaller when unequal latent distributions are considered (Fidalgo et al., 2007; Finch & French, 2007; Penfield, 2008), although power rates have also increased when impact was simulated (Finch & French, 2008; French & Maller, 2007; Penfield, 2007). In the scope of this simulation, by making it less likely for a focal group examinee to score 2 on the studied item, DIF detection techniques may have artificially interpreted this as evidence of DIF.

Finally, the results suggest that excluding the studied item from the anchor test scores led to an increase in statistical power. Although this goes against the statistical theory (Zwick, 1990; Zwick et al., 1993), it may be necessary to omit the studied item from the anchor test. If the studied item contains DIF, this may affect the capability of accurately estimating latent ability levels. It may have been the case that the anchor scores were contaminated by the inclusion of a studied item containing DIF.

This section presented some explanations why certain results were obtained in the simulation. Certainly, there are limitations in the simulation that may have affected the final results presented above. These limitations will be discussed later in this chapter. In the next section, practical implications of the results presented here will be developed.

Applied and Theoretical Implications

Given the results of the simulation, and assuming that the results are valid and accurate, there are a number of ways that these findings can inform polytomous DIF detection in practice. There are also a number of ways that these findings can inform

polytomous DIF detection theory. This section will describe the implications that these simulation results may have in theory and application.

For the analyst conducting a DIF analysis on polytomous items when examinee sample sizes are small, it is important to reduce the data into a contingency table of suitable size that allows for reasonable analysis. For example, the number of score categories on the studied item or the anchor test may need to be decreased in order to avoid large numbers of zero cells in the contingency table, leading to a violation of assumptions common to contingency table tests. This may require the analyst to combine score categories on the studied item to allow for analysis. For example, if the polytomous item's scale ranges from 1 to 5, it may be necessary to combine the middle three categories in order to have sufficient numbers of non-zero cells in the contingency table. Of course, the way that the analyst combines categories may affect the ability to detect DIF (Gelin & Zumbo, 2003).

For small sample sizes, the unmodified Mantel test and the unmodified Liu-Agresti test both provide the analyst with the ability to evaluate an item for DIF with the optimal, though not necessarily practical, level of statistical power under the assumption of constant or convergent DIF. Given that the Liu-Agresti statistic is slightly more complicated to compute and explain to a non-technical audience, the analyst may wish to resort to the computationally and conceptually simpler Mantel test when evaluating items for DIF. If the analyst wishes to use DIF detection tests with slightly higher statistical power, the analyst may wish to apply log-linear smoothing to the contingency table before computing the Mantel or Liu-Agresti tests. Although log-linear smoothing may improve power, the increase may only be a few percentage points.

The unmodified HW3 test is not recommended for the smallest sample sizes. Although the randomization-based HW3 test may improve the power of the HW3 test, the statistic may not be stable when samples sizes as small as 40 reference group and 40 focal group examinees are used. This does not imply that the HW3 is a poor DIF

detection method overall. The use of HW3 may still be valid for polytomous items when large sample sizes are available (Welch & Hoover, 1993; Welch & Miller, 1995).

Bayesian-based DIF detection methods are not recommended when used to conduct hypothesis tests for evaluating DIF, based on the implementation of Bayesian testing and the conditions used in this research. Type I error rates and statistical power, as defined using classical hypothesis testing methodology, were much lower for the Bayesian-based tests than the other DIF detection tests (modified or unmodified). However, as shown by Zwick et al. (1999), Bayesian-based DIF analysis is useful when the analyst wishes to compute the probability of an item being assigned to a series of categories. When applied to ETS's "ABC" method of evaluating items for DIF (Zwick et al., 1999), the Bayesian methodology is recommended.

With the exception of the HW3 test for extremely small sample sizes, randomization-based methods do not appear to be necessary when evaluating DIF using the Mantel or Liu-Agresti statistics. The asymptotically-based unmodified test statistics for the Mantel and Liu-Agresti tests worked well for the sample sizes considered in this simulation compared to modified Mantel and Liu-Agresti tests. Unfortunately, small sample sizes and small effect sizes yielded low power that may not be practical in application.

As suggested above, if the analyst wishes to alter a DIF detection test in an effort to increase statistical power, he or she could use log-linear smoothing on the cells of the contingency table. Log-linear smoothing does not guarantee that statistical power will be higher, nor does it guarantee that statistical power will increase by a large number of percentage points over its unmodified counterpart. However, it may prove useful for the polytomous case as it did with the dichotomous case (Puhan et al., 2009). Smoothing was able to reduce the number of non-calculable HW3 statistics, as shown in Table A5.

Both applied and theoretical psychometricians should exercise caution when divergent DIF is present in a polytomous item. The results of this study were highly

variable when considering the simulated studied item with divergent DIF. Furthermore, power rates were very low for some simulation treatments. As one of the future research items presented later in this chapter, it is recommended that theoretical psychometricians consider the divergent DIF case for further study.

If an applied psychometrician wishes to conduct DIF analyses on polytomously scored items, but cannot obtain a sufficiently large number of focal group examinees for item analysis, then it is recommended that he or she attempt to recruit reference group examinees, for whom it may be easier to obtain testing data. This may be the case during item tryouts when availability prevents the analyst from receiving data from a large number of focal examinees. If there is a large pool of reference group examinees available, the results presented here suggested that the analyst use these additional reference group examinees to improve statistical power.

Finally, the results of this research demonstrated that it is very difficult to obtain high power rates (i.e., power rates above 70% or 90%) with small examinee sample sizes unless there is a very large effect size present. The differences between the *b*-parameters for the focal and reference groups must be very large or the area between the item characteristic functions must be large for statistical power to be high.

Additional implications for DIF detection theory will be discussed in the final section of this chapter. Before discussing these additional theoretical implications, some attention will be given to the limitations of this study. As explained at the start of this section, the implications presented here are only valid if the reader accepts the validity of the simulation. In the next section, limitations that may have affected the simulation results and the generalizability of the study will be discussed.

Limitations of the Study

As with most simulation studies, there were decisions and concessions that were made that could have had an effect on the results. In this section, limitations related to

the design of the study and the generalizability of the study will be discussed. Not only does this section provide some caution while considering the results, but it also provides the reader with some examples of how the simulation could be adapted for future study.

First, as with all computer simulations, it is assumed that the computer code and calculations were accurate. This assumes that the computer code does not contain any programming errors. This also assumes that all mathematical calculations are free of rounding errors and numerical precision errors (Kincaid & Cheney, 2002). As a way of validating the simulation, the code is available in Appendix B. The reader is invited to study the code and confirm its validity. In preparing the DIF detection method functions, several sample contingency tables were produced. DIF detection statistics were calculated for these sample tables using the proposed *R* functions and by hand calculations. Calculations were accurate to at least 6 decimal places, so it is hoped that rounding errors and numerical precision errors were avoided.

As described in Chapter 4, there were some instances when a 2 x 3 x 4 table could not be formed by the algorithm used to produce the contingency tables. This usually occurred when no examinees in a particular subgroup *k* scored "2" on the studied item. When this occurred, the replication was thrown out and the Type I error rates and power rates were calculated based on the number of remaining replications. Though this occurrence did not happen often, these "missing values" may have led to a slight bias in the results.

A bigger problem occurred when HW3 could not be calculated for a given contingency table. This would occur when S^2_{Rk} and S^2_{Fk} equaled zero for a specific subtable *k*. Because the denominator of the effect size could not be calculated when this occurred, HW3 was recorded as a missing value. All Type I error rates and power rates were calculated using only the replications where a valid HW3 statistic was computed. This may have led to bias in the results, particularly in the case when impact was simulated for the divergent DIF case.

Occasionally, the 2 x 3 x 4 contingency table could not be computed for one of the anchor items when computing the Bayesian-based DIF detection tests. In cases where an anchor item's DIF statistic was missing in the Bayesian-based methods, the missing value was removed, and the prior distribution was produced using the non-missing DIF statistics.

During the log-linear smoothing methods, the algorithm used to generate the smoothed counts may not have produced results. This could have happened because there were not enough examinees (usually less than 5) to allow the algorithm to run. Occasionally, the algorithm would not converge, and epsilon (the variable used to evaluate algorithm convergence) would increase to infinity. In the rare case, a solution could not be found because the matrices used in the algorithm were singular. Any time the log-linear smoothing algorithm failed to produce a solution, the original (unsmoothed) data were used.

For the above situations, it may not have been possible to calculate a DIF test statistic. For these situations, it is unclear whether these replications would significantly increase or decrease Type I error rates or statistical power rates. Interpreting the Type I error rates and statistical power rates may be difficult, especially for those simulation conditions where a very large number of replications led to non-calculable test statistics.

One final design consideration that may be a limitation is the manner in which DIF was defined. In the simulation, DIF was defined by adding or subtracting a constant amount from the *b*-parameters of the reference group item characteristic functions. Some would argue that it is more appropriate to define DIF in terms of the area between the reference and focal group item characteristic functions. Because *a*-parameters were not altered by DIF, adding constants to the *b*-parameters and changing the area between the item characteristic functions should be equivalent. However, a third approach to incorporating DIF was not considered in this simulation. Some researchers define DIF as the addition of a second latent variable (dimension) that is not considered in subsequent

analyses (Ackerman, 1992; Camilli, 2006; Roussos & Stout, 1996a; Shealy & Stout, 1993b). Results obtained by defining DIF using 2-dimensional item response theory models may have been different from the results obtained in this simulation. Altered *b*-parameters may be more common in DIF simulation research, but the use of a second latent trait to define DIF may be a more realistic model of what occurs in assessment.

There are also a number of limitations that affected the generalizability of the results. The most obvious limitation is that only one anchor test was simulated. Different item parameters would likely have affected the anchor score. If the anchor test was more or less discriminating, power rates may have improved or suffered. Furthermore, the simulated test was designed with no DIF. This assumption is unlikely true in practice, even though there are ways to reduce the amount of DIF in the anchor test. Although the analyst could use DIF purification methods or use items in an item bank that have been known to be free of DIF, it is typically not possible to remove all evidence of DIF from an anchor test. Because the simulated anchor test was defined to not contain DIF, reducing the noise and variability in the anchor scores, this may have artificially increased statistical power rates.

Conclusions made about constant, convergent, and divergent DIF may not be generalizable across DIF patterns because only one example of each DIF pattern was simulated. This means that only one effect size was considered for each DIF pattern. Because statistical power is related to effect size, a better picture of statistical power could be produced by studying multiple effect sizes per pattern. Studying multiple effect sizes would have helped to better understand the variability of the results for the divergent DIF case. In fact, divergent DIF may be confounded with the scoring of the item. Studying multiple effect sizes may have provided more information about this confounding variable.

In the simulation, the anchor score was reduced from 31 (or 33, if the studied item was included) score levels to four (*K*) score levels using the quartiles of the observed

score distribution as boundaries. Not only does the size of *K* have implications on the validity of the results, but the use of quartiles as boundary criteria could also have implications on validity. Furthermore, because the quartiles were slightly different from one replication to the next, the criteria for belonging to ability level *K* changed from one replication to the next. In practice, the decision to use thick or thin matching and the decisions made to set the boundary criteria for each score level will have a large influence on the validity of DIF analysis.

Throughout the simulation, it was assumed that the item characteristic functions for a given item were parallel; in other words, for a given item, the *a*-parameters were identical regardless of examinee subgroup or score category. In practice, it may not be reasonable to assume that the item characteristic functions are parallel across score categories. Furthermore, DIF may be such that *a*-parameters differ by group. The results of this simulation did not adequately address this situation. Further research should continue to study the effects of differing *a*-parameters on DIF detection.

A number of other factors may have an effect on the statistical power of DIF procedures. As mentioned above, the effect size of DIF, the discrimination of the score levels, the incorporation of DIF in the anchor test, and the use of DIF purification techniques may lead to changes in statistical power of DIF detection tests. Although two different sample sizes were used for the reference group, additional sample sizes could have been used for the focal group to better understand the relationship between sample size and statistical power.

Many of the limitations described in this section suggest ways that this simulation could be adapted for further research. The final section of this chapter will discuss ways that this research can be extended, as well as discuss additional research questions that may be worthy of further study.

Future Research Directions

Based on some of the limitations present in the current study, there are a number of ways that researchers could alter or expand this simulation to better understand certain relationships in polytomous DIF for small sample sizes. For example, only one set of test parameters was simulated for the anchor test. Further research could use different sets of item parameters, especially item parameters based on a variety of realistic testing conditions, to see if the results of the simulation generalize across tests.

In this simulation, only one effect size was used for each DIF pattern studied. It may be beneficial for researchers to investigate a variety of effect sizes while keeping examinee sample size constant (and small) to better understand the effect sizes needed to achieve adequate power rates. This may be especially necessary for the case of divergent DIF. The results for the divergent DIF case varied greatly depending on treatment condition. Further study of this type of DIF may better explain why the simulation results were as varied as they were.

In addition to further study of the DIF patterns, it would be beneficial to study DIF patterns where the discrimination (*a*-) parameters differ between score levels on the studied item and between the focal and reference groups. This non-uniform DIF case would likely be difficult to detect properly, as is the case when non-uniform DIF is applied to dichotomous DIF detection techniques like the Mantel-Haenszel test. Furthermore, the case where the *a*-parameters differ may be more realistic in practice.

Other polytomous DIF detection statistics were not considered in the current study. The standardized mean difference (SMD) statistic (Zwick & Thayer, 1996) and the poly-SIBTEST procedure (Chang et al., 1996) are similar in approach to the Mantel test. Researchers may wish to study the relationship between these statistics and the Mantel, Liu-Agresti, and HW3 statistics for small examinee sample sizes. DIF detection methods based on regression modeling, including polytomous logistic regression (French & Miller, 1996), hierarchical logistic regression (Swanson et al., 2002; Vaughn, 2006),

and logistic discriminant function analysis (Miller & Spray, 1993) may provide higher power rates for small sample sizes, especially in the case where *a*-parameters are different across groups.

The results of the simulation showed that the HW3 statistic may not be appropriate when sample sizes are extremely small. The HW3 statistic, however, may still be a valid method of evaluating a polytomously-scored item for DIF (Welch & Hoover, 1995). Further simulation of the statistical properties of the HW3 statistic for various sample sizes and effect sizes should help understand the sample sizes for which HW3 is appropriate.

The methodology in this simulation used 2 x 3 x 4 contingency tables. This was done on purpose, in order to produce contingency tables with few cells with zero entries. A common assumption in contingency table-based hypothesis tests is that there are few zeroes in the cells of the contingency table. In practice, this may not be viable. For example, a performance test or psychological instrument may use items with five score categories, requiring 2 x 5 x *K* contingency tables. This may lead to having too many zero cells in the table. The easiest solution would be to reduce the number of score categories by combining the counts for adjacent categories. If there are five score categories, the analyst may need to combine the first two score categories and the last two score categories to form a 2 x 3 x *K* contingency table. One research article (Gelin & Zumbo, 2003) investigated DIF detection when score categories on polytomous items were collapsed into fewer categories, including collapsing the score categories to produce dichotomously scored items. Further research may be necessary to determine if this procedure has a major effect of DIF detection rates and if it should be recommended for small examinee sample sizes.

This simulation focused on DIF detection methods based on hypothesis testing. Not all DIF detection methods require the use of hypothesis testing. Educational Testing Service uses the magnitude of a test statistic, in addition to the significance of the

hypothesis test, as a guide for flagging items for DIF (Camilli, 2006; Penfield & Camilli, 2007; Zieky, 1993; Zwick et al., 1999). Based on the magnitude of the statistic, the studied item is assigned a value of "A", "B", and "C", along with a plus or minus to indicate whether the item favors the reference or focal group. There have also been DIF flagging procedures developed which do not consider hypothesis testing at all. For example, conditional *p*-value plots (Muñiz et al., 2001) require the analyst to compare the sum of the differences between difficulty *p*-values at the *K* anchor score levels to a pre-defined cut-score for dichotomous DIF detection. A similar procedure could be developed by computing the average studied item score for the focal and reference groups and each of the *K* subgroups and summing over the differences in the average scores. Delta plots (Angoff, 1982; Muñiz et al., 2001) rely on graphical methods to assess items for DIF. The cut-scores used to determine what is flagged for DIF usually have been set by the researchers who originally developed the DIF detection statistics, based on simulations and a particular set of real test data (Dorans & Kulick, 1983; Dorans, 1989; Muñiz et al., 2001). This creates a tradition of using cut-scores that may or may not be appropriate for a different testing setting. Hypothesis testing circumvents this by applying a standardized test statistic to a studied item, regardless of the assessment, using a cut-score defined by statistical theory. Future research may need to bridge the gap between DIF analysts who wish to use hypothesis testing and those who wish to use the magnitude of the DIF detection statistic and a pre-determined cut-score. Some testing companies, such as Educational Testing Service (Zieky, 1993), use both hypothesis testing and DIF statistic magnitudes to flag items for DIF. Future research may wish to study how much extra information is gained by using both a hypothesis test and DIF magnitude to flag items for DIF.

This section focused on a number of possible research directions that relate to the computational aspects of DIF detection methodology. Before concluding this section, however, it is necessary to consider one major component in DIF detection that cannot be

studied via simulation. In this final discussion, it is hoped that psychometricians who use DIF detection methods in practice put substantial thought into their definitions of focal and reference groups.

In practice, it is important for the analyst to develop focal and reference groups that are relevant to the test developers and users. The literature review in Chapter 2 showed that gender and race were most commonly used to define focal and reference groups. Blindly using race as a subgroup designator may be problematic. As an example, an analyst may use "Hispanic" as the focal group and "non-Hispanic" as the reference group. Hispanics, though, come from a wide variety of backgrounds and circumstances. Therefore, the Hispanic group would include fourth-generation Americans with Latin ancestry, as well as English as a second language (ESL) students whose parents emigrated into the United States. The Hispanic group would include represent a wide variety of cultures, including the United States, Mexico, Puerto Rico, Cuba, the Caribbean, Brazil, and other Latin American countries. Trying to understand DIF among such a diverse group of examinees may be problematic, as an item may exhibit DIF for some student backgrounds, but not others. Researchers should devote some attention to better understanding the use of broad ethnicity categories in DIF detection. New federal regulations regarding the collection of ethnicity data in student testing should help researchers obtain a more precise picture of the focal and reference groups that they intend to study.

In the post-Elementary and Secondary Education Act era, special emphasis has been given to assuring that those students in low economic backgrounds are given the same educational opportunities as those in those in higher economic backgrounds (Hess & Petrilli, 2006; McGuinn, 2006; Sadovnik, O'Day, Bohrnstedt, & Borman, 2008). It is surprising, then, that little DIF detection research is available using economic background as a focal group designation. Given the advances in data collection in large scale assessment, particularly the collection of variables that better address socio-economic

status (for example, if a K-12 examinee is eligible for free or reduced lunch), DIF analysis using socio-economic status may be useful in better understanding the performance of test items.

