THE EVALUATION OF NEW CRITERIA FOR POLYTOMOUS DIF IN THE DFIT FRAMEWORK

BY

KRISTEN A. FORTMANN-JOHNSON

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Psychology
in the Graduate College of the
Illinois Institute of Technology

Approved_____
Advisor

Chicago, IL
December 2007

ACKNOWLEDGEMENT

I would like to acknowledge a number of people without whom I could not have completed this process. First and foremost, I want to thank my advisor Dr. Scott Morris. He has been generous in sharing both his time and wisdom. I am grateful for his guidance and the patience he has shown in answering my countless questions. He has been a supportive mentor throughout my graduate studies, and I learned more through this process than I ever could have imagined.

I would also like to acknowledge the contribution of my former advisor Dr. Nambury Raju, who passed away during the conceptualization of this project. He shared with me his knowledge and passion for this topic and opened the door for me to pursue this line of research. I take great pride in completing the work we began together.

In addition, I would like to thank the members of my dissertation committee: Dr. Alan Mead, Dr. Annette Towler, Dr. David Arditi, and Dr. T. Chris Oshima. I am especially grateful to Dr. Alan Mead for his input and assistance in developing my methodology and to Dr. T. Chris Oshima for making herself available to me after the passing of Dr. Raju. Her input in this process has been invaluable.

Last but certainly not least, I want to thank my family. I am blessed each day by their love and support. I am especially grateful to my mom, Clara Fortmann, who has listened patiently and found ways to comfort me in times of stress. I am also grateful to my best friend and husband Ed. He encouraged me to pursue this degree and has been my rock throughout the ups and downs of this process. I share this accomplishment with him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Using a Monte Carlo research design, this study examined the efficacy of the item parameter replication (IPR) method (Oshima, Raju, & Nanda, 2006) for determining cutoff values for polytomous items within the differential functioning of items and tests (DFIT) framework (Raju, van der Linden, & Fleer, 1995). It was hypothesized IPR-based cutoffs would be more likely to detect differential item functioning (DIF) than previously recommended fixed cutoffs and would have false positive rates close to the nominal significance level. Results supported the efficacy of the IPR method. Further, the accuracy of DIF detection was compared between the IPR method and likelihood ratio (LR) test (Thissen, Steinberg, & Wainer, 1988). Across a number of conditions and items studied these methods demonstrated comparable power and Type 1 error rates. Differences between methods were observed in conditions with non-uniform DIF, where the LR test demonstrated more power to detect non-uniform DIF of small magnitudes. Sample size, focal group ability distribution, proportion of test-wide DIF, and direction of DIF had minimal effect on DIF detection across methods.

CHAPTER 1

INTRODUCTION

In recent decades, the use of standardized tests and questionnaires has grown tremendously within the business community. Personnel administrators rely heavily on the results of tests and questionnaires to make hiring and promotion decisions, to assess organizational attitudes such as job satisfaction, and to identify the need for organizational development initiatives. This often results in making comparisons on the ability or attitude of interest across groups (e.g., business units, work sites or demographic groups).

Using the information obtained from such tools in this way, however, assumes the test or questionnaire is an equally accurate measure of the ability or attitude of interest across groups. In other words, it is assumed there is measurement equivalence across groups and therefore any observed score differences reflect true group differences. For example, differences in measured ability between ethnic groups on a pre-employment test are assumed to reflect true differences in ability. What happens, however, when we do not have equivalent measurement?

When a test or questionnaire lacks measurement equivalence, the meaning of score differences becomes unclear because they reflect not only meaningful group differences but also systematic measurement error (Drasgow & Kanfer, 1985). Systematic measurement errors on a single item are referred to as differential item functioning (DIF). Differential test functioning (DTF) refers to systematic measurement errors of an entire test or scale. When DIF or DTF is present, inappropriate conclusions will be made potentially resulting in implementing unnecessary, costly organizational

interventions and/or making legally questionable decisions.

For instance, consider the use of a pre-employment cognitive ability test where observed scores are consistently lower for African American applicants than for Caucasian applicants. Based on the score differences alone, one would conclude that African American applicants have lower ability and are therefore less qualified than their Caucasian counterparts. Logically, one would then be more likely to hire Caucasian applicants. If DIF or DTF is present, however, score differences do not reflect true differences in cognitive ability but rather are the result of measurement error. In other words, an applicant's score on this pre-employment test is a function of both ability and group membership, and so two equally qualified applicants may not receive the same score. Any conclusions made regarding group differences in ability are now erroneous, and perhaps more importantly, will unnecessarily result in adverse impact in the hiring process. It is to a practitioner's advantage to avoid such circumstances. Therefore, it is of vital importance to assess measurement tools for DIF/DTF before use.

The differential functioning of items and tests (DFIT) framework (Raju, van der Linden, & Fleer, 1995) is an item response theory (IRT) based procedure for assessing DIF and DTF. While this framework has been shown to be an effective mechanism for detecting DIF/DTF (e.g., Flowers, Oshima, & Raju, 1999; Oshima, Raju, & Flowers, 1997; Raju et al.), past research has also indicated a need for caution in generalizing DFIT criteria across testing situations. Recently, a method to easily derive study-based criteria for DIF detection within the DFIT framework has been proposed (Oshima, Raju, & Nanda, 2006), and the current study aims to assess the efficacy of this new methodology within the DFIT framework.

The DFIT framework offers two indices for assessing DIF, noncompensatory differential item functioning (NCDIF) and compensatory differential item functioning (CDIF), as well as an index of DTF (Raju et al., 1995). Raju et al. recommended chi-square tests for assessing the statistical significance of DTF and NCDIF indices. The efficacy of the proposed chi-square tests was assessed in a Monte Carlo investigation (Fleer, 1993).

Fleer (1993) demonstrated the chi-square tests for DTF and NCDIF to be overly sensitive for large sample sizes. At the .01 level of significance, substantially greater than 1% of the items in the no-DIF condition were falsely identified as having DIF. Based on this finding, Fleer (1993) and Raju et al. (1995) recommended the use of empirically derived cutoff values to assess the practical significance of DIF. A cutoff value was derived by creating a frequency distribution of observed NCDIF values across 50 replications of a no-DIF condition. A cut-off value of .006 was associated with the 99^{th} percentile and so resulted in falsely identifying approximately 1% of items as exhibiting DIF. Based on this result, Fleer and Raju et al. recommended dichotomous items with NCDIF > .006 be designated as having DIF; DTF > .006 indicates DTF. Subsequent Monte Carlo investigations have recommended cutoff values for polytomous items as well (Bolt, 2002; Flowers et al., 1999; Meade, Lautenschlager, & Johnson, 2006).

These previous Monte Carlo investigations (Bolt, 2002; Fleer, 1993; Flowers et al., 1999; Meade et al., 2006; Raju et al., 1995) indicate the proposed cutoffs for dichotomous and polytomous items have worked well for assessing DIF and DTF. The cutoffs proposed, however, differed across data sets. This is to be expected, as Chamblee (1998) demonstrated that factors such as sample size and the IRT model influence the

cutoff values generated for dichotomously scored items. So although the previous cutoff values have worked well, they are probably not generalizable to other items and sample sizes. This is problematic because analyses conducted in practice are based on empirical data that span a variety of conditions (sample sizes, test lengths, etc.) and optimal cutoff values for a given testing situation are not known. Furthermore, practitioners may not have the time or the expertise to generate their own cutoffs using data simulated to match testing conditions. As a result, practitioners must currently rely upon previously generated cutoffs, which calls into question the accuracy of results based on these analyses. This reduces the practicality of assessing DIF and DTF using the DFIT framework.

As a possible solution to this problem, Oshima et al. (2006) recently proposed the item parameter replication (IPR) method for determining cutoff values for dichotomous items within the DFIT framework. This method provides cutoffs that are tailored to the data set and easy to use in practice.

The IPR method begins with estimates of item parameters and their variances and covariances. A large number of replications of item parameters are then generated from the initial set of item parameter estimates. Next, NCDIF values are computed for all replications of item parameters. The NCDIF values obtained are rank ordered to establish cutoff values across alpha levels. For example, the 99^{th} percentile rank corresponds to the cutoff value at the .01 level of significance. This cutoff value is used for assessing statistical significance of the initial NCDIF value obtained for the item. This process is repeated for all items in the test, thus potentially resulting in different cutoffs for different items. A Monte Carlo investigation by Oshima et al. (2006) showed

the IPR method is effective in maintaining acceptable Type 1 error and power rates. It therefore seems the IPR methodology is a promising improvement upon the DFIT framework when assessing measurement equivalence for dichotomous items.

Many tests and questionnaires used in practice, however, consist of polytomously scored items (e.g., Likert rating scales). It is therefore prudent that methods of assessing measurement equivalence be applicable to both dichotomous and polytomous data. The IPR methodology has been theoretically extended to the polytomous case (Raju, Oshima, Fortmann, Nering & Wonsuk, 2006), but empirical research examining its efficacy in detecting DIF for polytomous items has been very limited in scope. Fortmann, Raju, Oshima and Morris (2006) supported the efficacy of the IPR method with polytomous data, finding it produced results nearly identical to those based on the previously recommended fixed cutoff value used by Flowers et al. (1999). Their study, however, was limited to just one condition of DIF. The first purpose of the current study, therefore, is to conduct a more comprehensive assessment of its efficacy in detecting DIF for polytomous items. It is expected that IPR-based DIF analyses will be more likely to detect true DIF than previously recommended fixed cutoff values and will have false positive rates close to the nominal significance level.

The second purpose of this study is to compare the IPR method to the likelihood ratio test (LR; Thissen, Steinberg, & Wainer, 1988), another IRT-based DIF procedure. The LR test was chosen for comparison purposes because there has been growing interest in the recent literature on the use of these two procedures (Bolt, 2002; Braddy, Meade, & Johnson, 2006; Meade & Lautenschlager, 2004). This research has suggested the DFIT framework is less sensitive to detecting DIF than the LR test. It is important to note,

however, that this result is based on the use of a fixed cutoff value across items. With the introduction of the IPR method to the DFIT framework, different cutoff values may be derived across items. This leads one to question whether previous conclusions regarding the sensitivity of the DFIT framework in comparison to the LR test remain true. Potential factors moderating each statistics' ability to detect DIF are also examined. Since these aspects of the study are intended to be exploratory, no formal hypotheses are stated.

