2-12-2008

The Impact of Multidimensionality on the Detection of Differential Bundle Functioning Using SIBTEST.

Terris Raiford-Ross

ACCEPTANCE

This dissertation, THE IMPACT OF MULTIDIMENSIONALITY ON THE DETECTION OF DIFFERENTIAL BUNDLE FUNCTIONING USING SIBTEST, by TERRIS RAIFORD ROSS, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree Doctor of Philosophy in the College of Education, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chair, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty. The Dean of the College of Education concurs.

_____
Carolyn F. Furlow, Ph.D.
Committee Chair

_____
Phillip Gagne, Ph.D.
Committee Member

_____
T. Chris Oshima, Ph.D.
Committee Member

_____
Valerie A. Miller, Ph.D.
Committee Member

_____
Date

_____
Sheryl A. Gowen, Ph.D.
Chair, Department of Educational Policy Studies

_____
Ronald P. Colarusso, Ed.D.
Dean, College of Education

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education's director of graduate studies and research, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Terris Raiford Ross

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Terris Raiford Ross
209 Calibre Lake Parkway
Smyrna, GA 30082

The director of this dissertation is:

Dr. Carolyn F. Furlow
Department of Educational Policy Studies
College of Education
Georgia State University
Atlanta, GA 30303

VITA

Terris Raiford Ross

---

209 Calibre Lake Parkway
Smyrna, Georgia 30082
Day (678) 895-3486
Evening (678) 842-9592
Email: tross@doe.k12.ga.us

Education:

Ph.D.	2007	Georgia State University Research, Measurement and Statistics
Ed.S.	2001	Florida Atlantic University Educational Leadership
M.S.	1998	Clark Atlanta University Mathematics
B.S.	1998	Clark Atlanta University Mathematics

Professional Experience:

2005 – 2006	Instructor, Department of Educational Policy Studies Georgia State University, Atlanta, GA
2002 –2005	Visiting Instructor, Department of Mathematics and Statistics Georgia State University, Atlanta, GA
2001 –2002	Mathematics Teacher, Frederick Douglass High School, Atlanta, GA
1999 – 2001	Mathematics Teacher Suncoast Community High School, Riviera Beach, FL
2000 – 2001	Adjunct Instructor Palm Beach Community College, Palm Beach Gardens, FL

1999-2000 Curriculum Writer, Department of Equity in Education
The School District of Palm Beach County, West Palm Beach, FL

1998 -1999 Mathematics Teacher
Campbell High School, Smyrna, GA

1998 – 1999 Adjunct Instructor, Department of Developmental Studies
Georgia Perimeter College, Clarkston, GA

Summer 2003 Graduate Research Intern
Educational Testing Service, Princeton, New Jersey

Conference Presentations:

Raiford-Ross, T. & Furlow, C.F. (2005). Impact of Missing Data on Polytomous DIF Detection Methods. Presentation at 2005 Annual Meeting of the American Psychological Association, Washington, D.C.

Raiford-Ross, T., Furlow, C.F., & Gagné, P. (2007). The impact of multidimensionality on the detection of differential bundle functioning using SIBTEST. Presentation at the 2007 Annual Meeting of the American Educational Research Association. Chicago, IL.

Professional Organizations:

2003 – Present American Educational Research Association (AERA)
2003 – Present National Council on Measurement in Education (NCME)

ABSTRACT

THE IMPACT OF MULTIDIMENSIONALITY ON THE DETECTION OF DIFFERENTIAL BUNDLE FUNCTIONING USING SIBTEST
by
Terris Raiford Ross

In response to public concern over fairness in testing, conducting a differential item functioning (DIF) analysis is now standard practice for many large-scale testing programs (e.g., Scholastic Aptitude Test, intelligence tests, licensing exams). As highlighted by the Standards for Educational and Psychological Testing manual, the legal and ethical need to avoid bias when measuring examinee abilities is essential to fair testing practices (AERA-APA-NCME, 1999). Likewise, the development of statistical and substantive methods of investigating DIF is crucial to the goal of designing fair and valid educational and psychological tests.

Douglas, Roussos and Stout (1996) introduced the concept of item *bundle* DIF and the implications of differential bundle functioning (DBF) for identifying the underlying causes of DIF. Since then, several studies have demonstrated DIF/DBF analyses within the framework of "unintended" multidimensionality (Oshima & Miller, 1992; Russell, 2005). Russell (2005), in particular, examined the effect of secondary traits on DBF/DTF detection. Like Russell, this study created item bundles by including multidimensional items on a simulated test designed in theory to be unidimensional. Simulating reference group members to have a higher mean ability than the focal group

on the nuisance secondary dimension, resulted in DIF for each of the multidimensional items, that when examined together produced differential bundle functioning.

The purpose of this Monte Carlo simulation study was to assess the Type I error and power performance of SIBTEST (Simultaneous Item Bias Test; Shealy & Stout, 1993a) for DBF analysis under various conditions with simulated data. The variables of interest included sample size and ratios of reference to focal group sample sizes, correlation between primary and secondary dimensions, magnitude of DIF/DBF, and angular item direction. Results showed SIBTEST to be quite powerful in detecting DBF and controlling Type I error for almost all of the simulated conditions. Specifically, power rates were .80 or above for 84% of all conditions and the average Type I error rate was approximately .05. Furthermore, the combined effect of the studied variables on SIBTEST power and Type I error rates provided much needed information to guide further use of SIBTEST for identifying potential sources of differential item/bundle functioning.

THE IMPACT OF MULTIDIMENSIONALITY ON THE DETECTION OF DIFFERENTIAL BUNDLE FUNCTIONING USING SIBTEST

by
Terris Raiford Ross

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Research, Measurement & Statistics
in
the Department of Educational Policy Studies
in
the College of Education
Georgia State University

Atlanta, GA
2007

ACKNOWLEDGEMENTS

I am most grateful to the Lord Jesus Christ for his faithfulness throughout my life and during this humbling experience. I would like to thank my loving husband for all the encouragement and support he has provided from the moment we met. This journey would also not have been the same without the wisdom and mentoring of my twin sister, Dr. Jerris L. Raiford. Thank you for everything. I am also indebted to my beloved mother whose resilience in life has always been an inspiration to me. Many hugs and kisses are sent to my pastor and church family at Spread the Word Church Ministries who have been my longtime cheerleaders during this and every endeavor I have been blessed to share with them. I also offer many thanks and much appreciation to everyone who played a part in moving me along (academically) to the finish line. From my very capable and supportive dissertation advisory committee and department chair to the excellent technology support team at Assessment Systems Corporation (especially the ever helpful Eun Young Lim who searched high and low to provide me with the DOS version of SIBTEST). To Dr. Furlow and Dr. Gagne, thank you for not only being mentors and going out of your way to guide this research, but I appreciate your friendship as well. Finally, I am very grateful to all of my colleagues in the School Improvement Division at the Georgia Department of Education for their touching words and acts of kindness that helped make the last stretch even more enjoyable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CDIF	Compensatory Differential Item Functioning
DBF	Differential Bundle Functioning
DFIT	Differential Functioning of Items and Tests
DIF	Differential Item Functioning
DTF	Differential Test Functioning
SIBTEST	Simultaneous Item Bias Test
ICC	Item Characteristic Curve
IPR	Item Parameter Replication
IRT	Item Response Theory
MIRT	Multidimensional Item Response Theory
MMD	Multidimensional Model for DIF
NCDIF	Non-Compensatory Differential Item Functioning
PL	Parameter Logistic
TCC	Test Characteristic Curve

CHAPTER 1

INTRODUCTION

With the increased use of standardized test scores for making selection and promotion decisions, the issue of fairness has remained a crucial consideration in the fields of educational and psychological testing. Public scrutiny of test items and testing procedures that are potentially biased against various examinee subgroups (e.g., racial, ethnic, gender) has greatly influenced the development of standards of fairness in educational testing. However, deciding what is fair in testing is a complex process involving a variety of conditions that must be met to avoid biased items and tests.

The 1999 *Standards for Educational and Psychological Testing* manual delineates four definitions of fairness that should guide the testing process. First, tests and test items should not yield scores that lead to different score interpretations for different subgroups of examinees. That is, tests should be free of systematic bias that favors one group over another or results in scores that are less valid for one or more identifiable subgroups. In the past, focus has been placed on bias against minority subgroups defined by race, ethnicity, and gender. Today investigations of item/test bias are conducted on the behalf of examinees that differ from the majority based on other important characteristics such as disability or native language. For example, to avoid a certain form of language bias, a test designed to measure verbal analogical reasoning should use words in general use, not

those associated with specific vocations, ethnic groups, disciplines, or geographic locations.

The second definition of fairness provided in the *Standards* is that fair testing should involve equal treatment of examinees during the testing process. This includes activities that occur before, during and after the actual test administration. For example, *before* the test all examinees should be given equal opportunity to become familiar with the test format, study materials, etc. Additionally, all examinees should be provided with appropriate testing conditions. That is, for examinees to demonstrate their level of knowledge, skill, or ability, standard procedures should be used *during* the test to ensure fairness. However, appropriate conditions also include necessary accommodations for test takers, such as students with disabilities or English language learners, who would be disadvantaged by certain standardized procedures. Reasonable accommodations might include allowing for extended time or offering test forms written in multiple languages. Finally, fairness in the testing process also applies to how individual and group test results are reported *after* the test. For instance, when group achievement differences are reported efforts should be made to avoid misinterpretation of these results. That is, factors influencing these differences, such as unequal educational opportunity, should be included so that decisions based on test scores can be made wisely.

Third, it is assumed that if examinees are given a chance to demonstrate proficiency on the trait or construct being measured by the test, that the outcomes of testing will be equitable. However, this does not imply that group-level differences in test scores are an indication of bias since it is possible to design valid tests on which group ability distributions differ. Simply put, if testing is fair, then examinees of equal

ability (or level of the construct being measured by the test) should earn the same scores, regardless of group membership.

Finally, the *Standards* define fairness as providing every examinee with an equal opportunity to learn the test content. This is a concern particularly for achievement tests which are usually designed to measure knowledge, skills, and abilities that are based on formal instruction. If, for example, a state graduation test includes content that examinees from a particular school or school district have not had the opportunity to learn, then withholding a high school diploma based on those students' test scores may not be considered fair. On the other hand, credentialing exams such as high school graduation tests are meant to signify a specific level of proficiency. Thus, difficulties arise with this definition of fairness and depending on how high the stakes are (e.g., tests used for making graduation or promotion decisions are considered "high stakes") additional evidence of competence may need to be considered.

As the breadth of these definitions illustrate, fairness is a complex construct that serves many purposes, the ultimate of which is the achievement of equal opportunity in our society. Although inequalities at the individual and/or group level are inevitable, methods that minimize bias (e.g., standardized test administrations, sensitivity reviews, and statistical analysis of item characteristics) are available. Fortunately for test users, the development of psychometric methods for identifying biased items/tests has contributed greatly to improved test design and equitable score reporting. This study will focus on some of the methods used in the investigation of item/test bias.

Differential Item Functioning

The empirical evidence gathered in the investigation of bias is generally referred to as differential item functioning (DIF). In response to public concern over the bias that exists in some measures of aptitude and/or cognitive ability (e.g., Scholastic Aptitude Tests, intelligence tests, exams for licensure and promotion) conducting DIF analyses is now standard practice for most large-scale testing companies. As highlighted by the *Standards*, the legal and ethical need to avoid bias when measuring examinee ability is essential to fair testing practices (*Standards*, 1999). Likewise, the development of statistical and substantive methods of investigating DIF is crucial to the goal of designing fair and valid educational and psychological tests.

Until recently most investigations of test fairness involved examination of statistically flagged DIF items in an effort to identify item characteristics that lead to differential item functioning for two subgroups matched on ability. In the past, these groups have typically been represented by a majority group known as the reference group (e.g. whites, males) and a focal group comprised of minorities (e.g., blacks, females). After matching members of both groups on ability, items are submitted for DIF analysis. However the emphasis on statistical DIF analysis has not been successful in determining what causes items to exhibit DIF. Researchers argue that analyzing individual items one at a time may not be the most effective way to investigate sources of DIF since most tests are designed to assess skills or abilities using small groups of items (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001). Gierl et al. (2001) assert that "sources of DIF may be more apparent in patterns across multiple items rather than in performance characteristics associated with single items" (p. 27). As a result, more studies are being

conducted to examine the influence of groups of DIF items whose characteristics (e.g., content area or skill category) may point to sources of differential functioning.

Douglas, Roussos and Stout (1996) introduced the concept of item *bundle* DIF and the implications of differential bundle functioning (DBF) for identifying the underlying causes of DIF. A bundle is any dimensionally homogenous set of items that is not necessarily adjacent or related to a common text or passage (Douglas, Roussos & Stout, 1996). In DBF analyses, similar items are grouped together based on organizing principles (e.g., content, item type, etc.) believed to affect the performance of different groups of examinees. The basis for DBF analysis is the assertion that tests consist primarily of small bundles of items designed to measure a certain trait, skill or ability. Also, research suggests that methods for statistically identifying bundles or groups of items are more powerful than those which analyze items one at a time (Nandakumar, 1993). Hence, DBF analysis is preferable to DIF analysis when bundling permits small differences in group performance on individual items to be amplified. Furthermore, the DBF approach of examining potential sources of DIF by identifying suspect item bundles has great implications for improving test design and educational measurement.

Purpose of Study

The purpose of this study was to gather evidence in support of differential bundle functioning analysis for identifying potential causes of DIF in educational achievement tests. The phenomenon known as DIF amplification that may occur when item bundles, rather than individual items, are examined for bias against specific subgroups was investigated with the SIBTEST procedure. To accomplish this, the current study assessed

the performance of SIBTEST for DBF analysis under various conditions with simulated data.

Significance of Study

The importance of accurately identifying items with DIF cannot be overstated. In an effort to develop valid educational and psychological tests, it is essential that selected DIF/DBF methods are chosen for their ability to control Type I error while still maintaining adequate power to detect potentially biased items. In addition, more research is needed involving the substantive analysis of bias using differential bundle functioning analysis. In summary, the goal of this study is to improve the practice of educational measurement by identifying the strengths and weaknesses of a commonly used method of DIF/DBF detection as well as assist test developers in discovering potential sources of differential item functioning.

CHAPTER 2

REVIEW OF THE LITERATURE

The two primary frameworks for examining differential item functioning (DIF) are classical test theory (CTT) and item response theory (IRT). While CTT has a longer history in test development, classical indices such as item difficulty and discrimination suffer from their dependence on particular test items and examinees. That is, in classical test theory the proportion of examinees who answer an item correctly (item difficulty) depends on the particular group of test-takers. If the same test were given to another group of examinees, these item statistics could be very different. Another weakness of CTT is the requirement of parallel tests for the assessment of reliability. Since parallel tests are nearly impossible to obtain in practice, estimates of reliability are often imprecise and inconsistent in the CTT framework.

To overcome some of the shortcomings of classical test theory, item response theory models were developed and are currently being used by many large-scale testing agencies. IRT models the relationship between examinee performance on an item and the latent trait (e.g., ability) being measured by the test. Specifically, IRT postulates that examinee performance can be predicted or explained by a set of latent traits or abilities (θ) measured by the test. The mathematical model of this relationship is called an Item Characteristic Curve (ICC) and describes a monotonically increasing function that

specifies an increase in the probability of a correct response as the level of the latent trait increases.

IRT Assumptions

IRT models make two assumptions about tests and test items. First, it is assumed that a test is unidimensional (i.e., only one trait is being measured by items on a test). Second, after controlling for ability, the responses of examinees to any pair of items are statistically independent. That is, a response to one item should not be related to that of another item on the test. For example, the information provided in the stem of one test item should not contain information necessary to correctly answer another item on the same test.

IRT Models

Many models can be developed within the IRT framework. Each model accounts for item-level and examinee-level parameters. Three common IRT models are the 1-, 2-, and 3-Parameter Logistic (PL) Models. The 1-PL, commonly referred to as the Rasch Model assumes that the only item characteristic affecting examinee performance is item difficulty (*b*-parameter). However, the 2-PL model also accounts for item discrimination (i.e., the *a*-parameter). Finally, the 3-PL model includes item difficulty, discrimination and pseudo-guessing (i.e., the *c*-parameter).