Conclusion

Although there are differences in opinion regarding the utility of further research in differential item functioning (Wainer, 2010), it is believed that there is room for the study of DIF detection methodology under unique situations, such as the case when examinee sample sizes are small. The ability to conduct DIF analyses on performance assessments, particularly in the early item try-out phases of test development, requires the use of best practices to address small examinee sample sizes. The results of this simulation suggest that the Mantel and Liu-Agresti statistics, without modification, provide the analyst with the tools necessary to study DIF using analytical methods, although caution should be exercised when dealing with small examinee sample sizes because of suspect power values. Because the Mantel test is computationally simpler to implement and explain to a non-technical audience, the Mantel test may be preferred over the computationally-intensive Liu-Agresti statistic. Because high statistical power requires large effect sizes to be present in the studied item, DIF analysts must continue to supplement their analyses with the results available from fairness and sensitivity reviews. By using both fairness reviews and DIF detection methodology together, the test developer produces better evidence that the test items are or are not reacting differentially to different subgroups. This, in turn, adds to the validity evidence necessary to ensure that the assessment is unbiased and fair to examinees.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Ackerman, T. A. & Evans, J. A. (1992). An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and simultaneous item bias detection procedures. *Annual Meeting of the National Council on Measurement in Education,* San Francisco.

ACT. (2008). *Fairness report for the ACT tests*. Iowa City, IA: ACT.

Agresti, A. (2003). *Categorical data analysis* (2nd ed.) Wiley-Interscience.

Aguerri, M. E., Galibert, M. S., Attorresi, H. F., & Marañón, P. P. (2009). Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test. *Quality and Quantity, 43*, 35-44.

Allen, N. L., & Donoghue, J. R. (1996). Applying the Mantel-Haenszel procedure to complex samples of items. *Journal of Educational Measurement, 33*(2), 231-251.

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

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-96-116). Baltimore, MD: John Hopkins University Press.

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-106.

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.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement, 36*(4), 277-300.

Banks, K., & Walker, C. M. (2006). Performance of SIBTEST when focal group examinees have missing data. *Annual Meeting of the National Council on Measurement in Education,* San Francisco, CA.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*(4), 307-327.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 2*, 113-141.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement, 43*(4), 313-333.

Bolt, D. M., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Camilli, G. (1979). *A critique of the chi square method for assessing item bias*. Boulder, CO:

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-221-256). Westport, CT: Praeger Publishers.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 4*, 323-341.

Camilli, G., & Penfield, R. D. (1992). Differential test functioning: Definition and assessment with the Mantel-Haenszel chi-square. *Northeastern Educational Research Association Annual Meeting*, Ellenville, NY.

Camilli, G., & Smith, J. K. (1990). Comparison of the Mantel-Haenszel test with a randomized and a jackknife test for detecting biased items. *Journal of Educational Statistics, 15*, 53-67.

Carvajal, J., & Skorupski, W. P. (in press). The effects of small sample size on identifying polytomous DIF using the Liu-Agresti estimator of the cumulative common odds ratio. *Educational and Psychological Measurement,*

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd Ed. ed.). Pacific Grove, CA: Duxbury.

Chaimongkol, S. (2005). *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective.* (PhD Dissertation, Florida State University). , 143.

Chang, H., Mazzeo, J., & Roussos, L. A. (1996). Detecting DIF for polytomously scored items: an adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333-353.

Clauser, B. E. (1993). *Factors influencing the performance of the Mantel-Haenszel procedure in identifying differential item functioning.* (PhD Dissertation, The Univerisity of Massachusetts).

Clauser, B. E., & Mazor, K. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1991). An examination of item characteristics on Mantel-Haenszel detection rates. *Annual Meeting of the National Council on Measurement in Education,* Chicago.

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1993). The effect of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6,* 269-279.

Clauser, B. E., Mazor, K., & Hambleton, R. K. (1994). The effects of score group width on the Mantel-Haenszel procedure. *Journal of Educational Measurement, 31*(1), 67-78.

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

Cleary, T. A., & Hilton, T. L. (1968). An investigation into item bias. *Educational and Psychological Measurement, 8,* 61-75.

Cohen, A. S., & Kim, S. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17,* 39-52.

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

Cole, N. S., & Moss, P. A. (1989). Bias in Test Use. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 201-201-220). New York: American Council on Education.

Collins, W., Raju, N. S., & Edwards, J. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85,* 451-461.

Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: John Wiley & Sons.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B, 20*(2), 215-242.

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine, 23*, 241-256.

DeMars, C. E. (2008). Polytomous differential item functioning and violations of ordering of the expected latent trait by the raw score. *Educational and Psychological Measurement, 68*, 379-396.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*, 149-149-170.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedure for detecting DIF. *Journal of Educational and Behavioral Statistics, 18*, 131-154.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*, 217-233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

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.* Princeton, NJ: ETS.

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*(4), 355-368.

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the MMSE: An application of the Mantel Haenszel and Standardization procedures. *Medical Care, 44*(Suppl. 3), S107-S114.

Douglas, J. A., Stout, W., & DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics, 21*, 333-363.

Edelen, M. O., McCaffrey, D. F., Marshall, G. N., & Jaycox, L. H. (2008). Measurement of teen dating violence attitudes: an item response theory evaluation of differential item functioning according to gender. *Journal of Interpersonal Violence,*

Edelen, M. O., Thissen, D., Teresi, J., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. *Medical Care, 44*, S134-S142.

Educational Testing Service. (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service. (2009). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Educational Testing Service.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*, 54-77.

Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51*, 243-251.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004a). Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples. *Educational and Psychological Measurement, 64*, 925-936.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004b). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for Type I and Type II error rates. *Journal of Experimental Education, 73*, 23-39.

Fidalgo, A. M., Hashimoto, K., Bartram, D., & Muñiz, J. (2007). Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions. *Journal of Experimental Education, 75*, 293-314.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68*, 940-958.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification types on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5*, 43-53.

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

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: a comparison of four methods. *Educational and Psychological Measurement, 67,* 565-582.

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

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of polytomous-DFIT framework. *Applied Psychological Measurement, 23,* 309-326.

Forsyth, R. A., Ansley, T. N., Feldt, L. S., & Alnot, S. D. (2003). *Guide to research and development: Iowa Tests of Educational Development.* Iowa City, IA: The University of Iowa.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315-332.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67,* 373-393.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: an illustration with the Center for Epidemiologic Studies Depression Scale. *Educational and Psychological Measurement, 63*, 65-74.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman & Hall/CRC.

Gotzman, A. J., & Boughton, K. A. (2004). A comparison of Type I error and power rates for the Mantel-Haenszel and SIBTEST procedures when group difference are large and unbalanced. *Annual Meeting of the American Educational Research Association,* San Diego, CA.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12,* 211-235.

Hanson, B. A. (1998). Uniform DIF and DIF defined by differences in item response functions. *Journal of Educational and Behavioral Statistics, 23,* 244-253.

Harvey, A. L. (1990). The stability of the Mantel-Haenszel d-DIF statistic across populations differing in ability. *Annual Meeting of the National Council on Measurement in Education,* Boston, MA.

Hauck, W. W. (1979). The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics, 35*, 817-819.

Herrera, A., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality and Quantity, 42*, 739-755.

Hess, F. M., & Petrilli, M. J. (2006). *No child left behind primer*. New York: Peter Lang Publishing.

Hidalgo, M. D., & Gómez, J. (2006). Nonuniform DIF detection using discriminant logistic analysis and multinomial logistic regression: A comparison for polytomous items. *Quality and Quantity, 40*, 805-823.

Hidalgo-Montesinos, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*, 1-11.

Hidalgo-Montesinos, M. D., & López-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62*, 32-44.

Hidalgo-Montesinos, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

Hills, J. (1990). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*, 5-11.

Holland, P. W. (1985). On the study of differential item performance without IRT. *27th Annual Conference of the Military Testing Association* (pp. 282-287). San Diego, CA: Navy Personnel Research and Development Center.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133-133-183.

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.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *Guide to research and development: Iowa Tests of Basic Skills*. Iowa City, IA: The University of Iowa.

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

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Author.

Kim, S. (2000). An investigation of the likelihood ratio test, the Mantel test, and the generalized Mantel-Haenszel test of DIF. *Annual Meeting of the American Educational Research Association*, New Orleans, LA.

Kim, S., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51-66.

Kim, S., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355.

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93-116.

Kim, S., Cohen, A. S., & Kim, H. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*, 217-228.

Kincaid, D., & Cheney, W. (2002). *Numerical analysis: Mathematics of scientific computing* (3rd Ed. ed.). Pacific Grove, CA: Brooks/Cole.

Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Fort Worth, TX: Harcourt Brace.

Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement, 45*(3), 271-285.

Kok, F. G., Mellenbergh, G. J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22*, 295-303.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*, 935-953.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th Ed. ed., pp. 387-387-431). Westport, CT: American Council on Education.

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*(4), 21-27.

Lautenschlager, G. J., Flaherty, V. L., & Park, D. (1994). IRT differential item functioning: an examination of ability scale purifications. *Educational and Psychological Measurement, 54*, 21-31.

Lee, Y., & Klockars, A. J. (2005). The effects of sample size, test length, and inclusion criteria on the use of SIBTEST. *Annual Meeting of the American Educational Research Association,* Montréal, Quebec.

Lei, P., Chen, S., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*(4), 245-264.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: John Wiley and Sons.

Lewis, J., & Loftus, W. (2001). *Java software solutions: Foundations of program design* (2nd ed.). Reading, MA: Addison-Wesley.

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

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*, 164-174.

Liu, I., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH DIF across populations. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (pp. 171-196). Hillsdale, NJ: Erlbaum.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2008). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement,*

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortings (Ed.), *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.

Lu, S. (1996). *The relationship between item statistics and the Mantel-Haenszel and standardization DIF statistics when comparison groups differ in ability.* (PhD Dissertation, The University of Iowa).

Luppescu, S. (2002). DIF detection in HLM. *Annual Meeting of the American Educational Research Association,* New Orleans, LA.

Maller, S. J. (2001). Differential item functioning in the Wisc-III: item paramters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*, 793-817.

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.

Marañón, P. P., Garcia, M. I. B., & San Luis Costas, C. (1997). Identification of nonuniform differential item functioning: a comparison of Mantel-Haenszel and item rersponse theory analysis procedures. *Educational and Psychological Measurement, 57*, 559-568.

Mazor, K., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

Mazor, K., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*, 284-291.

Mazor, K., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analyses: the effects of matching on unidimensional subtest scores. *Applied Psychological Measurement, 22*, 357-367.

McGuinn, P. J. (2006). *No child left behind and the transformation of federal education policy, 1965-2005*. Lawrence, KS: University Press of Kansas.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31*, 430-455.

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

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement, 41*(4), 331-344.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*(2), 107-122.

Millsap, R. E., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement, 17*(4), 297-334.

Monahan, P. O., & Ankenmann, R. D. (2005). Effect of unequal variances in proficiency distributions on Type-I error of the Mantel-Haenszel chi-square test for differential item functioning. *Journal of Educational Measurement, 42*(2), 101-131.

Morales, L. S., Flowers, C. P., Gutiérrez, P., Kleinman, M., & Teresi, J. (2006). Item and scale differential functioning of the Mini-Mental Status Exam assessed using the DFIT methodology. *Medical Care, 44*, S143-S151.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115-135.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36*(3), 217-232.

Nandakumar, R., & Roussos, L. A. (1997). Validation of CATSIB to investigate DIF of CAT data. *Annual Meeting of the American Educational Research Association*, Chicago.

Nandakumar, R., & Roussos, L. A. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests.* Princeton, NJ: Law School Admission Council.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

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

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment, 18*, 9-15.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34*(3), 253-272.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT). *Journal of Educational Measurement, 43*(1), 1-17.

Park, D., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement, 14*, 163-173.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: a comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32*(3), 302-316.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3), 187-210.

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement, 45*(3), 247-269.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*(4), 353-370.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*(4), 295-312.

Penfield, R. D., Alvarez, K., & Lee, O. (2009). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*, 61-78.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Psychometrics* (pp. 125-168; 5). Amsterdam: Elsevier.

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice, 28*(1), 38-49.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: review and recommedations. *Educational Measurement: Issues and Practice, 19*(3), 5-15.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.

Puhan, G., Moses, T. P., Yu, L., & Dorans, N. J. (2009). Using log-linear smoothing to improve small-sample DIF estimation. *Journal of Educational Measurement, 46*(1), 59-83.

Quenouille, M. H. (1956). Note on bias in estimation. *Biometrika, 43*, 353-360.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied Measurement in Education, 2*, 1-13.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33*, 133-147.

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.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Lawrence Erlbaum Associates.

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: the case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement, 69*, 18-34.

Ross, T. R. (2007). *The impact of multidimensionality on the detection of differential bundle functioning using SIBTEST.* (PhD Dissertation, Georgia State University).

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24,* 293-322.

Roussos, L. A., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Roussos, L. A., & Stout, W. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230.

Ryan, K. E. (1991). The performance of the Mantel-Haenszel procedure across samples and matching criteria. *Journal of Educational Measurement, 28*(4), 325-337.

Sadovnik, A. R., O'Day, J. A., Bohrnstedt, G. W., & Borman, K. M. (Eds.). (2008). *No child left behind and the reduction of the achievement gap: Sociological perspectives on federal educational policy*. New York: Routledge.

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

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6,* 255-270.

Scheuneman, J. D. (1976). Validating a procedure for assessing bias in test items in the absence of an outside criterion. *Annual Meeting of the American Educational Research Association,* San Francisco, CA.

Scheuneman, J. D. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143-152.

Schmeiser, C. B., Geisinger, K. F., Johnson-Lewis, S., Roeber, E. D., & Schafer, W. D. (1995). *Code of professional responsibilities in educational measurement.* Washington, DC: National Council on Measurement in Education.

Shealy, R., & Stout, W. (1991). *A procedure to detect test bias present simultaneously in several items.* Unpublished manuscript.

Shealy, R., & Stout, W. (1993a). An item response theory model for test bias. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Erlbaum.

Shealy, R., & Stout, W. (1993b). 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*(2), 159-194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of six procedures for detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics, 6,* 317-375.

Shin, S. (1992). *An empirical investigation of the robustness of the Mantel-Haenszel procedure and sources of differential item functioning.* (PhD Dissertation, University of Illinois).

Sinharay, S., Dorans, N. J., Grant, M. C., & Blew, E. O. (2008). Using past data to enhance small sample DIF estimation; a Bayesian approach. *Journal of Educational and Behavioral Statistics,*

Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement, 66,* 305-317.

Spray, J. A., & Miller, T. R. (1992). *Performance of the Mantel-Haenszel statistic and the standardized difference in proportion correct when population distributions are incongruent.* Iowa City, IA: ACT.

Stephens-Bonty, T. A. (2008). *Using three different categorical data analysis techniques to detect differential item functioning.* (Ph.D., Georgia State University).

Stout, W., Li, H., Nandakumar, R., & Bolt, D. M. (1997). MULTISIB: a procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21,* 195-213.

Su, Y., & Wang, W. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning in polytomous items. *Applied Measurement in Education, 18,* 313-350.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*(53-75)

Teresi, J., Ocepek-Welikson, K., Kleinman, M., Cook, K., Crane, P. K., Gibbons, L. E., & Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Quality of Life Research, 16*, 43-68.

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detecting of differential item functioning using the parameters of item response models. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67-111). Hillsdale, NJ: Lawrence Erlbaum.

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

Van der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21*, 131-145.

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach.* (PhD Dissertation, Florida State University).

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.

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics, 35*, 5-5-25.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 61-100). Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: definitions and detection. *Journal of Educational Measurement, 28*(3), 197-219.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: a cognitive explanation for DIF. *Journal of Educational Measurement, 38*(2), 147-163.

Wang, M., & Russell, S. S. (2005). Measurement equivalence on the Job Descriptive Index across Chinese and American workers: Results from confirmatory factor analysis and item response theory. *Educational and Psychological Measurement, 65*, 709-732.

Wang, W., & Su, Y. (2004a). Effect of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17*, 113-144.

Wang, W., & Su, Y. (2004b). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450-480.

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

Welch, C. J. (2006). Item and prompt development in performance testing. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 303-303-327). Mahwah, NJ: Lawrence Erlbaum Associates.

Welch, C. J., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education, 6*, 1-19.

Welch, C. J., & Miller, T. R. (1995). Assessing differential item functioning in direct writing assessments: problems and an example. *Journal of Educational Measurement, 32*(2), 163-178.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement, 59*, 910-927.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*, 22-42.

Wood, S. W., & Ansley, T. N. (2008). An investigation of the validity of standardized achievement tests for predicting high school and first-year college GPA and college entrance examination scores. *Annual Meeting of the National Council on Measurement in Education,* New York.

Woods, C. M. (2008a). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement, 68*, 571-586.

Woods, C. M. (2008b). Likelihood-ratio DIF testing: effects of nonnormality. *Applied Psychological Measurement, 32*, 511-526.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33,* 42-57.

Wright, D. J. (1986). An empirical comparison of the Mantel-Haenszel and standardization methods of detecting differential item performance. *Annual Meeting of the National Council on Measurement in Education,* San Francisco, CA.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: a study evaluating a two-stage approach. *Educational and Psychological Measurement, 63,* 51-64.

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

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 359-376). Mahwah, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational and Behavioral Statistics, 15,* 185-197.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 221-244). Boston: Kluwer Academic Publishers.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Zwick, R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21,* 187-201.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test. *Applied Psychological Measurement, 26,* 57-76.

Zwick, R., & Thayer, D. T. (2003). *An empirical Bayes enhancement of Mantel-Haenszel DIF analysis for computer-adaptive tests.* Newtown, PA: Law School Admission Council.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1-28.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10,* 321-344.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121-140.

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32*(4), 341-363.

APPENDIX A

TABLES AND FIGURES

Table A1. Notation Used in DIF Detection Contingency Tables

	y_1	y_2	. . .	y_j	. . .	y_J	*Total*
Focal	n_{F1k}	n_{F2k}		n_{Fjk}		n_{FJk}	n_{F+k}
Reference	n_{R1k}	n_{R2k}		n_{Rjk}		n_{RJk}	n_{R+k}
Total	n_{+1k}	n_{+2k}		n_{+jk}		n_{+Jk}	n_{++k}

Note: The variable n_{ijk} represents the number of examinees in group i (focal or reference) that obtained a score of y_j on the studied item and a score of x_k on the anchor test. The above notation is used for anchor score category k.

Table A2. Hypothetical Example of 2 x *J* x *K* Contingency Table Used in a DIF Detection Analysis

		$y_1 = 0$	$y_2 = 1$	$y_3 = 2$
$k = 1$	Focal	1	3	9
	Reference	4	4	4
$k = 2$	Focal	4	0	7
	Reference	6	3	6
$k = 3$	Focal	2	0	11
	Reference	6	2	5
$k = 4$	Focal	3	3	7
	Reference	4	2	4

Table A3. Graded Response Model Item Parameters for Simulated Core Items

Item Number	a	b_2	b_3
1	1.46	0.16	1.46
2	1.81	-0.17	1.60
3	1.57	0.06	1.69
4	1.89	0.34	1.76
5	1.93	0.15	1.86
6	1.79	-0.07	1.77
7	2.35	0.53	1.85
8	2.12	0.38	1.70
9	2.19	0.19	1.69
10	1.79	0.13	1.62
11	1.75	0.18	1.66
12	1.86	0.31	1.65
13	2.18	0.16	1.75
14	2.14	0.38	1.56
15	2.12	0.39	1.61

Note: Under Samejima's graded response model, the parameter b_1 is not defined.