CHAPTER 2

LITERATURE REVIEW

Personnel administrators rely heavily on the use of standardized tests and questionnaires in making important organizational and administrative decisions. Before decisions are made, however, one must consider if the measures used provide equivalent measurement. Measurement equivalence is obtained when the relations between observed scores and latent constructs are identical across relevant subgroups (Drasgow & Kanfer, 1985). Without measurement equivalence, observed scores from different groups are in different scales and are therefore not comparable. In other words, the meaning of score differences is unclear because they reflect not only meaningful group differences but also item/test bias.

The term "bias" has commonly been understood to denote a lack of fairness in test results; although among psychometricians the terms "bias" and "fairness" convey distinct ideas. Item or test bias refers to systematic measurement error related to group membership. Bias is a characteristic of the item or test that is defined statistically. Fairness, on the other hand, is a socially defined concept. A test is said to be fair if the decisions made and opportunities presented based on test results are in accordance with accepted principles, such as equal treatment or equal opportunity. This paper will focus solely on the psychometric issue of item/test bias.

Investigations of item (or test) bias involve gathering empirical evidence of differential performance on an item (or test) between members of a minority group of interest and members of the majority group. Based on this empirical evidence alone, however, one cannot conclude that an item (or test) is biased (Hambleton, Swaminathan,

& Rogers, 1991). This conclusion involves an inference that goes beyond the data. The need to distinguish between the empirical evidence obtained in such research and the resulting conclusions has led to a shift in terminology. The term differential item functioning (DIF) has been adopted to describe the statistical concept of item bias (i.e., the empirical evidence obtained). The term differential test functioning (DTF) similarly describes the statistical concept of test bias. These terms will be used throughout this paper.

Research concerning DIF/DTF began in the 1960s as a result of civil rights legislation and primarily focused on understanding observed group differences on tests of cognitive ability (Raju & Ellis, 2002). To this day, aptitude and achievement testing continues to be one area in which measurement equivalence across groups is of extreme importance. If some groups (e.g., minority groups) systematically receive lower mean test scores than other groups (e.g., Caucasians), one must determine whether these differences are due to true differences on the latent construct or measurement error. If score differences at the item or test-level are a reflection of systematic measurement error, an item or test is said to exhibit DIF or DTF (Drasgow & Kanfer, 1985). DTF could potentially cause group differences in the selection rates (i.e., adverse impact), even when both groups are equally talented.

Similarly, measurement equivalence should be examined for attitudinal questionnaires (Drasgow & Kanfer, 1985). Questionnaires are used to assess organizational attitudes and identify the need for training and organizational development initiatives. Score differences among individuals from different groups, (e.g., men vs. women, Caucasians vs. minority groups, managers vs. non-managers, etc.) are typically

assumed to reflect true group differences on the attitude being measured. Costly organizational interventions are then planned to address these group differences. If DIF or DTF is present in the questionnaires used, however, observed group differences may be erroneous and the conclusions drawn based on these results will be inappropriate. In light of the potential consequences, it is therefore of vital importance that practitioners have a readily available means by which to assess differential functioning of items and tests. To date, methods for examining DIF/DTF are rooted in two popular measurement frameworks: Classical test theory and item response theory (IRT).

2.1 Classical Test Theory

Classical test theory rests on four fundamental assumptions. First, classical test theory postulates an individual's observed score on a single administration of a test is equivalent to the sum of his or her true score and measurement error. The true score represents the true level of the ability or trait that the individual possesses, and measurement error is defined as the discrepancy between the observed and true scores. Second, within a person, it is assumed that the mean of errors across all possible replications of a test equals zero. A person's true score can never be known, however based on this second assumption, the true score can be operationally defined as the average observed score over an infinite number of test replications. Third, classical test theory assumes that true scores are uncorrelated with error scores. Finally, error scores from one replication of a test are assumed to be uncorrelated with true scores from another replication of the same test. Weaknesses in this classical model led psychometricians to seek alternative measurement models (Hambleton et al., 1991).

Perhaps the largest deficiency of classical test theory is the inability to separate

examinee and test characteristics (Hambleton et al., 1991). Within classical test theory the true score represents an examinee's true level of ability. This true score, however, is defined in terms of observed performance on the test. In other words, ability is not independent of the specific test items. Difficult test items will result in the examinee appearing to have low ability, while easy test items will result in the examinee appearing to have high ability. Similarly, item characteristics are not independent of examinee characteristics. Classical test theory defines item difficulty as the proportion of examinees in a group of interest who answer the item correctly (Hambleton et al.). Examinee ability will determine the proportion that answer the item correctly and therefore influence interpretation of an item as easy or difficult. This inability to separate examinee and test characteristics makes it difficult to compare items completed by different groups of examinees, as well as to compare examinees completing different tests.

A second deficiency of classical test theory is the assumption of equal standard errors of measurement (SEM) for each examinee (Hambleton et al., 1991). This assumption is not plausible, as test scores for examinees of different ability contain different amounts of error. In other words, scores on a test are not equally precise measures for different levels of ability. As a result, a test may be good at distinguishing between people with high levels of ability but fail to adequately make distinctions between people with low levels of ability.

Finally, classical test theory is test-oriented rather than item-oriented (Hambleton et al., 1991). This is a limitation in that this model does not allow one to consider examinee performance at the item-level. Investigations of bias are therefore limited to

examination of overall test results. An alternative theory, IRT, addresses the limitations of classical test theory.

2.2 Item Response Theory

At the foundation of IRT is the basic idea that an examinee's performance on a test item can be explained by a set of latent abilities. IRT postulates that the probability of answering an item correctly for an examinee of a given level of ability, $P(\theta)$, is a function of the examinee's ability (θ) and the characteristics of the test item. The relationship between item performance, $P(\theta)$, and this set of latent abilities (θ) is assumed to be a monotonically increasing function. This function is referred to as the item response function (IRF) or item characteristic curve (ICC; see Figure 1). In other words, the IRF illustrates that as the examinee's level of ability increases so does his or her probability of answering an item correctly.

A key advantage of IRT is that ability and item parameters are invariant. Unlike classical test theory, estimates of examinee ability are not dependent on the set of test items and item indices are not dependent on the group of examinees' abilities. This means that estimates of ability for a given examinee will be the same regardless of the items used. Similarly, item indices, such as item difficulty, will be the same regardless of the group of examinees. Because of this property of invariance, IRT is well suited for addressing item bias. A second advantage of IRT is that it provides a means for computing the SEM for individual ability estimates. This overcomes classical test theory's implausible assumption of equal errors for each examinee. The IRT framework and its fundamental assumptions will be described in more detail in the sections to follow.

Assumptions of IRT. Two assumptions are made regarding the data to which the IRT model will be applied. First, it is assumed the items in a given test constitute one dominant dimension. That is, the test measures one ability. When this assumption is met local independence is also obtained (Hambleton et al., 1991). Local independence means that when the abilities influencing test performance are held constant, an examinee's responses to any pair of items are uncorrelated. Mathematically this implies the probability of a response pattern for an examinee on a set of items is equal to the product of the probabilities of the individual item responses. For a unidimensional test this again suggests the ability specified in the IRT model is the only factor influencing examinee responses to test items. It should be noted that multidimensional IRT models have been developed, but these are beyond the scope of the current paper.

The second assumption made in IRT is that the IRF specified in the model represents the true relationship between the latent ability and item responses. In other words, it is assumed the IRT model chosen is appropriate for the available data.

Unidimensional Dichotomous IRT Models. IRT models differ with regards to the mathematical function used to model the data, as well as the number of specified item parameters. Early work in IRT modeled data using the normal ogive. The standard normal ogive represents the area to the left of any *z*-score on the normal curve as a function of the *z*-score. *Z*-scores are represented along the *x*-axis, while the area to the left of any z-score is represented as a proportion along the *y*-axis. Applied to IRT, values on the *x*-axis represent the latent ability measured and the height of the curve above a value represents the proportion of examinees of a given level of ability that can be expected to answer the item correctly. The normal ogive model, however, has commonly

been replaced by the logistic function because logistic models are more mathematically tractable (Hambleton et al., 1991). Variations of both the normal ogive and logistic models are determined by the number of item parameters used to model the data. Since logistic models are more commonly used and provide essentially the same interpretation, these will be described in further detail. Three of the most common logistic models are referred to as the one, two and three-parameter logistic models and are appropriate for dichotomous item response data. Hambleton et al. indicate the one-parameter logistic (1-PL) model to be the most widely used.

The 1-PL model, also known as the Rasch model in honor of its developer, assumes that examinee ability and item difficulty are the only parameters influencing examinee performance. Mathematically the IRF for a given item in the 1-PL model is defined as follows:

$$P(\theta) = \frac{e^{\theta - b}}{1 + e^{\theta - b}}, \tag{1}$$

where $P(\theta)$ represents the probability of answering the item correctly for a given level of ability, θ represents the examinee ability continuum, *e* is a constant equal to 2.718, and *b* is the item difficulty parameter. Specifically the *b*-parameter is defined as the point on the ability continuum at which the probability of answering the item correctly is equal to .5. The larger the value of the *b*-parameter, the greater the ability required for an examinee to have a 50% chance of producing a correct answer and so the harder the item.

One potential limitation of this model is that it does not allow for differently discriminating items. Item discrimination refers to the ability of an item to differentiate among examinees on the basis of the trait or ability being examined. Practically speaking, it is not possible for the IRFs to cross. So, for example, within the 1-PL model it is not possible for item 1 to be more difficult than item 2 for Examinee A but vice versa for Examinee B. The two-parameter logistic (2-PL) model addresses this limitation.

The 2-PL model is defined by the following equation:

$$P(\theta) = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}, \tag{2}$$

where D is a constant equal to 1.7; a is the item discrimination index; and $P(\theta)$, θ, e and b are defined as in Equation 1. The a-parameter, or item discrimination index, is proportional to the slope of the IRF at the point b on the ability scale. The bigger the a-parameter, the stronger the relationship between the item and the underlying construct; such items are more useful for discriminating among examinees near a certain ability level. The addition of this item parameter to the model allows the IRFs for different items to cross (see Figure 2). In other words, it allows for differently discriminating items. Still ignored by both the 1 and 2-PL models, however, is the impact of guessing on the probability of answering an item correctly.