The most salient feature of IRT that distinguishes it from CTT is the property of invariance of item and ability parameters (Hambleton, Swaminathan & Rogers, 1991). Invariance means that item parameters (e.g., difficulty, discrimination and guessing) are not dependent on the ability distribution of any particular group of examinees and the examinee ability parameter (θ) is not dependent on a specific set of test items. This also

implies that for a correctly specified IRT model, the ICC for two subpopulations of examinees will be the same regardless of the groups' ability distributions (Hambleton, Swaminathan & Rogers, 1991). This property makes IRT an attractive framework for examining DIF since the occurrence of non-coinciding ICCs is an indicator of differential item functioning between two groups.

In summary, IRT is useful in overcoming the weaknesses of classical test theory for the following reasons. IRT is not item or sample-dependent. That is, the assumption of local independence and the property of invariance allow item parameters to be estimated for any group of examinees and permits estimations of examinee ability that are not test-dependent. Also, IRT provides a mathematical model for predicting item success, conditional on ability level. Finally, test reliability in IRT is based only on the items chosen for a test rather than on a specific group of test takers and the availability of parallel forms such as are required for classical test theory.

Shealy-Stout's Multidimensional Model for DIF

The IRT framework is not only a useful tool for investigating differential item functioning; it has also led to theories about what may cause DIF to occur. Over the last decade or so, researchers have argued that the leading cause of DIF is the inclusion of multidimensional test items. That is, many tests thought to be unidimensional – an important assumption in IRT – are in fact measuring additional latent traits other than the primary trait of interest (e.g., Oshima & Miller, 1992; Shealy & Stout, 1993a; Roussos & Stout, 1996; Russell, 2005). These *secondary* dimensions represent either intentionally or unintentionally measured traits. According to Shealy and Stout's (1993a)

Multidimensional Model for DIF (MMD), secondary dimensions require further examination. If a secondary dimension is unintentionally assessed as part of the construct of interest (e.g., verbal ability necessary for an item designed to measure math ability) then it is termed *nuisance,* but if the test is designed to measure a primary trait (e.g., math ability) *and* a secondary trait (e.g., logical reasoning) then the secondary dimension may be considered *auxiliary*. Under the MMD differential item functioning that results from auxiliary dimensions is *benign* (indicating impact), whereas nuisance traits produce *adverse* DIF (indication of bias). It is important to note the difference between impact and bias. Impact is a reflection of true ability differences between groups on a relevant or intentional construct. On the other hand, bias is a term reserved for items that measure an irrelevant secondary trait for which examinees from one group are systematically advantaged or disadvantaged. Furthermore, it is important to distinguish between intentionally multidimensional tests and tests that unintentionally measure multiple dimensions since distributional differences on nuisance abilities are potential sources of DIF. On the other hand, different ability distributions on an auxiliary trait (i.e., an intentionally measured secondary dimension) would not be considered as an indication of DIF between examinee subpopulations (Oshima et al., 1997).

Violations of Unidimensionality Assumption

As stated earlier, Item Response Theory assumes that tests are unidimensional. However, many experts agree that most achievement and aptitude tests are actually multidimensional with perhaps a dominant primary dimension and several secondary dimensions (e.g., Lord, 1980; Kok, 1988; Shealy & Stout, 1993a; Camilli & Shepard,

1994; Ackerman, 1992; Oshima et al., 1997; Embretson & Reise, 2000). Hence, the use of unidimensional IRT models with multidimensional test data violates the unidimensionality assumption and poses a potentially serious threat to item and examinee parameter estimation.

Many studies (e.g., Reckase, 1979; Ansley & Forsyth, 1985; Reckase, Ackerman & Carlson, 1988; Kirisci, Hsu & Yu, 2001) have assessed the effects of this violation on measurement equivalence and the results have been used both to support the continued use of unidimensional IRT and to encourage development of Multidimensional IRT (MIRT) models as well. For example, in an early study Reckase (1979) found that good θ estimates can be obtained for multidimensional data that have a dominant dimension accounting for as little as 10% of variance and several minor dimensions. He noted however, that to obtain stable estimates, the dominant dimension should explain at least 20% of variance in item responses. In a study examining the effects of two-dimensional data on unidimensional parameter estimates, Ansley & Forsyth (1985) showed that as the correlation between two dimensions increased, the effects of violating the assumption of unidimensionality decreased using the modified three parameter logistic (3 PL) model with fixed *c*-parameter (i.e., *c* was fixed to 0.2). Later, Reckase, Ackerman and Carlson (1988) used multidimensional IRT to show that a test with items that are designed to measure the same composite of abilities – rather than a single ability – is sufficient to satisfy the unidimensionality assumption.

More recently, Kirisci, Hsu and Yu (2001) investigated (among other factors) the robustness of three item parameter estimation programs (BILOG, MULTILOG, and XCALIBRE) to unidimensionality assumption violations. Results for BILOG supported

findings from previous studies (e.g., Doody, 1985; Drasgow & Parsons, 1983; Harrison, 1986) that showed for correlated dimensions, the effect of multidimensionality on parameter estimates was small. That is, of the three estimation programs, BILOG was most robust to multidimensionality. Furthermore, Kirisci et al. (2001) presented several guidelines for practitioners wanting to use unidimensional IRT models to estimate multidimensional test parameters. Specifically, Kirisci et al. make three suggestions. First, it is necessary to determine the multidimensional structure of the data since unidimensional IRT models may be permissible if there is only one dominant dimension with several minor dimensions. However, as Reckase (1979) suggested, the dominant factor may need to account for at least 20% of variance in order to obtain stable item parameter estimates. Second, when there are several dimensions of approximately equal dominance, the magnitude of the correlations between dimensions should be assessed. Citing earlier studies (e.g., Ackerman, 1989; Drasgow &Parsons, 1983; Harrison, 1986), Kirisci and colleagues (2001) suggest proceeding with unidimensional IRT models if the multiple dimensions are highly correlated ($r > .4$) and those correlations are approximately equal across dimensional pairs. Finally, in the event that dimensions are not highly correlated ($r \leq .4$) and/or the correlations among them vary to a large degree, then multidimensional IRT should be applied.

Although multidimensionality can seriously affect the performance of unidimensionally-based procedures such as BILOG and LOGIST, these studies provide evidence that correlated dimensions can mediate that influence, making such procedures adequate for item and ability estimation. Additionally, results from those investigations have corroborated Stout's theory and test of *essential unidimensionality* which challenges

the traditional unidimensionality assumption of IRT. Stout (1990) argued that the presence of exactly one dominant dimension (i.e., *essential unidimensionality* or d_E =1)) is a sufficient and more psychologically appropriate requirement, considering the nature of achievement testing. Therefore, by validating that a test measures one primary dimension and one or more minor secondary dimensions, support is gathered for the use of standard IRT calibration programs such as BILOG or LOGIST. However, if secondary dimensions influence item responses to a large degree, Stout's test of *essential unidimensionality* should conclude that $d_E > 1$ and practitioners would be advised against the use of standard IRT data analysis programs (Nandakumar, 1991). A simulation study by Nandakumar (1991) assessed the performance of Stout's hypothesis test of *essential dimensionality* in terms of Type I error and verified its reliability and power to detect essential unidimensionality in test data where traditional dimensionality may exceed one (i.e. $d_E > 1$). As a result, unidimensionally-based IRT procedures continue to be used in applications with real data that meet the newer standard of *essential unidimensionality*. A discussion on traditional DIF analysis follows.

Traditional DIF Analysis

DIF occurs when focal and reference group members of equal ability on the latent trait have different probabilities of answering an item correctly. DIF procedures generally begin by identifying a reference group (usually comprised of examinees in the majority; e.g., males, whites, etc.) and focal group (typically a minority subgroup; e.g., Hispanics, females, etc.) of interest. After test administration and scoring, members of the reference and focal groups are matched on some measure of ability related to the test (usually total

test score). From there, statistical procedures are used to identify which items yield group differences [see Millsap and Everson (1993) for an extensive review of statistical methods used to detect DIF].

In the event an item is functioning differently for one of the groups, a decision must be made about whether to retain the item or delete it from the test. If the item proves to be biased against a subgroup, its magnitude is strong enough to bias test results, and a rationale exists for why it may exhibit DIF, then the item should be deleted (Camilli & Shepard, 1994). However, without a substantive review of the item to understand the reason it resulted in DIF, test developers do not actually know if the source of DIF is due to a construct-relevant or irrelevant dimension being measured by the test. Therefore, it is important to accurately interpret the nature of DIF so that differences between the groups' cognitive skills or opportunities to learn can be appropriately addressed.

Issues with Traditional DIF Analysis

Typically DIF analyses are conducted in two steps: statistical identification (including effect size measures of practical significance) of items that favor particular groups followed by a substantive review of potentially biased items to locate the sources of differential item functioning. Although many advances have been made in the statistical analysis of DIF items, much remains to be learned about how to accurately pinpoint why DIF occurs. During the substantive analysis of DIF, items are usually reviewed by subject-area experts (e.g., curriculum specialists or item writers) in an attempt to interpret the factors contributing to differential performance between specific

subgroups of examinees. Even though this is an important step in the process of eliminating bias and ensuring test fairness, studies indicate that this method of substantive analysis has met with limited success (see Bond, 1993; Camilli & Shepard, 1994; Engelhard, Hansche, & Rutledge, 1990; Gierl, Rogers & Klinger, 1999; O'Neill & McPeek, 1993; *Standards for Educational and Psychological Testing*, 1999; Sudweeks & Tolman, 1993). For example, Bond (1993) noted that "theories about why items behave differentially across groups can be described only as primitive" (p. 279). Roussos and Stout (1996a) also highlight the failure of substantive analyses of statistically identified DIF items to interpret the sources of DIF. Remarking on the task of predicting DIF items without empirical evidence, Engelhard, Hansche, and Rutledge (1990) state that, "the agreement between the judgmental and empirical indices of DIF are very low and usually not better than what would be expected by chance" (p. 358). In many cases, the judgments made by item reviewers tend to either disagree with DIF statistics or with one another (Engelhard, Hansche, & Rutledge, 1990). These inconsistencies may be attributable to the many possible hypotheses about why an item displays DIF *after* it has been identified statistically. As a result, definitive conclusions about the sources of DIF are rarely drawn. Some researchers believe that this is an inherent problem with single-item DIF analyses and argue that more can be learned from studying groups of items simultaneously rather than one at a time (Douglas, Roussos & Stout, 1996; Boughton, Gierl & Khaliq, 2000). Although the current primary method of DIF analyses (statistical evaluation followed by substantive review of individual items) does not typically provide more conclusive answers regarding DIF sources, a related procedure known as differential bundle functioning (DBF) holds promise for addressing this problem.

Differential Bundle Functioning

Because substantive DIF analyses following the statistical identification of DIF items have yielded little information in understanding the sources of DIF, methods have been developed that use the results of substantive analyses (item review) to statistically investigate items believed to function differentially. Instead of seeking to interpret DIF statistics for substantive meaning, the process (and therefore logic) is reversed by first forming substantive hypotheses regarding potential DIF items and then testing those items statistically. In particular, Douglas, Roussos and Stout (1996) introduced the concept of item *bundle* DIF and the implications of differential bundle functioning (DBF) for identifying the underlying causes of DIF. A bundle is any dimensionally homogenous set of items that is not necessarily adjacent or related to a common text or passage (Douglas, Roussos & Stout, 1996). In DBF analyses, similar items are grouped together based on organizing principles (e.g., content, item type, etc.) believed to affect the performance of different groups of examinees. The basis for DBF analysis is the assertion that tests consist primarily of small bundles of items designed to measure a certain trait, skill or ability. Similarly, Gierl, Bisanz, Bisanz, Boughton, and Khaliq (2001) assert that "sources of DIF may be more apparent in patterns across multiple items rather than in performance characteristics associated with single items" (p. 27). Also, research has shown that methods for statistically identifying bundles or groups of items are more powerful than those which analyze items one at a time (Nandakumar, 1993). Hence, DBF analysis is preferable to DIF analysis when bundling permits small differences in group performance on individual items to be amplified. Furthermore, the

DBF approach of examining potential sources of DIF by identifying suspect item bundles has great implications for improving test design and educational measurement. The next section discusses different organizing principles used to investigate differential bundle functioning.

Bundle Formation

Although any number of organizing principles can be used to identify items suspected of measuring multiple abilities, prior studies have used four methods in particular. While different, each is focused on the identification of items that appear to measure a common secondary trait in addition to the target ability. The methods can be classified as exploratory versus confirmatory depending on the manner in which they are employed and/or the rationale for using a specific bundling strategy.

First, a test's dimensionality structure can be assessed using test specifications that outline both the content area and cognitive skill categories that the test is designed to measure. A list of test specifications serves as the blueprint to guide item writers when sampling items from the achievement domain. These items are also designed to measure specific cognitive skills and processes. Therefore, a detailed analysis of test specifications may highlight subsets of items that measure a number of different dimensions associated with certain content and skill areas (Gierl, et al., 2005).

For example, Oshima, Raju, Flowers and Slinde (1998) demonstrated DBF analysis using the cognitive dimensions measured by the reading comprehension portion of the Metropolitan Achievement Test as well as by bundling the items associated with reading passages. Although cognitive classifications did not appear to elicit differential

functioning, Oshima et al. did find large DBF in favor of boys for a reading passage titled "The Roadrunner: A Strange Bird". They were able to interpret the potential cause of the differential functioning by reasoning that boys were possibly more familiar with the context of the passage which described characteristics of the Roadrunner such as his diet and speed.

Second, the dimensionality structure of a test can be uncovered through the use of subject area experts who use their experience to identify specific dimensions through a thorough analysis of test content. A content analysis may be conducted during either an item review session with content specialists or a review of the literature for judgments regarding the content of well-known tests such as the SAT (e.g., O'Neill & McPeek, 1993; Douglas et al., 1996; Gierl & Bolt, 2003). For example, Douglas, Roussos and Stout (1996) used a panel of experts to select item bundles from a test deemed to be essentially unidimensional. They argued that this method is especially appropriate in cases where a test is so dominated by a primary dimension that many statistical dimensionality assessment tools would be unable to detect minor secondary dimensions. On the other hand, O'Neill and McPeek (1993) demonstrated how previously flagged DIF items could be used in a post hoc analysis to investigate the interaction of item content and item type on a measure of reading ability. That is, they used previously reported empirical evidence that sentence completion items associated with science content tended to produce DIF in favor of males to investigate DIF based on content area and item type.

Third, cognitive psychology can be used to identify dimensions on which certain groups are hypothesized to differ in ability. Although examples in the literature are

scarce, research is underway that uses cognitive theory to predict group differences on specific item types. For example, Gallagher et al. (2000) used high school and college students to investigate gender differences in advanced mathematical problem solving on SAT-M and GRE-Q, respectively, by dividing math problems into two main types (conventional and unconventional) based on the cognitive processes associated with answering the items correctly. Using "think aloud" problem solving with a group of high ability high school students, Gallagher et al. (2000) found that males tended to be more flexible in their use of problem-solving strategies, while females tended to employ conventional problem-solving algorithms to solve advanced mathematical problems.

In addition, Gallagher et al. (2000) developed a revised taxonomy for classifying mathematics problems by item context, verbal and spatial demands, content mastery, and strategy selection, based on the findings of Halpern (1992; 1997) that certain cognitive processes are associated with tasks that favor males versus those that favor females. Using this taxonomy, Gallagher et al. successfully predicted gender performance on the GRE-Q among college students with different majors (i.e., arts and humanities, social sciences, and technical sciences). Although the investigation would be classified as an impact study rather than an examination of item bias, the results have important implications for improving test development, classroom instruction, and measurement specialists' understanding of factors that influence differential group performance on standardized achievement tests. Fortunately, plans are underway to use the cognitive categories in the Gallagher et al. taxonomy to statistically test for differential bundle functioning (Gierl et al., 2001; Boughton, Gierl & Khaliq, 2004). The results of that

research are expected to inform and advance the current process of assessing test dimensionality using psychological analysis.

Finally, secondary dimensions can be identified using statistical dimensionality assessment tools. Using a mixed exploratory and confirmatory approach, Douglas, Roussos, and Stout (1996) combined hierarchical agglomerative cluster analysis (HCA), a cluster analysis procedure, and DIMTEST, a nonparametric statistical dimensionality test based on Stout's (1987) concept of essential unidimensionality, to identify suspect item bundles. First, an exploratory dimensionality analysis was performed so that DIF hypotheses could be developed and then confirmed with a cross-validation sample. Results revealed a six-item bundle measuring a secondary dimension interpreted as "knowledge of some important documents in early American history" (Douglas et al., 1996, p.477). When tested for DIF, the bundle was found to favor females.