Table A4. Graded Response Model Item Parameters for Simulated Studied Item

Condition	a	b_2	b_3
Reference Group	1.89	0.01	1.86
Focal Group			
No DIF	1.89	0.01	1.86
Pervasive/Constant	1.89	0.46	2.31
Pervasive/Convergent	1.89	0.46	2.61
Pervasive/Divergent	1.89	0.46	1.41

Note: Under Samejima's graded response model, the parameter b_1 is not defined.

Table A5. Number of Replications Where a DIF Test Could Not Be Computed

				Mantel				Liu-Agresti				HW3			
DP	SS	Im	IA	U	B	R	S	U	B	R	S	U	B	R	S
No	40	Yes	Yes	4	4	4	7	5	5	4	7	289	289	289	77
			No	11	11	11	12	11	11	11	12	103	103	103	70
		No	Yes	0	0	0	0	0	0	0	0	7	7	7	0
			No	0	0	0	0	0	0	0	0	2	2	2	0
	400	Yes	Yes	0	0	0	0	4	4	0	0	35	35	35	0
			No	0	0	0	0	3	3	0	0	32	32	32	0
		No	Yes	0	0	0	0	0	0	0	0	0	0	0	0
			No	0	0	0	0	0	0	0	0	1	1	1	0
C	40	Yes	Yes	12	12	12	12	12	12	12	12	437	437	437	132
			No	9	9	9	9	10	10	9	9	212	212	211	124
		No	Yes	0	0	0	0	0	0	0	0	17	17	17	0
			No	0	0	0	0	0	0	0	0	5	5	5	0
	400	Yes	Yes	0	0	0	0	9	9	0	0	36	36	36	1
			No	0	0	0	0	2	2	0	0	37	37	37	2
		No	Yes	0	0	0	0	0	0	0	0	0	0	0	0
			No	0	0	0	0	0	0	0	0	0	0	0	0
Co	40	Yes	Yes	19	19	19	22	22	22	19	22	450	450	450	148
			No	21	21	21	21	21	21	21	21	207	207	206	157
		No	Yes	0	0	0	0	0	0	0	0	7	7	7	0
			No	0	0	0	0	0	0	0	0	5	5	5	0
	400	Yes	Yes	0	0	0	0	6	6	0	0	45	45	45	4
			No	0	0	0	0	8	8	0	0	36	36	36	2
		No	Yes	0	0	0	0	0	0	0	0	0	0	0	0
			No	0	0	0	0	0	0	0	0	0	0	0	0

(Table A5 continued)

				Mantel				Liu-Agresti				HW3			
DP	SS	Im	IA	U	B	R	S	U	B	R	S	U	B	R	S
Di	40	Yes	Yes	0	0	0	2	0	0	0	2	440	440	440	27
			No	0	0	0	2	0	0	0	2	195	195	195	31
		No	Yes	0	0	0	0	0	0	0	0	26	26	26	0
			No	0	0	0	0	0	0	0	0	10	10	10	0
	400	Yes	Yes	0	0	0	0	116	116	0	0	383	383	383	0
			No	0	0	0	1	146	146	0	3	418	418	418	4
		No	Yes	0	0	0	0	0	0	0	0	1	1	1	0
			No	0	0	0	0	0	0	0	0	0	0	0	0

Note: Cell contents are the number of replications (out of 1,000) where a DIF detection statistic could not be calculated. U=Unmodified, B=Bayesian, R=Randomization-based, S=Log-linear Smoothing; DP=DIF Pattern, SS=Reference group sample size, Im=Impact present, IA=Is studied item included in anchor?; No=No DIF, C=Constant DIF, Co=Convergent DIF, Di=Divergent DIF.

Table A6. Type I Error Rates for Three Polytomous DIF Detection Methods (α = .05)

Sample Size	Impact	In Anchor?	MH	LA	HW3
40/40	No	Yes	.042	.041	**.022**
		No	.051	.051	**.023**
	Yes	Yes	***.068***	.063	**.028**
		No	.051	.050	**.033**
400/40	No	Yes	.062	***.069***	.061
		No	.061	.058	.055
	Yes	Yes	.056	.057	.052
		No	.056	.064	.054

Note: In Anchor? = Is the studied item included in the anchor score? MH = Mantel/Cox's β test, LA = Liu-Agresti statistic, HW3 = HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A7. Statistical Power for Three Polytomous DIF Detection Methods for an Item with Constant Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	MH	LA	HW3
40/40	No	Yes	.185	.171	**.129**
		No	.200	.200	**.131**
	Yes	Yes	***.333***	.337	**.263**
		No	.380	.389	**.282**
400/40	No	Yes	.317	***.320***	.294
		No	.338	.342	.324
	Yes	Yes	.508	.486	.441
		No	.577	.575	.549

Note: In Anchor? = Is the studied item included in the anchor score?, MH = Mantel/Cox's β test, LA = Liu-Agresti statistic, HW3 = HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A8. Statistical Power for Three Polytomous DIF Detection Methods for an Item with Convergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	MH	LA	HW3
40/40	No	Yes	.324	.333	**.249**
		No	.360	.362	**.266**
	Yes	Yes	***.392***	.396	**.307**
		No	.409	.431	**.308**
400/40	No	Yes	.560	***.585***	.547
		No	.566	.590	.561
	Yes	Yes	.577	.584	.508
		No	.653	.666	.594

Note: In Anchor? = Is the studied item included in the anchor score?, MH = Mantel/Cox's β test, LA = Liu-Agresti statistic, HW3 = HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A9. Statistical Power for Three Polytomous DIF Detection Methods for an Item with Divergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	MH	LA	HW3
40/40	No	Yes	.046	.044	**.025**
		No	.045	.045	**.031**
	Yes	Yes	***.136***	.110	**.104**
		No	.189	.147	**.171**
400/40	No	Yes	.084	***.071***	.075
		No	.076	.062	.064
	Yes	Yes	.184	.137	.199
		No	.424	.300	.404

Note: In Anchor? = Is the studied item included in the anchor score?, MH = Mantel/Cox's β test, LA = Liu-Agresti statistic, HW3 = HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A10. Type I Error Rates for Four Variations of the Mantel Test/Cox's β (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.042	**.011**	.057	.038
		No	.051	**.006**	.052	.043
	Yes	Yes	***.068***	**.029**	.054	.052
		No	.051	**.026**	.040	.049
400/40	No	Yes	.062	**.011**	***.070***	.050
		No	.061	**.018**	.058	.052
	Yes	Yes	.056	**.016**	.048	.052
		No	.056	**.034**	.048	.060

Note: In Anchor? = Is the studied item included in the anchor score?, Original = Unmodified Mantel/Cox's β test, Bayesian = Empirical Bayes Mantel Test, Random = Randomization test, Smoothed = Log-Linear smoothing based Mantel test. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A11. Type I Error Rates for Four Variations of the Liu-Agresti Statistic (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.041	**.016**	.046	.038
		No	.051	**.005**	.050	.044
	Yes	Yes	.063	**.028**	.058	.053
		No	.050	**.027**	.052	.055
400/40	No	Yes	***.069***	**.016**	***.072***	.058
		No	.058	**.019**	.053	.049
	Yes	Yes	.057	**.018**	.060	.058
		No	.064	**.035**	.064	.062

Note: In Anchor? = Is the studied item included in the anchor score?, Original = Unmodified Liu-Agresti statistic, Bayesian = Empirical Bayes Liu-Agresti statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based Liu-Agresti statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A12. Type I Error Rates for Four Variations of the HW3 Statistic (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	**.022**	**.008**	.046	**.015**
		No	**.023**	**.000**	.053	**.021**
	Yes	Yes	**.028**	**.011**	.060	**.034**
		No	**.033**	**.013**	.056	.038
400/40	No	Yes	.061	**.012**	***.067***	.041
		No	.055	**.017**	.061	.047
	Yes	Yes	.052	**.018**	***.069***	.043
		No	.054	**.022**	.060	.054

Note: In Anchor? = Is the studied item included in the anchor score?, Original = Unmodified HW3 statistic, Bayesian = Empirical Bayes HW3 statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A13. Statistical Power for Four Variations of the Mantel Test/Cox's β for an Item with Constant Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.185	**.074**	.144	.183
		No	.200	**.076**	.154	.197
	Yes	Yes	***.333***	**.190**	.241	.346
		No	.380	**.230**	.281	.396
400/40	No	Yes	.317	**.145**	***.260***	.306
		No	.338	**.151**	.293	.335
	Yes	Yes	.508	**.301**	.426	.517
		No	.577	**.422**	.515	.596

Note: In Anchor? = Is the studied item included in the anchor score?, Original = Unmodified Mantel/Cox's β test, Bayesian = Empirical Bayes Mantel Test, Random = Randomization test, Smoothed = Log-Linear smoothing based Mantel test. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A14. Statistical Power for Four Variations of the Liu-Agresti Statistic for an Item with Constant Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.171	**.082**	.169	.182
		No	.200	**.084**	.196	.204
	Yes	Yes	.337	**.183**	.322	.358
		No	.389	**.228**	.358	.402
400/40	No	Yes	***.320***	**.157**	***.303***	.312
		No	.342	**.175**	.330	.338
	Yes	Yes	.486	**.285**	.476	.502
		No	.575	**.408**	.573	.590

Note: In Anchor? = Is the studied item included in the anchor score? Original = Unmodified Liu-Agresti statistic, Bayesian = Empirical Bayes Liu-Agresti statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based Liu-Agresti statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A15. Statistical Power for Four Variations of the HW3 Statistic for an Item with Constant Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	**.129**	**.044**	.177	**.111**
		No	**.131**	**.032**	.195	**.127**
	Yes	Yes	**.263**	**.153**	.345	**.256**
		No	**.282**	**.157**	.361	.322
400/40	No	Yes	.294	**.133**	***.305***	.283
		No	.324	**.144**	.337	.315
	Yes	Yes	.441	**.232**	***.483***	.477
		No	.549	**.387**	.580	.564

Note: In Anchor? = Is the studied item included in the anchor score? Original = Unmodified HW3 statistic, Bayesian = Empirical Bayes HW3 statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A16. Statistical Power for Four Variations of the Mantel Test/Cox's β for an Item with Convergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.324	**.164**	.266	.331
		No	.360	**.151**	.285	.362
	Yes	Yes	***.392***	**.242**	.302	.408
		No	.409	**.275**	.313	.420
400/40	No	Yes	.560	**.299**	***.486***	.563
		No	.566	**.306**	.502	.554
	Yes	Yes	.577	**.374**	.502	.595
		No	.653	**.485**	.591	.666

Note: In Anchor? = Is the studied item included in the anchor score?, Original = Unmodified Mantel/Cox's β test, Bayesian = Empirical Bayes Mantel Test, Random = Randomization test, Smoothed = Log-Linear smoothing based Mantel test. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A17. Statistical Power for Four Variations of the Liu-Agresti Statistic for an Item with Convergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.333	**.170**	.312	.337
		No	.362	**.172**	.336	.372
	Yes	Yes	.396	**.239**	.371	.422
		No	.431	**.275**	.391	.451
400/40	No	Yes	***.585***	**.345**	***.558***	.584
		No	.590	**.350**	.561	.583
	Yes	Yes	.584	**.363**	.574	.606
		No	.666	**.506**	.650	.672

Note: In Anchor? = Is the studied item included in the anchor score? Original = Unmodified Liu-Agresti statistic, Bayesian = Empirical Bayes Liu-Agresti statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based Liu-Agresti statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A18. Statistical Power for Four Variations of the HW3 Statistic for an Item with Convergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	**.249**	**.095**	.311	**.236**
		No	**.266**	**.077**	.336	**.254**
	Yes	Yes	**.307**	**.182**	.358	**.318**
		No	**.308**	**.175**	.389	.346
400/40	No	Yes	.547	**.301**	***.546***	.528
		No	.561	**.308**	.550	.525
	Yes	Yes	.508	**.279**	***.563***	.519
		No	.594	**.424**	.630	.617

Note: In Anchor? = Is the studied item included in the anchor score? Original = Unmodified HW3 statistic, Bayesian = Empirical Bayes HW3 statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A19. Statistical Power for Four Variations of the Mantel Test/Cox's β for an Item with Divergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.046	**.014**	.052	.045
		No	.045	**.013**	.058	.040
	Yes	Yes	***.136***	**.067**	.094	.139
		No	.189	**.110**	.138	.177
400/40	No	Yes	.084	**.025**	***.088***	.076
		No	.076	**.026**	.080	.073
	Yes	Yes	.184	**.127**	.153	.172
		No	.424	**.391**	.366	.463

Note: In Anchor? = Is the studied item included in the anchor score?, Original = Unmodified Mantel/Cox's β test, Bayesian = Empirical Bayes Mantel Test, Random = Randomization test, Smoothed = Log-Linear smoothing based Mantel test. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A20. Statistical Power for Four Variations of the Liu-Agresti Statistic for an Item with Divergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	.044	**.015**	.047	.045
		No	.045	**.017**	.049	.047
	Yes	Yes	.110	**.040**	.109	.116
		No	.147	**.080**	.157	.140
400/40	No	Yes	***.071***	**.017**	***.068***	.062
		No	.062	**.020**	.070	.060
	Yes	Yes	.137	**.095**	.173	.133
		No	.300	**.260**	.398	.329

Note: In Anchor? = Is the studied item included in the anchor score? Original = Unmodified Liu-Agresti statistic, Bayesian = Empirical Bayes Liu-Agresti statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based Liu-Agresti statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A21. Statistical Power for Four Variations of the HW3 Statistic for an Item with Divergent Differential Item Functioning (α = .05)

Sample Size	Impact	In Anchor?	Original	Bayesian	Random	Smoothed
40/40	No	Yes	**.025**	**.008**	.048	**.025**
		No	**.031**	**.002**	.053	**.023**
	Yes	Yes	**.104**	**.059**	.163	**.133**
		No	**.171**	**.078**	.207	.159
400/40	No	Yes	.075	**.019**	***.079***	.064
		No	.064	**.022**	.075	.067
	Yes	Yes	.199	**.136**	***.208***	.184
		No	.404	**.436**	.452	.459

Note: In Anchor? = Is the studied item included in the anchor score? Original = Unmodified HW3 statistic, Bayesian = Empirical Bayes HW3 statistic, Random = Randomization test, Smoothed = Log-Linear smoothing based HW3 statistic. Values in boldface represent conditions where Type I error rates were significantly less than 0.05. Values in boldface and italics represent conditions where Type I error rates were significantly larger than 0.05.

Table A22. ANOVA Results for Statistical Power Rates Using the Mantel/Cox's β Test

Source	SS	df	MS	F	*p*-value
DIF Pattern	8933.1	3	2977.7	187.4	< **0.001**
Sample Size	1113.6	1	1113.6	70.1	< **0.001**
Impact	878.1	1	878.1	55.3	< **0.001**
Inclusion in Anchor	98.3	1	98.3	6.19	0.027
Pattern x Size	504.7	3	168.2	10.6	**0.001**
Pattern x Impact	483.6	3	161.2	10.1	**0.001**
Pattern x Inclusion	54.2	3	18.1	1.1	0.371
Size x Impact	22.1	1	22.1	1.4	0.260
Size x Inclusion	18.8	1	18.8	1.2	0.297
Impact x Inclusion	51.7	1	51.7	3.3	0.094
Residuals	206.6	13	15.9		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A23. ANOVA Results for Statistical Power Rates Using the Liu-Agresti Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	10032.4	3	3344.1	434.7	< **0.001**
Sample Size	1009.0	1	1009.0	131.2	< **0.001**
Impact	655.2	1	655.2	85.2	< **0.001**
Inclusion in Anchor	87.1	1	87.1	11.3	**0.005**
Pattern x Size	566.1	3	188.7	24.5	< **0.001**
Pattern x Impact	388.0	3	129.3	16.8	< **0.001**
Pattern x Inclusion	33.9	3	11.3	1.5	0.269
Size x Impact	2.9	1	2.9	0.4	0.548
Size x Inclusion	8.9	1	8.9	1.2	0.301
Impact x Inclusion	44.2	1	44.2	5.7	0.032
Residuals	100.0	13	7.7		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A24. ANOVA Results for Statistical Power Rates Using the HW3 Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	6801.7	3	2267.2	150.4	< **0.001**
Sample Size	1815.1	1	1815.1	120.4	< **0.001**
Impact	648.0	1	648.0	43.0	< **0.001**
Inclusion in Anchor	93.6	1	93.6	6.2	0.027
Pattern x Size	678.9	3	226.3	15.0	< **0.001**
Pattern x Impact	485.8	3	162.0	10.7	**0.001**
Pattern x Inclusion	45.2	3	15.1	1.0	0.424
Size x Impact	12.5	1	12.5	0.8	0.379
Size x Inclusion	29.9	1	29.9	2.0	0.183
Impact x Inclusion	60.6	1	60.6	4.0	0.066
Residuals	196.0	13	15.1		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A25. ANOVA Results for Statistical Power Rates Using the Bayesian Mantel Test

Source	SS	df	MS	F	*p*-value
DIF Pattern	3291.6	3	1097.2	62.2	< **0.001**
Sample Size	659.6	1	659.6	37.4	< **0.001**
Impact	1045.5	1	1045.5	59.3	< **0.001**
Inclusion in Anchor	124.5	1	124.5	7.1	0.020
Pattern x Size	258.8	3	86.3	4.9	**0.017**
Pattern x Impact	301.8	3	100.6	5.7	**0.010**
Pattern x Inclusion	53.2	3	17.7	1.0	0.422
Size x Impact	80.9	1	80.9	4.6	0.052
Size x Inclusion	60.2	1	60.2	3.4	0.088
Impact x Inclusion	121.3	1	121.3	6.9	0.021
Residuals	229.3	13	17.6		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A26. ANOVA Results for Statistical Power Rates Using the Bayesian Liu-Agresti Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	3936.6	3	1313.2	148.1	< **0.001**
Sample Size	618.5	1	618.5	69.7	< **0.001**
Impact	621.9	1	621.9	70.1	< **0.001**
Inclusion in Anchor	109.5	1	109.5	12.3	**0.004**
Pattern x Size	330.3	3	110.1	12.4	< **0.001**
Pattern x Impact	196.3	3	65.4	7.4	**0.004**
Pattern x Inclusion	33.0	3	11.0	1.2	0.335
Size x Impact	34.2	1	34.2	3.9	0.071
Size x Inclusion	41.0	1	41.0	4.6	0.051
Impact x Inclusion	92.4	1	92.4	10.4	**0.007**
Residuals	115.3	13	8.9		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A27. ANOVA Results for Statistical Power Rates Using the Bayesian HW3 Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	2065.5	3	688.5	21.4	< **0.001**
Sample Size	1007.0	1	1007.0	31.3	< **0.001**
Impact	742.2	1	742.2	23.0	< **0.001**
Inclusion in Anchor	115.33	1	115.33	3.6	0.081
Pattern x Size	357.9	3	119.3	3.7	0.040
Pattern x Impact	314.0	3	104.7	3.2	0.057
Pattern x Inclusion	62.2	3	20.7	0.6	0.601
Size x Impact	54.0	1	54.0	1.7	0.218
Size x Inclusion	133.9	1	133.9	4.2	0.062
Impact x Inclusion	128.7	1	128.7	4.0	0.067
Residuals	418.8	13	32.2		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A28. ANOVA Results for Statistical Power Rates Using the Randomization Based Mantel Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	6030.9	3	2010.3	165.6	< **0.001**
Sample Size	1195.1	1	1195.1	98.4	< **0.001**
Impact	454.8	1	454.8	37.5	< **0.001**
Inclusion in Anchor	88.0	1	88.0	7.2	0.018
Pattern x Size	571.1	3	190.4	15.7	< **0.001**
Pattern x Impact	330.2	3	110.1	9.1	**0.002**
Pattern x Inclusion	54.0	3	18.0	1.5	0.265
Size x Impact	54.5	1	54.5	4.5	0.054
Size x Inclusion	29.9	1	29.9	2.5	0.141
Impact x Inclusion	53.2	1	53.2	4.4	0.056
Residuals	157.8	13	12.1		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A29. ANOVA Results for Statistical Power Rates Using the Randomization Based Liu-Agresti Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	8566.6	3	2855.5	198.4	< **0.001**
Sample Size	1200.3	1	1200.3	83.4	< **0.001**
Impact	766.5	1	766.5	53.3	< **0.001**
Inclusion in Anchor	101.5	1	101.5	7.1	0.020
Pattern x Size	547.1	3	182.4	12.7	< **0.001**
Pattern x Impact	416.4	3	138.8	9.6	**0.001**
Pattern x Inclusion	57.7	3	19.2	1.3	0.306
Size x Impact	36.1	1	36.1	2.5	0.137
Size x Inclusion	21.1	1	21.1	1.5	0.247
Impact x Inclusion	57.7	1	57.7	4.0	0.066
Residuals	187.1	13	14.4		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A30. ANOVA Results for Statistical Power Rates Using the Randomization Based HW3 Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	7850.0	3	2616.7	161.7	< **0.001**
Sample Size	1137.7	1	1137.7	70.3	< **0.001**
Impact	953.1	1	953.1	58.9	< **0.001**
Inclusion in Anchor	99.7	1	99.7	6.16	0.027
Pattern x Size	505.5	3	168.5	10.4	**0.001**
Pattern x Impact	560.0	3	186.7	11.5	**0.001**
Pattern x Inclusion	58.4	3	19.5	1.2	0.347
Size x Impact	29.1	1	29.1	1.8	0.203
Size x Inclusion	25.1	1	25.1	1.5	0.235
Impact x Inclusion	51.7	1	51.7	3.2	0.097
Residuals	210.3	13	16.2		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A31. ANOVA Results for Statistical Power Rates Using the Log-Linear Smoothing Modification for the Mantel Test