The three-parameter logistic (3-PL) model incorporates this additional item parameter. The mathematical equation for this model is as follows:

$$P(\theta) = c + (1-c)\left(\frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}}\right), \quad (3)$$

where *c* represents the probability of examinees with low ability answering the item correctly and all other notations are defined as in Equation 2. The *c*-parameter allows for a nonzero lower asymptote for the IRF. The addition of this item parameter takes into account performance at the low end of the ability continuum, where guessing may be a factor in test performance. As such, the *c*-parameter is often referred to as the pseudo-guessing parameter.

Although commonly used, the above models are only applicable to dichotomous response data. Many item types used in psychological research, however, allow for multiple response options (e.g., Likert-type rating scales). Therefore it is necessary to also understand the polytomous IRT framework.

Unidimensional Polytomous IRT Models. Polytomous IRT models are an extension of the dichotomous case. A key difference between polytomous and dichotomous IRT, though, relates to the function used to describe the relationship between $P(\theta)$ and θ. Dichotomous IRT models this relationship with a single response function for each item that marks the boundary between the two possible response categories. As discussed, the IRF graphically displays the probability of a correct response, or in other words, the boundary between the correct and incorrect response. Polytomous IRT models are more complex in that a given item contains multiple response categories, and so at least one response category is always defined by two boundaries (the boundary between that

response category and each adjacent response category).

In the case of items consisting of ordered response categories (i.e., ordinal data), polytomous IRT is focused on two probabilities: The probability of responding positively rather than negatively at a given boundary between response categories and the probability of responding in a given response category. The boundary response functions (BRFs; see Figure 3) illustrate this first probability, the probability of responding positively rather than negatively at a given boundary between response categories. The number of BRFs will be equal to the number of response categories minus one. Polytomous models handle ordinal data by creating multiple dichotomies and modeling each BRF separately using a dichotomous model. The information obtained from the BRFs is then used to derive the probability of responding in a given response category.

The probability of selecting a particular response category as a function of θ is graphically displayed by the category response functions (CRFs; see Figure 4). Unlike the dichotomous IRF, polytomous IRT yields as many response functions as there are response categories for the item. This marks a clear distinction from the dichotomous IRT models; within polytomous IRT there is no single equivalent of the dichotomous IRF. Further, the CRFs are not monotonically increasing. The sum of these probabilities for an examinee at a given level of ability will equal one.

It should be noted that not all polytomous IRT models are designed to handle ordinal data. Several polytomous models exist (e.g., Bock, 1972; Masters, 1982; Muraki, 1990) which vary with respect to the assumptions made regarding response categories. The current research however, is based on Samejima's (1969) graded response model

(GRM). Samejima's GRM has been used in past DFIT research and is one of the most commonly used polytomous models (Zickar, 2002), making it an appropriate choice for the current investigation. Therefore this model will be described in detail in the section to follow. Interested readers are referred to Ostini and Nering (2006) for further discussion on polytomous models.

Samejima's (1969) Graded Response Model. Graded response models assume item response categories can be rank ordered, or in other words represent ordinal data (Zickar, 2002). Samejima's (1969) GRM extends the 2-PL model to the polytomous case.

The GRM breaks polytomous data into multiple dichotomies, such that the number of dichotomies is equal to the number of response categories minus one. A 5-category item, for instance, is represented by four dichotomies, and $P^*(\theta)$ is computed for each:

$$P_1^*(\theta) = \frac{e^{Da(\theta - b_1)}}{1 + e^{Da(\theta - b_1)}}, \tag{4}$$

$$P_2^*(\theta) = \frac{e^{Da(\theta - b_2)}}{1 + e^{Da(\theta - b_2)}}, \tag{5}$$

$$P_3^*(\theta) = \frac{e^{Da(\theta - b_3)}}{1 + e^{Da(\theta - b_3)}}, \tag{6}$$

$$P_4^*(\theta) = \frac{e^{Da(\theta - b_4)}}{1 + e^{Da(\theta - b_4)}}, \tag{7}$$

where $P_1^*(\theta)$, $P_2^*(\theta)$, and $P_3^*(\theta)$ represent the probabilities of responding in or above the 2nd, 3rd, and 4th response categories, respectively, and $P_4^*(\theta)$ represents the probability of responding in the 5th response category. As shown in Equations 4-7, a five-category item is represented by one *a*-parameter and four *b*-parameters. The number of *b*-parameters for an item is always equal to the number of response categories minus one. The relationship between θ and $P^*(\theta)$ for each dichotomy is illustrated by the boundary response functions (BRFs; see Figure 3).

Three things should be noted regarding the BRFs. First, similar to the IRF, BRFs are monotonically increasing. Second, because the item is represented by only one *a*-parameter, the BRFs for an item never cross. Finally, the *b*-parameters are always rank ordered such that $b_1 \leq b_2 \leq b_3 \leq b_4$. Related to the BRFs are the category response functions (CRFs).

The CRF for a given response category represents the probability of choosing that response category as a function of θ. This probability is equivalent to the difference between the probabilities of responding in or above category *n* and responding in or above category *n*+1. The probability of responding to an item at all is set equal to one. As such, the CRFs are computed based on knowing the dichotomy probabilities defined above. Equations 8-12 define the CRFs for a 5-category item.

$$P_1(\theta) = 1 - P_1^*(\theta) \tag{8}$$

$$P_2(\theta) = P_1^*(\theta) - P_2^*(\theta) \tag{9}$$

$$P_3(\theta) = P_2^*(\theta) - P_3^*(\theta) \quad (10)$$

$$P_4(\theta) = P_3^*(\theta) - P_4^*(\theta) \quad (11)$$

$$P_5(\theta) = P_4^*(\theta) \quad (12)$$

So, $P_1(\theta)$, $P_2(\theta)$, $P_3(\theta)$, $P_4(\theta)$ and $P_5(\theta)$ represent the probability of choosing the 1st, 2nd, 3rd, 4th and 5th response categories, respectively. Again, there will be as many CRFs for an item as there are response categories (see Figure 4).

Definition of the True Score. Recall that in classical test theory the true score is operationally defined as the average observed score over an infinite number of test replications. Although this value can never be known, an estimate of the true score for a set of dichotomously scored items is obtained by determining the number of correct responses. Within IRT, estimation of the true score is not quite as simple.

As previously described, examinee responses to a set of items are used to estimate an ability score θ. The scale of θ is arbitrary, with values ranging from $-\infty$ to ∞. However because ability is invariant with respect to the items and item parameters are invariant with respect to the sample, the θ-scale may be transformed provided the item parameter values are also transformed (Hambleton et al., 1991). The most important transformation of the θ-scale is to the true score scale.

The true score is defined as the expected value of the score on a test. For a dichotomously scored item, the expected score on an item for an examinee of a given level of ability may be defined as:

$$E(X_i) = (1)P_i(\theta) + (0)Q_i(\theta) = P_i(\theta), \quad (13)$$

where *i* represents the item number, $P_i(\theta)$ represents the probability of a correct response (i.e., the IRF) and $Q_i(\theta)$ represents the probability of an incorrect response. The expected value of a test score defined as the total number correct may then be expressed as:

$$T(\theta) = E(\sum_{i=1}^{n} X_i). \quad (14)$$

Equation 14 may be re-written such that:

$$T(\theta) = \sum_{i=1}^{n} P_i(\theta), \quad (15)$$

where *n* represents the total number of test items. That is, the true score of an examinee of a given ability level is equal to the sum of the IRFs. This true score is commonly referred to as the test characteristic curve. Because the value of $P_i(\theta)$ is between 0 and

1, *T* will range between 0 and *n*. An important implication of this transformation is that it yields an estimate of the true score that is on a readily interpretable scale. The same logic can be extended to determination of the true score for polytomously scored items.

Polytomous IRT yields multiple CRFs, and this information may be used to arrive at the expected item true score for an examinee of a given level of ability. Following is a general equation for the item true score (Raju, Laffitte, & Byrne, 2002), assuming the response categories are scored 1, 2…*m*:

$$t_i(\theta) = (1)P_{i1} + (2)P_{i2} + \ldots + (m_i - 1)P_{i(m_i-1)} + (m_i)P_{i(m_i)}, \quad (16)$$

where P_{i1}, P_{i2}, etc. represent the probabilities of responding in a given response category, and the values 1 through *m* represent the response category scores or weights. The total true score (i.e., expected test score function) can then be computed, such that:

$$T(\theta) = \sum_{i=1}^{n} t_i(\theta), \quad (17)$$

where *n* again represents the number of items in the test, and *i* represents the item number. In other words, Equation 17 indicates the total true score is equivalent to the

sum of the item true score functions. DIF can be conceptualized in terms of comparison of the true score across groups.

One can compare the probability of answering an item correctly for examinees of the same level of ability across groups, and DIF occurs when this probability differs (i.e., when the probability of answering an item correctly depends not only on ability but also group membership). Examination of DIF is essential in order to determine whether scores differences across groups are due to true differences in the trait or ability being measured, or to systematic measurement error related to group membership. As mentioned previously, the use of tests or questionnaires containing DIF may result in costly or legally questionable decisions. This analysis is therefore essential to the effective use of tests and questionnaires in practice. Before comparison of true scores across groups is possible, however, parameter estimates for the focal and reference groups must be placed on a common metric in order for meaningful comparisons to be made. Two methods for doing this are described next.

Linking versus Concurrent Estimation. Parameter estimates may be placed on a common metric using either linking or concurrent estimation procedures. Using the former, parameters are independently estimated for different groups of examinees, thus requiring separate computer runs for each group. During estimation a scale is arbitrarily assigned to theta, such that the mean is equal to zero and the standard deviation is equal to one. The specific estimates obtained from these independent calibrations will therefore be different because the metric defined by each calibration is arbitrary. Therefore, estimates must be placed on a common metric, and this process is known as linking. In other words, linking involves transforming the metric from the first

calibration to the metric of the second calibration. The presence of biased items is known to distort computation of the linking coefficients (Millsap & Everson, 1993) and so iterative linking procedures are typically used.