A recent investigation by Gierl, Tan and Wang (2005) using empirical dimensionality assessment tools was conducted to identify content and cognitive dimensions on the SAT. Although DBF analysis was not a focus of the study, the finding that distinct, yet correlated, dimensions exist on the math and critical reading subtests of the SAT could be used to investigate differential bundle functioning among specific subgroups of examinees. That is, using the dimensions identified in the study, DIF hypotheses could be developed and tested using currently available DIF methods. The authors are currently exploring this line of research in hopes that it will enhance the diagnostic value of the SAT for identifying examinees' cognitive strengths and weaknesses.

In summary, differential bundle functioning can be examined by confirmatory methods such as using test specifications or content analyses or through the exploratory approach of allowing statistical procedures to identify the number of dimensions being measured by a test. It should be noted that all of the methods described above include assumptions and limitations that make them more or less desirable in certain contexts. For example, cognitive classifications listed in a table of test specifications are developed by item writers who try to anticipate steps in the cognitive process that examinees typically follow in arriving at correct answers. However, item writers are often content experts, such as teachers and curriculum specialists, who are not usually trained to identify these mental processes. Furthermore, cognitive skill categories are often based on the Taxonomy of Educational Objectives developed by Bloom, Englehart, Furst, Hill, and Krathwohl (1956) which has been shown to be inadequate for classifying or predicting students' cognitive processes on tests of math achievement, for example (Gierl, 1997). As a result, care must be taken in deciding whether to rely on a single organizing principle or whether the research goal is best met by using a combination of strategies.

Regardless of the type of organizing principle used, creating bundles is only a first step in the two-stage Roussos-Stout DIF analysis paradigm (Roussos & Stout, 1996a). This stage can be described as a substantive analysis of the dimensional structure of a test. It is substantive to the extent that the dimensions are actually interpretable. The dimensions must then be identified as primary or secondary and then a decision must be made about whether secondary dimensions are measuring auxiliary or nuisance abilities. This leads to the formulation of hypotheses that will guide the DIF/DBF study. The

second stage involves statistically testing the dimensionality-based DIF hypotheses for differential bundle functioning.

SIBTEST

SIBTEST, short for "simultaneous item bias" test, was developed by Shealy and Stout (1993a, 1993b) as an outgrowth of their Multidimensional Model for DIF (MMD). SIBTEST is an item response theory (IRT)-based method that models the relationship between item performance and the latent trait(s) measured by a test. Equipped to handle dichotomously- or polytomously-scored items, this method tests for significant DIF amplification that occurs when a group of DIF items act together to produce differential bundle functioning.

Under the MMD, the latent (e.g., ability) space is considered a multidimensional continuum measuring primary (θ) and secondary (η) traits. The SIBTEST method uses a paramter estimate ($\widehat{\beta}_{UNI}$) to indicate the magnitude of DIF in an item. For large samples, $\widehat{\beta}_{UNI}$ has a normal distribution with mean 0 and standard deviation 1 under the null hypothesis of no DIF. The statistical hypothesis tested by SIBTEST is

$$H_o : \beta_{UNI} = 0 \ \ vs. \ \ H_1 : \beta_{UNI} \neq 0 . \tag{1}$$

Here β_{UNI} is defined as

$$\beta_{UNI} = \int [P(\theta, R) - P(\theta, F)\theta] f_F(\theta) d\theta , \tag{2}$$

where $P(\theta, R)$ and $P(\theta, F)$ are the probabilities of correct response (conditional on θ) for examinees in the reference and focal groups, respectively. The expression $f_F(\theta)$ is the density function for θ in the focal group. β_{UNI} is integrated over θ and yields a weighted

expected score difference between reference and focal group examinees of equal ability on a specific item or bundle. A statistically significant positive value of β_{uni} represents DIF against the focal group and a statistically significant negative value indicates DIF in favor of the focal group.

SIBTEST requires that test items be divided into a "studied" subtest (of items believed to exhibit DIF) and a matching (or "valid") subtest (of items believed to be DIF-free). That is, the studied subtest contains the items or bundle believed to measure the primary and secondary dimensions whereas the matching subtest contains the items believed to measure only the primary dimension. Examinee performance on items is compared by placing reference and focal group members into subgroups at each score level on the matching subtest.

β_{UNI} is unknown and therefore must be estimated from data. An unbiased estimate ($\hat{\beta}_{UNI}$) of β_{UNI} is obtained using the weighted mean difference between the reference and focal groups on the studied item/bundle across the *K* matched ability subgroups, or

$$\hat{\beta}_{UNI} = \sum_{k=0}^{K} p_k \left(\overline{Y^*_{Rk}} - \overline{Y^*_{Fk}} \right), \tag{3}$$

where the proportion of focal group examinees in subgroup *k* is represented by p_k and $\overline{Y^*_{Rk}} - \overline{Y^*_{Fk}}$ is the difference in the adjusted means on the studied subtest item or bundle for examinees in the reference and focal groups, respectively, in each subgroup *k*. Shealy and Stout (1993) added a regression correction to adjust the means on the studied subtest item or bundle to account for any differences in the ability distributions of the reference and focal groups.

The following classification scheme was proposed by Roussos and Stout (1996b, p.220) for assistance in interpreting the practical significance of a statistically significant value of $\widehat{\beta}_{UNI}$ on an single item: (a) Negligible or A-level DIF: Null hypothesis is rejected and the absolute value of $\widehat{\beta}_{UNI} < 0.059$, (b) Moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq \widehat{\beta}_{UNI} < 0.088$, and (c) Large or C-level DIF: Null hypothesis is rejected and $\left|\widehat{\beta}_{UNI}\right| \geq 0.088$. Unfortunately, no comparable guidelines exist to interpret or classify values of $\widehat{\beta}_{UNI}$ for item bundles.

DIF/DBF Research

As previously mentioned, most of the research on test bias has focused on the detection of differential functioning at the item level (see Millsap & Everson, 1993; Clauser & Mazor, 1998 for an extensive review). However, recent developments have extended DIF analyses to include investigations of differential bundle and test functioning (DBF/DTF). The first study involving DBF was conducted by Douglas, Roussos and Stout (1996) who demonstrated two methods for selecting item bundles suspected of exhibiting gender DIF amplification. Method 1 used a panel of judges to identify bundles of items that appeared to measure secondary abilities in addition to the target ability (i.e., logical reasoning) using data from a standardized administration of the logical reasoning subtest of the Law School Admission Test (LSAT). Using this method, the panel was able to form eight suspect item bundles, thus eight DIF hypotheses, to submit for statistical analysis. Method 2 was a mixed exploratory-confirmatory approach to identifying suspect bundles. This portion of the study involved a statistical IRT

dimensionality analysis followed by the use of expert opinion (i.e., panel of judges) to develop DIF hypotheses based on the number and type of dimensions that were identified statistically.

In both methods, SIBTEST was used to analyze the bundles for differential bundle/test functioning. For the example used with Method 1, although only four of the eight DIF hypotheses were statistically significant, Douglas et al. found that for seven of the eight bundles, the direction of DIF hypothesized by the panel of judges agreed with statistical results obtained with SIBTEST. Method 2 was illustrated using a 36-item National Assessment of Educational Progress (NAEP) history examination. The statistical dimensionality analysis augmented by expert opinion yielded a statistically significant six-item bundle representing a secondary dimension that Douglas et al. (1996) refer to as "knowledge of some important documents in early American history" (p. 477). The authors note that three of the six items in this bundle found to favor females did not reach statistical significance at the .05 level when a standard one-at-a-time DIF analysis was conducted. Their claim that test fairness should be investigated at the bundle level rather than at the item level is supported by this example of DIF amplification at the bundle level. Furthermore, that a secondary dimension could be identified as the source of DBF on the NAEP history exam illustrates the usefulness of these methods in identifying the causes of DIF/DBF in general.

In a later study, Oshima, Raju, Flowers and Slinde (1998) described how the framework of differential functioning of items and tests (DFIT) could be used to identify potential sources of DIF through the analysis of item bundles. Using an empirical data set, Oshima et al. (1998) used gender and socioeconomic status (SES; low vs. high) to

detect DBF on a reading comprehension test. Although no large bundle NCDIF values were obtained for passages in the SES analysis, the gender comparison did identify bundles that favored males over females due to item content. This finding highlighted the potential of DBF analysis as a mechanism for identifying possible sources of differential functioning.

Research has also demonstrated the use of DBF analysis for enhancing the interpretability of DIF results and how using different "organizing principles" for bundling items can improve the substantive review of flagged DIF items (see Gierl et al., 2001; Gierl & Khaliq, 2001). Other studies have compared the Type I error and power performance of procedures capable of detecting differential item, test, or bundle functioning (see Bolt, 2002; Russell, 2005). For example, Bolt (2002) studied DIF by testing the robustness to model misfit of two parametric methods (DFIT and the Likelihood Ratio Test) and one non-parametric method (Poly-SIBTEST) capable of detecting DIF for polytomously scored items. That is, by fitting one IRT model (Graded Response Model; Samejima, 1969) to data that were generated from two different IRT models – Muraki's (1992) generalized partial credit model and Mellenbergh's (1995) two-parameter sequential response model – Bolt defined DIF as model misfit and compared the power of each method to detect DIF at the item level. Results showed that there are advantages to using parametric methods when the appropriate IRT model is known. Additionally, while DFIT was more robust to model misfit (i.e., displayed lower Type I error rates) than the Likelihood Ratio test, DFIT also exhibited lower power in detecting DIF. Other notable findings emphasized the benefits of using different methods

(parametric vs. nonparametric) depending on the manipulated conditions (e.g., sample size and mean ability differences between groups).

A recent study by Russell (2005) assessing the Type I error and power performance of DFIT and SIBTEST for dichotomously-scored items is an extension of the work completed by Bolt (2002) with polytomous DIF detection methods. In his investigation, Russell subjected DFIT and SIBTEST to independent DTF and DBF analyses, respectively, using simulated and real data under a variety of conditions. Although intended to inform organizational research and practice, Russell's (2005) selection of conditions (sample size, test length, mean ability differences) include those common to educational achievement testing as well.

The Type I Error study conducted by Russell (2005) included 50 replications per condition and evaluated DTF and DBF detection at the .01 level of significance. Across conditions, the number of DTF Type I errors ranged from 0 (in two different conditions) to 24 (i.e., roughly 50 % of all statistical decisions were wrong in this condition). Given the 50 replications for each of 24 conditions and $\alpha = .01$, the average DTF Type I error rate with DFIT was .227. Using the same number of replications and level of significance, SIBTEST DBF Type I errors in a single condition ranged from 0 (in three different conditions) to 47 (i.e., almost all statistical decisions to reject the null hypothesis of no-DIF were erroneous in this condition). Moreover, the Type I error rate for DBF analysis with SIBTEST was .261 across all conditions.

Overall, DFIT and SIBTEST performed similarly across the various manipulated conditions. However, results cannot be directly compared because DFIT and SIBTEST were subjected to different levels of analysis (i.e., DTF and DBF, respectively). The

variables included in the Type I error study were dimensionality, test length, sample size, and target ability differences. As far as the impact of studied variables on false DTF and DBF detection with DFIT and SIBTEST, respectively, Russell observed large errors for each manipulated condition.

Dimensionality. Russell hypothesized that DFIT and SIBTEST would perform best within their intended domains. That is, the unidimensional IRT framework for DFIT would lead to fewer errors when analyzing unidimensional tests. On the other hand, SIBTEST was expected to perform better when analyzing multidimensional tests since it is based on Shealy-Stout's Multidimensional Model for DIF (MMD; 1993a). Although his hypotheses were confirmed, the Type I error rates for both methods were still much higher than the expected rate of .01. For instance, when DFIT was used to analyze unidimensional tests, the average DTF Type I error rate across all conditions was .188. Meanwhile, the average DBF Type I error rate across all conditions when SIBTEST was used for multidimensional tests was .168.

Test Length and Sample Size. Two test lengths (10 items versus 20 items) were simulated in the study. Although no condition was associated with acceptable Type I error rates for either method, DFIT made fewer Type I errors (error rate = .163) when analyzing longer, unidimensional tests while SIBTEST did not perform as well (error rate = .433) in this context. The study also compared group sample sizes of 500 (each for the reference and focal groups) and 1000 (each for the reference and focal groups) to examine the effect on the error rates of DFIT and SIBTEST. Both methods were hypothesized to exhibit fewer errors with smaller sample sizes (e.g., $N = 500$) than with larger sample sizes ($N = 1000$). This expectation was met for DBF analyses with

SIBTEST but not for DTF analyses with DFIT. In fact, substantial errors were reported for DFIT when used to analyze two-dimensional tests with small sample sizes (error rate = .370). However, SIBTEST performed at its best within this context (error rate = .140) and at its worst with one-dimensional tests and large samples sizes (error rate = .413)

Target Ability Differences. Russell studied the influence of target ability differences, or impact, between the reference and focal groups. Impact was simulated in one-direction (against the focal group) at three levels: 0.0, 0.5, and 1.00 to represent no impact, moderate impact and large impact, respectively. Russell's expectation that errors would increase as differences in target ability increased, were supported for both methods. However, the observed Type I error rate for SIBTEST (when used with one-dimensional tests) climbed as high as .715 when target ability differences, or impact, between the reference and focal group reached the highest levels in the study. Russell concluded that when subgroups differ greatly on the target ability, SIBTEST is more likely to mistakenly identify the impact as DIF/DBF.

The power study included 24 unique combinations of sample size (500 and 1000), test length (10 items vs. 20 items), target ability differences, or impact (0.0, 0.5, 1.00) and nuisance ability differences (DTF or DBF; 0.5 and 1.00). Using the same α level of .01, the average DTF power rate across all conditions with DFIT was .421. Russell concluded that the levels of DTF simulated in the study were too small and would often not be detected by DFIT (when these levels exist in real datasets). Furthermore, the number of times (out of 50) that DFIT accurately identified DTF in a single condition ranged from as low as 6 (i.e., for this condition test-level DIF was rarely detected) to 37.

On the other hand, SIBTEST demonstrated an overall power rate (across all conditions) of .725 with the number of true-positives in a single condition ranging from 11 to 50 (observed for 5 different conditions). Thus SIBTEST was found to be better at detecting differential bundle functioning than DFIT at identifying differential test functioning. Results for the individual variables fared as poorly (for both methods) as was observed for the Type I Error study. For example, low DTF and DBF power rates were associated with both test lengths although SIBTEST showed greater power to detect differential bundle functioning with longer tests. In the case of sample size differences, Russell's hypothesis that DFIT would perform best with large samples was rejected while SIBTEST demonstrated power near the acceptable level of .80. For conditions involving target ability differences, both methods performed best when impact was minimal. Finally, nuisance or secondary ability differences – the variable used to simulate DBF and DTF in the Russell study – exhibited a large influence on power rates for both techniques. According to Russell, DFIT did not reach acceptable power rates for large amounts of DIF but may potentially identify true-DTF at the target level (of .80) for tests with a greater percentage of items measuring the nuisance ability. On the other hand, DBF power rates for SIBTEST reached as high as 100% for various conditions involving large nuisance ability differences.

In summary, Russell's study examined the effect of secondary traits on DIF detection. That is, Russell (2005) used multidimensionality to produce differential item/bundle/test functioning in simulated item responses. Overall, DFIT and SIBTEST demonstrated poor Type I error control in the presence of multidimensionality. Also both methods performed best for large sample sizes (N=1000 vs. N=500), minimal target

ability differences, and shorter tests (10 items vs. 20 items), although SIBTEST adhered to the nominal alpha level better than DFIT for multidimensional tests and smaller sample size (N=500). However, interpretation of Russell's (2005) findings must be made with caution considering DFIT and SIBTEST were subjected to different levels of analysis.

Additionally, two power studies were conducted to assess DFIT and SIBTEST's ability to detect simulated bias at the test and bundle level, respectively. Results showed SIBTEST to be more powerful (average power of .725 across all conditions) at detecting DBF than DFIT (average power of .421 across all conditions) for assessing DTF. As noted earlier, no direct comparisons could be made due to the disparate levels of analysis. Nonetheless, attempts were made to explain the stark differences in the two methods. For example, the SIBTEST procedure is advantaged since only user-specified bundles of items are submitted for DBF analysis, whereas the DFIT procedure made use of all test items during the assessment of DTF.