Source	SS	df	MS	F	*p*-value
DIF Pattern	9469.9	3	3156.6	145.6	< **0.001**
Sample Size	1133.3	1	1133.3	52.3	< **0.001**
Impact	1081.8	1	1081.8	49.9	< **0.001**
Inclusion in Anchor	115.9	1	115.9	5.3	0.038
Pattern x Size	469.4	3	156.5	7.2	**0.004**
Pattern x Impact	529.9	3	176.7	8.1	**0.003**
Pattern x Inclusion	64.0	3	21.3	1.0	0.431
Size x Impact	41.5	1	41.5	1.9	0.190
Size x Inclusion	33.6	1	33.6	1.6	0.235
Impact x Inclusion	72.3	1	72.3	3.3	0.091
Residuals	281.8	13	21.68		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A32. ANOVA Results for Statistical Power Rates Using the Log-Linear Smoothing Modification for the Liu-Agresti Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	10660.0	3	3553.3	325.5	< **0.001**
Sample Size	936.1	1	936.1	85.8	< **0.001**
Impact	835.6	1	835.6	76.5	< **0.001**
Inclusion in Anchor	88.1	1	88.1	8.1	0.014
Pattern x Size	508.4	3	169.5	15.5	< **0.001**
Pattern x Impact	411.7	3	137.2	12.6	< **0.001**
Pattern x Inclusion	33.5	3	11.2	1.0	0.414
Size x Impact	9.8	1	9.8	0.9	0.362
Size x Inclusion	13.2	1	13.2	1.2	0.292
Impact x Inclusion	43.4	1	43.4	4.0	0.067
Residuals	141.9	13	10.9		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A33. ANOVA Results for Statistical Power Rates Using the Log-Linear Smoothing Modification for the HW3 Statistic

Source	SS	df	MS	F	*p*-value
DIF Pattern	6946.9	3	2315.6	105.9	< **0.001**
Sample Size	1756.6	1	1756.6	80.4	< **0.001**
Impact	1058.7	1	1058.7	48.4	< **0.001**
Inclusion in Anchor	140.9	1	140.9	6.4	0.025
Pattern x Size	674.0	3	224.7	10.3	**0.001**
Pattern x Impact	506.9	3	169.0	7.7	**0.003**
Pattern x Inclusion	49.6	3	16.5	0.8	0.539
Size x Impact	20.3	1	20.3	0.9	0.353
Size x Inclusion	37.4	1	37.4	1.7	0.214
Impact x Inclusion	84.3	1	84.3	3.9	0.071
Residuals	284.2	13	21.9		

Note: ANOVA model includes all main effects and two-way interactions. *P*-values in boldface represent effects significant at the $\alpha = 0.01$ level.

Table A34. Means and Standard Deviations for Statistical Power Rates Using the Mantel/Cox's β Test

Source	N	Mean	SD
DIF Pattern			
No DIF	8	5.6	0.8
Divergent DIF	8	14.8	12.5
Constant DIF	8	35.5	13.5
Convergent DIF	8	48.0	12.2
Sample Size			
$n_R = 40$	16	20.1	14.4
$n_R = 400$	16	31.9	23.3
Impact			
No Impact	16	20.7	18.2
Impact Present	16	31.2	20.8
Inclusion in Anchor			
Included in Anchor	16	24.2	19.1
Excluded in Anchor	16	27.7	21.3

Table A35. Means and Standard Deviations for Statistical Power Rates Using the Liu-Agresti Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	5.7	0.9
Divergent DIF	8	11.4	8.5
Constant DIF	8	35.3	13.4
Convergent DIF	8	49.3	12.7
Sample Size			
$n_R = 40$	16	19.8	15.0
$n_R = 400$	16	31.0	23.8
Impact			
No Impact	16	20.9	19.1
Impact Present	16	29.9	21.3
Inclusion in Anchor			
Included in Anchor	16	23.8	19.6
Excluded in Anchor	16	27.1	21.7

Table A36. Means and Standard Deviations for Statistical Power Rates Using the HW3 Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	4.1	1.6
Divergent DIF	8	13.4	12.6
Constant DIF	8	30.2	14.3
Convergent DIF	8	41.8	14.8
Sample Size			
$n_R = 40$	16	14.8	11.5
$n_R = 400$	16	29.9	21.7
Impact			
No Impact	16	17.9	17.9
Impact Present	16	26.9	19.0
Inclusion in Anchor			
Included in Anchor	16	20.6	17.6
Excluded in Anchor	16	24.1	20.2

Table A37. Means and Standard Deviations for Statistical Power Rates Using the Bayesian Mantel Test

Source	N	Mean	SD
DIF Pattern			
No DIF	8	1.9	1.0
Divergent DIF	8	9.7	12.7
Constant DIF	8	19.9	11.8
Convergent DIF	8	28.7	10.9
Sample Size			
$n_R = 40$	16	10.5	9.2
$n_R = 400$	16	19.6	16.9
Impact			
No Impact	16	9.3	10.0
Impact Present	16	20.7	15.6
Inclusion in Anchor			
Included in Anchor	16	13.1	12.1
Excluded in Anchor	16	17.0	16.2

Table A38. Means and Standard Deviations for Statistical Power Rates Using the Bayesian Liu-Agresti Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	2.1	0.9
Divergent DIF	8	6.8	8.4
Constant DIF	8	20.0	10.8
Convergent DIF	8	30.3	11.2
Sample Size			
$n_R = 40$	16	10.4	9.3
$n_R = 400$	16	19.2	16.8
Impact			
No Impact	16	10.4	11.5
Impact Present	16	19.2	15.3
Inclusion in Anchor			
Included in Anchor	16	12.9	12.3
Excluded in Anchor	16	16.6	15.8

Table A39. Means and Standard Deviations for Statistical Power Rates Using the Bayesian HW3 Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	1.3	0.7
Divergent DIF	8	9.5	14.5
Constant DIF	8	16.0	11.2
Convergent DIF	8	23.0	11.8
Sample Size			
$n_R = 40$	16	6.8	6.6
$n_R = 400$	16	18.1	15.8
Impact			
No Impact	16	7.6	10.0
Impact Present	16	17.3	14.5
Inclusion in Anchor			
Included in Anchor	16	10.6	10.1
Excluded in Anchor	16	14.4	15.8

Table A40. Means and Standard Deviations for Statistical Power Rates Using the Randomization Based Mantel Test

Source	N	Mean	SD
DIF Pattern			
No DIF	8	5.3	0.9
Divergent DIF	8	12.9	10.2
Constant DIF	8	28.9	12.7
Convergent DIF	8	40.6	12.7
Sample Size			
$n_R = 40$	16	15.8	10.5
$n_R = 400$	16	28.0	20.3
Impact			
No Impact	16	18.2	15.2
Impact Present	16	25.7	18.5
Inclusion in Anchor			
Included in Anchor	16	20.3	15.8
Excluded in Anchor	16	23.6	18.6

Table A41. Means and Standard Deviations for Statistical Power Rates Using the Randomization Based Liu-Agresti Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	5.7	0.8
Divergent DIF	8	13.4	11.7
Constant DIF	8	34.1	13.4
Convergent DIF	8	46.9	13.0
Sample Size			
$n_R = 40$	16	18.9	13.7
$n_R = 400$	16	31.1	23.0
Impact			
No Impact	16	20.1	17.9
Impact Present	16	29.9	20.6
Inclusion in Anchor			
Included in Anchor	16	23.2	18.7
Excluded in Anchor	16	26.8	21.0

Table A42. Means and Standard Deviations for Statistical Power Rates Using the Randomization Based HW3 Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	5.9	0.7
Divergent DIF	8	16.1	13.5
Constant DIF	8	34.8	13.5
Convergent DIF	8	46.0	12.4
Sample Size			
$n_R = 40$	16	19.7	13.4
$n_R = 400$	16	31.7	22.6
Impact			
No Impact	16	20.2	17.5
Impact Present	16	31.2	19.9
Inclusion in Anchor			
Included in Anchor	16	23.9	18.1
Excluded in Anchor	16	27.5	20.7

Table A43. Means and Standard Deviations for Statistical Power Rates Using the Log-Linear Smoothing Modification for the Mantel Test

Source	N	Mean	SD
DIF Pattern			
No DIF	8	4.9	0.7
Divergent DIF	8	14.8	13.8
Constant DIF	8	35.9	14.3
Convergent DIF	8	48.7	12.2
Sample Size			
$n_R = 40$	16	20.2	15.1
$n_R = 400$	16	32.1	24.1
Impact			
No Impact	16	20.3	18.4
Impact Present	16	31.9	21.8
Inclusion in Anchor			
Included in Anchor	16	24.2	19.8
Excluded in Anchor	16	28.0	22.0

Table A44. Means and Standard Deviations for Statistical Power Rates Using the Log-Linear Smoothing Modification for the Liu-Agresti Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	5.2	0.8
Divergent DIF	8	11.7	9.4
Constant DIF	8	36.1	13.8
Convergent DIF	8	50.3	12.3
Sample Size			
$n_R = 40$	16	20.4	15.9
$n_R = 400$	16	31.2	24.4
Impact			
No Impact	16	20.7	19.2
Impact Present	16	30.9	22.1
Inclusion in Anchor			
Included in Anchor	16	24.2	20.4
Excluded in Anchor	16	27.5	22.1

Table A45. Means and Standard Deviations for Statistical Power Rates Using the Log-Linear Smoothing Modification for the HW3 Statistic

Source	N	Mean	SD
DIF Pattern			
No DIF	8	3.7	1.3
Divergent DIF	8	13.9	14.2
Constant DIF	8	30.7	15.6
Convergent DIF	8	41.8	14.6
Sample Size			
$n_R = 40$	16	15.1	12.1
$n_R = 400$	16	29.9	22.5
Impact			
No Impact	16	16.8	17.3
Impact Present	16	28.3	20.0
Inclusion in Anchor			
Included in Anchor	16	20.4	18.0
Excluded in Anchor	16	24.6	20.9

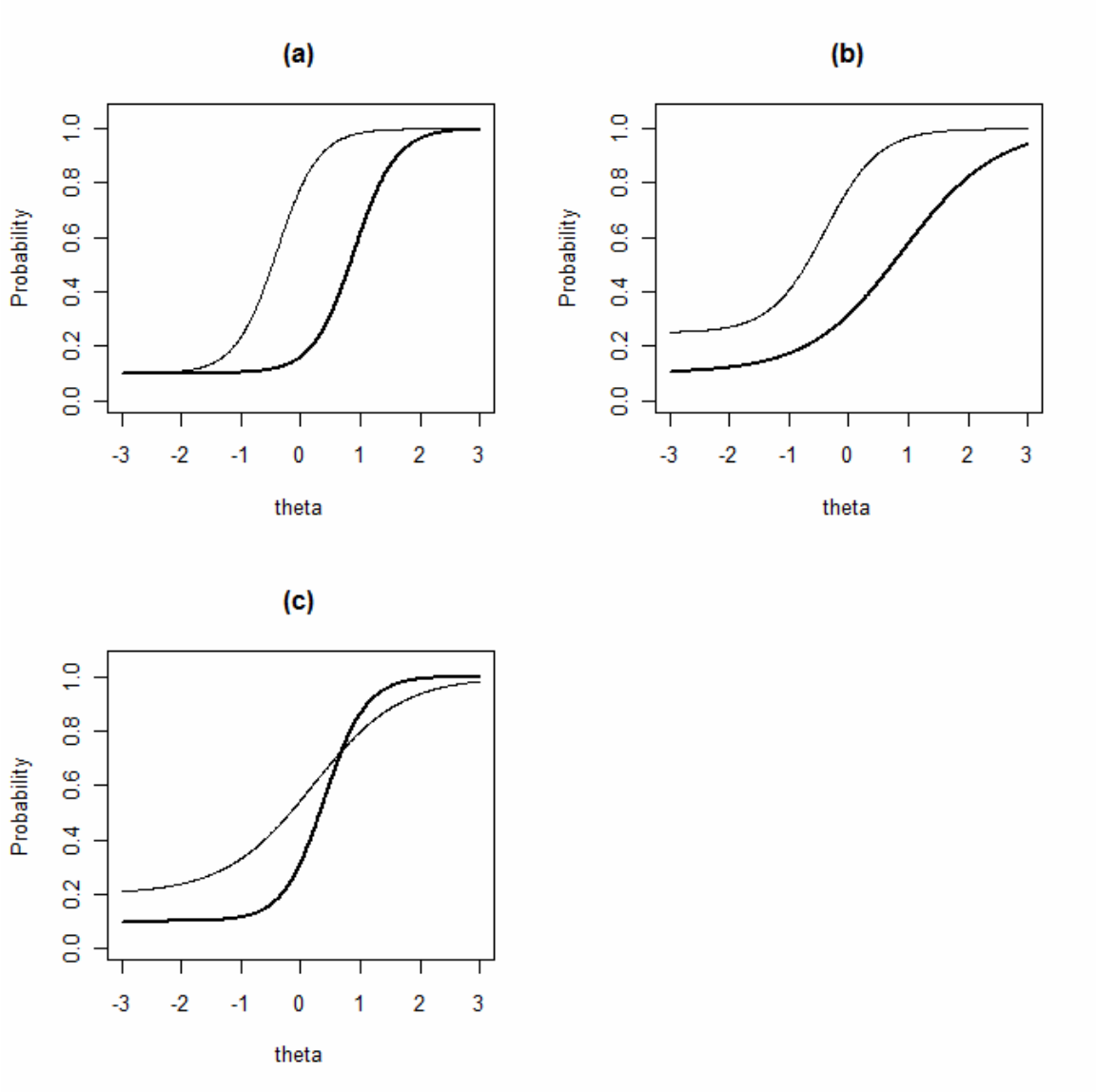

Figure A1. Patterns of Differential Item Functioning for Dichotomous Items. The item characteristic curve for the reference group is the thick line; the item characteristic curve for the focal group is the thin line. Patterns include (a) uniform DIF, (b) unidirectional DIF, and (c) crossing DIF.

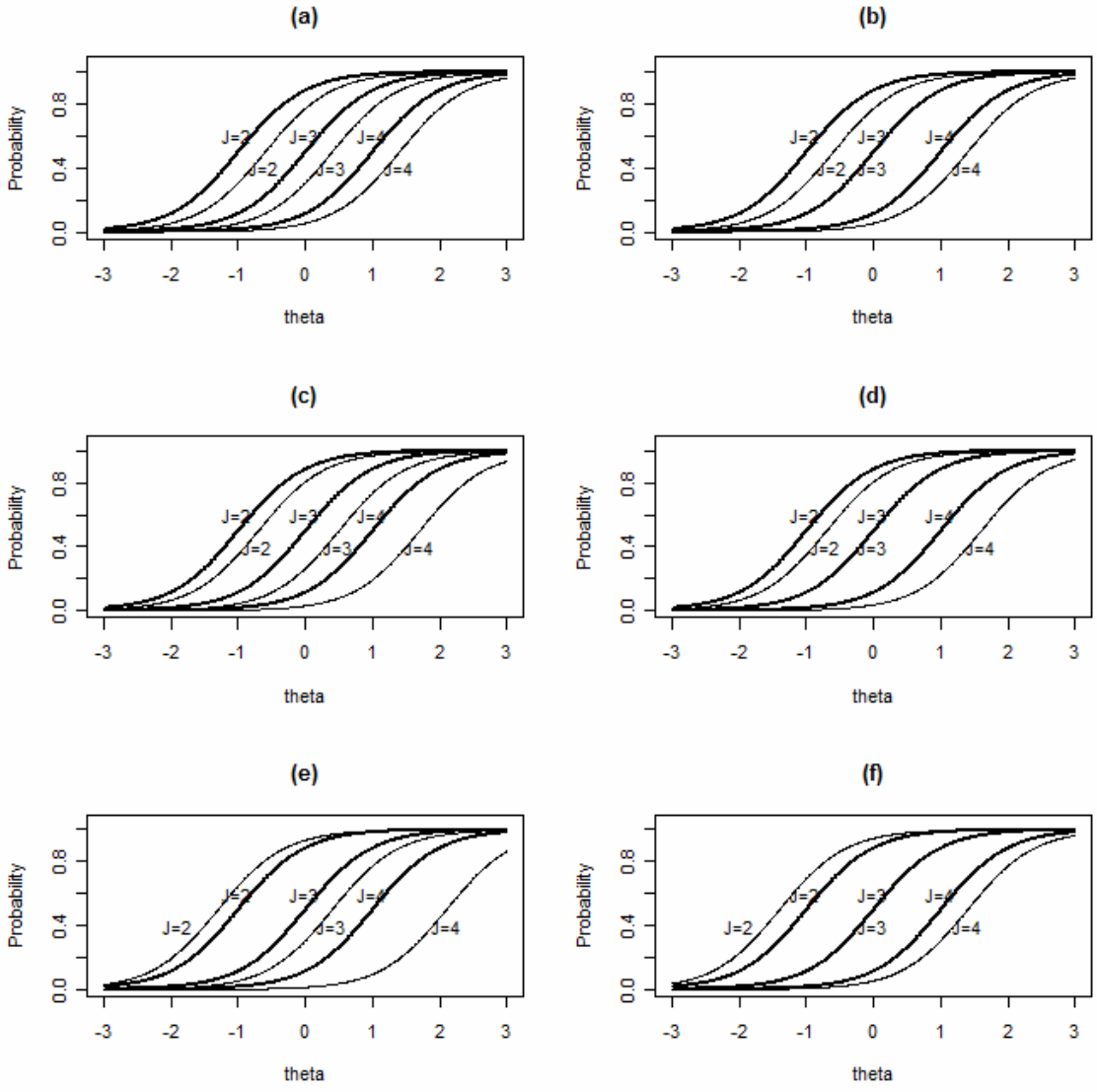

Figure A2. Patterns of Differential Item Functioning for Polytomous Items. The cumulative category response functions for an item where $J = 4$ are shown. The item characteristic curves for the reference group are thick lines; the item characteristic curves for the focal group are thin lines. Patterns include (a) pervasive constant DIF, (b) non-pervasive constant DIF, (c) pervasive convergent DIF, (d) non-pervasive convergent DIF, (e) pervasive divergent DIF, and (f) non-pervasive divergent DIF. Note that the cumulative category response function for $J = 1$ is a horizontal line through probability 1. Also note that the cumulative response function for $J = 3$ is identical for the focal and reference groups in graphs (b), (d), and (f).