Using an iterative approach, linking coefficients are first computed based on all test items. A DIF analysis is then conducted to identify items exhibiting DIF. Using only the items deemed DIF-free from this previous step, second-stage linking coefficients are then computed. These DIF-free items are referred to as the anchor items. DIF analysis is then conducted again for all items using the second-stage linking coefficients, and items exhibiting DIF are identified. While concurrent estimation also requires a set of DIF-free anchor items, an iterative process such as this is not used.

Concurrent estimation combines data from both groups and simultaneously estimates item and ability parameters within a single computer run. Using this approach, parameter estimates for at least one item must be fixed across groups. This item (or set of items) is referred to as the anchor item (or anchor set) and links the metric for parameter estimation. This ensures that all parameter estimates are on the same scale. Using this approach, the anchor set is typically defined by items known to be DIF-free. Different approaches to DIF analysis tend to use either linking or concurrent estimation. For example, the LR test uses concurrent estimation procedures; while DFIT tends to use separate estimation with linking.

Comparison of linking versus concurrent parameter estimation, however, suggests comparable recovery of item and ability parameters across methods (Hanson & Beguin, 2002; Kim & Cohen, 2002). Based on these results, in the current study it is felt appropriate to apply concurrent estimation procedures to all DIF analyses. Further,

although differences between linking and concurrent estimation are seemingly small, the use of linking versus concurrent estimation procedures across DIF analyses would still introduce a potential confounding variable into the research. DIF analyses would be based on slightly different parameter estimates and this could potentially contribute to observed differences in test results. Use of a common estimation method across DIF procedures strengthens the methodology by removing this potential confound. Methods for examining DIF will be described next.

2.3 Methods for Examining Differential Item Functioning (DIF)

Methods for examining DIF may be placed into two general categories: Unobserved conditional invariance (UCI) methods and observed conditional invariance (OCI) methods (Millsap & Everson, 1993). UCI methods are based on assumed measurement models that relate an observed measure to the latent construct being assessed. Bias is then examined by evaluating whether the measurement model remains invariant across populations. IRT-based methods for examining DIF fall within this category. OCI methods, on the other hand, do not attempt to relate an observed measure to the latent construct. Instead, some other observed variable is used as a proxy for the latent construct. For instance, the total test score may be used as a proxy for an examinee's true ability. Methods of assessing DIF rooted in classical test theory fall within this category. Because such methods result in sample-dependent indices of DIF they are not as well suited to this purpose as the IRT-based methods (Hambleton et al., 1991) and will therefore not be addressed in the current investigation. Interested readers, however, may refer to Millsap and Everson for further discussion.

Of interest to the current investigation are the IRT-based methods of examining

DIF. Several IRT-based methods exist but all are rooted in the previously described measurement model relating the probability of a correct response, or in the case of polytomous IRT the probability of responding in a particular response category to a person's true ability and other characteristics of the test item. Further, all IRT-based methods require estimation of item parameters for each subgroup under investigation. Commonly, one group is referred to as the reference group, while the other is referred to as the focal group. In the case of comparisons across ethnic groups, the group representing the majority is referred to as the reference group, and the group representing the minority is referred to as the focal group. Statistically, DIF may be defined in terms of either the estimated item parameters or the item response functions (Hambleton et al., 1991; Millsap & Everson, 1993), and different methods of examining DIF are rooted in these different definitions.

Recall that IRT assumes item parameters are invariant across groups of examinees drawn from the same population (Hambleton et al., 1991). One would therefore expect that item parameter estimates for the focal group would be identical to those for the reference group minus differences due to estimation error. An item is said to exhibit DIF when the two sets of item parameter estimates are not the same (Hambleton et al.; Millsap & Everson, 1993). Based on this definition, some IRT-based methods of examining DIF directly test the equality of individual item parameters across groups and finding of statistically significant differences provides evidence of DIF.

Because item parameters dictate the shape of the item response function, differences in parameters between the focal and reference groups will result in differences in the item response functions. Alternatively, then, an item may be said to

exhibit DIF when the item response functions are not identical across groups (Hambleton et al., 1991; Millsap & Everson, 1993). Defined in this way, other IRT-based measures examine DIF by computing either the total area between item response functions or the average distance between item response functions over a selected interval of the ability continuum. The greater the area or average distance, the greater the magnitude of DIF.

Defining DIF in terms of the item response functions, however, is restrictive in the sense that it is only applicable to dichotomously scored items. A more general definition suggests an item be considered to exhibit DIF when the item true score functions are not identical across groups (Cohen, Kim & Baker, 1993; Kim, Cohen & Park, 1995). In the dichotomous case, the item true score function is identical to the item response function and so these two definitions are one in the same. In the polytomous case, the item true score functions of the focal and reference groups will be identical when the boundary response functions are equivalent, or when the item parameter estimates are equivalent. The IRT-based methods of examining DIF under examination in the current study will now be described in more detail.

Likelihood Ratio Test. The likelihood ratio (LR) test (Thissen, Steinberg, & Gerrard, 1986; Thissen et al., 1988) examines the equality of item parameters between groups. DIF analysis using the LR method involves comparing a series of nested models, wherein item parameters are fixed or freed across groups. One model is commonly referred to as the compact model, and all other models are referred to as the augmented models. A separate augmented model is estimated for each studied item, and each item is tested one at a time for DIF. Item parameters for each model are estimated using maximum likelihood estimation procedures, resulting in the likelihood value of model fit known as

the fit function (Meade & Lautenschlager, 2004). The fit function is an index of how well the given IRT model fits the data as a result of the maximum likelihood estimation procedures used. The fit of the nested models is simultaneously compared via the LR test.

The LR is computed as follows:

$$LR_j = \frac{L_C}{L_{Aj}}, \tag{18}$$

where L_C represents the likelihood function of the compact model and L_{Aj} represents the likelihood function of the augmented model for studied item j. Taking the natural log of this function results in a test statistic that is distributed as a chi-square,

$$G^2 = \chi^2(df) = -2\ln(LR) = -2\ln L_C + 2\ln L_{Aj}, \tag{19}$$

with degrees of freedom (*df*) equal to the difference in the number of item parameters estimated in the compact model versus the augmented model. A significant χ^2 value indicates the augmented model fits the data better than the compact model. The implied assumption here is that one of the models must fit the data. It should also be noted that

this approach was made viable due to the development of concurrent estimation procedures, and another inherent assumption of this approach is that the anchor items are unbiased.

Different approaches to using the LR test exist that differ with regards to how the anchor items are defined. Traditionally, in the compact model all item parameters are constrained to be equal across the reference and focal groups. Estimation of the compact model provides a baseline likelihood value for item parameter fit for the model, and therefore this model is also referred to as the baseline model. In each augmented model then, item parameters for the studied item are allowed to vary, while all other item parameter estimates are constrained to be equal across groups. In other words, the set of anchor items is defined as "all other items" in the test or questionnaire.

An alternative approach is to identify a set of core items modeled to be DIF-free (Ankenmann, Witt, & Dunbar, 1999). The core items serve as the common anchor items. Other researchers (e.g., Stark, Chernyshenko, & Drasgow, 2006; Wang & Yeh, 2003) have also used this approach varying the number of anchor items. Using this approach, in the baseline model the designated anchor items are constrained to be equal across groups and all other item parameters are free to vary. A separate model is then estimated for each item such that the studied item, in addition to the anchor items, is constrained to be equal across groups. In other words, using this approach, the baseline model is now an augmented model in which item parameters are free to vary, and a separate compact model is estimated for each item. Regardless of which of these approaches to defining the anchor items is used, estimation of each model yields a likelihood value of model fit, and LR values and test statistics are computed as in Equations 18 and 19.

Differential Functioning of Items and Tests. The LR test (Thissen et al., 1986; Thissen et al., 1988) provides an item-level index of differential functioning. It does not, however, provide a means by which to assess differential functioning at the test level. It is simply assumed the removal of items with significant DIF will result in a test that is unbiased (Raju et al., 1995). Raju et al. noted the desirability of having a psychometric measure of DTF and proposed the DFIT framework.

The DFIT framework was originally proposed for dichotomous items and was later extended to the polytomous case (Flowers et al., 1999). Flowers et al. note the only difference between dichotomous and polytomous DFIT is in the computation of the item and test true scores. Once these values are known, computations of indices of DIF/DTF are identical.

DFIT (Raju et al., 1995) begins by defining differential functioning at the test level and then decomposing it into differential functioning at the item level. For a polytomously scored item, let $t_i(\theta_s)$ represent the item true score for examinee *s* with ability level θ on item *i*. Assume the test consists of *n* items, and item parameters have been estimated separately for the focal group (F) and reference group (R). Further assume these two sets of item parameters have been placed on a common metric. We may then compute the item true score on the i^{th} item for examinee *s* as a member of the focal group $[t_{iF}(\theta_s)]$ and as a member of the reference group $[t_{iR}(\theta_s)]$. If $t_{iF}(\theta_s) \neq t_{iR}(\theta_s)$, then the item is said to exhibit DIF. Recall, for polytomously scored items the test true score for a given examinee is equivalent to the sum of the item true scores. Therefore, it is also possible to compute two test true scores: One treating the examinee as a member of the focal group (T_{sF}) and one as a member of the reference group (T_{sR}). If $T_{sF} \neq T_{sR}$,

the examinee's true score is not independent of group membership; the test is said to exhibit DTF. The greater the difference between these two true scores, then the greater the magnitude of DTF.

Raju et al. (1995) defined a measure of DTF at the examinee level as $(T_{sF} - T_{sR})^2$. Across examinees, DTF may therefore be defined as follows:

$$DTF = E(T_{sF} - T_{sR})^2, \tag{20}$$

where the expectation (E) may be taken over the reference group or the focal group. Assuming the expectation is taken over the focal group, Equation 20 may be rewritten such that:

$$DTF = \underset{F}{E}(T_{sF} - T_{sR})^2. \tag{21}$$

If we let D_s represent the difference between the two true scores, then Raju et al. showed Equation 21 may be rewritten as follows:

$$DTF = \underset{F}{E} D_s^2 = \int_{\theta} D_s^2 f_F(\theta) d\theta = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2, \quad (22)$$

where $f_F(\theta)$ is the density function of θ in the focal group, and μ_{TF} and μ_{TR} represent the mean true score of examinees in the focal and reference groups, respectively.