Conditions That Impact DIF Detection

A number of studies have examined the impact of various conditions on the performance of DIF detection techniques often used in educational achievement testing (e.g., Miller & Oshima, 1992; Nandakumar, 1993; Raju, van der Linden & Fleer, 1995; Chamblee, 1998; Bolt, 2002; Russell, 2005). Research in this area has guided measurement specialists in the selection of DIF methods that perform best under specific conditions. Most of the articles reviewed for this study concentrated on the investigation or manipulation of conditions such as test length, sample size, item parameters,

proportion of test-wide DIF, group ability distributions, multidimensionality, type of IRT model, as well as the magnitude, direction and uniformity of DIF and the impact of these conditions on DIF detection. Although the list is long, it is important to systematically study the effect of these conditions that commonly occur in many testing applications.

A review of Monte Carlo studies that manipulated different combinations of the previously mentioned conditions yielded consistent, as well as diverging conclusions, regarding the impact on prominent DIF detection methods. For example, in manipulating sample size, proportion of test-wide DIF, and direction and uniformity of DIF for four IRT-based DIF methods (i.e., Lord's χ^2, Exact Signed Area, Exact Unsigned Area, and DFIT) Raju, van der Linden and Fleer (1995) found fewer detection errors with larger sample sizes and smaller proportions of test-wide DIF (i.e., number of DIF items). However, results of the bidirectional DIF condition (i.e., DIF that favors reference groups for some items and focal group for others) and nonuniform DIF (i.e., unequal discrimination and equal or unequal difficulty parameters across groups) conflicted with those obtained from an earlier study by Cohen and Kim (1993). In Raju et al. (1995), bidirectional and nonuniform DIF caused higher detection errors for the Exact Signed Area (ESA) method, whereas the Cohen and Kim study found no differences between the DIF methods. Another study, comparing single versus two-stage procedures for estimating DIF found that the decision should be based on the magnitude and percentage of DIF items on the test (Miller & Oshima, 1992). In that investigation, Miller and Oshima concluded that a single-stage procedure may be more useful for examining race, ethnic, or gender bias when there are few DIF items (e.g., 5% or 10%) and the magnitude, or strength, of DIF is weak.

Another variable previously shown to effect DIF detection is related to item discrimination and is known as angular item direction. As shown by Reckase and McKinley (1991), the discrimination power of an item is related to a specific direction in the latent space. For multidimensional items, this angular direction can be calculated with respect to the latent axes. The value of α can vary from 0° to 90° based on the degree to which an item measures each trait. For the case of two dimensions, if an item measures only the first ability, θ, then the item direction (α) is 0°. If an item measures only the second ability, η, then the item direction (α) is 90°. When α = 45° an item measures two abilities (θ and η) equally. Therefore, calculating an item's angular direction (α) provides information about what items are really measuring.

Reckase and McKinley (1991) introduced formulas for computing α that require knowledge of other parameters such as multidimensional discrimination (MDISC) and multidimensional item difficulty (MID). To investigate the impact of simulated item directions on DBF detection with SIBTEST, these formulas are unnecessary. Since the angle formed by the discrimination vector and the horizontal axis for the first dimension (θ) can be calculated with simple trigonometry of right triangles, a simplified version of the formulas provided by Reckase and McKinley (1991) for computing item direction follows the example below.

Figure 1 shows the discrimination vector for an item in a two-dimensional space in which the horizontal axis represents measurement of the first dimension θ and the vertical axis represents measurement of the second dimension η. This item has discrimination $a_1 = 1.40$ and $a_2 = 0.40$. Therefore, the angular direction (i.e., the angle the

discrimination vector forms with the θ axis) of this item (which discriminates more on θ than η) can be calculated using the following formula:

$$\alpha = \tan^{-1}\frac{a_2}{a_1} = \tan^{-1}\frac{0.4}{1.4} = 15.95° \quad (4)$$

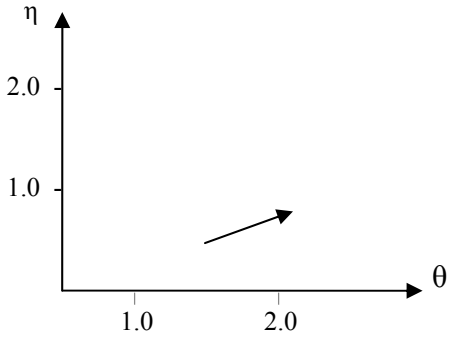

Figure 1. Discrimination Vector for an Item in a Two-Dimensional Ability Space

Since the item's angular direction is less than 45°, the item measures the θ dimension to a greater degree than the η dimension. Few studies have examined the impact of item direction on the detection of DIF and none have investigated its effect on DBF analysis. One study by Oshima and Miller (1992) observed increases in power (with four item bias methods - signed area, unsigned area, signed sum of squares and unsigned sum of squares) when multidimensional items embedded with DIF measured the secondary dimension more than the first. The results of that study highlighted the usefulness of item dimensionality assessment for enhancing DIF detection.

The literature on differential item/bundle functioning also includes studies that specifically examine the performance of DBF methods (capable of detecting DIF amplification and cancellation) under conditions similar to those reviewed above. One of the earliest such investigations was conducted by Nandakumar (1993) who used

SIBTEST to study simultaneous DIF amplification and cancellation. By varying sample size, test length, percent of DIF items, and direction of DIF with simulated data, Nandakumar's (1993) study helped to establish SIBTEST as an effective procedure for studying DIF at the item and test level. In the study, SIBTEST was compared to the Mantel-Haenszel technique and found to perform similarly across conditions in the assessment of DIF at the item level. At the test level, SIBTEST successfully estimated the cumulative effect of DIF. That is, whether DIF was amplified to produce differential test functioning (DTF) or cancelled out due to bidirectional DIF in individual items, SIBTEST was capable of assessing both.

In her dissertation, Chamblee (1998) investigated conditions that impact Type I error rates of DFIT including test length, sample size, choice of IRT model (1-, 2-, or 3-PL) and linking method for four different alpha levels. These generated data were used to empirically determine critical values for the NCDIF and DTF indices associated with DFIT. Chamblee concluded that Type I error rates decreased with increasing sample sizes and when number of estimated parameters (i.e. IRT model) decreased. Also, shorter tests (20 items) were associated with larger critical values. Few differences were observed between the two linking methods (Test Characteristic Curve and modified Test Characteristic Curve) used in the study.

Recent investigations have examined the impact of various conditions while comparing the performance of DIF/DBF detection methods. As previously mentioned, Bolt (2002) compared parametric and nonparametric polytomous DIF methods under a number of simulated conditions. In his study, DIF was generated using a traditional approach (i.e., specifying different reference and focal group parameters) and using a

somewhat novel technique. The second approach operationalized DIF as model misfit by using data generated from Muraki's (1992) generalized partial credit model and Mellenbergh's (1995) two-parameter sequential response model to estimate item parameters for Samejima's (1969) graded response model. The conditions manipulated in the study included the model used to generate the data, sample size, and mean ability differences between the focal and reference groups. Across conditions, DFIT exhibited greater power and was prone to more Type I error than poly-SIBTEST, particularly for the smaller sample size condition (N=300).

Although Bolt's (2002) study was one of the first to compare the DFIT and SIBTEST procedures across various conditions, his investigation was limited to an item-level analysis. As a result, in 2005 Russell examined Type I error and power performance of DFIT and SIBTEST at the test and bundle level, respectively. As mentioned earlier, Russell's selection of conditions (e.g., test length, sample size, mean ability differences) included those that are common to educational testing.

Overall, Russell's results both corroborated and conflicted with Bolt's (2002) findings but it is important to note that the Russell (2005) study compared DFIT and SIBTEST across different levels of analysis without comparing the two techniques within a specific domain (e.g., comparing DFIT and SIBTEST performance for specific bundles of items). Also, despite manipulating such a large number of variables (i.e. sample size, test length, mean ability differences, and presence/absence of a secondary trait), Russell (2005) did not explore more than two levels for most of the simulated conditions.

Similar to Russell (2005) this study does not include results for differential *bundle* functioning analysis with the DFIT procedure. Although the Russell (2005) study

examined differential *test* functioning with DFIT, results indicated low power and high Type I error rates leading to the conclusion (like many previous studies) that more reliable cutoff values are needed to accurately assess DIF/DTF with DFIT. To address this long-standing issue, Oshima, Raju and Nanda (2006) introduced the Item Parameter Replication (IPR) method that determines cutoff values for DFIT based on particular datasets and individual items (see Appendix A for more detailed information on DFIT). Although the new DFIT procedure using the IPR method has been demonstrated within a simulation study (Oshima, Raju & Nanda, 2006) involving one replication, its capacity to examine differential item/bundle/test functioning with more than a few replications is severely limited by current computer processing time. In fact, this study was originally designed to include an evaluation of DBF analysis with the SIBTEST *and* DFIT procedures. However, in preparing data for analysis with DFIT, many complications arose. Therefore it was determined that at the current time, while it is a potentially useful tool for applied researchers, DFIT presents too many limitations for use in a simulation study. More information on DFIT is provided in Chapter 5 and Appendix A.

As a result of the challenges posed by DFIT, this study examines only the SIBTEST procedure for DBF analysis. The purpose of this study was to assess the performance of SIBTEST for DBF analysis under various conditions with simulated data. Therefore the simulation study examined how Type I error and power rates of SIBTEST are impacted by multiple levels of various manipulated conditions that applied researchers often encounter. Additionally, suggestions are provided for increasing the potential of DFIT for use with simulation research.

CHAPTER 3

METHODOLOGY

In this Monte Carlo simulation study differential bundle functioning was assessed using the SIBTEST procedure under various conditions that practitioners often encounter with real data. The variables of interest in the study included sample size and sample size ratios, correlation between primary and secondary dimensions, magnitude of DIF/DBF, and angular item direction. The impact of each of these variables and their combined effect on accurate and false identification of DBF was the primary focus of the investigation. To evaluate the power and Type I error performance of SIBTEST the following study design was used.

Study Design

The item parameters used to generate unidimensional data across all of the conditions were selected from Raju, van der Linden and Fleer's (1995) study of differential item functioning. These parameters were reported to have item characteristics typically found in real testing situations. Parameters for the multidimensional bundle were held constant for all ten items (within each level of item direction) so that DIF/DBF would not be confounded with item difficulty and discrimination. Appendix B (Tables 7 – 9) contains a complete list of unidimensional and multidimensional item parameter values.

Sample Size and Sample Size Ratio. Assessing the impact of sample size on DIF detection was one of the conditions investigated in this study. Specifically, two levels of overall sample size (i.e., 2000 and 5000) as well as two different sample size ratios of reference to focal group members (i.e., 50/50 and 90/10) were used to compare the Type I error and power performance of SIBTEST. Although sample size is one of the most commonly studied conditions simulated in DIF research, it is an important factor to include here as a control for absolute sample size when sample size ratios are also being manipulated. As for the sample size ratio conditions, these proportions (i.e., 50/50 and 90/10) were chosen to represent other characteristics of real test data not examined in previously mentioned research [e.g., Raju et al., 1995; Russell, 2005].

Instances in which the ratio of reference to focal group members in the population might be expected to be approximately equal would include studies of gender DIF in which there are relatively equal numbers of boys and girls (who constitute the reference and focal groups, respectively). An example of the second instance (unequal proportions) might be when a large number of white students take an exam compared to a smaller number of black or Hispanic students. In such cases, the examinee population (and hence representative sample) might reveal a sharp contrast of maybe 90% whites to 10% blacks, or 90/10 ratio.

Correlation between Dimensions. Another condition of interest in this study is the impact of different correlations between primary and secondary dimensions, or abilities. Since prior research (e.g., Kirisci, Hsu, & Yu, 2001) has suggested the use of unidimensional IRT methods when the correlation between dominant dimensions is moderate to large (e.g., $r > .40$), this study examined the effect of correlations between

dimensions on Type I error and power rates of SIBTEST. The higher the correlation is between dominant dimensions (e.g., $r = 0.8$) the more similar the dimensions are and may be considered as one primary dimension. On the other hand, dimensions become more distinct as the correlation between them decreases (e.g., $r = 0.3$). Therefore, three levels are proposed for this condition: $r = 0.316$, 0.632 and 0.837 (corresponding to $r^2 = 0.1$, 0.4, and 0.7, respectively). Selecting this range of values was guided by a couple of considerations. First, a zero correlation between dominant dimensions on say, an achievement test, is very unlikely. Second, it is unusual to observe correlations higher than 0.8, although the dimensions on some tests of math ability (e.g., SAT-Math) have been shown to have higher correlations (i.e., $.946 < r < 1.000$).

Magnitude of DIF/DBF. This study also manipulated DIF magnitude which was expressed as differences in secondary ability (η) distributions for the focal and reference groups. Examining the influence of varying magnitudes of DIF in item bundles may help to determine the minimum amounts of multidimensionality required to detect differential bundle functioning with SIBTEST. Therefore four levels of DIF magnitude (i.e., nuisance ability differences): $d_\eta = 0.25$, $d_\eta = 0.5$, $d_\eta = 0.75$, $d_\eta = 1.0$ were included to reflect small to large amounts of DIF.

Angular Item Direction. Finally, angular item direction was manipulated to examine its effect on DBF detection with SIBTEST. Recall the earlier example in which the item direction was 15.95° indicating that the item measured the primary dimension (θ) to a greater degree than the secondary dimension (η). What if this item was intended to measure only θ? What if subgroups of examinees differed in mean ability on the second dimension η? Would DIF result? What would happen if the angular direction

were greater for this item? That is, what would be the effect of increasing the degree to which this item measured the nuisance secondary trait? These are all questions that were explored in this study. More specifically, the influence of item direction on DBF detection with SIBTEST was investigated for two levels of α: 26.1° (higher discrimination on the primary trait) and 58.5° (higher discrimination on the nuisance trait). The item directions were simulated using discrimination parameters of $a_1 = 1.00$ and $a_2 = 0.49$ for α = 26.1° and $a_1 = 0.30$ and $a_2 = 0.49$ for α = 58.5°.

Overview of Study Design

Power. A power study was conducted to investigate the impact of several manipulated variables on the ability of SIBTEST to detect differential bundle functioning on a simulated 40-item test. The variables studied include overall sample size, reference to focal group sample size ratios, DIF magnitude, correlations between primary and secondary abilities, and item directions. The power study included 2 (overall sample sizes) x 2 (sample size ratios) x 4 (magnitudes of DIF) x 3 (correlations between primary and secondary traits) x 2 (item directions) = 96 fully crossed conditions (see Table 1). This analysis used 1000 replications of each condition.

Type I error. A Type I error study was also conducted to examine false detection of differential bundle functioning with SIBTEST. Conditions thought to impact Type I error rates were examined for the simulated 40 item test (see Table 1). To study Type I error performance, the presence/absence of a secondary ability was partially crossed with sample size, ratio of reference to focal group sample size, correlation between primary (θ) and secondary (η) abilities and item direction. Hence, the design of the Type I error

Table 1

Study Conditions

Sample Size
1. 2000
2. 5000

Sample Size Ratio
1. 50:50
2. 90:10

Correlation between Dimensions
1. .316
2. .632
3. .837

DIF/DBF Magnitude
1. 0.25
2. 0.50
3. 0.75
4. 1.00

Angular Item Direction
1. 26.1°
2. 58.5°

---

study included 2 (overall sample sizes) x 2 (sample size ratios) x 2 (presence/absence of secondary trait) = 8 fully crossed conditions, and 2 (overall sample sizes) x 2 (sample size ratios) x 3 (correlation between primary and secondary traits) x 2 (item directions) minus the need to examine the baseline Type I error condition (i.e., absence of a secondary trait) again for the second item direction = 20 partially crossed conditions. The "baseline" Type I error study was represented by conditions involving the 10 multidimensional items when there was no DIF (i.e., the distribution of ability on the nuisance secondary dimension was equal for the reference and focal groups). Finally, the analysis was conducted using 1000 replications of each condition so that the effect of individual manipulated variables (e.g., sample size ratio) on Type I error rates could be examined.