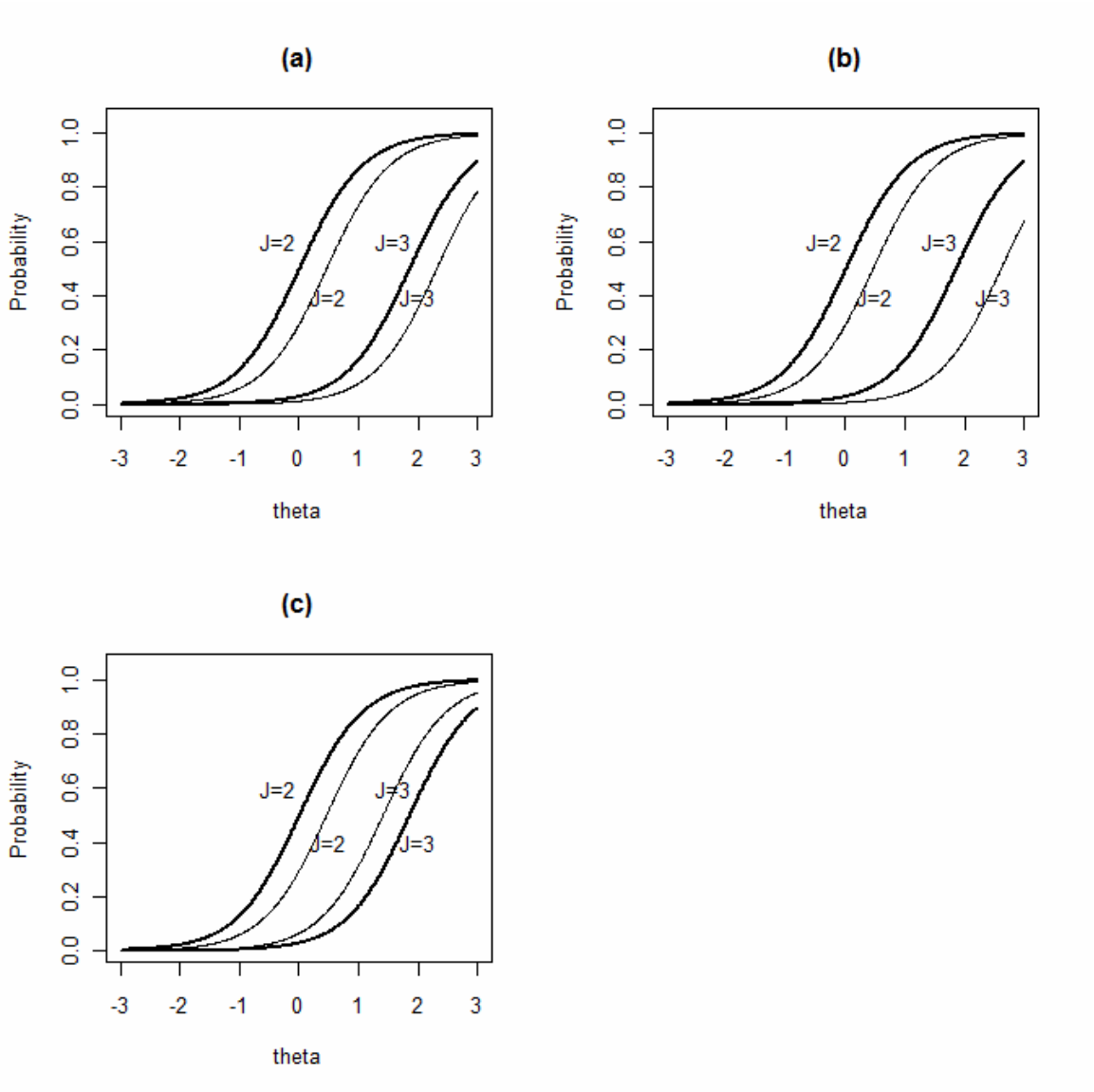

Figure A3. Cumulative Category Response Functions for the Studied Item. The item characteristic curves for the reference group are thick lines; the item characteristic curves for the focal group are thin lines. Patterns include (a) pervasive constant DIF, (b) pervasive convergent DIF, and (c) pervasive divergent DIF. Note that the cumulative category response function for $J = 1$ is a horizontal line through probability 1.

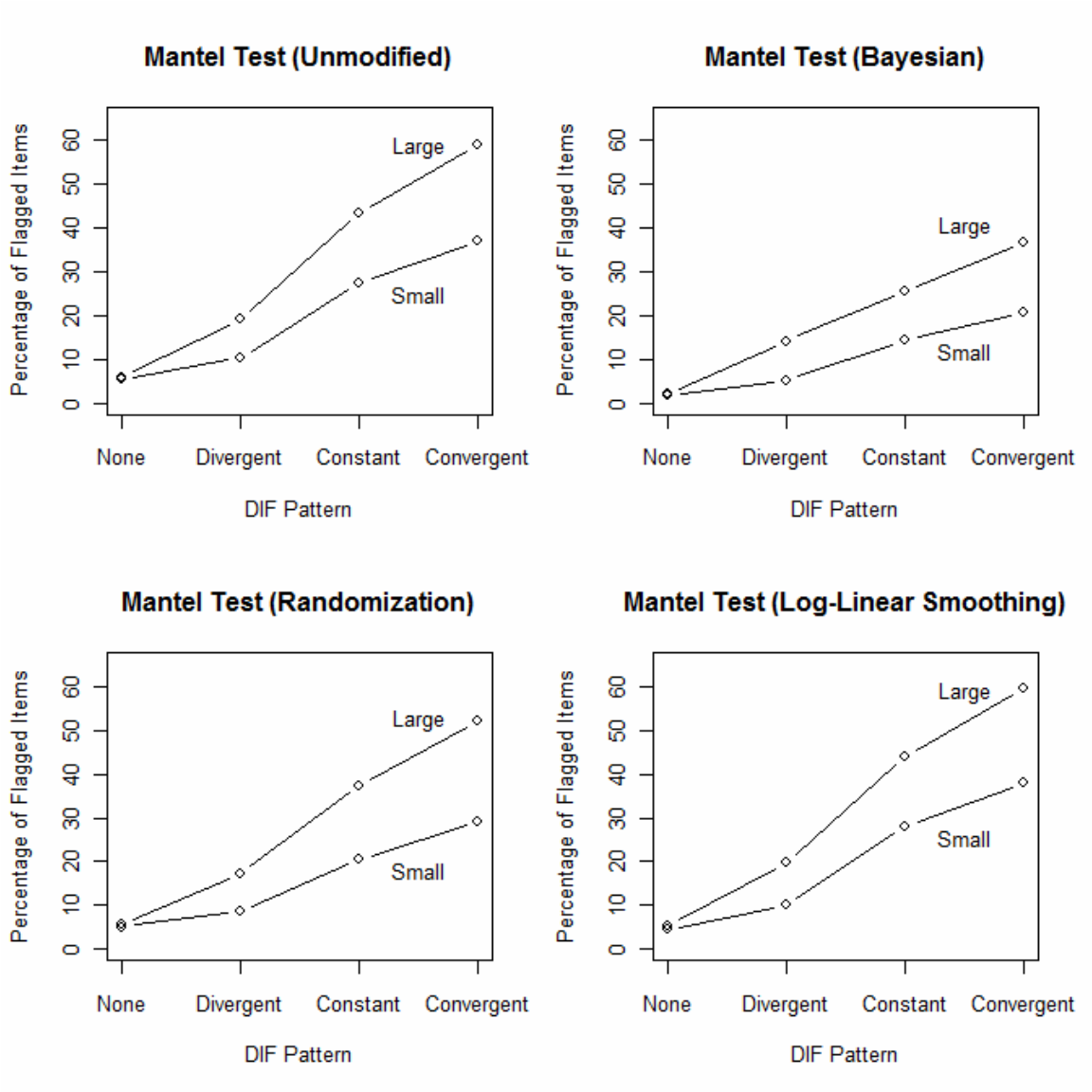

Figure A4. DIF Pattern by Reference Group Sample Size Interaction Plots for the Mantel Tests. Within each graph, the upper line represents the treatments where 400 reference group examinees were used; the lower line represents the treatments where 40 reference group examinees were used.

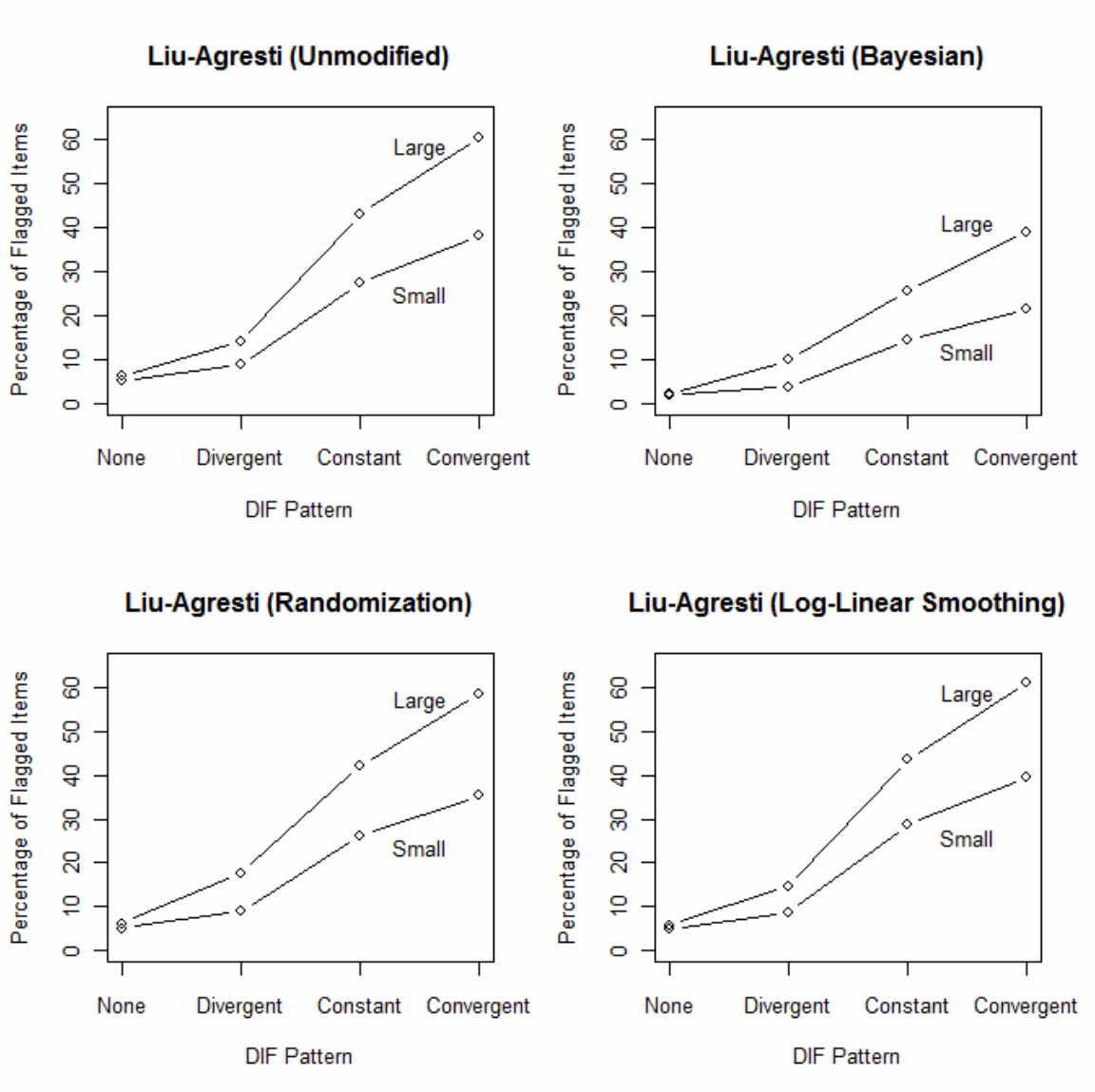

Figure A5. DIF Pattern by Reference Group Sample Size Interaction Plots for the Liu-Agresti Tests. Within each graph, the upper line represents the treatments where 400 reference group examinees were used; the lower line represents the treatments where 40 reference group examinees were used.

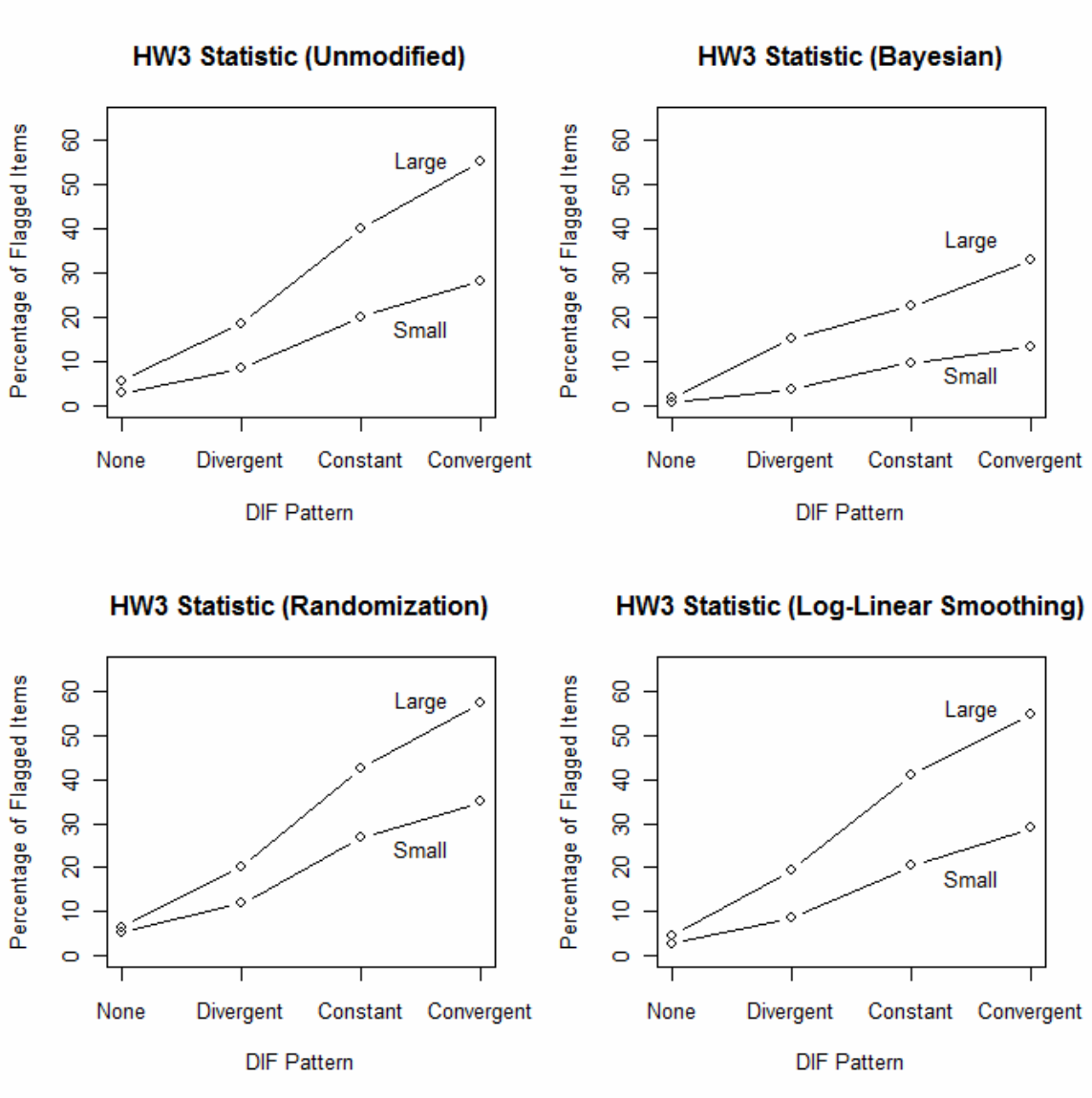

Figure A6. DIF Pattern by Reference Group Sample Size Interaction Plots for the HW3 Tests. Within each graph, the upper line represents the treatments where 400 reference group examinees were used; the lower line represents the treatments where 40 reference group examinees were used.