Based on the definition of DTF offered in Equation 22, Raju et al. (1995) derived an index of compensatory differential item functioning (CDIF). Equation 20 can be rewritten such that:

$$DTF = E\left[\left(\sum_{i=1}^{n} d_{is}\right)^2\right], \quad (23)$$

where d_{is} represents the difference between item true scores. This can be rewritten as:

$$DTF = \sum_{i=1}^{n}\left[Cov(d_i, D) + \mu_{d_i}\mu_D\right], \quad (24)$$

where $Cov(d_i, D)$ represents the covariance between the difference in item true scores and the difference in total test true scores, μ_{d_i} represents the mean difference in item true

scores, and μ_D represents the mean difference in test true scores. From this, CDIF is defined as:

$$CDIF_i = E(d_i D) = Cov(d_i, D) + \mu_{d_i}\mu_D . \quad (25)$$

The CDIF index has several advantages over other measures of DIF. First, the CDIF index is additive such that differential functioning at the test level is equal to the sum of the differential functioning at the item level. In other words,

$$DTF = \sum_{i=1}^{n} CDIF_i . \quad (26)$$

This index is also compensatory in the sense that it takes into account compensating bias across items. The CDIF index may yield either a positive or negative value indicating an item to be in favor of either the focal or the reference group. Practically speaking, this means that if one item favors one group while another item favors the other group, the bias present in these items may cancel each other out. As such, the items together may not contribute to DTF. These properties of the CDIF index allow practitioners to gain a sense of each item's contribution to DTF and aid in decisions regarding which items to

delete in order to have the largest overall impact on reducing DTF.

The DFIT framework (Raju et al., 1995) further offers an index of noncompensatory differential item functioning (NCDIF). NCDIF is non-directional, meaning it does not take into account compensating bias across items. The NCDIF index instead reflects only the differential functioning of the item under review. In essence, it is assumed all other items in a test are DIF-free. Based on this assumption, it must be true that $d_j = 0$ for all *j* items, where $j \neq i$. Equation 25 may then be rewritten as:

$$NCDIF_i = Ed_i^2 = \sigma_{d_i}^2 + \mu_{d_i}^2 . \quad (27)$$

This means items with significant NCDIF do not necessarily have significant CDIF (Raju et al.). If one items favors the reference group and another favors the focal group, both items will have significant NCDIF while CDIF indices may not be significant due to cancellation at the test level. If the assumption that no other items in the test exhibit DIF is true, the values obtained for NCDIF and CDIF will be equal. Further, NCDIF will equal zero when the item parameters for item *i* are identical across the focal and reference groups.

It is worth mentioning that in addition to cancellation at the test level, Flowers et al. (1999) note polytomous response data allows for cancellation within an examinee at the item level. In the polytomous case, multiple probabilities (i.e., BRFs) are computed for each item, and these are combined to arrive at the item true score. So, for a given

examinee, one response category can cancel the effects in another category when computing the difference between item true scores. For example, for a given item if the probability of responding in Category 1 is greater for the reference group than it is for the focal group and vise versa for Category 2, these differences will cancel each other out and the item true score difference will remain close to zero. This indicates no differential functioning at the item level within an examinee.

Raju et al. (1995) recommended significance tests to be used with the DFIT indices. Beginning with DTF, if we assume that D_s is normally distributed with a mean of μ_D and a standard deviation of σ_D a *z*-score can be computed for examinee *s* as follows:

$$Z_s = \frac{D_s - \mu_D}{\sigma_D}. \tag{28}$$

Z_s^2 is known to have a χ^2 distribution with one degree of freedom. The sum of Z_s^2 across *N* examinees has a χ^2 distribution with *N* degrees of freedom:

$$\chi_N^2 = \sum_{s=1}^{N} Z_s^2 = \frac{\sum_{s=1}^{N} (D_s - \mu_D)^2}{\sigma_D^2}. \tag{29}$$

Under the null hypothesis of no DIF, $\mu_D = 0$. Substituting $\mu_D = 0$ into Equation 29 yields:

$$\chi^2_N = \frac{\sum_{s=1}^{N} D_s^2}{\sigma_D^2}. \tag{30}$$

According to the definition of DTF provided in Equation 22, Raju et al. indicate Equation 30 can be expressed as:

$$\chi^2_N = \frac{N(DTF)}{\sigma_D^2}. \tag{31}$$

If we substitute a sample-based estimate for the variance of D, then

$$\chi^2_N = \frac{N(DTF)}{\hat{\sigma}_D^2}. \tag{32}$$

A significant χ^2 value indicates one or more items function differentially (i.e., there is significant DTF). Because CDIF indices sum to DTF, one may use this information to then identify the items causing the significant χ^2. Raju et al. (1995) recommended deletion, one at a time, of items with large, positive CDIF values. Following deletion of a single item, the χ^2 test of DTF is re-computed based on the remaining items. This process is repeated until DTF becomes nonsignificant. When one deletes an item, however, not only is the CDIF of that item removed from the DTF index but also its contribution to other CDIF indices (Raju, 1999a). Therefore, identification of the item with the largest CDIF may not be the best strategy for identifying an item for deletion. Raju recommended computing A = 2CDIF – NCDIF for each of the items in a test and removing the item with the largest A value. Regardless of the strategy used, deleted items are labeled as having significant CDIF, and therefore no separate significance test of the CDIF index was proposed.

Raju et al. (1995) did propose a χ^2 test similar to the one for DTF for testing the significance of NCDIF:

$$\chi^2_N = \frac{N(NCDIF)}{\hat{\sigma}^2_{d_i}}. \tag{33}$$

Monte Carlo examination of the proposed indices suggested the χ^2 tests for DTF and NCDIF to be overly sensitive for large sample sizes (Fleer, 1993).

In theory, DTF and NCDIF indices are defined in terms of true item and person parameters. In practice, however, these values are not known. One must compute estimates of DTF and NCDIF based on estimated item and person parameters. As a result, the estimates of DTF and NCDIF used in practice have two distinct sources of error: Sampling error resulting from drawing a sample from a population of examinees and estimation error resulting from the estimation of parameter values (Raju et al., 1995). The proposed χ^2 significance tests do not fully account for the error associated with the estimation of parameter values (Fleer, 1993; Raju et al.). According to Raju et al., when using estimated parameter values in the computation of DTF and NCDIF it is highly unlikely $D_s = 0$ for all examinees, even when there is no differential functioning (i.e., the null condition). Therefore, in a simulated no-DIF condition, one would expect some items to be falsely identified as exhibiting DIF due to this error. Specifically, with an alpha level of .01, one would expect 1% of items to be falsely identified as exhibiting DIF. In Fleer's no-DIF condition, though, the proportion of items falsely identified as exhibiting DIF was substantially greater than 1%. This finding highlighted the fact that with large sample sizes the proposed χ^2 tests will tend to be statistically significant even when the observed NCDIF values are extremely small, resulting from error and not reflecting DIF. Items that do not function differentially across groups will thus be falsely identified as exhibiting DIF. This suggested the need to establish empirically derived cutoff values for the DTF and NCDIF indices. These cutoff values provide a means to identify findings of differential functioning that are not only statistically significant but also practically nontrivial.

Cutoff values were established by creating a frequency distribution of observed

NCDIF values across 50 replications of the no-DIF condition (Fleer, 1993). That is, for 50 separate no-DIF data sets, item and person parameters for the focal and reference groups were estimated and linked to a common metric. For each of the 50 pairs of parameter estimates, NCDIF was then computed. The resulting 50 values of NCDIF were rank ordered to identify the value of NCDIF associated with the 99th percentile. A cutoff value of .006 was associated with the 99th percentile and so resulted in falsely identifying approximately 1% of items as exhibiting DIF. Based on this result from Fleer's Monte Carlo study, Raju et al. (1995) recommended items with NCDIF > .006 and a statistically significant χ^2 be designated as exhibiting DIF; DTF > .006 and a significant χ^2 suggests differential functioning at the test level. Using the recommended cutoff value, subsequent analysis of this data (Fleer; Raju et al.) provided support for the DFIT framework and showed it to produce results comparable to other available measures of DIF. A separate Monte Carlo investigation (Chamblee, 1998) extended these results.

Chamblee (1998) manipulated the IRT model used to generate dichotomous item data, as well as the sample size. Recommended cutoff values associated with the 95th, 99th, 99.5th, and 99.9th percentiles were identified from the distribution of NCDIF values obtained across 50 replications. Interestingly, Chamblee did not replicate the cutoff value recommended by Fleer (1993). Tables of empirically derived cutoff values across simulated conditions and alpha levels were provided. It is interesting to note the cutoff values obtained from the distribution of observed NCDIF values ranged from .003 to .018. Whereas Fleer recommended a cutoff value of .006, under the same conditions, the results of Chamblee suggested a cutoff value of .008. Findings further suggested that as

sample size increases and the number of parameters in the IRT model decreases, recommended NCDIF cutoff values tend to become smaller. In summary, this study supported the idea that optimal cutoff values are related to sample size and the specific IRT model used in the investigation and suggests that cutoff values may not be generalizable across data sets. Subsequent research seeking to establish recommended cutoff values for polytomous data (Flowers et al., 1999; Raju, 1999b) echoes this concern.

Based on the cutoff values established for dichotomous data, Raju (1999b) provided a strategy for identifying cutoff values for polytomous data that does not require an extensive Monte Carlo investigation. For an item with *k* response-categories, the item true scores vary between 1 and *k*. This polytomous 1-*k* scale may be transformed into a dichotomous scale, where item true scores vary between 0 and 1 using a transformation originally proposed by Raju, Burke and Normand (1990). According to Raju et al., the item true score on the 0-1 scale (t^*) is computed as follows:

$$t^* = \frac{t-1}{k-1}, \tag{34}$$

where *t* represents the item true score on the 1-*k* scale and *k* represents the number of response categories. The item true score difference may be written as:

$$d^* = t_F^* - t_R^* = \left(\frac{t_F - 1}{k-1}\right) - \left(\frac{t_R - 1}{k-1}\right) = \frac{t_F - t_R}{k-1} = \frac{d}{k-1}. \quad (35)$$

Based on Equation 27, the NCDIF index associated with the 1-*k* scale may be computed as:

$$NCDIF_i^* = Ed_i^{*2} = \frac{Ed_i^2}{(k-1)^2} = \frac{NCDIF_i}{(k-1)^2}, \quad (36)$$

or

$$NCDIF_i = (k-1)^2 (NCDIF_i^*), \quad (37)$$

where $NCDIF_i^*$ represents the index associated with the 0-1 scale and $NCDIF_i$ represents the index associated with the 1-*k* scale. Raju (1999b) suggested using the previously established .006 cutoff value for $NCDIF_i^*$. As such, the NCDIF cutoff for an item with *k* response categories is equal to $(k-1)^2(.006)$. Using this equation, the recommended cutoff for a 5-category item, for example, is .096. Raju (1999c) suggested this procedure may be conservative in assessing instances of practically significant DTF, however. In other words, this cutoff may suggest practically nontrivial DTF too often. Separate cutoffs for DTF were subsequently recommended.