Data Generation

Unidimensional and multidimensional data for the Type I error and power studies were generated according to the conditions described above. IRTGEN (Whitaker, Fitzpatrick, Williams, & Dodd, 2003), a SAS macro program, was used to simulate unidimensional items. However, multidimensional items for the reference and focal group were simulated using the SAS/IML program GENMIRT (Kromrey, Parshall, Chason, & Yi, 1999). For items measuring only the primary ability (θ), dichotomous data (i.e., item responses of 0 and 1 representing incorrect and correct answers, respectively) were generated using the three-parameter logistic model (3PL) as shown in Equation 5.

$$P\left(u_i = 1 \mid \theta\right) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}} \tag{5}$$

where $P(u_i=1 \mid \theta)$ is the probability that an examinee correctly answers an item (and $u_i=0$ for an incorrect answer) and e represents the base of natural logarithms ($e \approx 2.718$). However, a two-dimensional 3PL model (see Equation 6) was used to simulate items measuring the primary and secondary (i.e., nuisance) dimensions.

$$P\left(u_i = 1 \mid \theta, \eta\right) = c_i + \frac{1 - c_i}{1 + e^{\left\{-1.7\left(a_{\theta_i}\left(\theta - b_{\theta_i}\right) + a_{\eta_i}\left(\eta - b_{\eta_i}\right)\right)\right\}}} \tag{6}$$

For conditions involving no-DIF, reference and focal group ability on the primary (θ) and secondary (η) traits was generated according to a multivariate normal distribution with mean 0 and standard deviation of 1. In the power study, DIF was added to each multidimensional item by increasing the mean ability of the reference group above that of the focal group by 0.25, 0.50, 0.75 and 1.00 standard deviations on the second dimension, η. After generating item response data, SAS was used to call the DOS-based version of SIBTEST and to produce output files for the DBF analyses. A copy of the program can be found in Appendix C.

Analyses

This study involved simulating 25% of test items to contain DIF. That is, 10 items that measured primary and secondary abilities formed a "bundle" that was submitted to SIBTEST for differential bundle functioning analyses. Conducting DBF analyses using SIBTEST is a one-step procedure that does not require parameter estimation or equating steps. The SIBTEST program requests information from the user such as, the identification of items that will form the valid and studied subtests and

whether the test permits guessing and, if so, what the *c*-parameter value (for the 3PLM) is. For the simulated 40-item test, the last 10 items formed the multidimensional bundle. Thus for the power and Type I error studies, the valid subtest consisted of the first 30 items and the studied subtest consisted of the last 10 items.

SIBTEST produces a parameter estimate, $\widehat{\beta}_{UNI}$, to indicate the magnitude of DIF in an item or bundle. A statistically significant positive value of $\widehat{\beta}_{UNI}$ represents DIF against the focal group and a statistically significant negative value indicates DIF in favor of the focal group. Power and Type I error rates were calculated for each condition by tallying the number of times (out of 1000 replications) that SIBTEST detected statistically significant DBF in the 10-item bundle.

CHAPTER 4

RESULTS

Tables 2 – 6 display the results for conditions in which differential item functioning was embedded in a group of ten multidimensional items on a 40-item test, producing differential bundle functioning (Power Study) and conditions for which no-DIF was imbedded in these multidimensional items (Type I Error study). The results begin with a summary of the overall effect of each manipulated variable on Power and Type I Error. Overall results are followed by detailed descriptions of study effects for simulations involving item direction $\alpha = 26.1°$ and item direction $\alpha = 58.5°$, respectively.

Power. In this study, DBF was created in a ten-item bundle of multidimensional items by simulating the reference group examinees to have higher ability than the focal group on the secondary (nuisance) dimension across all DBF conditions. Manipulated variables included total sample size and sample size ratio of reference to focal group, magnitude of DIF/DBF, correlation between the primary and secondary dimension, and item direction. All conditions in the power study were replicated 1000 times with the SIBTEST procedure.

Item Direction. Two item directions were simulated for the study. The first scenario ($\alpha = 26.1°$) represented multidimensional items that discriminated more on the primary dimension compared with results that were obtained in the second scenario ($\alpha = 58.5°$) where multidimensional items discriminated more on

Table 2

Power Rates for Item Direction α = 26.1°. Percentage of Times (out of 1000 replications per condition) that SIBTEST Accurately Identified DBF.

				Nuisance Ability Difference			
				d_η =.25	d_η =.50	d_η =.75	d_η =1.00
Correlation between dimensions	Ntot	Nref	Nfoc	DBF	DBF	DBF	DBF
r = 0.316	5000	4500	500	**52.4**	**96.6**	**100.0**	**100.0**
		2500	2500	**89.5**	**100.0**	**100.0**	**100.0**
	2000	1800	200	**28.4**	**70.0**	**94.2**	**99.6**
		1000	1000	**58.4**	**97.7**	**100.0**	**100.0**
r = 0.632	5000	4500	500	**48.0**	**94.1**	**99.7**	**100.0**
		2500	2500	**86.7**	**100.0**	**100.0**	**100.0**
	2000	1800	200	**23.5**	**64.6**	**91.0**	**99.2**
		1000	1000	**49.4**	**97.5**	**100.0**	**100.0**
r = 0.837	5000	4500	500	**42.7**	**92.7**	**99.7**	**100.0**
		2500	2500	**84.0**	**100.0**	**100.0**	**100.0**
	2000	1800	200	**21.7**	**59.6**	**86.4**	**98.0**
		1000	1000	**48.7**	**96.2**	**100.0**	**100.0**

Table 3

Power Rates for Item Direction α = 58.5° Percentage of Times (out of 1000 replications per condition) that SIBTEST Accurately Identified DBF.

				Nuisance Ability Difference			
				d_η =.25	d_η =.50	d_η =.75	d_η =1.00
Correlation between dimensions	Ntot	Nref	Nfoc	DBF	DBF	DBF	DBF
r = 0.316	5000	4500	500	**91.3**	**100.0**	**100.0**	**100.0**
		2500	2500	**100.0**	**100.0**	**100.0**	**100.0**
	2000	1800	200	**55.5**	**98.6**	**100.0**	**100.0**
		1000	1000	**94.1**	**100.0**	**100.0**	**100.0**
r = 0.632	5000	4500	500	**88.7**	**100.0**	**100.0**	**100.0**
		2500	2500	**99.8**	**100.0**	**100.0**	**100.0**
	2000	1800	200	**51.8**	**96.8**	**100.0**	**100.0**
		1000	1000	**90.0**	**100.0**	**100.0**	**100.0**
r = 0.837	5000	4500	500	**84.8**	**100.0**	**100.0**	**100.0**
		2500	2500	**100.0**	**100.0**	**100.0**	**100.0**
	2000	1800	200	**48.9**	**94.9**	**99.9**	**100.0**
		1000	1000	**87.6**	**100.0**	**100.0**	**100.0**

Table 4

Baseline Type I Error Rates. Percentage of Times (out of 1000 replications per condition) that SIBTEST Falsely Identified DBF in the Unidimensional Item Bundle.

Sample Size			
Ntot	Nref	Nfoc	DBF
5000	4500	500	**0.05**
	2500	2500	**0.05**
2000	1800	200	**0.06**
	1000	1000	**0.06**

Table 5

Type I Error Rates for Item Direction α = 26.1°. Percentage of Times (out of 1000 replications per condition) that SIBTEST Falsely Identified DBF in the Multidimensional Item Bundle.

	Total Sample Size					
	Ntot=5000			Ntot=2000		
Correlation between dimensions	Nref	Nfoc	DBF	Nref	Nfoc	DBF
0.316	4500	500	**0.05**	1800	200	**0.04**
	2500	2500	**0.06**	1000	1000	**0.04**
0.632	4500	500	**0.04**	1800	200	**0.05**
	2500	2500	**0.06**	1000	1000	**0.04**
0.837	4500	500	**0.04**	1800	200	**0.06**
	2500	2500	**0.04**	1000	1000	**0.05**

Table 6

Type I Error Rates for Item Direction α = 58.5°. Percentage of Times (out of 1000 replications per condition) that SIBTEST Falsely Identified DBF in the Multidimensional Item Bundle.

	Total Sample Size					
	Ntot=5000			Ntot=2000		
Correlation between dimensions	Nref	Nfoc	DBF	Nref	Nfoc	DBF
0.316	4500	500	**0.05**	1800	200	**0.06**
	2500	2500	**0.06**	1000	1000	**0.05**
0.632	4500	500	**0.05**	1800	200	**0.07**
	2500	2500	**0.06**	1000	1000	**0.05**
0.837	4500	500	**0.07**	1800	200	**0.05**
	2500	2500	**0.06**	1000	1000	**0.06**

the nuisance trait. For α = 26.1°, a_1 = 1.00 and a_2 = 0.49 whereas for α = 58.5°, a_1 = 0.30 and a_2 = 0.49.

SIBTEST power results are displayed for each item direction in Tables 2 and 3. Overall, item direction was very influential in DBF detection. Comparing the values in the tables, it is evident that power to detect differential bundle functioning across all conditions with SIBTEST was much greater for the larger item direction (α = 58.5°). For example, observed power rates for α = 26.1° ranged from 21.7% to 100% whereas rates for α = 58.5° ranged from 48.9% to 100% across all conditions. Therefore, when multidimensional items measure the nuisance trait more than the primary trait and the average ability level of the focal group is lower than the reference group on that nuisance dimension, then DIF/DBF is more easily identified by SIBTEST.

Sample size. Sample size, defined as the sum of the number of examinees in the reference and focal group, heavily influenced the DBF detection power of SIBTEST. Two overall sample sizes (N = 2000 and N = 5000) were simulated for this study. As might be expected, higher power rates were associated with larger sample sizes (especially for α = 26.1°). In particular, in conditions with N = 2000 examinees, the average SIBTEST power rate failed to reach 80% for α = 26.1° (actual rate = 78.5%) whereas the average rate of DBF detection was over 90% for conditions where N = 5000 examinees (actual rate = 91.1%). On the other hand, average power rates for α = 58.5° were 92.4% and 98.5% with N = 2000 and N = 5000, respectively.

Sample size ratio. The ratio of reference to focal group members was also investigated. Two ratios (90:10 and 50:50) were examined for their impact on DBF detection rates. Across all conditions, SIBTEST demonstrated greater detection power for equal reference and focal group sizes (i.e., 50:50) than for sharply unequal group sizes (i.e., 90:10). Specifically, for $\alpha = 26.1°$, 9 out of 12 (75%) low power conditions (i.e., power rate less than 80%) involved the 90:10 sample size ratio. On the other hand, all three low power conditions observed for $\alpha = 58.5°$ involved 90:10 ratios with focal groups of only 200 examinees and small amounts of DBF ($d_\eta = .25$).

Magnitude of DIF/DBF. Tables 2 and 3 display the four levels of DIF/DBF magnitude (.25, .50, .75 and 1.00) that were simulated and the resulting power rates. Not surprisingly, DIF bundles were correctly identified by SIBTEST more often for higher magnitudes of DIF/DBF. That is, DIF amplification was observed with greater frequency for maximum differences between reference and focal group ability on the secondary dimension. For example, the average power rates for DIF/DBF magnitude of .25 were 52.8% (for item direction $\alpha = 26.1°$) and 82.7% (for item direction $\alpha = 58.5°$). Whereas the average power rates for DIF/DBF magnitude of .50 were 89.1% (for item direction $\alpha = 26.1°$) and 99.2% (for item direction $\alpha = 58.5°$). The average power rates for conditions involving DIF/DBF magnitude greater than .50 (i.e., .75 and 1.00) were 98.7% (for item direction $\alpha = 26.1°$) and 100% (for item direction $\alpha = 58.5°$).

Correlation between primary and nuisance traits. This study examined the effect of correlated dimensions on DBF detection at three levels: 0.316, 0.632,

and 0.837 (corresponding to $r^2 = 0.1$, 0.4, and 0.7, respectively). Although detection power was lost as dimensions became more correlated, the average power rate when $r_{\theta\eta} = .837$ was 83.1% (for item direction $\alpha = 26.1°$) and 94.8% (for item direction $\alpha = 58.5°$). Therefore, the ability of SIBTEST to identify DIF/DBF caused by multidimensionality is considerably high even when the nuisance trait (η) is strongly correlated (e.g., $r = 0.837$) with the targeted primary ability dimension (θ).

Type I Error. As illustrated by Tables 4 - 6, SIBTEST adhered closely to the nominal alpha level of .05 when the test contained only unidimensional items and for non-DIF multidimensional items. That is, the hypothesis of No-DBF was falsely rejected at the .05 level of significance roughly 5% of the time. Across all study conditions, the average Type I error rate was 4.7% which is an acceptable amount that can be attributed to random sampling error.

For the "baseline" Type I error study, error rates (as shown in Table 4) ranged from .05 to .06 across the four conditions. In particular, the average error rates for N = 2000 and N = 5000 were .06 and .05, respectively. Thus, Type I error rates showed a slight decrease as sample size increased.

For the analysis of DBF with non-DIF multidimensional items, the error rates (as shown in Tables 5 and 6) ranged from .04 to .06 for $\alpha = 26.1°$ and from .05 to .07 for $\alpha = 58.5°$. Overall, Type I error was slightly higher for the larger item direction. The same relationship was observed when investigating the influence of correlated dimensions and sample size on DBF detection with SIBTEST. In fact, the average Type I error rate across all correlations and across

both sample sizes was .05 for α = 26.1° and .06 for α = 58.5°. Finally, a notable finding was that the smallest error rate (.04) was observed for conditions involving either the highest total sample size (N = 5000) and strongest correlation ($r_{\theta\eta}$ = .837) *or* the smallest total sample size (N = 2000) and the weakest correlation ($r_{\theta\eta}$ = .316). However, this relationship was only observed for the item direction conditions (α = 26.1°) in which all of the multidimensional items discriminated more on the target, or primary, dimension (θ) than the nuisance secondary dimension (η)

The results presented so far represent a summary of influences exerted by the variables investigated in the study. However, the effect of each manipulated variable was, in most cases, mediated by the influence of other study variables. Hence, a description of the combined effects of all conditions on the power and Type I error performance of SIBTEST follows.

For both item directions (α = *26.1*° and α = 58.5°) the lowest power rates were associated with DIF/DBF magnitude (d_η) of .25, r = .837, and 90:10 ratio of reference to focal group members. This was especially true when the condition involved 1800 reference group members and 200 focal group members. For α = 26.1° and d_η = .25 acceptable power (i.e., power ≥ .80) was observed only for large reference *and* focal group sizes (i.e., 50:50 ratio with n_{ref} = 2500 and n_{foc} = 2500). As medium amounts of DIF (i.e., d_η = .50) are introduced acceptable power rates are reached for all sample sizes except 1800:200 ratio.

On the other hand, for α = 58.5° acceptable power is observed across all conditions when DIF magnitude reaches at least .50. For α = 26.1°, the

correlation between dimensions greatly affected DBF detection when ability differences on the nuisance trait were small to medium (i.e., .25 to .50). As shown in Table 2, as the correlation between dimensions increased from .316 to .837, power to detect DBF decreased across all sample size ratios and for every level of DBF simulated. For $\alpha = 58.5°$, once the amount of DIF/DBF reached .50 this pattern was only observed for the 90:10 ratio with the small focal group size of 200.

Not surprisingly, the highest power rates were associated with the largest simulated level of DIF magnitude ($d_\eta = 1.00$). Higher amounts of DIF were required before acceptable power was reached for every combination of sample size, sample size ratio and correlation between the two dimensions. For $\alpha = 26.1°$, the required amount was $d_\eta = .75$ whereas for $\alpha = 58.5°$ a mean difference as large as .50 in reference and focal group ability on the nuisance secondary dimension was sufficient for SIBTEST to detect DBF with at least 95% accuracy.

CHAPTER 5

DISCUSSION

This study investigated factors that effect DBF detection with SIBTEST, a well-known IRT-based method. DIF/DBF was created in ten multidimensional items that formed a "bundle" on a simulated 40-item test designed to measure only one dimension. Power and Type I error performance was assessed by calculating DBF detection rates for the ten-item bundle of DIF items and the 30-item bundle of non-DIF items, respectively. The interaction of sample size, sample size ratio, DIF/DBF magnitude, correlation between dimensions, and angular item direction provided much needed information to guide further use of SIBTEST for identifying potential sources of differential item/bundle functioning.

Power Performance for true-DIF/DBF conditions

Sample size and sample size ratio. As with many DIF studies involving sample size (see Shealy & Stout, 1993b; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996b), results here indicated an increase in power to detect DIF/DBF as sample size increased (e.g., from increasing from 2000 to 5000 total examinees). The effect of sample size was further examined by comparing detection rates for equal and unequal reference and focal groups. When the number of reference group members was nine times that of the focal group (i.e., 1800:200 and 4500:500 ratios), SIBTEST exhibited less power than in the equal sample size conditions (i.e., 1000:1000 and 2500:2500).