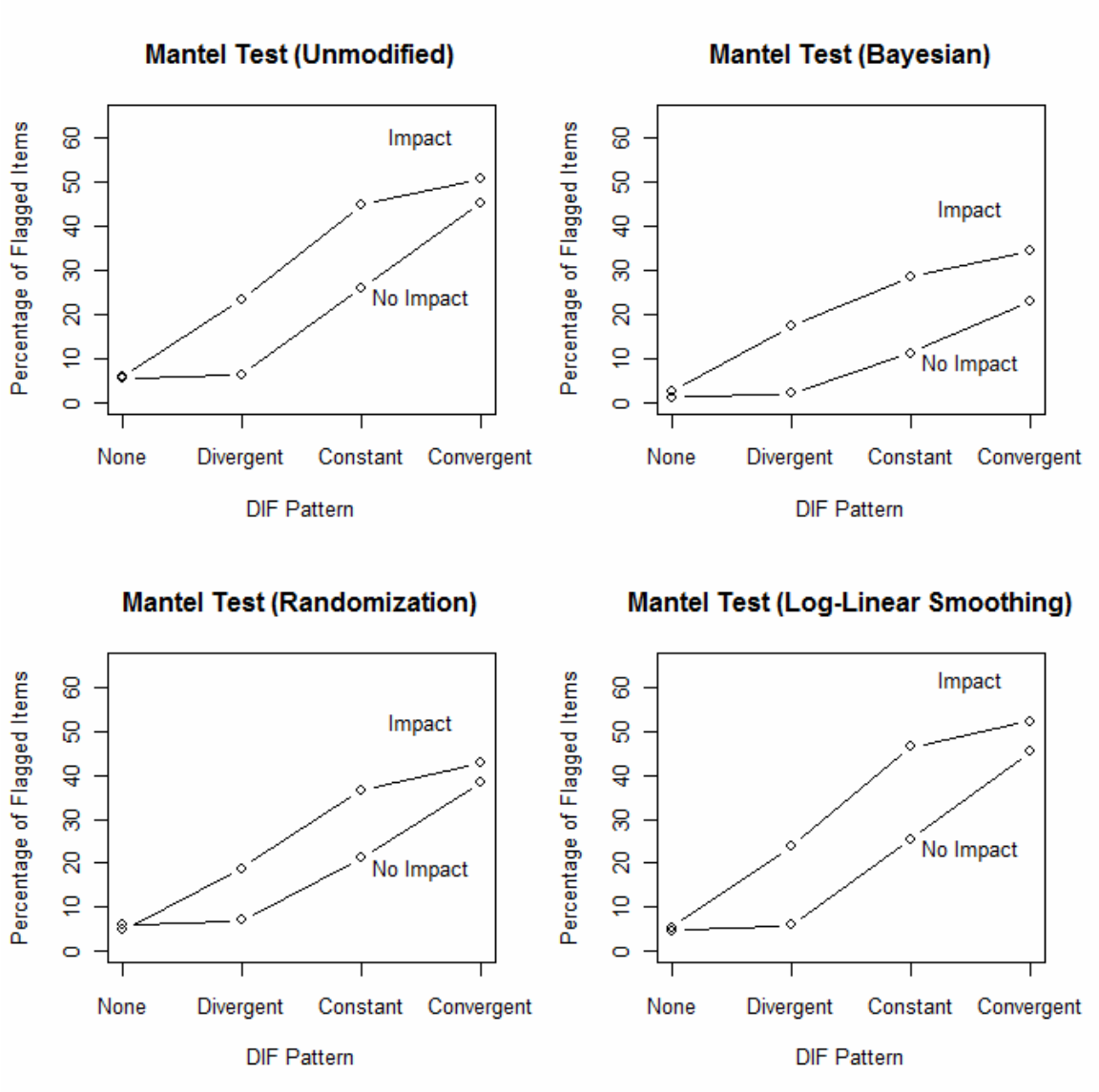

Figure A7. DIF Pattern by Impact Interaction Plots for the Mantel Tests. Within each graph, the upper line represents the treatments where impact was simulated for examinees; the lower line represents the treatments where impact was not simulated for examinees.

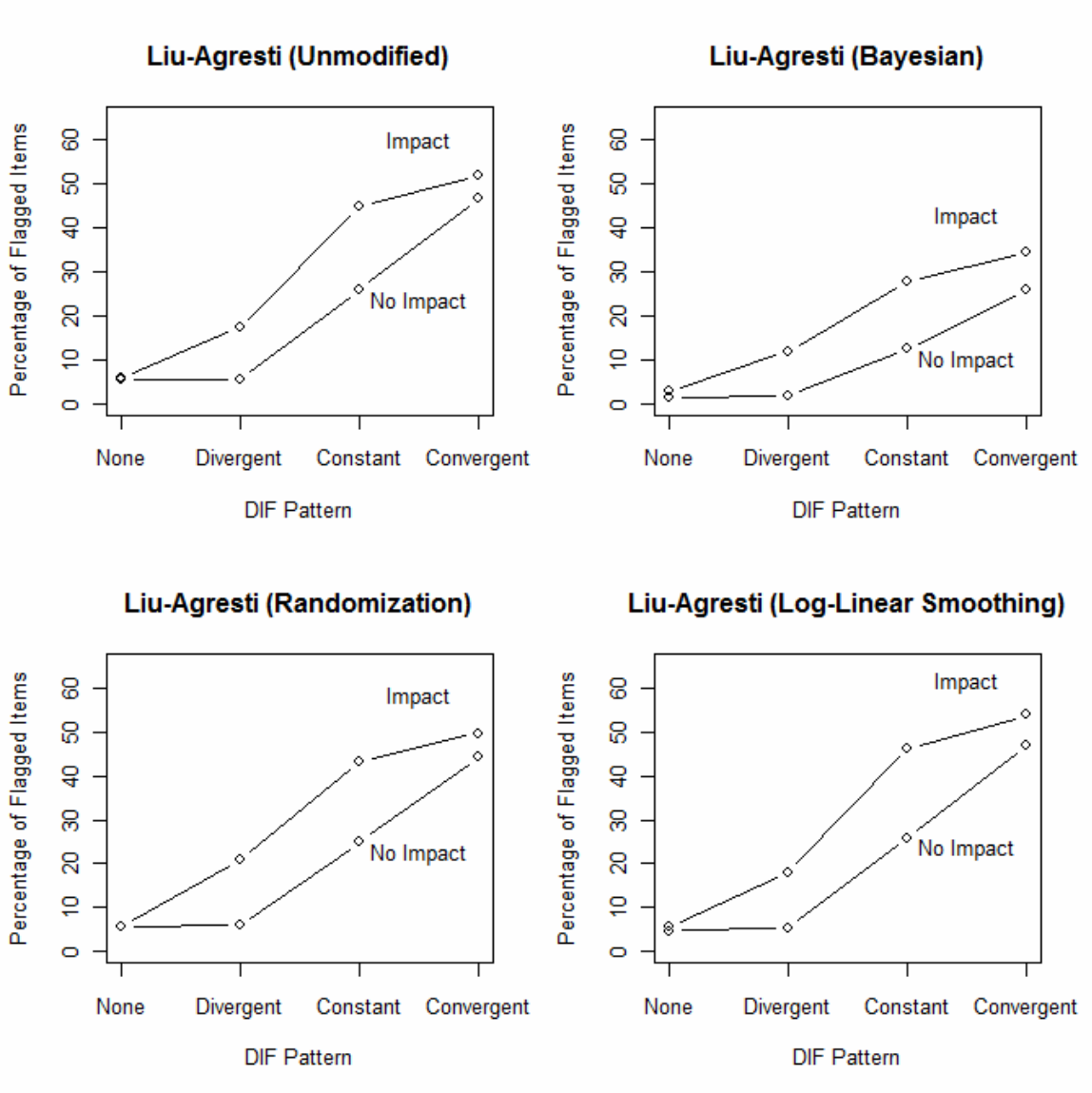

Figure A8. DIF Pattern by Impact Interaction Plots for the Liu-Agresti Tests. Within each graph, the upper line represents the treatments where impact was simulated for examinees; the lower line represents the treatments where impact was not simulated for examinees.

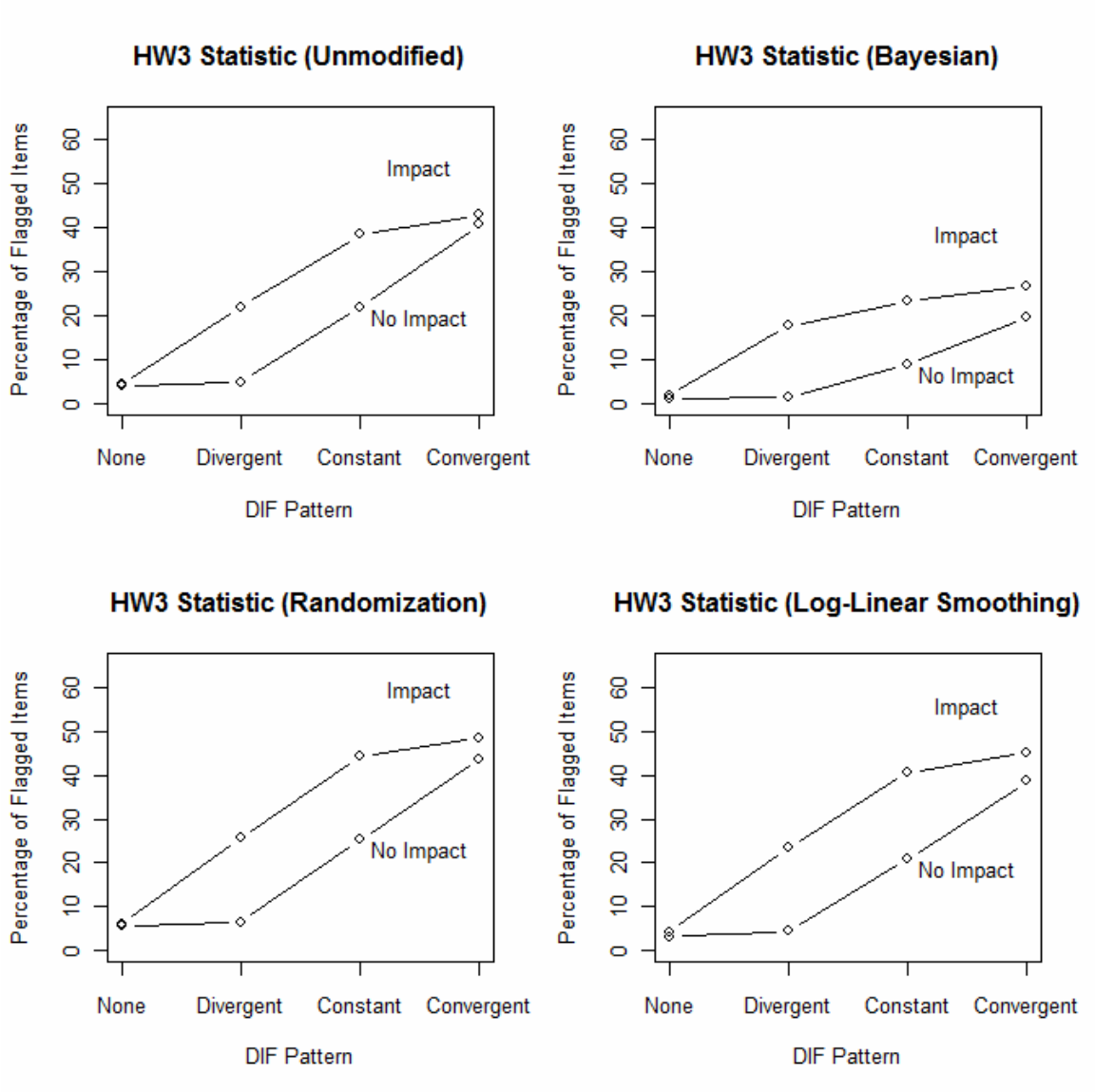

Figure A9. DIF Pattern by Impact Interaction Plots for the HW3 Tests. Within each graph, the upper line represents the treatments where impact was simulated for examinees; the lower line represents the treatments where impact was not simulated for examinees.

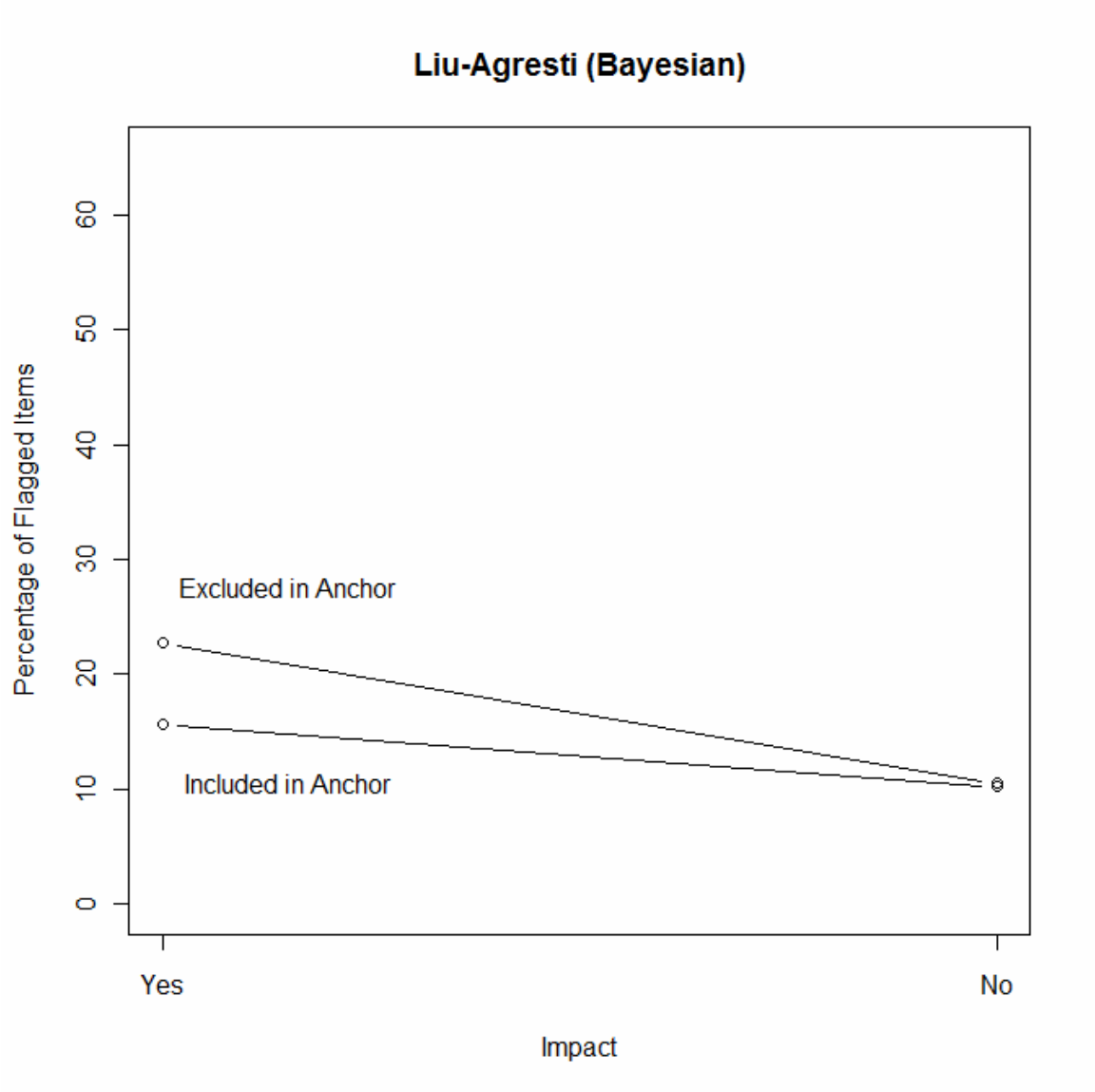

Figure A10. Impact by Inclusion/Exclusion of Studied Item in Anchor Interaction Plot for the Bayesian Liu-Agresti Test. Within each graph, the upper line represents the treatments where the studied item was excluded from the anchor test; the lower line represents the treatments where the studied item was included in the anchor test. This interaction term was significant at the α = .01 level.

APPENDIX B

R FUNCTIONS OF POLYTOMOUS DIF DETECTION METHODS

Table B1. *R* Function for Computing the General Mantel Test for a 2 x 3 x 4 Contingency Table

```
#  Input: test.data   : 2 x 3 x 4 table for the studied item
#       : item.scores : potential scores of the studied item
# Output: X2.value    : chi-square (Mantel) statistic
#       : p.value     : p-value for the Mantel hypothesis test

Mantel <- function(test.data, item.scores) {
   T.vector  <- rep(0, times=4)
   ET.vector <- rep(0, times=4)
   VT.vector <- rep(0, times=4)
   n.plusjk  <- rep(0, times=3)

   for(k in 1:4) {
      T.vector[k] <- sum(item.scores*test.data[1,,k])
      for(j in 1:3) {n.plusjk[j] <- sum(test.data[,j,k])}
      n.Fplusk     <- sum(test.data[1,,k])
      n.Rplusk     <- sum(test.data[2,,k])
      n.plusplusk  <- sum(test.data[,,k])

      ET.vector[k] <- (n.Fplusk/n.plusplusk)*sum(item.scores*n.plusjk)
      part.1 <- (n.Rplusk*n.Fplusk)/(n.plusplusk*n.plusplusk*(n.plusplusk -
1))
      part.2 <- n.plusplusk * sum(item.scores * item.scores * n.plusjk)
      part.3 <- sum(item.scores * n.plusjk)
      VT.vector[k] <- part.1 * (part.2 - part.3*part.3)
   }
   X2.value <- ((sum(T.vector) - sum(ET.vector))**2)/sum(VT.vector)

   return(list(X2.value=X2.value, p.value=pchisq(X2.value, df=1, lower.tail =
FALSE)))
}
```

Table B2. *R* Function for Computing Cox's Beta for a 2 x 3 x 4 Contingency Table

```
#  Input: test.data   : 2 x 3 x 4 table for the studied item
#       : item.scores : potential scores of the studied item
# Output: beta        : Cox's beta
#       : var.beta      : Variance of Cox's beta
#       : p.value     : p-value for Cox's beta hypothesis test

Cox.Beta <- function(test.data, item.scores) {
   T.vector  <- rep(0, times=4)
   ET.vector <- rep(0, times=4)
   VT.vector <- rep(0, times=4)
   n.plusjk  <- rep(0, times=3)

   for(k in 1:4) {
      T.vector[k] <- sum(item.scores*test.data[1,,k])
      for(j in 1:3) {n.plusjk[j] <- sum(test.data[,j,k])}
      n.Fplusk     <- sum(test.data[1,,k])
      n.Rplusk     <- sum(test.data[2,,k])
      n.plusplusk  <- sum(test.data[,,k])

      ET.vector[k] <- (n.Fplusk/n.plusplusk)*sum(item.scores*n.plusjk)
      part.1 <- (n.Rplusk*n.Fplusk)/(n.plusplusk*n.plusplusk*(n.plusplusk -
1))
      part.2 <- n.plusplusk * sum(item.scores * item.scores * n.plusjk)
      part.3 <- sum(item.scores * n.plusjk)
      VT.vector[k] <- part.1 * (part.2 - part.3*part.3)
   }

   beta     <- (sum(T.vector) - sum(ET.vector))/sum(VT.vector)
   X2.value <- (beta**2)/(1/sum(VT.vector))
   return(list(beta=beta, var.beta=(1/sum(VT.vector)),
p.value=pchisq(X2.value, df=1, lower.tail = FALSE)))
}
```