According to Raju (1999c), a proposed cutoff for DTF should take into account

the NCDIF cutoff for a single item as well as the number of items in the test:

$$\text{DTF Cutoff} = n\,(\text{NCDIF Cutoff}), \tag{38}$$

where n represents the number of items in the test. Using Equation 38, the proposed cutoff for a 10-item dichotomously scored test would be 10 (.006) = .060, and the proposed cutoff for a test with 10 five-response category items would be 10 (.096) = .960. Empirically derived cutoff values yield different recommendations (Flowers et al., 1999; Meade et al., 2006).

Based on simulated data using Samejima's (1969) graded response model, Flowers et al. (1999) recommended a cutoff value of .016 for a 5-category item ($\alpha = .01$). This same cutoff value is used in assessing NCDIF and DTF. Also for a 5-category item, Meade et al. (2006) recommended cutoff values ranging from .006 to .0115 depending on sample size, and Bolt (2002) recommended cutoffs ranging from .009 to .010 depending on sample size and group differences in ability. As can be seen, these values are all substantially smaller than the cutoff values suggested using Equations 37 and 38. This again demonstrates that cutoff values may not generalize well across situations. Bolt (2002) also noted inflated Type 1 errors on individual items and stated there may be need for caution in applying the same empirical cutoff across all items in a test.

This dilemma concerning appropriate cutoff values for NCDIF and DTF indices poses an obstacle to using the DFIT framework in practice. Using Monte Carlo methods,

researchers have achieved success in generating cutoff values (Bolt, 2002; Flowers et al., 1999; Raju et al., 1995) that are appropriate for the specific conditions simulated in a given study. Based on these Monte Carlo investigations, there seems to be general support for the DFIT framework as a viable means for detecting differential functioning at the item and test levels. These cutoff values, however, are most likely not generalizable to other conditions found in practice. In determination of appropriate cutoff values, one may need to consider the effects of various combinations of factors, such as test length, sample size, IRT model, and the method of linking (Chamblee, 1998), that are likely to be found in practice. This is discouraging from a practical standpoint because typical practitioners may not have the expertise or time to produce empirically derived cutoff values using simulated data. As such, in practice Equations 37 and 38 have proven useful because they do not necessitate conducting a Monte Carlo study. Unfortunately these values are also not likely to be optimal for the above reasons. As such, across the literature, there has been a call for better procedures for assessing the statistical significance of DIF and DTF indices (Bolt, 2002; Flowers et al., 1999; Raju et al., 1995). In response to this call, Oshima et al. (2006) proposed the item parameter replication (IPR) method. The IPR method will be discussed in the pages to follow; however, it is first important to discuss how the above IRT based methods for examining DIF/DTF compare.

2.4 Comparison of the DFIT Framework and LR Test

Recent research (Bolt, 2002; Braddy et al., 2006; Meade & Lautenschlager, 2004), has shown interest in comparison of the DFIT framework and LR test. Given the current attention being given to these two methods, the LR test was chosen for comparison

purposes in the present study.

Prior research comparing DFIT and the LR test has concluded the LR test is more sensitive to detecting DIF than is the DFIT framework. For example, Meade and Lautenschlager (2004) compared these approaches using simulated 6-item, polytomous response data across three conditions of sample size (150, 500, and 1000) and three amounts of DIF (0, 2, or 4 items). The .096 fixed NCDIF cutoff proposed by Raju (1999b) was used as the criterion for DIF. Across conditions, results suggested DFIT tended to be more conservative and rarely identified items as exhibiting DIF. Subsequent research has also found the LR test to be more sensitive to detecting DIF (Bolt, 2002; Braddy et al., 2006), particularly with larger sample sizes. As Braddy et al. point out, this result is not surprising when one considers the different decision rules these two methods use for detecting DIF.

The LR test uses a chi-square distribution to examine the statistical significance of differences in item parameters. The chi-square statistic is known to show increased power at larger sample sizes. DFIT, on the hand, considers practical significance via the use of cutoff scores. In fact the reason Raju et al. (1995) incorporated the use of cutoff values into the DFIT framework was to address the over sensitivity of the chi-square tests to DIF. Use of this decision rule therefore increases the amount of DIF needed in an item before the item should be considered to exhibit DIF (Braddy et al., 2006). As a result, items containing a small amount of DIF may not be detected using DFIT. If this small amount of DIF is not practically significant, this is not problematic. However, if practically significant DIF is not detected due to the use of an inappropriate cutoff value for the given data, this may be problematic to the extent that such items are used in tests

and questionnaires and impact decisions made based on these measures.

Considering this, along with the difficulty in establishing appropriate empirically derived cutoff values, it would seem DFIT may not be advantageous from a practitioner's stand point. Introduction of the IPR method, however, once again provides a means for examining the statistical significance of DIF. Further, it gives practitioners a means of generating cutoff values that is easy to use in practice and allows for different cutoff values across items. This may increase the sensitivity of DFIT relative to the LR test. It is therefore important to reassess the efficacy of these two methods in light of these advances to DFIT. The IPR method will now be described in detail.

2.5 The Item Parameter Replication (IPR) Method

The IPR method (Oshima et al., 2006) provides a means for deriving study-based cutoff values for use in assessing differential functioning within the DFIT framework. This approach is different from the earlier approach in that the cutoff values established using the IPR method focus solely on testing statistical significance, whereas the earlier approach focused on practical significance. Both are important but might be considered separately.

The IPR method begins with estimates of item parameters and their variances and covariances. The item parameters and variances can be obtained from the output of an IRT calibration program, such as PARSCALE or MULTILOG. Unfortunately the covariances among item parameters are not available as part of the standard output. These values, however, can be derived using available information. Based on these initial estimates, a large number of replications of item parameters are then generated with the restriction that the expectation of the newly generated item parameters equals the

initial estimates of item parameters with the same variance and covariance structure. That is, any differences in the sets of estimates must be due to sampling error. This new method consists of nine major steps that will be described here for a single polytomous item *i*. The IPR method is identical for all items in a test.

1. Let the item parameter estimates be denoted by a column vector M_i. In the case of Samejima's (1969) GRM, a polytomous item with five response categories will be represented by one *a*-parameter and four *b*-parameters. Therefore, M_i will consist of five elements:

$$M_i = \begin{bmatrix} a_i \\ b_{i1} \\ b_{i2} \\ b_{i3} \\ b_{i4} \end{bmatrix}. \qquad (39)$$

Each item is also associated with a matrix consisting of the sampling variances and covariances of the item parameter estimates. Let this be represented as:

$$V_i = \begin{bmatrix} \sigma_a^2 & \sigma_{ab_{i1}} & \sigma_{ab_{i2}} & \sigma_{ab_{i3}} & \sigma_{ab_{i4}} \\ \sigma_{b_{i1}a} & \sigma_{b_{i1}}^2 & \sigma_{b_{i1}b_{i2}} & \sigma_{b_{i1}b_{i3}} & \sigma_{b_{i1}b_{i4}} \\ \sigma_{b_{i2}a} & \sigma_{b_{i2}b_{i1}} & \sigma_{b_{i2}}^2 & \sigma_{b_{i2}b_{i3}} & \sigma_{b_{i2}b_{i4}} \\ \sigma_{b_{i3}a} & \sigma_{b_{i3}b_{i1}} & \sigma_{b_{i3}b_{i2}} & \sigma_{b_{i3}}^2 & \sigma_{b_{i3}b_{i4}} \\ \sigma_{b_{i4}a} & \sigma_{b_{i4}b_{i1}} & \sigma_{b_{i4}b_{i2}} & \sigma_{b_{i4}b_{i3}} & \sigma_{b_{i4}}^2 \end{bmatrix}. \quad (40)$$

The correlation matrix (R_i) for the item parameters can then be derived from V_i :

$$R_i = \begin{bmatrix} 1 & \rho_{ab_{i1}} & \rho_{ab_{i2}} & \rho_{ab_{i3}} & \rho_{ab_{i4}} \\ \rho_{b_{i1}a} & 1 & \rho_{b_{i1}b_{i2}} & \rho_{b_{i1}b_{i3}} & \rho_{b_{i1}b_{i4}} \\ \rho_{b_{i2}a} & \rho_{b_{i2}b_{i1}} & 1 & \rho_{b_{i2}b_{i3}} & \rho_{b_{i2}b_{i4}} \\ \rho_{b_{i3}a} & \rho_{b_{i3}b_{i1}} & \rho_{b_{i3}b_{i2}} & 1 & \rho_{b_{i3}b_{i4}} \\ \rho_{b_{i4}a} & \rho_{b_{i4}b_{i1}} & \rho_{b_{i4}b_{i2}} & \rho_{b_{i4}b_{i3}} & 1 \end{bmatrix}. \quad (41)$$

Assuming R_i is positive definite, it can be expressed as the product of a triangular matrix (T_i) and its transpose (T_i') (Graybill, 1969). This is the Cholesky decomposition (Press, Flannery, Teukolsky, & Vetterling, 1992):

$$R_i = T_i' T_i . \quad (42)$$

2. Let *m* represent the number of response categories. Let X_{1i} represent a column vector of *m* elements, with each element drawn at random from one of *m* independent, standardized, and normally distributed populations. Let X_{2i} represent a second vector of *m* elements similarly drawn.
3. Using the T_i matrix in Equation 42, transform the two *X* vectors into two *Z* vectors as follows:

$$Z_{1i} = T_i' X_{1i}, \tag{43}$$

$$Z_{2i} = T_i' X_{2i}. \tag{44}$$

Each *Z* vector now represents a random element from an *m*-dimensional standardized multivariate normal distribution with a correlation structure for the *m* dimensions conforming to the correlation structure in the R_i matrix.