This relationship was magnified in the conditions involving only 200 focal group members (compared to 1800 in the reference group). In practice, ratios like these might be observed in gender DIF studies involving an equal number of male and female examinees (50:50) or when ethnic DIF is investigated for a group of White and Hispanic examinees (90:10).

Magnitude of DIF/DBF. The influence of various examinee characteristics were also investigated in this study. In particular, the difference in mean reference and focal group ability on the "unintended" secondary dimension was manipulated at four levels or magnitudes of DIF/DBF. For a standardized difference of .25, .50, .75, and 1.00 in mean ability between the groups, DIF detection was examined with SIBTEST and found to increase with increasing levels of DIF/DBF magnitude. The impact of DIF/DBF on DIF detection was not surprising as higher amounts of DIF should be detected with greater frequency than smaller amounts. Still, the degree to which SIBTEST was able to identify DIF when only small to medium amounts were simulated was impressive. The highest power rates were observed for DIF magnitudes of .50 or greater combined with high sample sizes and low correlation between the primary and secondary dimension. These results agree with those obtained in earlier studies on the effect of DIF magnitude on detection rates. For example, Oshima and Miller (1992) observed high power to detect item bias in a simulation study involving an essentially unidimensional 40-item test that included varying numbers of multidimensional items for which there was a difference in mean ability for the reference and focal groups on the nuisance trait. Their results showed higher detection rates for four bias indexes (signed area, unsigned area, signed sum of

squares and unsigned sum of squares) when the reference group had a mean ability that was 0.5 higher than the focal group on the secondary dimension.

Correlation between dimensions. The inclusion of a secondary dimension (η) in this investigation was necessary for the creation of item bundles to be tested for differential bundle functioning. Simulating item bundles is a relatively new concept and thus requires sufficient description. In this study, DIF was produced in ten multidimensional items for which the reference and focal group differed in mean ability. This allowed the ten items to form a "bundle" that was subsequently submitted to SIBTEST for DBF analysis. Additionally, it was of interest to this study to examine the power of SIBTEST to detect DIF/DBF when the two dimensions exhibited small, medium and large degrees of correlation (i.e., 0.316, 0.632, and 0.837 respectively).

Previous research by Lee (2004) investigated DIF detection with correlated dimensions of magnitude 0.4 and 0.6. Although Lee (2004) hypothesized that lower power would be associated with a higher correlation, no difference in detection rates was observed as the correlation between dimensions increased to 0.6. While the current study did observe a negative relationship between correlation and power, the trend was slight given the impact of other study variables on power performance. That is, although power decreased slightly as the correlation increased, the effect was almost negligible as DIF/DBF magnitude and sample size increased. For example, when the dimensions were correlated as high as 0.837, power rates between 96% and 100% were observed for all sample size conditions (except the 1800:200 ratio) involving DIF/DBF magnitude of at least .50 (regardless of angular item direction).

Item direction. The final variable of interest in this study was the degree to which the multidimensional items containing DIF measured the primary and secondary dimensions. Two item directions, 26.1° and 58.5°, were simulated to examine the effect of having higher discriminating power of items on the secondary dimension versus the primary dimension. Since a test that unintentionally includes multidimensional items (for which subgroups of examinees differ in mean ability on the secondary trait) is certainly not designed to measure a secondary or nuisance trait, it is reasonable to not expect those items to be a better measure of the secondary dimension than the first. However, in practice this may be more commonplace than previously thought. Mathematical story or word problems are item types that typically exhibit this behavior. Take for instance, the following set of multiple choice items (Reckase, 1985, p. 411) designed to measure mathematical ability:

9. $|-5|+|6|+(-5)+6=?$
   A. -22
   B. -10
   C. 2
   D. 10
   E. 12

34. Which line is parallel to *y = 3x +1* and intersects *y = 6x-1* on the *y*-axis?
   A. $y=3x-1$
   B. $y=2x-1$
   C. $y=\frac{1}{3}x-1$
   D. $y=\frac{1}{3}x+1$
   E. $y=\frac{1}{2}x-1$

20. A serving of a certain cereal, with milk, provides 35% of the potassium required daily by the average adult. If a serving of this cereal with milk contains 112

milligrams of potassium, how many milligrams of potassium does the average adult require each day?

A. 35

B. 39

C. 147

D. 320

E. 392

While Item 9 seems to measure only math ability, Items 34 and 20 appear to be measuring both math and reading ability with item 20 requiring more reading comprehension skills than Item 34. What if Item 20 has a higher discrimination value for the reading dimension than the math dimension? That is, if this item differentiates between proficient and non-proficient readers to a greater degree than it separates students with high and low math ability, then any estimate of math ability using this item would be inaccurate using traditional IRT methods. Thus, the inferences and conclusions reached about the mathematics abilities of examinees would not be valid. Assuming this can and does happen in real testing situations, what is the impact of multidimensional item discrimination on DIF/DBF detection? Using angular item directions of $\alpha = 26.1°$ and $\alpha = 58.5°$ (corresponding to discrimination pairs of $a_1 = 1.00$, $a_2 = 0.49$ and $a_1 = 0.30$, $a_2 = 0.49$, respectively) DBF detection rates were explored using SIBTEST.

Similar to the findings of Oshima and Miller (1992), this study observed an increase in power to detect DIF/DBF when the items in the multidimensional bundle had higher discrimination parameters for the secondary trait than for the first. Thus, as the items measured the nuisance dimension more than the primary dimension or target ability, reference and focal group differences in mean ability on the nuisance trait were more easily identified by SIBTEST. Oshima and Miller (1992) investigated item directions and power to detect differential functioning in *individual* items. However,

their conclusion that higher item directions (e.g., 45° and 60°) lead to better DIF detection than with lower angles (e.g., 15° and 30°) mirrors the findings in this study for differential functioning in an item *bundle*.

Type I Error Performance

Type I error was assessed on two levels – when items only measured one dimension and when multidimensionality was present in an item bundle. Therefore, a baseline was established from which SIBTEST's ability to control false flagging of non-DIF multidimensional items could be evaluated. Thus, a unidimensional test of 40 items was simulated and the last ten items were submitted to SIBTEST as a bundle so that the "baseline" Type I error rate could be determined. This was followed by an assessment of the SIBTEST Type I error rate when the ten-item bundle was composed of multidimensional items. For both the baseline condition and the conditions involving multidimensional items, SIBTEST adhered closely to the nominal alpha level of .05. The fact that inflated Type I Error was not observed in this study is a testament to SIBTEST's ability to separate multidimensionality from multidimensionality that leads to DIF/DBF. These findings confirm those from a previous DIF study by Oshima & Miller (1992) who reported small Type I error rates when mean ability on the secondary trait was equal for the reference and focal group.

Summary of the Effects of Manipulated Variables on DBF Detection

This study examined the influence of several variables on DBF detection with SIBTEST. Although each studied variable impacted power and Type I error, it was their

combined effect that provides the most information to guide test development and practitioners in the field. Differences in mean ability on the secondary trait heavily impacted the rate of accurate DBF detection with noted increases in power observed for higher angular item directions. As with most simulation studies that include sample size as a variable, the larger the overall sample size and the less unequal the ratio of reference to focal group members is, the greater the probability that true-DIF/DBF will be detected. Finally, the smaller the correlation is between the dimensions measured by a bundle of items on a test, the more power there will be to detect differential functioning in that bundle. These results provide insight into the ideal conditions for detecting individual DIF as well as amplified DIF exhibited by a group, or bundle, of items.

Comparison of the Present Findings to Russell (2005)

As previously described in Chapter 2, Russell's (2005) Monte Carlo study examined the impact of several variables on the DTF and DBF detection rates of DFIT and SIBTEST, respectively. The current study only reports results for SIBTEST but the process of simulating differential bundle functioning was the same and is thus comparable to Russell's investigation. Like Russell, this study created item bundles by including multidimensional items on a simulated test designed in theory to be unidimensional. Simulating reference group members to have a higher mean ability than the focal group on the nuisance secondary dimension, resulted in DIF for each of the multidimensional items, that when examined together produced differential bundle functioning. The results of both studies give support to the claim that DIF is caused by

"unintentional" multidimensionality for which groups differ in ability on the dimension not intended to be measured.

A comparison of the Russell (2005) study to the present findings highlights some similarities and differences. First, although the factors impacting DBF detection with SIBTEST varied in the two studies, a few of the variables were the same. For example, like the current study, Russell looked at the influence of dimensionality, sample size and nuisance ability differences (or magnitude of DIF) on DBF power and Type I error rates with SIBTEST.

However, Russell's findings for SIBTEST fell somewhat short of his expectations. For instance, in the study of Type I error, Russell hypothesized that SIBTEST would not exceed the nominal α of .01 when used to analyze multidimensional item bundles for which there were no ability differences on the nuisance trait. Also, the average "baseline" (i.e., when the test is comprised of only unidimensional items) DBF Type I error rate across all conditions was .353. These results would indicate that SIBTEST is incapable of avoiding false positives even when there is only one dimension being measured by a test. However, upon closer inspection it is evident that these errors increased greatly as target ability differences increased from moderate to large. In fact, in conditions with no primary ability differences, the average Type I error rate was .015 for conditions with the unidimensional bundle and .030 for conditions with the multidimensional bundle. On the other hand, the current study found that SIBTEST controlled Type I errors very well across all simulated conditions. In fact, the average "baseline" Type I error rate was .055 and the average error rate with two dimensions was .048 (for $\alpha = 26.1°$) and .058 (for $\alpha = 58.5°$) – all close to the nominal alpha level of .05.

With regard to power performance, this study manipulated several of the variables investigated by Russell. For example, both studies found SIBTEST to be more powerful in detecting DBF as sample size increased. However, the magnitude of this finding varied across the two studies. In Russell (2005) for instance, the average SIBTEST power rate for the largest sample size (N_{ref} = 1000 and N_{foc} = 1000) was .775 compared to a power rate of .675 for the smallest reference and focal group sizes ($N_{ref} = N_{foc}$ = 500). On the other hand, the current study observed slightly higher power rates for the sample size conditions involving 1000 reference and 1000 focal group members. That is, the average DBF power rates for this sample size level were .873 (for $\alpha = 26.1°$) and .976 (for $\alpha = 58.5°$). Furthermore, the SIBTEST power rate reported by Russell across all conditions was .725 compared with the average rate of .848 (for $\alpha = 26.1°$) and .955 (for $\alpha = 58.5°$) observed in the present study. It seems that the presence of target ability differences also negatively influenced power results in Russell's study. In Russell (2005) SIBTEST demonstrated an overall empirical power of .725, with the proportion of true-positives in a single condition ranging from .22 to 1.0. For conditions involving primary ability differences, SIBTEST performed best when impact was minimal.

The Type I error and power rate differences observed between Russell (2005) and the current study may be explained by differences in study design. First, Russell evaluated the Type I error and power performance of DFIT and SIBTEST at the alpha level of .01 which is more conservative than the level of significance (.05) used in this study. Russell explains early in his study that $\alpha = .01$ will be used as the criteria for rejecting the no-DIF hypothesis because it is the level recommended by Raju, van der Linden and Fleer (1995) when analyzing items with DFIT. However, Raju, et al's (1995)

rationale for using this significance level was based on a study by Fleer (1993) that found, after many replications with simulated data, that approximately 1% of non-DIF items were falsely flagged with DIF using a cutoff-value of .006 for the DFIT indices. Raju et al (1995) and others have since questioned the use of this somewhat arbitrary cutoff value and whether it or an empirically-determined cutoff value is ideal for DIF/DTF analyses.

Second, the number of replications per condition varied between the two studies. That is, compared with the 1000 replications per condition in the current study, Russell examined DTF and DBF performance using only 50 replications for each simulated condition. Although some research (e.g., Harwell, Stone, Hsu, & Kirisci, 1996) has supported the use of fewer replications with Monte Carlo studies, the probability of making Type I and Type II errors increases when minimal replications are used. With this in mind, it may be more important to examine fewer variables if it will increase the number of replications that can be conducted.

Although the amount of simulated DIF between the two studies (20% vs. 25%) is negligible, the difference in test length was not. As previously mentioned, Russell (2005) examined DBF as a function of the number of test items and observed a positive relationship between the two – longer tests yielded smaller Type I error rates. Perhaps there is an ideal combination of test length and percentage of DIF items that minimizes error and increases power. The results of the current study would indicate that minimal (but meaningful) amounts of DIF (e.g., magnitudes of 0.5 or greater) can be detected with great accuracy by SIBTEST when a reasonable proportion of items exhibit DIF (20% –

25%) and the test length (e.g., 40 items) maximizes the information that can be gathered about examinee ability on the latent trait.

One of the major differences in study design between this and Russell's study was the presence or absence of target ability differences in the investigation of DBF power and Type I error performance with SIBTEST. Russell observed inflated Type I errors and decreased power for conditions where reference and focal groups differed greatly on target, or primary, dimension. However, for comparable conditions, results from this study were similar to those of Russell. For instance, with no impact present, Type I error rates closely adhered to the nominal alpha level specified in each study. In Russell's study, however, Type I error rates dramatically increased when ability differences between the reference and focal group existed on the primary dimension, with the largest rates occurring with the largest ability difference.

With regard to power performance, since Russell (2005) used one set of discrimination parameters for his multidimensional items where the second dimension had a higher discrimination value than the first ($a_1 = 0.30$; $a_2 = 1.00$), it would be expected that his results would follow more closely with the power results in this study from conditions where $a_1 = 0.30$. Indeed, both studies found that with DIF magnitude of .5 or greater, power was close to 100%. The only exception to this occurred in Russell's study in conditions with large impact and 10 items. With moderate DIF, comparable sample sizes, and no impact, power was also similar for both studies, with rates also close to 100%. In both studies, SIBTEST was found to be more powerful in detecting DBF as sample size increased. While power was very high in most of this study's conditions, Russell explored the impact of primary ability differences among his conditions and

found that with a large difference, power was greatly reduced, particularly with shorter tests. The findings from both studies support the strength of SIBTEST in identifying DBF, particularly when the second dimension has a higher discrimination than the first, but Russell's study also shows a weakness of SIBTEST in detecting DBF when ability differences are present on the primary dimension.

Limitations and Directions for Future Research

The generalizability of these findings is limited by the specific simulated conditions chosen for the study. The available choices for selecting generating item parameters, sample size and sample size ratios, trait distributions for the reference and focal groups, correlations between the primary and nuisance dimensions, item directions, and dimensionality structure were wide and varied and thus not fully captured by this study. Here, unidimensional item parameters were based on realistic values used in previous research (Raju, van der Linden, Fleer, 1995), simulated DIF/DBF was unidirectional (i.e., favored the reference group) across all conditions, three levels of correlation were explored to represent weak, moderate and strong relationships between ability dimensions, item directions were limited to two levels – low angle and high angle, and items were simulated to measure only two dimensions.

These limitations suggest directions for future research. First, it would be interesting to observe the effects that other variables, such as percent of test items that form a DIF bundle, have on DIF detection. Since previous studies suggest that DIF/DBF caused by multidimensionality is more likely to be detected if the test is essentially unidimensional (Oshima & Miller, 1992) assessing the performance of SIBTEST as the hypothesis of essential unidimensionality becomes less tenable could result in

information related to the minimal degree of unidimensional structure that is required to detect DIF/DBF with adequate power while controlling Type I error.

Second, this study's use of identical correlations between the first and second dimension for the reference and focal groups may not reflect real-test conditions. Perhaps the strength of correlation between multiple abilities (such as reading and math) is not the same for two subgroups of examinees (such as boys and girls). For example, in the case of math word problems, it may be possible that for boys' reading ability is more highly correlated with math ability than it is for girls. Results from the current study illustrate the adverse effects that small sample sizes and large ratios of reference to focal group have on DIF detection. Recall, that DIF detection was poorest across all magnitudes of DIF/DBF when the reference group (size of 1800) was nine times larger than the focal group (size of 200). This could have a significant effect on the ability of DIF/DBF methods to accurately identify DIF items and bundles for subgroups of similar sizes with real test data. However, to test this hypothesis and others, future research should consider simulating different correlations between the primary and secondary dimension for the reference and focal groups.