Table B3. *R* Function for Computing the Liu-Agresti Common Odds Ratio for a 2 x 3 x 4 Contingency Table

```
#  Input: test.data   : 2 x I x J table for the studied item
#       : item.scores : potential scores of the studied item
# Output: LA.stat     : Liu-Agresti common odds ratio
#       : var.LA        : variance of Liu-Agresti statistic
#       : p.value     : p-value of Liu-Agresti test (LA=0)

Liu.Agresti <- function(test.data, item.scores) {
   n.Flow <- 0; n.Fhigh <- 0; n.Rlow <- 0
   n.Rhigh <- 0; LA.numerator <- 0; LA.denominator <- 0

   for(k in 1:4) {
   for(j in 1:2) {
      n.Flow  <- sum(test.data[1,1:j,k])
      n.Fhigh <- sum(test.data[1,,k])-n.Flow
      n.Rlow  <- sum(test.data[2,1:j,k])
      n.Rhigh <- sum(test.data[2,,k])-n.Rlow
      LA.numerator   <- LA.numerator + (n.Flow*n.Rhigh/sum(test.data[,,k]))
      LA.denominator <- LA.denominator + (n.Rlow*n.Fhigh/sum(test.data[,,k]))
   }
   }
   LA.stat <- LA.numerator/LA.denominator
   V.LA.denominator <- LA.stat * LA.stat * LA.denominator * LA.denominator
   V.LA.numerator <- 0
   for(k in 1:4) {
   for(j in 1:2) {
      part1 <-
sum(test.data[2,,k])*sum(test.data[1,,k])/(sum(test.data[,,k])**2)
      part2 <- LA.stat * sum(test.data[1,1,k]) * (sum(test.data[2,,k])-
sum(test.data[2,1:2,k]))/sum(test.data[2,,k])
      part3 <- 1 + (LA.stat-1)*(sum(test.data[1,1:2,k]))/sum(test.data[1,,k])
      part4 <- sum(test.data[2,1,k])*(sum(test.data[1,,k])-
sum(test.data[1,1:2,k]))/sum(test.data[1,,k])
      part5 <- LA.stat-(LA.stat-
1)*(sum(test.data[2,1:2,k])/sum(test.data[2,,k]))
      V.LA.numerator <- V.LA.numerator + part1*(part2*part3+part4*part5)
   }
   }
   for(k  in 1:4) {
      part1 <-
sum(test.data[2,,k])*sum(test.data[1,,k])/(sum(test.data[,,k])**2)
      part2 <- LA.stat * sum(test.data[1,1,k]) * (sum(test.data[2,,k])-
sum(test.data[2,1:2,k]))/sum(test.data[2,,k])
      part3 <- 1 + (LA.stat-1)*(sum(test.data[1,1:2,k])/sum(test.data[1,,k]))
      part4 <- sum(test.data[2,1,k])*(sum(test.data[1,,k])-
sum(test.data[1,1:2,k]))/sum(test.data[1,,k])
      part5 <- LA.stat-(LA.stat-
1)*(sum(test.data[2,1:2,k])/sum(test.data[2,,k]))
      V.LA.numerator <- V.LA.numerator + 2*part1*(part2*part3+part4*part5)
   }
   V.LA <- V.LA.numerator / V.LA.denominator
   X2.value <- (log(LA.stat)**2)/V.LA
   Z.value <- log(LA.stat)/sqrt(V.LA)
return(list(LA.stat=LA.stat, var.LA=V.LA, p.value=pchisq(X2.value, df=1,
lower.tail = FALSE)))
}
```

Table B4. *R* Function for Computing the HW3 Statistic and Hypothesis Test for a 2 x 3 x 4 Contingency Table

```
#  Input: test.data   : 2 x I x J table for the studied item
#       : item.scores : potential scores of the studied item
# Output: HW3.statistic : HW3
#       : p.value     : p-value of HW3 test (HW3=0)

HW3 <- function(test.data, item.scores) {
   F.means <- rep(0, times=4)
   R.means <- rep(0, times=4)
   F.vars  <- rep(0, times=4)
   R.vars  <- rep(0, times=4)
   F.sizes <- rep(0, times=4)
   R.sizes <- rep(0, times=4)

   for(k in 1:4) {
      F.means[k] <- sum(item.scores*test.data[1,,k])/sum(test.data[1,,k])
      R.means[k] <- sum(item.scores*test.data[2,,k])/sum(test.data[2,,k])
      F.vars[k]  <- (1/(sum(test.data[1,,k])-
1))*sum(test.data[1,,k]*(item.scores-F.means[k])**2)
      R.vars[k]  <- (1/(sum(test.data[2,,k])-
1))*sum(test.data[2,,k]*(item.scores-R.means[k])**2)
      F.sizes[k] <- sum(test.data[1,,k])
      R.sizes[k] <- sum(test.data[2,,k])
    }

   degrees <- F.sizes + R.sizes - 2
   t.stats <- (F.means-
R.means)/sqrt(((F.vars*F.sizes+R.vars*R.sizes)/degrees)*
((1/F.sizes)+(1/R.sizes)))
   effect.sizes <- sqrt((1/R.sizes)+(1/F.sizes))*t.stats
   correction.f <- 1-(3/(4 * degrees - 1))
   dk.vector    <- correction.f*effect.sizes
   S2k.vector   <- (correction.f**2)*(degrees/(degrees-
2))*((1/R.sizes)+(1/F.sizes))+(dk.vector**2)*(correction.f*correction.f*degree
s/(degrees-2)-1)
   D.statistic <- sum(dk.vector/S2k.vector)/sum(1/S2k.vector)
   Sd.statistic <- 1/sqrt(sum(1/S2k.vector))
   HW3.statistic <- D.statistic/Sd.statistic

   return(list(HW3.statistic=HW3.statistic, p.value=pchisq(HW3.statistic**2,
df=1, lower.tail = FALSE)))

}
```

Table B5. *R* Function for the Standardized Mean Difference Statistic and Hypothesis Test for a 2 x 3 x 4 Contingency Table

```
#  Input: test.data    : 2 x I x J table for the studied item
#       : item.scores  : potential scores of the studied item
# Output: SMD.stat     : SMD statistic
#       : V.SMD        : Variance of SMD (multivar. Hypergeometric assumption)
#       : p.value      : p.value

       SMD <- function(test.data, item.scores) {
          n.Fplusk     <- rep(0, times=4)
          n.Rplusk     <- rep(0, times=4)
          n.Fplusplus  <- rep(0, times=4)
          weight.sum.F <- rep(0, times=4)
          weight.sum.R <- rep(0, times=4)
          for(k in 1:4) {
             n.Fplusk[k]      <- sum(test.data[1,,k])
             n.Rplusk[k]      <- sum(test.data[2,,k])
             n.Fplusplus[k]   <- sum(test.data[1,,])
             weight.sum.F[k]  <- sum(item.scores*test.data[1,,k])
             weight.sum.R[k]  <- sum(item.scores*test.data[2,,k])
          }

          SMD.stat <- sum((n.Fplusk/n.Fplusplus)*(weight.sum.F/n.Fplusk)) -
sum((n.Fplusk/n.Fplusplus)*(weight.sum.R/n.Rplusk))
          VF.vector <- rep(0, times=4)

          for(k in 1:4) {
             part0 <- (n.Rplusk[k] *
n.Fplusk[k])/(sum(test.data[,,k])**2*(sum(test.data[,,k])-1))
             part1 <-
sum(item.scores*item.scores*(test.data[1,,k]+test.data[2,,k]))
             part2 <- sum(item.scores*(test.data[1,,k]+test.data[2,,k]))
             VF.vector[k] <- part0*((sum(test.data[,,k])*part1)-(part2*part2))
          }

          V.SMD <-
sum((n.Fplusk/n.Fplusplus)**2*((1/n.Fplusk)+(1/n.Rplusk))**2*VF.vector)

          return(list(SMD.stat=SMD.stat, V.SMD=V.SMD,
p.value=pchisq(SMD.stat*SMD.stat/V.SMD, df=1, lower.tail = FALSE)))

}
```

Table B6. *R* Function for SIBTEST for a 2 x 3 x 4 Contingency Table

```
#  Input: test.data       : 2 x I x J table for the studied item
#       : item.scores     : potential scores of the studied item
#       : group.vector    : vector of group memberships (focal/reference)
#       : anchor.test     : a data frame containing the item scores on anchor
#       : anchor.total    : vector of total scores on anchor test
#       : anchorcat.vector: vector of anchor categories (k) on anchor test
# Output: B.adjusted      :  SIBTEST statistic using corrected Y-bar values
#       : p.value         : p-value of SIBTEST hypothesis test (B=0)
#   NOTE: Requires the function “cronbach.alpha”.

cronbach.alpha <- function(anchor.test, anchor.total) {
   no.items  <- ncol(anchor.test)
   item.vars <- rep(0, times=no.items)
   for(i in 1:no.items) { item.vars[i] <- var(anchor.test[,i]) }
   return((no.items/(no.items-1))*(1 - sum(item.vars)/var(anchor.total)))
}

SIBTEST <- function(test.data, item.scores, group.vector, anchor.test,
anchor.total, anchorcat.vector) {
   alpha.F<-cronbach.alpha(anchor.test[group.vector=="F",], anchor.total)
   alpha.R<-cronbach.alpha(anchor.test[group.vector=="R",], anchor.total)
   Xbar.F  <- mean(anchor.total[group.vector=="F"])
   Xbar.R  <- mean(anchor.total[group.vector=="R"])

   Xbarvector.F <- rep(0, times=4)
   Xbarvector.R <- rep(0, times=4)
   Tvector.F    <- rep(0, times=4)
   Tvector.R    <- rep(0, times=4)
   Tvector      <- rep(0, times=4)
   Ybarvector.F <- rep(0, times=4)
   Ybarvector.R <- rep(0, times=4)
   Mvector.F    <- rep(0, times=4)
   Mvector.R    <- rep(0, times=4)
   n.Fplusk     <- rep(0, times=4)
   n.Rplusk     <- rep(0, times=4)
   F.means      <- rep(0, times=4)
   R.means      <- rep(0, times=4)
   F.vars       <- rep(0, times=4)
   R.vars       <- rep(0, times=4)

   for(k in 1:4) {
      Xbarvector.F[k] <- mean(anchor.total[group.vector=="F" &
anchorcat.vector==k])
      Xbarvector.R[k] <- mean(anchor.total[group.vector=="R" &
anchorcat.vector==k])
      Tvector.F[k]    <- Xbar.F + alpha.F*(Xbarvector.F[k]-Xbar.F)
      Tvector.R[k]    <- Xbar.R + alpha.R*(Xbarvector.R[k]-Xbar.R)
      Tvector[k]      <- (Tvector.F[k] + Tvector.R[k])/2
      Ybarvector.F[k] <- sum(item.scores*test.data[1,,k])/sum(test.data[1,,k])
```

(Table B6 continued)

```
      Ybarvector.R[k] <- sum(item.scores*test.data[2,,k])/sum(test.data[2,,k])
      n.Fplusk[k]     <- sum(test.data[1,,k])
      n.Rplusk[k]     <- sum(test.data[2,,k])
      F.means[k] <- sum(item.scores*test.data[1,,k])/sum(test.data[1,,k])
      R.means[k] <- sum(item.scores*test.data[2,,k])/sum(test.data[2,,k])
      F.vars[k]  <- (1/(sum(test.data[1,,k])-
1))*sum(test.data[1,,k]*(item.scores-F.means[k])**2)
      R.vars[k]  <- (1/(sum(test.data[2,,k])-
1))*sum(test.data[2,,k]*(item.scores-R.means[k])**2)
   }
   Mvector.F[1] <- (Ybarvector.F[2]-Ybarvector.F[1])/(Tvector.F[2]-
Tvector.F[1])
   Mvector.R[1] <- (Ybarvector.R[2]-Ybarvector.R[1])/(Tvector.R[2]-
Tvector.R[1])
   Mvector.F[2] <- (Ybarvector.F[3]-Ybarvector.F[1])/(Tvector.F[3]-
Tvector.F[1])
   Mvector.R[2] <- (Ybarvector.R[3]-Ybarvector.R[1])/(Tvector.R[3]-
Tvector.R[1])
   Mvector.F[3] <- (Ybarvector.F[4]-Ybarvector.F[2])/(Tvector.F[4]-
Tvector.F[2])
   Mvector.R[3] <- (Ybarvector.R[4]-Ybarvector.R[2])/(Tvector.R[4]-
Tvector.R[2])
   Mvector.F[4] <- (Ybarvector.F[4]-Ybarvector.F[3])/(Tvector.F[4]-
Tvector.F[3])
   Mvector.R[4] <- (Ybarvector.R[4]-Ybarvector.R[3])/(Tvector.R[4]-
Tvector.R[3])

   Pstarvector.F <- Ybarvector.F + Mvector.F * (Tvector - Tvector.F)
   Pstarvector.R <- Ybarvector.R + Mvector.R * (Tvector - Tvector.R)
   B.adjusted    <- (1/sum(test.data[1,,]))*sum(n.Fplusk*(Pstarvector.R-
Pstarvector.F))
   V.B.adjusted  <-
(1/sum(test.data[1,,])**2)*sum(n.Fplusk**2*((R.vars/n.Rplusk)+(F.vars/n.Fplusk
)))

   return(list(B.adjusted=B.adjusted,
p.value=pchisq(B.adjusted**2/V.B.adjusted, df=1, lower.tail = FALSE)))

}
```

Table B7. *R* Function for Bayesian DIF Detection Methods

```
#  Input: examinee.group   : vector of F and R group designations
#       : response.strings : simulated item response strings
#       : include.in.anchor: Y/N, does the anchor score include studied item
# Output: p.value1         : p-value for the Bayesian Cox.Beta Procedure
#       : p.value2         : p-value for Bayesian Liu-Agresti Procedure
#       : p.value3         : p-value for Bayesian HW3 Procedure

Bayesian.Tests <-function(examinee.group, response.strings,
include.in.anchor){
   no.items        <- ncol(response.strings)
   beta.statistics <- rep(0.0, items=no.items)
   LA.statistics   <- rep(0.0, items=no.items)
   HW3.statistics  <- rep(0.0, items=no.items)
   anchor.scores   <- rowSums(response.strings)

   for(y in 1:no.items){
      if(include.in.anchor=="N") anchor.scores <- rowSums(response.strings[-
y])
      anchor.cat    <- create.anchor.cat(anchor.scores)
      current.table <- table(examinee.group, response.strings[,y], anchor.cat)
      if((dim(current.table)[1] != 2) || (dim(current.table)[2] != 3) ||
(dim(current.table)[3] != 4))
      {
         beta.statistics[y] <- NaN
         LA.statistics[y]   <- NaN
         HW3.statistics[y]  <- NaN
      }
      else
      {
      beta.statistics[y] <- Cox.Beta(current.table, c(0,1,2))$beta
      LA.statistics[y]   <- log(Liu.Agresti(current.table, c(0,1,2))$LA.stat)
      HW3.statistics[y]  <- HW3(current.table, c(0,1,2))$HW3.statistic
      }
}
    if(include.in.anchor=="N") anchor.scores <- rowSums(response.strings[,-
16])
      anchor.cat    <- create.anchor.cat(anchor.scores)
      current.table <- table(examinee.group, response.strings[,16],
anchor.cat)
      beta.value <- Cox.Beta(current.table, c(0,1,2))$beta
      LA.value   <- log(Liu.Agresti(current.table, c(0,1,2))$LA.stat)
      HW3.value  <- HW3(current.table, c(0,1,2))$HW3.statistic

   beta.statistics <- beta.statistics[!is.nan(beta.statistics)]
   LA.statistics <- LA.statistics[!is.nan(LA.statistics)]
   HW3.statistics <- HW3.statistics[!is.nan(HW3.statistics)]

   prior.mean  <- mean(beta.statistics)
   prior.var   <- var(beta.statistics)
   data.mean   <- beta.value
   data.var    <- Cox.Beta(current.table, c(0,1,2))$var.beta
   W.value     <- prior.var/(prior.var+data.var)
   posterior.mean <- W.value*data.mean + (1.0-W.value)*prior.mean
   posterior.var  <- W.value * data.var
   Z.stat         <- (0.0-posterior.mean)/sqrt(posterior.var)
   if(!is.nan(posterior.mean) && !is.na(posterior.mean)) {
      if(posterior.mean < 0) p.value1 <- 2*pnorm(0, mean = posterior.mean, sd
= sqrt(posterior.var), lower.tail = FALSE)
```

(Table B7 continued)

```
else p.value1 <- 2*pnorm(0, mean=posterior.mean, sd=sqrt(posterior.var))
   }
   else
      p.value1 <- NaN   prior.mean  <- mean(LA.statistics)
   prior.var   <- var(LA.statistics)
   data.mean   <- LA.value
   data.var    <- Liu.Agresti(current.table, c(0,1,2))$var.LA
   W.value     <- prior.var/(prior.var+data.var)
   posterior.mean <- W.value*data.mean + (1.0-W.value)*prior.mean
   posterior.var  <- W.value * data.var
   Z.stat         <- (0.0-posterior.mean)/sqrt(posterior.var)
   if(!is.nan(posterior.mean) && !is.na(posterior.mean)) {
      if(posterior.mean < 0) p.value2 <- 2*pnorm(0, mean = posterior.mean, sd
= sqrt(posterior.var), lower.tail = FALSE)
      else p.value2 <- 2*pnorm(0, mean=posterior.mean, sd=sqrt(posterior.var))
   }
   else
      p.value2 <- NaN

   prior.mean  <- mean(HW3.statistics)
   prior.var   <- var(HW3.statistics)
   data.mean   <- HW3.value
   data.var    <- 1.0
   W.value     <- prior.var/(prior.var+data.var)
   posterior.mean <- W.value*data.mean + (1.0-W.value)*prior.mean
   posterior.var  <- W.value * data.var
   Z.stat         <- (0.0-posterior.mean)/sqrt(posterior.var)
   if(posterior.mean < 0) p.value3 <- 2*pnorm(0, mean = posterior.mean, sd =
sqrt(posterior.var), lower.tail = FALSE)
   else                   p.value3 <- 2*pnorm(0, mean = posterior.mean, sd =
sqrt(posterior.var))

   prior.mean  <- mean(SMD.statistics)
   prior.var   <- var(SMD.statistics)
   data.mean   <- SMD.value
   data.var    <- SMD(current.table, c(0,1,2))$V.SMD
   W.value     <- prior.var/(prior.var+data.var)
   posterior.mean <- W.value*data.mean + (1.0-W.value)*prior.mean
   posterior.var  <- W.value * data.var
   Z.stat         <- (0.0-posterior.mean)/sqrt(posterior.var)
   if(!is.nan(posterior.mean) && !is.na(posterior.mean)) {
      if(posterior.mean < 0) p.value3 <- 2*pnorm(0, mean = posterior.mean, sd
= sqrt(posterior.var), lower.tail = FALSE)
      else p.value3 <- 2*pnorm(0, mean=posterior.mean, sd=sqrt(posterior.var))
   }
   else
      p.value3 <- NaN

return(list(p.value1=p.value1, p.value2=p.value2, p.value3=p.value3))

}
```

Table B8. *R* Function for Randomization-Based DIF Detection Methods

```
#  Input: contingency.table   : 2 x 3 x 4 contingency table of scores
#  Output: p.value1         : p-value for the Randomized Cox.Beta Procedure
#       : p.value2         : p-value for Randomized Liu-Agresti Procedure
#       : p.value3         : p-value for Randomized HW3 Procedure

Randomized.Tests <- function(contingency.table) {
   replications <- 200
   beta.stats <- rep(0.0, times=replications)
   LA.stats   <- rep(0.0, times=replications)
   HW3.stats  <- rep(0.0, times=replications)
   scores.all <- NULL; k.all <- NULL

   for(k in 1:4){
      row.totals <- rowSums(contingency.table[,,k])
      col.totals <- colSums(contingency.table[,,k])
      scores     <- c(rep(0, times=col.totals[1]), rep(1,
times=col.totals[2]), rep(2, times=col.totals[3]))
      scores.all <- c(scores.all, scores)
      k.vector   <- rep(k, times=sum(contingency.table[,,k]))
      k.all      <- c(k.all, k.vector)
   }

   for(q in 1:replications){
      FR.all <- NULL
         for(k in 1:4) {
            FR.vector  <- rep("R", times=sum(contingency.table[,,k]))
            to.be.F    <- sample(1:sum(contingency.table[,,k]),
rowSums(contingency.table[,,k])[1])
            FR.vector[to.be.F] <- "F"
            FR.all     <- c(FR.all, FR.vector)
         }
      sample.table  <- table(FR.all, scores.all, k.all)
      beta.stats[q] <- Cox.Beta(sample.table, c(0,1,2))$beta
      LA.stats[q]   <- Liu.Agresti(sample.table, c(0,1,2))$LA.stat
      HW3.stats[q]  <- HW3(sample.table, c(0,1,2))$HW3.statistic
   }

   percentile1 <- length(beta.stats[beta.stats <= Cox.Beta(contingency.table,
c(0,1,2))$beta])/replications
   percentile2 <- length(LA.stats[LA.stats <= Liu.Agresti(contingency.table,
c(0,1,2))$LA.stat])/replications
   percentile3 <- length(HW3.stats[HW3.stats <= HW3(contingency.table,
c(0,1,2))$HW3.statistic])/replications
   if(percentile1 > 0.50) p.value1 <- 2*(1-percentile1)
      else p.value1 <- 2*percentile1
   if(percentile2 > 0.50) p.value2 <- 2*(1-percentile2)
      else p.value2 <- 2*percentile2
   if(percentile3 > 0.50) p.value3 <- 2*(1-percentile3)
      else p.value3 <- 2*percentile3

   # If HW3 is not calculatable...
   if(length(sort(HW3.stats)) == 0)  p.value3 <- NaN
   return(list(p.value1=p.value1, p.value2=p.value2, p.value3=p.value3))

}
```