4. By definition, each element in the *Z* vectors is standardized in that its expectation and variance are 0 and 1, respectively. Each *Z* vector is now transformed to a *Y* vector so that the elements in the new vector will have the appropriate mean and variance as shown in the M_i and V_i matrices above. To achieve this transformation, let D_i represent a diagonal matrix consisting of the variances contained in V_i. Now, let

$$Y_{1i} = \sqrt{D_i} Z_{1i} + M_i , \quad (45)$$

$$Y_{2i} = \sqrt{D_i} Z_{2i} + M_i . \quad (46)$$

5. Vectors Y_{1i} and Y_{2i} represent two estimates of item parameters from two populations with identical item parameters. In other words, these vectors may represent item parameter estimates for the focal and reference groups when there is no DIF. Any differences in these estimates must be due to sampling error. Therefore, an NCDIF index for item *i* can be obtained with the help of the two *Y* vectors and the estimates of thetas for the focal group using the equations previously described.
6. Steps 1-5 can be replicated as many times as desired.
7. NCDIF values obtained from all replications can be rank ordered, and the 90th, 95th, 99th, 99.5th and 99.9th percentile rank scores are recorded to establish the cutoff values for alpha levels at .10, .05, .01, .005 and .001, respectively.
8. Once the alpha level is chosen, the cutoff associated with it is used as the cutoff for assessing statistical significance of the initial NCDIF value obtained for item *i*.
9. This process is repeated for all items in the test, thus potentially resulting in different cutoff values for different items.

Oshima et al. (2006) note several distinctions between the new IPR method and the method used to generate cutoff values in previous research (e.g., Bolt, 2002; Fleer, 1993; Flowers et al., 1999; Raju et al., 1995). First, a large number of replications of

item parameters are generated from the initial set of estimates obtained from the IRT calibration program. This eliminates the need for extra calibrations of item parameters, which is one of the most time consuming aspects of the previous method. Further, this offers a theoretical advantage in the sense that it is tailored to a particular dataset. As such, other unknown factors that may influence the error associated with parameter estimation are taken into account. Second, the distribution of NCDIF values is obtained for each item on a test and so it is possible to generate a cutoff value for each item. Lastly, the IPR method is easier to use from a practitioners standpoint because the procedure is implemented within a computer program. The only task practitioners have is to provide estimates of item parameters, their variances and covariances, and ability estimates for the focal (or reference) group. Providing this information is something one would need to do anyway to conduct a DFIT analysis. A Monte Carlo study was conducted to examine the efficacy of the IPR method in generating cutoff values for dichotomous items (Oshima et al.)

Oshima et al. (2006) manipulated the IRT model used to generate the simulated data, the proportion of test-wide DIF, the presence of uniform versus nonuniform DIF, and sample size. Further, ability distributions for the reference and focal groups were simulated to reflect conditions of matched ability and impact (i.e., the mean theta in the focal group was .50 standard deviations below the reference group). NCDIF cutoff values were obtained from 1,000 replications. Overall results suggest the IPR method performed well and provides a practical means of assessing differential functioning within the DFIT framework. Obtained IPR-based NCDIF cutoff values ranged from .0026 to .0118. The proposed cutoffs tended to be larger with smaller sample sizes,

as well as shorter test lengths. Further, more parameters in the IRT model, as well as larger standard errors of the item parameters resulted in larger cutoff values. Results further suggested the new IPR method outperforms the previously proposed cutoff (.006).

Looking at a condition with 40-items, a sample size of 1,000, and 10% DIF, the new method was able to correctly identify all DIF-items and did not incorrectly identify any non-DIF items (Oshima et al., 2006). The old cutoff (.006) missed one DIF item. This again emphasizes the need for study-based cutoff values and supports the IPR method. Research is needed to examine the IPR method for use with polytomous data.

2.6 Summary and Statement of Purpose

The IRT framework is well suited to detecting differential functioning of items or tests across sub-populations because of the property of invariance. That is, estimates of item characteristics are independent of the group of examinees, and estimates of examinee ability are independent of specific test items. As such, researchers have proposed several IRT-based methods for examining DIF. Two such methods under investigation in the current study are the DFIT framework (Raju et al., 1995) and LR test (Thissen et al., 1986; Thissen et al., 1988).

The DFIT framework offers several advantages to other DIF detection methods (Flowers et al., 1999). First, it offers a means by which to assess differential functioning at both the item and test-levels. Second, this method can be applied to both dichotomous and polytomous data, as well as unidimensional and multidimensional data. Third, DFIT offers two indices for assessing DIF: NCDIF and CDIF. Use of the NCDIF index assumes all items in the test except for the item under consideration contain no DIF (Raju et al., 1995). Although this assumption is commonly made by other measures of DIF as

well, it may not be plausible. CDIF, on the other hand, does not make this assumption. CDIF is additive in the sense that DTF equals the sum of the CDIF indices across the items in a test. This is advantageous in the sense that it provides a means by which to assess the overall effect of removing an item from a test.

Although the DFIT framework has been shown to be an effective mechanism for detecting DIF/DTF in IRT-based tests and questionnaires (e.g., Flowers et al., 1999; Oshima et al., 1997; Raju et al., 1995), researchers have indicated a need for better procedures for assessing the statistical significance of DIF and DTF indices. As a possible solution to this problem, Oshima et al. (2006) recently proposed the IPR method for determining NCDIF cutoff values for dichotomous items within the DFIT framework. A Monte Carlo investigation by Oshima et al. showed the IPR method is effective in maintaining acceptable Type 1 error and power rates. Although Raju et al. (2006) have described how the IPR-based NCDIF significance test extends to polytomous data, empirical research has been limited in scope. Fortmann et al. (2006) found support for the IPR method with polytomous data, but this was based on examination of just one condition of DIF. The first purpose of this study is to conduct a more comprehensive assessment of its efficacy in detecting DIF for polytomous items compared to previously recommended fixed cutoff values. Specifically, the following hypotheses are made:

Hypothesis 1: The IPR-based NCDIF test is more likely to detect true DIF than the NCDIF test with previously recommended fixed cutoffs.

Hypothesis 2: The IPR-based NCDIF test will have false positive rates close to the nominal significance level.

Further, the IPR method is compared to the LR test (Thissen et al., 1986; Thissen et al.,

1988). The LR test was chosen for comparison purposes because there has been growing interest in the recent literature on the use of these two procedures (Bolt, 2002; Braddy et al., 2006; Meade & Lautenschlager, 2004). This research has suggested the DFIT framework is less sensitive to detecting DIF than the LR test. It is important to note, however, that this result is based on the use of a fixed cutoff value across items. With the introduction of the IPR method to the DFIT framework, different NCDIF cutoff values may be derived across items and then used to examine statistical significance. This leads one to question whether previous conclusions regarding the sensitivity of the DFIT framework in comparison to the LR test remain true. Specifically, this research will address the question is true DIF more likely to be detected using the LR test or the IPR-based NCDIF test? Factors influencing each statistic's ability to detect DIF will also be examined. Since these aspects of the study are designed to be exploratory in nature, no formal hypotheses are stated.

CHAPTER 3

METHOD

3.1 Overview of the Monte Carlo Methodology

Monte Carlo research methods have become the preferred approach in DIF research aiming to assess the accuracy of different DIF detection methods. In this approach, data sets are simulated under a specified IRT model by generating examinee responses to a pre-determined number of test items using parameters obtained from either previously analyzed empirical test data or randomly selected computer generated distributions. Data sets are simulated for both a reference and focal group. The simulated data is then subjected to DIF analysis. This process is repeated many times to simulate the performance of the statistical test over repeated experiments.

The use of simulated test data in the current research application is highly desirable for two reasons. First, in DIF analyses conducted based on empirical test data it is not possible to know how many items are truly biased because true item parameters are not known. Using the estimated parameter values DIF analyses are conducted, but the results of this analysis only allow one to assess the degree of congruence across DIF detection methods. It is not possible to know if any of the methods accurately identified true DIF items. With Monte Carlo methods, on the other hand, the researcher knows the true item parameters that were used to simulate the data and thereby controls which items contain DIF. This allows one to compare the results of the DIF analysis to the true characteristics of the data. As a result, one can assess both the accuracy of a single DIF detection method and compare the congruence across methods. Second, Monte Carlo methods allow the researcher to manipulate other characteristics of the data as well, such

as sample size. One can then examine the factors moderating a statistics ability to detect DIF.

The Monte Carlo approach is not without limitations however. First, the data simulated is unrealistically clean. For example, it can be simulated to represent perfect unidimensionality, while in practice, test and questionnaire data is seldom perfectly unidimensional. In other words, while there may be one dominant factor being measured, there are often other factors accounting for some degree of the variability in performance. This provides reason to question whether results from Monte Carlo IRT analyses can be generalized to real test data containing mild violations of the statistical assumptions. Second, because the researcher determines what characteristics of the data to manipulate, it is possible the conditions tested are not realistic to practice. It is therefore important one models realistic situations. If not, this causes further concern regarding the generalizability of results. In the current research application, the strengths of this approach are felt to outweigh these limitations.

Because this study aims to assess the efficacy of the IPR methodology within the DFIT framework, a Monte Carlo research design is necessary as it provides the only means possible to identify true DIF items and measure the accuracy of DIF detection. Further, the conditions chosen for inclusion in this study were selected to represent conditions commonly found in practice, thereby minimizing this potential limitation.

3.2 Data Simulation

Polytomous item response data was generated based on Samejima's (1969) GRM using a Fortran 90 program developed for this purpose. Item parameters used to simulate the data were adapted from Flowers et al. (1999). Modifications to the parameters used

by Flowers et al. were made to avoid instances of extremely low response frequencies in the top two response categories. The modified item parameters are depicted in Tables 1 and 2. These items represent a wide range of magnitudes of DIF, as indicated by the true NCDIF values in Table 2. True NCDIF was computed according to Equation 27 using the known population parameters for each item, assuming an ability distribution with mean equal to zero and standard deviation equal to one.