A third direction for future research could be to include bundle DIF that favors the focal group in addition to that which advantages the reference group. By balancing the direction of differential bundle functioning, DIF amplification and DIF cancellation can be explored. It is conceivable that some items on a test would favor one subgroup (e.g., reading comprehension items in which the context is football might advantage male examinees) while other items favor another subgroup (e.g., those with context involving interpersonal relationships might advantage female examinees). It was interesting to

examine the ability of SIBTEST to detect the accumulation of DIF resulting in significant DBF against a particular subgroup. However, assessing DIF cancellation (i.e., when DIF/DBF that favors the reference group for one set of items is cancelled by DIF/DBF in another set of items that favors the focal group) with SIBTEST would also be very informative.

Another suggestion is to examine the interaction between item difficulty and item discrimination and its effect on DBF detection with SIBTEST. Are easier items that measure the secondary trait more than the first, but which contain no bias, more likely to be identified as DIF items or bundles? Other questions regarding the ability of SIBTEST to control the inflation of Type I error while demonstrating sufficient power to detect true-DIF/DBF could be answered by varying item difficulty and discrimination parameters.

Finally, another limitation of this study was the exclusion of other DBF methods. SIBTEST demonstrated excellent power and Type I error control when used to examine differential bundle functioning under the specific conditions simulated in this study. However, SIBTEST is not the only procedure capable of analyzing bundles of items for DIF amplification. The other method designed to do this is known as DFIT (Differential Functioning of Items and Tests; Raju, van der Linden, and Fleer, 1992) and it uses a three-step procedure to test for differential item/bundle/test functioning (see Appendix A for more details on DFIT). However, the separate phases of analysis with DFIT make it less attractive for use in a simulation study. Although Russell (2005) used DFIT to analyze differential test functioning with simulated items, the results were less than stellar. For example, in Russell's study only 50 replications were conducted for each

condition (presumably because of the large amount of computing time required to analyze a single condition). Russell also observed very low power rates (i.e., average of .425 across all conditions) and largely inflated Type I error rates (i.e., average of .227 across all conditions) with DFIT. When compared with SIBTEST, power was much lower and Type I error rates were higher with DFIT. These unfavorable findings for DFIT highlight some of the disadvantages to using this method. On the other hand, the version of DFIT available at the time of Russell's study made use of an arbitrary cutoff value of .006 for the statistical significance test. A newer version now exists that establishes an empirical cutoff using the Item Parameter Replication (IPR) method. The IPR method is discussed in more detail in Appendix A.

The current study originally set out to examine DBF analysis with SIBTEST *and* DFIT. Since these are the only two procedures capable of analyzing item bundles for DIF amplification, it was logical to try and examine them together. However the overall goal of this study was to investigate conditions that impact DBF detection *and* further explore the potential of DBF analysis in identifying the source of differential item functioning. Hence, once it was determined that the current version of DFIT is unable to realistically evaluate DBF in a simulation study with the minimal number of replications required to obtain stable results, it was eliminated from further analysis. However, details on the efforts to include DFIT in the current study may provide useful information to guide future research. Therefore, a discussion of the DFIT experience follows (a detailed description of the DFIT procedure and indices can be found in Appendix A). Note that the following is an attempt to describe the challenges that arose due to IRT calibration and item parameter linking as separate issues from those associated with DBF

analysis with the DFIT program. However, since DFIT relies on IRT parameter estimates and linking coefficients from external programs, its performance will invariably be dependent on the quality of the item response data and the external programs (e.g., IRT calibration and linking programs) used to analyze those data.

In the process of preparing the data for DFIT analysis many complications arose. First, the current procedure for analyzing items, bundles and tests with DFIT makes use of the Item Parameter Replication method (Oshima, Raju & Nanda, 2006) to produce cutoff values for DFIT based on particular datasets and individual items. While this is very helpful to practitioners in the field, it presents some challenges to simulation research with DFIT. For example, DFIT uses item and ability estimates from unidimensional IRT programs such as BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2003) and PARSCALE (Muraki & Bock, 1997). When multidimensional items in the present study were submitted to BILOG-MG3 and PARSCALE, lack of convergence was a major problem. On average, roughly four out of 10 replications failed to yield acceptable parameter estimates. While the exact reason for lack of convergence is unknown, it is likely that the degree of multidimensionality present in the data was too high for the simulated test to be considered essentially unidimensional. As a result, the two IRT programs (BILOG and PARSCALE) were unable to produce stable estimates of item parameters. Athough no test of essentially dimensionality was conducted, it is suggested that future studies investigate degree of dimensionality before attempting to estimate item parameters with unidimensional IRT calibration programs.

The results of parameter estimation in this study contradict those obtained by Kirisci, Hsu and Yu (2001) in their investigation of the robustness of unidimensional IRT

calibration programs (specifically, BILOG, MULTILOG and XCALIBRE). In their study, Kirisici, Hsu and Yu concluded that BILOG was robust to multidimensionality when root mean squared error (RMSE; see equation 10) of approximation was used to compare true parameter values to estimated values.

$$RMSE_X = \sqrt{\frac{\sum_{j}^{n} (\hat{X}_{ij} - X_i)^2}{n}}$$

(10)

where

i and *j*, respectively, represent items and replications,
n is the total number of replications, and
$\hat{X}$ is the estimate of parameter *X (a, b, or c)*.

Using the results of previous research (see Ansley & Forsyth, 1985; De Ayala, 1994; Wang, 1986) that suggested unidimensional item and ability estimates approach the average of their multidimensional parameter values, Kirisci, Hsu and Yu used the average of the true parameter values for the three simulated latent traits as the value for *X* in equation 10. The study conducted 10 replications and calculated the mean RMSE for a set of conditions involving a three-dimensional test structure and for a set of conditions representing a unidimensional test structure. Therefore, the calibration programs were categorized as robust against the violation of unidimensionality if there was no statistically significant difference between the three-dimensional mean RMSE and the unidimensional mean RMSE. Kirisci, Hsu and Yu (2001) do not suggest that essential unidimensionality is a pre-requisite for using these IRT calibration programs. Hence, this study did not make the assumption that essentially unidimensionality was a factor in the robust estimates obtained from using BILOG with multidimensional data.

In the current study preliminary results with BILOG were unstable, yielding (in a most random fashion) unidimensional estimates of discrimination that approached either the discrimination parameter value for one of the simulated dimensions or the average of the two discrimination parameters. By the standards established for this study, these inconsistencies were unacceptable and it was decided that BILOG would not be a feasible option for item calibration. Subsequently, PARSCALE was consulted but abandoned after similar behavior was observed. The reason for these unfavorable results is unknown but one explanation may be that the amount of variance accounted for by the target dimension was neither fixed nor varied in the present study. It was not examined at all. That is, Reckase's (1979) suggestion that stable item parameter estimates using these programs are more likely to occur when the dominant dimension accounts for at least 20% of variance in test scores, was not evaluated in this study. The influence of target ability on test performance was not specifically investigated and therefore its contribution to the instability of BILOG and PARSCALE estimates of multidimensional discrimination parameters was not determined.

Another difference between Kirisci, Hsu and Yu (2001) and this study is that the previous work suggested use of unidimensional IRT programs such as BILOG when there are several highly correlated (i.e., $r > .4$) dimensions of approximately equal dominance and for which the correlations are approximately equal (in magnitude) across dimension pairs. Their study generated data for a three-dimensional test, yielding three pairs of correlations (i.e., $r_{\theta_1\theta_2}, r_{\theta_1\theta_3}, r_{\theta_2\theta_3}$). However, the current study only examined two dimensions, yielding one pair of correlated dimensions ($r_{\theta\eta}$). Perhaps, this factor disqualified BILOG (and all unidimensional IRT programs) for use in parameter

estimation for this study. Since the primary focus of this investigation was DBF detection rates – not parameter recovery – the decision to exclude the DFIT program (since its performance would certainly be influenced by the quality of parameter estimates from BILOG) was deemed appropriate.

Leaving the issues with BILOG, the use of DFIT also requires that parameter estimates be submitted to an external program such as EQUATE (Baker, 1993) or IPLINK (Lee & Oshima, 1996) to calculate linking coefficients that will convert estimates to a common scale for the reference and focal group since they are calibrated separately in BILOG-MG3. There are complications involved with using either of these linking programs. For example, EQUATE outputs linking coefficients to the computer screen instead of to a file where the linking coefficients could be retrieved for later use by DFIT. While this is not a concern for real test data that will be analyzed one time, it was certainly an obstacle for the current simulation study that attempted to analyze 104 conditions with 1000 replications per condition. The other linking program, IPLINK, suffered from the opposite limitation; although output is directed to a file, the program cannot be called from DOS.

Finally, DFIT is limited by current computer processing time. To run 1000 replications of one condition would require approximately 250 hours of non-stop processing. When multiplied by 104 conditions, the total time required is almost three years! The reason is that DIF/DBF is analyzed within DIFCUT, a SAS program that creates a null distribution of 1000 item parameter pairs (for the reference group and focal group) using the estimates from BILOG-MG3. From this empirical distribution of item parameters, cutoff values for DIF/DBF are computed at four levels of significance (.10,

.05, .01, and .001). As a result, with DFIT it is not currently feasible to run even the minimal number of replications that are currently acceptable for Monte Carlo research.

Simply put, extensive research and software development may be required before DFIT using the new Item Parameter Replication method can realistically be used in simulation research. Among the necessary changes are: the need for DOS-based linking programs that direct output to files instead of the computer screen, reduced processing time within the DIFCUT program, and better convergence during parameter estimation when using IRT calibration programs such as BILOG-MG3 with multidimensional data (although this may be mediated by simulating essentially unidimensional tests). The current version of DFIT is written in SAS which appears to hinder processing time. There is a FORTRAN version of DFIT that is scheduled for release in Summer 2007. The expectation is that DFIT will run much faster and that may be an ideal time to examine DFIT for DBF analysis with essentially unidimensional data and for a minimal (but acceptable) number of replications.

Implications and Conclusions

One advantage of DBF analysis over traditional DIF analysis is the potential for DIF to be amplified and increase the probability of detection. This study demonstrated the use of SIBTEST to detect DIF in multidimensional item bundles and the findings illustrate the usefulness of Shealy-Stout's (1993a) Multidimensional Model for DIF (MMD). The ability to identify "suspect" items or item bundles thought to measure a secondary trait in addition to the target ability can lead to substantive DIF/DBF hypotheses *a priori* and hence provide information on the potential cause of differential item, bundle, or test functioning. As with all simulation studies, knowledge of the "truth"

provides an opportunity to examine the degree to which a statistical procedure such as SIBTEST is able to recover the "truth". Similarly, the ability to identify which test items measure multiple dimensions increases not only the power to locate DIF/DBF but also increases the accuracy of DIF hypotheses used to explain the cause of DIF/DBF.

Of course, the MMD assumes that a meaningful secondary dimension can be identified prior to statistical DIF/DBF analysis and that there exists subgroups of examinees that differ in ability on that dimension. This study suggested several organizing principles that can be used to develop DIF/DBF hypotheses before statistically analyzing items or bundles for DIF/DBF. Whether using test specifications or a statistical dimensionality tool such as DIMTEST, the assignment of items to bundles is basically an *a priori* assessment of dimensionality.

The examples used in this study highlight an issue that is becoming more pervasive in testing and which complicates use of unidimensional IRT models to estimate examinee ability on the target dimension. However, since it is still popular to use unidimensional models in the presence of multidimensional items, the importance of knowing which items are measuring more than one dimension cannot be overstated. Hopefully this study will help advance the use of DBF analysis to identify potentially biased items and develop fair tests for all examinees.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*, 113-127.

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.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*, 37-48.

Baker, F. B. (1993). EQUATE 2.0: A Computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement*, *17*(1), 20.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *A taxonomy of educational objectives, handbook 1: The cognitive domain*. New York: David McKay.

Bolt, D. M. (2002). A monte carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113-141.

Bond, L. (1993). Comments on the O'Neill and McPeek paper. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 277-279). Hillsdale, NJ:Erlbaum.

Boughton, K., Gierl, M. J., & Khaliq, S. N. (2000). Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding

differential performance. *Paper presented at the annual meeting of the Canadian Society for Studies in Education*. Edmonton, Alberta, Cananda.

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

Chamblee, M. C. (1998). A monte carlo investigation of conditions that impact type I error rates of DFIT (Doctoral dissertation, Georgia State University, 1990). *Dissertation Abstracts International*, (UMI No. 9910356).

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

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. Applied Psychological Measurement, *18*, 155-170.

Doody, E. N. (1985, April). *Examining the effects of multidimensional data on ability and item parameter estimation using the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465-484.

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

Embretson, S.E., & Reise, S. P. (2000). *Item response theory for psychologists*.

Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, *3*, 347-360.

Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias (Doctoral dissertation, Illinois Institute of Technology, 1993). Dissertation Abstracts International, 54-04, 2266B.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-DeLisi, A.V., Morley, M., Cahalan, C. (2000). Gender differences in advanced mathematical problem-solving. *Journal of Experimental Child Psychology*, *75*, 165-190.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematical test with Bloom's taxonomy. *Journal of Educational Research*, *91*(1), 26-32.

Gierl, M.J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analysis to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, 20, 26-36.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of*

Educational Measurement, *38*(2), 164-187.

Gierl, M. J., & Bolt, D. (2003, April). *Implications of the multidimensionality-based DIF analysis framework for selecting a matching and studied subtest*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT*. (No. 2005-11). New York, NY: The College Board.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist*, *52*, 1091-1102.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. (Vol. 2). Newbury Park, CA: Sage Publications.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations on the unidimensionality assumption. *Journal of Educational Statistics*, *11*, 91-115.

Harwell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. Applied Psychological Measurement, *20*(2), 101-125.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*(2), 146-162.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine, & J. Rost (Eds), *Latent trait and latent class models* (pp. 263-274). New York: Plenum Press.

Kromrey, J.D., Parshall, C. G., Chason, W. M., & Yi, Q. (1999). *Generating item responses based on multidimensional item response theory*. Retrieved July 25,

2006 from http://www2.sas.com/proceedings/sugi24/Posters/p241-24.pdf.

Lee, Y. (2004). The impact of a multidimensional item on differential item functioning (DIF). *Dissertation Abstracts International,* (UMI No. 3139494).

Lee, K., & Oshima, T. C. (1996). IPLINK: Multidimensional and undimensional item parameter linking in item response theory. *Applied Psychological Measurement*, *20*(3), 230.

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

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91-100.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, *16*(4), 381-388.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Muraki, E. & Bock, R. D. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data. Chicago: Scientific Software International.

Nanda, A. O., Oshima, T. C., Gagne, P. (2006). DIFCUT: A SAS/IML program for conducting significance tests for differential functioning of items and tests. *Applied Psychological Measurement*, *30*(2), 150-151.

Nandakumar, R. (1991). Traditional versus essential unidimensionality. *Journal of Educational Measurement*, *28*, 99-117.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *30*, 293-311.

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.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Erlbaum.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement*, *16*, 237-248.

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*, 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) framework. *Journal of Educational Measurement*, *43*(1), 1-17.

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*(4), 353-368.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412.

Reckase, M.D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *215*, 193-203.

Reckase, M. D., & McKinley, R. L. (1991). The Discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*(4), 361-373.

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

Roussos, L. A., & Stout, W. F. (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*, 215-230.

Russell, S. S. (2005). Estimates of Type I Error and power for indices of differential bundle and test functioning. *Dissertation Abstracts International,* (UMI No. 3175804).

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph* Supplement, 34(4), 100-114.

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

Shealy, R., & Stout, W. F. (1993b). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential*

item functioning (pp. 197-239). Hillsdale, NJ: Erlbaum.

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

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

Stout, W. F. (1987). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika*, *55*, 293-326.

Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching*, *30*, 3-19.

Wang, M. M. (1986). Fitting a unidimensional model to multidimensional item response data (ONR Report No. 042286). Iowa City, IA: University of Iowa.

Whitaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, *27*(4), 299-300.

Yoonsun, L. (2004). The Impact of a Multidimensional Item on Differential Item Functioning (DIF). *Dissertation Abstracts International*, (UMI No. 3139494).

Zimowski, M,F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG3 [computer program]. Chicago, IL: Scientific Software.

APPENDIXES

APPENDIX A

DFIT

The first method capable of analyzing bundles of items was developed by Raju, van der Linden, and Fleer (1992; 1995) for detecting differential functioning of items and tests (DFIT). The DFIT framework uses three indices to measure differential performance between subpopulations of interest: 1) NCDIF (noncompensatory differential item functioning); 2) CDIF (compensatory differential item functioning); and 3) DTF (differential test functioning). As with other DIF indices, NCDIF assumes that all items except the studied item are DIF-free. However, the CDIF index does not make this assumption. Rather, its values total the overall differential functioning of the test (DTF), allowing the DIF of individual items to either add up or cancel each other out. This is generally referred to as DIF amplification and cancellation. The compensating ability of the CDIF index aids in test design by estimating the net effect of removing items from a test (Raju, 1995; Chamblee, 1998; Oshima, 1998).