Table B9. *R* Function for Log-Linear Smoothing-Based DIF Detection Methods

```
#  Input: anchor.vector    : vector of F and R group designations
#       : possible.scores  : simulated item response strings
#       : c.value          : degree of log-linear polynomial (default=3)
#       : rho.value        : smoothing parameter (default = 0.8)
#       : stop.crit        : determines when algorithm has converged (.001)
# Output: smoothed.table   : The smoothed contingency table

loglinear.smoothing <- function(anchor.vector, possible.scores, c.value = 3,
rho.value = 0.8, stop.crit = 0.001) {

   t.value  <- length(possible.scores)
   n.vector <- rep(0.0, times = t.value)
   b.matrix <- NULL
   BSaB     <- matrix(0.0, nrow = c.value, ncol = c.value)
   BSlogA   <- matrix(0.0, nrow = c.value, ncol = 1)
   BprimeN  <- NULL
   BprimeM  <- NULL

   for(w in 1:t.value) { n.vector[w] <- length(anchor.vector[anchor.vector ==
possible.scores[w]]) }

if(sum(n.vector) > 3) {
   for(c in 1:c.value) {
    temp.vector <- possible.scores**c
    b.matrix <- cbind(b.matrix,(temp.vector-
mean(temp.vector))/sd(temp.vector))
   }

   a.vector <- rho.value*n.vector + (1-rho.value)*(sum(n.vector)/t.value)
   for(r in 1:c.value) {
      for(s in 1:c.value) {
         part.1 <- sum(b.matrix[,r]*b.matrix[,s]*a.vector)
         part.2 <- sum(b.matrix[,r]*a.vector)*sum(b.matrix[,s]*a.vector)
         BSaB[r,s]   <- part.1-(sum(n.vector)**-1)*part.2
      }
   }
   for(r in 1:c.value) {
      part.1 <- sum(b.matrix[,r]*a.vector*log(a.vector))
      part.2 <- sum(b.matrix[,r]*a.vector)*sum(a.vector * log(a.vector))
      BSlogA[r,1] <- part.1 - (sum(n.vector)**-1)*part.2
   }
   beta.new <- solve(BSaB, BSlogA)
   epsilon <- 1.0

   while((epsilon > stop.crit) && (!is.nan(epsilon))) {
      for(r in 1:c.value) {
         for(s in 1:c.value) {
            part.1 <- sum(b.matrix[,r]*b.matrix[,s]*a.vector)
            part.2 <- sum(b.matrix[,r]*a.vector)*sum(b.matrix[,s]*a.vector)
            BSaB[r,s]   <- part.1-(sum(n.vector)**-1)*part.2
         }
      }
      BprimeN <- t(b.matrix) %*% n.vector
      BprimeM <- t(b.matrix) %*% a.vector
      delta.n1 <- solve(BSaB, BprimeN-BprimeM)
      beta.old <- beta.new
      beta.new <- beta.old + delta.n1
      likelihood.old <- sum(n.vector*log(a.vector))
```

(Table B9 continued)

```
      p.vector <- exp(b.matrix%*%beta.new)/sum(exp(b.matrix%*%beta.new))
      a.vector <- sum(n.vector)*p.vector
      likelihood.new <- sum(n.vector*log(a.vector))
      epsilon <- abs((likelihood.new-likelihood.old)/likelihood.old)

   if(is.nan(epsilon)) {
      print("Yikes!  Epsilon is not a number!")
      a.vector <- n.vector
      epsilon <- 0.0
      }
   }
}
else {
      print("Yikes! Sample size too small!")
      a.vector <- n.vector
}
   return(a.vector)
}

create.smoothed.table <- function(examinee.group, item.score, anchor.score,
max.score) {
   cell.size   <- rep(0.0, times = 6)
   data.to.use <- anchor.score[examinee.group=="F" & item.score==0]
   cell.size[1]<- length(data.to.use)
   smooth1     <- loglinear.smoothing(data.to.use, 0:max.score)
   data.to.use <- anchor.score[examinee.group=="R" & item.score==0]
   cell.size[2]<- length(data.to.use)
   smooth2     <- loglinear.smoothing(data.to.use, 0:max.score)
   data.to.use <- anchor.score[examinee.group=="F" & item.score==1]
   cell.size[3]<- length(data.to.use)
   smooth3     <- loglinear.smoothing(data.to.use, 0:max.score)
   data.to.use <- anchor.score[examinee.group=="R" & item.score==1]
   cell.size[4]<- length(data.to.use)
   smooth4     <- loglinear.smoothing(data.to.use, 0:max.score)
   data.to.use <- anchor.score[examinee.group=="F" & item.score==2]
   cell.size[5]<- length(data.to.use)
   smooth5     <- loglinear.smoothing(data.to.use, 0:max.score)
   data.to.use <- anchor.score[examinee.group=="R" & item.score==2]
   cell.size[6]<- length(data.to.use)
   smooth6     <- loglinear.smoothing(data.to.use, 0:max.score)

   quants      <- c(0, 0, 0, max.score)
   cum.percent <- 0.0

   if(cell.size[1] == 0) cell.size[1] <- 1
   if(cell.size[2] == 0) cell.size[2] <- 1
   if(cell.size[3] == 0) cell.size[3] <- 1
   if(cell.size[4] == 0) cell.size[4] <- 1
   if(cell.size[5] == 0) cell.size[5] <- 1
   if(cell.size[6] == 0) cell.size[6] <- 1

   for(q in 1:(max.score+1)) {
      percent.to.add <-
((smooth1[q]/cell.size[1])+(smooth2[q]/cell.size[2])+(smooth3[q]/cell.size[3])
+(smooth4[q]/cell.size[4])+(smooth5[q]/cell.size[5])+(smooth6[q]/cell.size[6])
)/6
```

(Table B9 continued)

```
      if(cum.percent < 0.25 && (percent.to.add + cum.percent) >= 0.25)
quants[1] <- (q-1)
      if(cum.percent < 0.50 && (percent.to.add + cum.percent) >= 0.50)
quants[2] <- (q-1)
      if(cum.percent < 0.75 && (percent.to.add + cum.percent) >= 0.75)
quants[3] <- (q-1)
      cum.percent <- cum.percent + percent.to.add
   }
   temp.value <- NULL
   temp.value <- c(temp.value, sum(smooth1[1:(quants[1]+1)]))
   temp.value <- c(temp.value, sum(smooth2[1:(quants[1]+1)]))
   temp.value <- c(temp.value, sum(smooth3[1:(quants[1]+1)]))
   temp.value <- c(temp.value, sum(smooth4[1:(quants[1]+1)]))
   temp.value <- c(temp.value, sum(smooth5[1:(quants[1]+1)]))
   temp.value <- c(temp.value, sum(smooth6[1:(quants[1]+1)]))
   temp.value <- c(temp.value, sum(smooth1[(quants[1]+2):(quants[2]+1)]))
   temp.value <- c(temp.value, sum(smooth2[(quants[1]+2):(quants[2]+1)]))
   temp.value <- c(temp.value, sum(smooth3[(quants[1]+2):(quants[2]+1)]))
   temp.value <- c(temp.value, sum(smooth4[(quants[1]+2):(quants[2]+1)]))
   temp.value <- c(temp.value, sum(smooth5[(quants[1]+2):(quants[2]+1)]))
   temp.value <- c(temp.value, sum(smooth6[(quants[1]+2):(quants[2]+1)]))
   temp.value <- c(temp.value, sum(smooth1[(quants[2]+2):(quants[3]+1)]))
   temp.value <- c(temp.value, sum(smooth2[(quants[2]+2):(quants[3]+1)]))
   temp.value <- c(temp.value, sum(smooth3[(quants[2]+2):(quants[3]+1)]))
   temp.value <- c(temp.value, sum(smooth4[(quants[2]+2):(quants[3]+1)]))
   temp.value <- c(temp.value, sum(smooth5[(quants[2]+2):(quants[3]+1)]))
   temp.value <- c(temp.value, sum(smooth6[(quants[2]+2):(quants[3]+1)]))
   temp.value <- c(temp.value, sum(smooth1[(quants[3]+2):(quants[4]+1)]))
   temp.value <- c(temp.value, sum(smooth2[(quants[3]+2):(quants[4]+1)]))
   temp.value <- c(temp.value, sum(smooth3[(quants[3]+2):(quants[4]+1)]))
   temp.value <- c(temp.value, sum(smooth4[(quants[3]+2):(quants[4]+1)]))
   temp.value <- c(temp.value, sum(smooth5[(quants[3]+2):(quants[4]+1)]))
   temp.value <- c(temp.value, sum(smooth6[(quants[3]+2):(quants[4]+1)]))
   smoothed.table <- array(temp.value, dim=c(2,3,4))

   if((quants[2]-quants[1] == 0) || (quants[3]-quants[2] == 0) || (quants[4]-
quants[3] == 0)) {
         smoothed.table <- array(temp.value, dim=c(4,3,2))
   }
   return(smoothed.table)
}
```

Table B10. Java Function for Simulating Theta Values and Item Response Strings

```
import java.io.*;
import java.util.Random;

class DataGeneration
{
   public static void main(String[] args) throws IOException
   {
      int Fsize = 40;   // 40
      int Rsize = 40;   // 40 or 400
      int testLength = 16;
             int response = 0;
      double aTheta = 0.0;
      Random generator = new Random();
      // From Lewis and Loftus Example
      FileWriter fw = new FileWriter("C:/temp/responses.txt");
      BufferedWriter bw = new BufferedWriter(fw);
      PrintWriter outFile = new PrintWriter(bw);

      double[][] testParams = { {1.46,  0.16, 1.46},
                                {1.81, -0.17, 1.60},
                                {1.57,  0.06, 1.69},
                                {1.89,  0.34, 1.76},
                                {1.93,  0.15, 1.86},
                                {1.79, -0.07, 1.77},
                                {2.35,  0.53, 1.85},
                                {2.12,  0.38, 1.70},
                                {2.19,  0.19, 1.69},
                                {1.79,  0.13, 1.62},
                                {1.75,  0.18, 1.66},
                                {1.86,  0.31, 1.65},
                                {2.18,  0.16, 1.75},
                                {2.14,  0.38, 1.56},
                                {2.12,  0.39, 1.61},
                                {1.89,  0.01, 1.86} };

      for(int z = 1; z <= 1000; z++)
      {
             testParams[testLength - 1][0] = 1.89;
             testParams[testLength - 1][1] = 0.46;
             testParams[testLength - 1][2] = 2.61;

             for(int i = 1; i <= Fsize; i++)
             {
                aTheta = generator.nextGaussian(); // Subtract 0 or 0.5
                for(int j = 0; j < testLength; j++)
                {
                   response=produceResponse(aTheta, testParams[j]);
                   outFile.print(response);
                   if(j != (testLength - 1)) outFile.print(",");
                     }
                outFile.println();
             }
             testParams[testLength - 1][0] = 1.89;
             testParams[testLength - 1][1] = 0.01;
             testParams[testLength - 1][2] = 1.86;
```

(Table B10 continued)

```
            for(int i = 1; i <= Rsize; i++)
            {
                  aTheta = generator.nextGaussian();
                  for(int j = 0; j < testLength; j++)
                  {
                       response=produceResponse(aTheta, testParams[j]);
                       outFile.print(response);
                       if(j != (testLength - 1)) outFile.print(",");
                  }
                  outFile.println();
            }
      }
      outFile.close();
   }

      // Method to compute the item response based on Samejima's model

   public static int produceResponse(double theta, double[] params)
   {
      int theResponse = 9;
              int noCategories = params.length;
              double[] probVector = new double[noCategories];
              double[] cumProbVector = new double[noCategories];

        cumProbVector[0] = 1.0;
        cumProbVector[1] = Math.exp(params[0]*(theta - params[1]))/(1 +
Math.exp(params[0]*(theta - params[1])));
        cumProbVector[2] = Math.exp(params[0]*(theta - params[2]))/(1 +
Math.exp(params[0]*(theta - params[2])));

        probVector[0] = cumProbVector[0] - cumProbVector[1];
        probVector[1] = cumProbVector[1] - cumProbVector[2];
        probVector[2] = cumProbVector[2];

        double uVariate = Math.random();

        if(uVariate <= probVector[0])
             theResponse = 0;
    else
     {
       if((uVariate > probVector[0]) && (uVariate <= (probVector[0] +
probVector[1])))
      theResponse = 1;
       else
       {
      if((uVariate > (probVector[0] + probVector[1])) && (uVariate <= 1.0))
        theResponse = 2;
       }
      }

   return theResponse;
   }
}
```

Table B11. *R* Program for Running Simulation Defined in Chapter 3

```
source("C:\\Users\\Woodrow\\Documents\\Simulations\\CoreX.txt")
source("C:\\Users\\Woodrow\\Documents\\Simulations\\BayesianX.txt")
source("C:\\Users\\Woodrow\\Documents\\Simulations\\RandomXX.txt")
source("C:\\Users\\Woodrow\\Documents\\Simulations\\SmoothingX.txt")
my.alpha <- 0.05
F.size <- 40; R.size <- 40     # (40,40) or (40,400)
examinee.group <- c(rep("F", times = F.size), rep("R", times = R.size))
item.no <- 16
include.in.anchor <- "Y"       # "Y" or "N"

simulation.results <- rep(0.0, times = 24)
bugs.results       <- rep(0.0, times = 24)
HW3.results        <- rep(NA, times = 1000)

for(no.rep in 1:1000) {
   response.strings  <- read.table(file = "C:\\temp\\responses.txt", sep=",",
nrows=80, skip=80*(no.rep-1))

   if(include.in.anchor == "Y")    {anchor.scores <-
rowSums(response.strings)}
   else                {anchor.scores <- rowSums(response.strings[,-
item.no])}

   anchor.cat        <- create.anchor.cat(anchor.scores)
   contingency.table <- table(examinee.group, response.strings[,item.no],
anchor.cat)

if((dim(contingency.table)[1] == 2) && (dim(contingency.table)[2] == 3) &&
(dim(contingency.table)[3] == 4) && (no.rep != 873)) {

   smoothed.table  <- create.smoothed.table(examinee.group,
response.strings[,item.no], anchor.scores, 2*item.no)

   # Mantel ##################
   the.test <- Mantel(contingency.table, c(0,1,2))

   if(!is.nan(the.test$p.value))
      { if(the.test$p.value < my.alpha) simulation.results[1] <-
simulation.results[1] + 1 }
   else
      {bugs.results[1] <- bugs.results[1] + 1 }

   # LA   ####################
   the.test <- Liu.Agresti(contingency.table, c(0,1,2))

   if(!is.nan(the.test$p.value))
      { if(the.test$p.value < my.alpha) simulation.results[2] <-
simulation.results[2] + 1 }
   else
      {bugs.results[2] <- bugs.results[2] + 1 }

   # HW3  ####################
   the.test <- HW3(contingency.table, c(0,1,2))
   HW3.results[no.rep] <- the.test$HW3.statistic
   if(!is.nan(the.test$p.value))
      { if(the.test$p.value < my.alpha) simulation.results[3] <-
simulation.results[3] + 1 }
   else
      {bugs.results[3] <- bugs.results[3] + 1 }
```

(Table B11 continued)

```
   ######################################################
   # Bayesian Tests #####################################
   the.test <- Bayesian.Tests(examinee.group, response.strings,
include.in.anchor)

   # Bayesian Cox's Beta ################
   if(!is.nan(the.test$p.value1))
    {  if(the.test$p.value1 < my.alpha) simulation.results[4] <-
simulation.results[4] + 1 }
   else
      {   bugs.results[4] <- bugs.results[4] + 1 }

   # Bayesian Liu-Agresti ###############
   if(!is.nan(the.test$p.value2))
    {  if(the.test$p.value2 < my.alpha) simulation.results[5] <-
simulation.results[5] + 1 }
   else
      {   bugs.results[5] <- bugs.results[5] + 1 }

   # Bayesian HW3        ################
   HW3.results[no.rep] <- the.test$Z.stat

   if(!is.nan(the.test$p.value3))
      {   if(the.test$p.value3 < my.alpha) simulation.results[6] <-
simulation.results[6] + 1   }
   else
      {   bugs.results[6] <- bugs.results[6] + 1 }

   ######################################################
   # Randomized Tests ###################################
   the.test <- Randomized.Tests(contingency.table)

   # Random Cox's Beta ################
   if(the.test$p.value1 < my.alpha) simulation.results[7] <-
simulation.results[7] + 1

   # Random Liu-Agresti ###############
   if(the.test$p.value2 < my.alpha) simulation.results[8] <-
simulation.results[8] + 1

   # Random HW3        ################
   if(!is.nan(the.test$p.value3))
      {   if(the.test$p.value3 < my.alpha) simulation.results[9] <-
simulation.results[9] + 1   }
   else
      {   bugs.results[9] <- bugs.results[9] + 1 }

   ####################################
   # Mantel ###################
  if((dim(smoothed.table)[1] == 2) && (dim(smoothed.table)[2] == 3) &&
(dim(smoothed.table)[3] == 4)) {
   the.test <- Mantel(smoothed.table, c(0,1,2))

   if(!is.nan(the.test$p.value) && !is.na(the.test$p.value))
      {if(the.test$p.value < my.alpha) simulation.results[10] <-
simulation.results[10] + 1 }
   else
      {bugs.results[10] <- bugs.results[10] + 1 }
```

(Table B11 continued)

```
   # LA   ####################
   the.test <- Liu.Agresti(smoothed.table, c(0,1,2))

   if(!is.nan(the.test$p.value))
      {if(the.test$p.value < my.alpha) simulation.results[11] <-
simulation.results[11] + 1 }
   else
      {bugs.results[11] <- bugs.results[11] + 1 }

   # HW3  ####################
   the.test <- HW3(smoothed.table, c(0,1,2))

   if(!is.nan(the.test$p.value))
      {if(the.test$p.value < my.alpha) simulation.results[12] <-
simulation.results[12] + 1 }
   else
      {bugs.results[12] <- bugs.results[12] + 1 }
  }
  else {
     bugs.results[10] <- bugs.results[10] + 1
     bugs.results[11] <- bugs.results[11] + 1
     bugs.results[12] <- bugs.results[12] + 1
     print("######################PROBLEM WITH DIMENSIONS OF SMOOTHED TABLE")
     }

}
else {
      bugs.results <- bugs.results + 1
      print("BAD DIMENSIONS ON CONTINGENCY TABLE")
      }

}
print(no.rep)
print(simulation.results)
print(bugs.results)
```