Differential Test Functioning. In Item Response Theory (IRT), an examinee's expected proportion correct (EPC), or true score (T_j) on a unidimensional dichotomously scored test can be expressed as

$$T_j = \sum_{i=1}^{n} P_i\left(\theta_j\right), \tag{1}$$

where $P_i(\theta_j)$ represents the probability of success for examinee *j* with trait level (e.g., ability) θ on item *i*. P_i can be represented either by a one-, two-, or three parameter logistic model or the normal ogive model (Raju, 1995; Lord, 1980) with item parameters *a* (discrimination), *b* (difficulty), and *c* (pseudo-guessing). In the DFIT framework, each focal group examinee has a true score (T_{jF}) that is compared to the true score he or she would have if he/she were a member of the reference group (T_{jR}). If an examinee's (T_{jF}) and (T_{jR}) scores are equal, then the examinee's true score is independent of group membership. That is, differential test functioning (DTF) is indicated by differences between (T_{jF}) and (T_{jR}). As differences between (T_{jF}) and (T_{jR}) increase, so does the amount of DTF. At the examinee level, DTF is defined as $(T_{jF} - T_{jR})^2$. Hence, DTF across examinees, is defined as

$$DTF = \varepsilon\left(T_{jF} - T_{jR}\right)^2 \tag{2}$$

where the expectation ε can be taken over the focal or reference group.

Bundle CDIF. Raju's DFIT framework also includes a compensatory DIF (CDIF) index, the values of which add up to the total DTF value for a test. CDIF considers other items in the calculation of DIF, and its extension bundle CDIF, does the same with bundles of items. The CDIF index (see Equation 3) allows test developers to remove items or item bundles that contribute significantly to differential test functioning.

$$\sum_{i=1}^{n} CDIF_i = DTF \tag{3}$$

Bundle NCDIF. Bundle NCDIF, an extension of NCDIF, examines whether two matched ability groups respond differently to a bundle of items. DBF analysis within the DFIT framework involves calculating a DTF value for each bundle of items. However,

in bundle NCDIF no other bundles are considered in calculating DTF for the bundle under study. That is, NCDIF and bundle NCDIF assume that all but the studied item/bundle is DIF-free. Additionally, ability parameter estimates are based on all items in a test or bundle and are only obtained once. Hence the calculation of bundle NCDIF differs from bundle CDIF. Since the DTF value depends on the number of items in a bundle, bundles containing different numbers of items should not be directly compared. Instead the average DTF for each bundle (a.k.a. Bundle NCDIF) obtained by dividing the total DTF by the number of items in the bundle can be used to compare the amount of DBF in multiple bundles. A formulaic definition of NCDIF follows in Equation 4.

$$NCDIF_i = \sigma_{d_i}{}^2 + \mu^2{}_{d_i}, \qquad (4)$$

where $\sigma_{d_i}{}^2$ is the variance in item probabilities for item *i* and μ_{d_i} is the mean difference in item probabilities for item *i*.

Previously, an empirically-determined cutoff value (.006) was used with the DTF and NCDIF indices. Raju, van der Linden, and Fleer (1995) raised questions regarding the adequacy of this value and suggested that future research focus on establishing precise critical values for different alpha levels. In 2006, Oshima, Raju and Nanda introduced the Item Parameter Replication (IPR) method that determines cutoff values for DFIT based on particular datasets and individual items. The IPR algorithm is completed using the SAS (SAS Institute, 2002) program DIFCUT which uses a single file of item parameters to simulate many item parameter pairs resulting in the distribution of DIF/DTF indices under the condition of no DIF/DTF (Nanda, Oshima, & Gagne, 2006).

Figure 2 illustrates the three-step procedure DFIT uses to test for differential item/test functioning. First, item parameters (*a, b, c*) and person parameters (θ) are

estimated using BILOG-MG3 (Mislevy & Bock, 1990), followed by an assessment of model-data fit. Second, reference and focal group parameters are placed on a common scale using the program IPLINK (Lee & Oshima, 1996). This is required when parameters for each group are estimated separately. The third and final step is calculation of DBF values for each item bundle. DFIT includes NCDIF and DTF indices for differential item/test functioning, respectively. Therefore, a single bundle (consisting of multidimensional items) is submitted to DFIT and the DTF index is used to quantify the amount of differential bundle functioning. This step involves submitting three BILOG-MG3 output files and the linking coefficient file from step two to calculate NCDIF and DTF cutoff scores for the multidimensional bundles. The output is a list of NCDIF and DTF values and their significance levels.

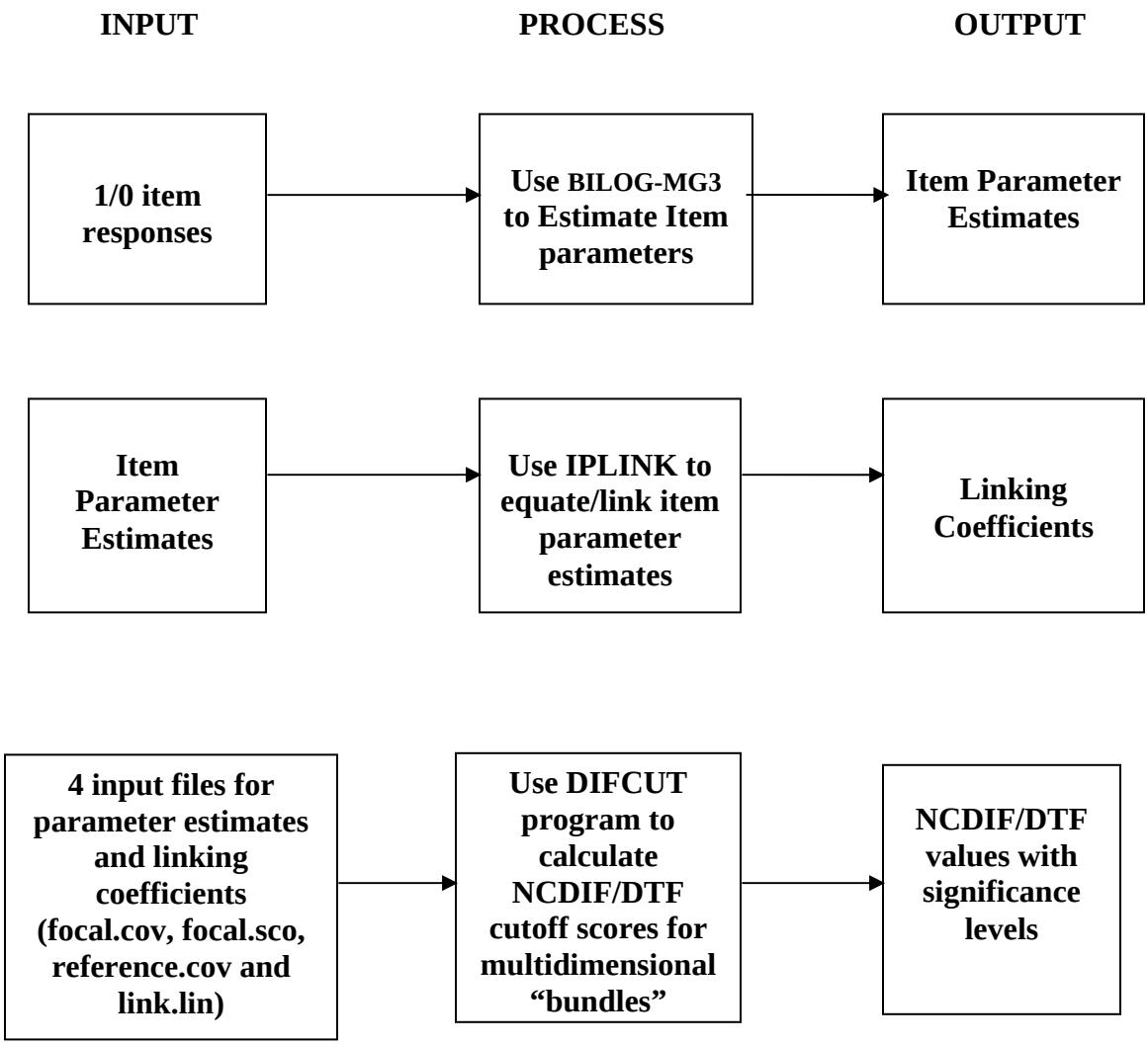

Figure 2. A Graphical Representation of the three-step DFIT procedure.

APPENDIX B

Table 7

RefParam.txt. Item Parameters Used to Generate Unidimensional Item Response Data for the Reference Group and Focal Group (before DIF was embedded).

a	*b*	*c*
0.55	0.00	0.2
0.55	0.00	0.2
0.73	-1.04	0.2
0.73	-1.04	0.2
0.73	0.00	0.2
0.73	0.00	0.2
0.73	0.00	0.2
0.73	0.00	0.2
0.73	0.00	0.2
0.73	1.04	0.2
0.73	1.04	0.2
1.00	-1.96	0.2
1.00	-1.96	0.2
1.00	-1.04	0.2
1.00	-1.04	0.2
1.00	-1.04	0.2
1.00	-1.04	0.2
1.00	0.00	0.2
1.00	0.00	0.2
1.00	0.00	0.2
1.00	0.00	0.2
1.00	0.00	0.2
1.00	0.00	0.2
1.00	0.00	0.2
1.00	0.00	0.2
1.00	1.04	0.2
1.00	1.04	0.2
1.00	1.04	0.2
1.00	1.04	0.2
1.00	1.96	0.2
1.00	1.96	0.2

Table 8

GenMirt Param.txt. Item Parameters Used to Generate Multidimensional Item Response Data for the Reference Group and Focal Group (before DIF was embedded for α = 26.1°).

a_1	a_2	b	c
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2
1.00	0.49	0.0	0.2

Table 9

GenMirt_Param_2.txt. *Item Parameters Used to Generate Multidimensional Item Response Data for the Reference Group and Focal Group (before DIF was embedded for* $\alpha = 58.5°$.

a_1	a_2	b	c
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2
0.30	0.49	0.0	0.2

APPENDIX C

SAS program

Uni_Multi.sas

```
Options ls= 182  ps=32767  pageno=1  formdlim='=';

%MACRO TR;


proc printto  LOG='c:\SIBdat\Logstuff.txt' NEW;


proc printto print='C:\SIBdat\POWER\2500_3_25_Outputstuff.txt' NEW;


%DO rep = 1 %TO 1000;


*2500_3_25;

%let r = 0.316;
%let dn = 0.25;
%let sampref = 2500;
%let sampfoc = 2500;
%let seedref = 62733;
%let seedfoc = 58015;
%let uni_i=30;
%let multi_i=10;




/* Generate uni_i unidimesional items for the reference group */
DATA D1;
  INFILE 'C:\SIBTEST\Refparam.txt' MISSOVER;
  INPUT A B C;
run;
%INCLUDE 'C:\SAS class\project\refTRGEN.SAS';
```

```
*%INCLUDE 'C:\dissertations\terris\refTRGEN.SAS';
%refTRGEN(DIST= 'NORMAL', MODEL=L3, DATA=D1, OUT=D3, NI=&uni_i,
NE=&sampref);

/* Generate uni_i unidimesional items for the focal group */
DATA E1;
  INFILE 'C:\SIBTEST\Refparam.txt' MISSOVER; *use same parametes as
REFERENCE group since this is theta dimension;
  INPUT A B C;
run;
%INCLUDE 'C:\SAS class\project\focTRGEN.SAS';
*%INCLUDE 'C:\dissertations\terris\focTRGEN.SAS';
%focTRGEN(DIST= 'NORMAL', MODEL=L3, DATA=E1, OUT=E3, NI=&uni_i,
/*NE=&sampfoc*/NE=&sampfoc);

data params;
  infile 'c:\SIBTEST\GenMIRT param.txt' lrecl=124 missover;
  input itemnum a1 a2 b c;
run;

proc iml;
items=&uni_i+&multi_i;

use d3; read all into ref_uni;
use e3; read all into foc_uni;
megamtx=repeat(0,&sampref+&sampfoc,items);
refmtx=repeat(0,&sampref,items);
focmtx=repeat(0,&sampfoc,items);
do i = 1 to &uni_i; * This makes the (sampref+sampfoc) x uni_i matrix of
unidimensional item responses;
  do n = 1 to &sampref;
    megamtx[n,i]=ref_uni[n,1+i];
  end;
  do n = 1 to &sampfoc;
    megamtx[&sampref+n,i]=foc_uni[n,1+i];
  end;
end;

use params;
  read all var {a1 a2} into popa;
  read all var {b} into popb;
  read all var {c} into popc;;
nitms = nrow(popa);
/*****************************************************************
  Define the subroutine to analyze each examinee response vector.
*****************************************************************/
```

```
start irtscore (nitems, simulees, idn2, rrv, popa, popb, popc, score);
  factnorm=PROBNORM(popb+(popa*simulees));
  prob_i = (popc+((1-popc)#factnorm))`;
/*****************************************************************
  The following yields the score vector (1,0)
*****************************************************************/
  score = prob_i > rrv;
finish;

/*****************************************************************
  This loop generates two theta values for each examinee
  and a set of NITMS random numbers.
*****************************************************************/

do n = 1 to &sampref;
  seed1 = &seedref + &rep*n;
/*****************************************************************
  Generation of theta (and eta) values from N(0,1) distribution
*****************************************************************/
  sim1 = rannor(seed1);  *my note: this generates a single value;
  sim2 = (&r*sim1+sqrt(1-&r**2)*rannor(seed1*2))+&dn;

  simulees = sim1//sim2;
/*****************************************************************
  Generation of uniform random numbers for each person
  and each test item.  These are used to determine item
  response correctness.
*****************************************************************/
  rrv = J(1,nitms,0);
  do k = 1 to nitms;
    rrv[1,k] = RANUNI(seed1+2*&rep);
  end;
/*****************************************************************
  Call the scoring subroutine
*****************************************************************/
  run irtscore (nitms, simulees, idn2, rrv, popa, popb, popc, score);

  do i = 1 to &multi_i;
    megamtx[n,&uni_i+i]=score[1,i];
  end;
end; *Ends sampref loop;
do n = 1 to &sampfoc;
  seed1 = &seedfoc + &rep*n;
  sim1 = rannor(seed1);
  sim2 = (&r*sim1+sqrt(1-&r**2)*rannor(seed1*2));
  simulees = sim1//sim2;
```

```
  rrv = J(1,nitms,0);
  do k = 1 to nitms;
    rrv[1,k] = RANUNI(seed1+2*&rep);
  end;
  run irtscore (nitms, simulees, idn2, rrv, popa, popb, popc, score);
  do i = 1 to &multi_i;
    megamtx[&sampref+n,&uni_i+i]=score[1,i];
  end;
end; *Ends sampfoc loop;
*print megamtx;
do i = 1 to items;
  do n = 1 to &sampref;
    refmtx[n,i]=megamtx[n,i];
  end;
end;
do i = 1 to items;
  do n = 1 to &sampfoc;
    focmtx[n,i]=megamtx[&sampref+n,i];
  end;
end;

create ref from refmtx;
append from refmtx;

create foc from focmtx;
append from focmtx;

quit;

data ref1; set ref;
file 'c:\um_ref.txt';
put col1-col40;

data foc1; set foc;
file 'c:\um_foc.txt';
put col1-col40;

proc iml;
start system(command);
  call push(" x '",command,"'; resume;");
  pause;
  finish;
run system('C:\sibtest\runsibt');

quit;
```

```
data sibtest;
infile "c:\SIBdat\2500_3_25.out";
input
///////////////
///////////////
///////////////
///////////////
///////////////
///////////////
///////////////
///////////////
///////////////
//////////////

@31 buni 6.3  @48 pvalue 5.3;

proc append base=final data=sibtest;
run;

%END;
%MEND TR;

%TR;

data tally; set final;
file 'c:\SIBdat\2500_3_25_results.txt';
put @1 buni 6.3  @8 pvalue 5.3;
if pvalue < .05 then tally =1;
          else tally = 0;
run;
 title '2500_3_25';
proc freq;
tables tally; run
```