To begin data simulation, the true ability of *N* examinees was generated from a normal distribution with mean and standard deviation specified by the condition. Ability scores were then generated using the International Mathematical and Statistical Library (IMSL; 1984) pseudo-random number generator DRNNOR. Next, a probability distribution was generated for each examinee on each item, indicating the probability of responding in each response category. The probability distribution was generated as a function of ability according to Samejima's (1969) GRM, using the item parameters specified by the condition. Based on this probability distribution, a random response was then generated for each examinee for each item using the IMSL DRNGDT routine.

In Monte Carlo DIF research, multiple replications of the data are simulated for each condition of the study. DIF analyses are then conducted separately for each replication of the data and results are averaged across replications to arrive at an overall index of accuracy for each condition. In the present study, each condition was replicated 100 times. This number of replications is consistent with much previous research (e.g., Bolt, 2002; Cohen, Kim & Wollack, 1996; Kim & Cohen, 1998b; Meade et al., 2006) and marks the maximum number of replications found in previous research.

3.3 Manipulated Factors

The current investigation included two sample sizes (500 and 1,000). Sample size was equal across the reference and focal groups. These sample sizes were chosen to replicate those used by Fleer (1993) and Raju et al. (1995). The smaller sample size reflects the minimum sample size recommended for accurate recovery of item parameters in the GRM (Ankenmann & Stone, 1992; Reise & Yu, 1990). A sample size of 1,000 is thought to represent a fairly large sample size relative to what is typically found in practice.

Past research suggests test length has minimal to no effect on DIF detection errors (Chamblee, 1998; Flowers et al., 1999; Oshima et al., 2006). Therefore, test length was held constant at 40-items, representing a moderate test or questionnaire length (Fleer, 1993). All items consisted of five response categories.

Further, to assess the effect of group differences in theta levels (impact) on detection of DIF, two different θ distributions were simulated for the focal group. In the first condition the focal and reference groups were sampled from populations with equal θ distributions. Specifically, they were randomly sampled from a normal distribution with mean equal to zero and standard deviation equal to one [N(0,1)]. This condition is referred to as the no impact condition. In the second condition, the impact condition, the focal group was sampled from a normal distribution with mean equal to negative one and standard deviation equal to one [N(-1,1)]. This resulted in the focal group having lower θ values than the reference group and reflects the size of the Black-White racial group difference (one standard deviation) commonly found on tests of general cognitive ability.

The generating item parameters (Flowers et al., 1999) result in simulating four

proportions of test-wide DIF (0%, 5%, 10%, and 20%) and two conditions of direction of DIF (Unidirectional and Balanced-Bidirectional). DIF was modeled by adding a constant to the *a* and/or *b* parameters of the focal group. The differences between focal and reference group item parameters (computed as focal group value minus reference group value) are presented in Table 2. So, for the 40-item test, 0, 2, 4, or 8 items were embedded with DIF. In the unidirectional conditions all items favored the reference group. In the balanced-bidirectional conditions, items favoring the reference group were balanced with items favoring the focal group. The parameters used further resulted in items generated to simulate uniform DIF ($a_{iR} = a_{iF}$ and $b_{iR} \neq b_{iF}$) and nonuniform DIF ($a_{iR} \neq a_{iF}$, either with $b_{iR} \neq b_{iF}$ or $b_{iR} = b_{iF}$). Nonuniform DIF items were only embedded in the 20% DIF condition. Figure 5 illustrates the simulation design. This resulted in 28 conditions. Each condition was replicated 100 times, resulting in 2,800 datasets per group.

3.4 Data Analysis

Parameter Estimation. Concurrent estimation of item and ability parameters for the reference and focal groups using Samejima's (1969) GRM was completed prior to DFIT and LR DIF analyses. Parameters were estimated using marginal maximum likelihood estimation as implemented within the software package MULTILOG (Thissen, 1991). A prior distribution, with mean equal to zero and standard deviation equal to 1.5, was imposed on the *b*-parameters. The prior distribution was set such that all plausible parameter values were included within 3 SDs of the mean. Four known non-DIF items (Items 1, 11, 21, and 31) were chosen as anchor items and linked the metric for parameter estimation. The four anchor items were chosen to represent a range of *a*- and *b*-

parameter values.

The number of anchor items is a topic that has received much attention in the literature, particularly as it relates to the LR test. Traditionally, one item is studied at a time and all other items serve as the anchor set. According to Wang and Yeh (2003) however, the appropriateness of this approach, in terms of power and Type 1 error, decreases as the average signed area between groups (i.e., DIF) increases. Their results supported use of 1, 4, or 10-anchor items over the "all other items" approach. Stark et al. (2006) recommended use of a single anchor item and also found this to be more effective in detecting DIF than the "all other items" approach. Other research has shown that the recovery of item parameters from concurrent estimation becomes more accurate as the number of anchor items increases (Kim & Cohen, 1998a, 2002). Similarly, the power of DIF detection is higher with larger numbers of anchor items (Wang & Yeh). Therefore, we decided to use four anchor items. Four anchor items allowed reasonable parameter recovery and DIF detection (Wang & Yeh), while still maintaining a decent number of studied items (36).

DFIT. Parameter estimates used for DFIT analyses were taken from the compact (or baseline) model of each replication of each condition. The standard MULTILOG output provides only the parameter estimates and their standard errors, and does not provide the parameter covariances required for the IPR method. Asymptotic estimates of the covariance matrix of IRT parameter estimates were obtained from the inverse of the Fisher Information matrix, however. A Fortran program computed parameter covariances using the method described in Li and Lissitz (2004) and Morris, Fortmann, and Oshima (2007).

The information matrix is equal to -1 times the expected value of the Hessian (i.e., the matrix of second partial derivatives) of the likelihood function. For a 3-category polytomous model, the information matrix would have elements corresponding to the *a*-parameter and 2 *b*-parameters, e.g.

$$\mathbf{I} = \begin{bmatrix} I_{aa} & I_{ab_1} & I_{ab_2} \\ I_{ab_1} & I_{b_1b_1} & I_{b_1b_2} \\ I_{ab_2} & I_{b_1b_2} & I_{b_2b_2} \end{bmatrix}. \tag{47}$$

Let P_k be used as shorthand for the item response function indicating the probability of a response in category k as a function of ability (θ) and a set of item parameters ξ, and let $Q_k = 1\text{-}P_k$. Li and Lissitz (2004) show that, for a m-category polytomous model, the element of **I** related to any two parameters ξ_s and ξ_t is given by,

$$I_{\xi_s\xi_t} = -E\left(\frac{\partial^2 L}{\partial\xi_s\partial\xi_t}\right) = N\int\left\{\sum_{k=1}^{m}\left(\frac{1}{P_kQ_k}\right)\left(\frac{\partial P_k}{\partial\xi_s}\frac{\partial P_k}{\partial\xi_t}\right)\right\}\phi(\theta)d\theta\,. \tag{48}$$

A numerical approximation of the integral can be computed by the Gauss-Hermite quadrature,

$$I_{\xi_s\xi_t} = N\sum_{q=1}^{Q}\left[\left\{\sum_{k=1}^{m}\left(\frac{1}{P_{kq}Q_{kq}}\right)\left(\frac{\partial P_{kq}}{\partial \xi_s}\frac{\partial P_{kq}}{\partial \xi_t}\right)\right\}A\left(X_q\right)\right], \tag{49}$$

where X_q is one of Q quadrature points, $A(X_q)$ is the associated quadrature weight, and P_{kq} represents the item response function evaluated at $\theta=X_q$.

The derivatives of the boundary response functions with respect to a and b_k are given by:

$$\frac{\partial P_k^*}{\partial a} = P_k^* Q_k^* D\left(\theta - b_k\right) \tag{50}$$

and

$$\frac{\partial P_k^*}{\partial b_k} = P_k^* Q_k^* \left(- Da\right) \quad . \tag{51}$$

The category response function is the difference between two adjacent boundary response functions,

$$P_k = P_{k-1}^* - P_k^* \quad . \tag{52}$$

Therefore, the derivatives of the category *k* response function are

$$\frac{\partial P_k}{\partial a} = P^*_{k-1}Q^*_{k-1}D(\theta - b_{k-1}) - P^*_k Q^*_k D(\theta - b_k), \tag{53}$$

$$\frac{\partial P_k}{\partial b_{k-1}} = P^*_{k-1}Q^*_{k-1}(-Da), \tag{54}$$

$$\frac{\partial P_k}{\partial b_k} = P^*_k Q^*_k(Da), \tag{55}$$

The derivatives will be 0 for all other boundary location parameters (e.g., b_{k+1}).

Substituting these values for the derivatives in Equation 49 yields,

$$I_{aa} = N\sum_{q=1}^{Q}\left\{\sum_{k=1}^{m}\left(\frac{1}{P_{kq}Q_{kq}}\right)\left[P^*_{k-1,q}Q^*_{k-1,q}D(X_q - b_{k-1}) - P^*_{kq}Q^*_{kq}D(X_q - b_k)\right]^2\right\}A(X_q), \tag{56}$$

$$I_{ab_k} = N\sum_{q=1}^{Q}\left\{\begin{array}{l}\left(\frac{1}{P_{kq}Q_{kq}}\right)\left[P^*_{k-1,q}Q^*_{k-1,q}D(X_q - b_{k-1}) - P^*_{kq}Q^*_{kq}D(X_q - b_k)\right]\left[P^*_{kq}Q^*_{kq}(Da)\right] \\ +\left(\frac{1}{P_{k+1,q}Q_{k+1,q}}\right)\left[P^*_{kq}Q^*_{kq}D(X_q - b_k) - P^*_{k+1,q}Q^*_{k+1,q}D(X_q - b_{k+1})\right]\left[P^*_{kq}Q^*_{kq}(-Da)\right]\end{array}\right\}A(X_q), \tag{57}$$

$$\begin{aligned} I_{b_k b_k} &= N\sum_{q=1}^{Q}\left\{\left(\frac{1}{P_{kq}Q_{kq}}\right)\left[P^*_{kq}Q^*_{kq}(Da)\right]^2 + \left(\frac{1}{P_{k+1,q}Q_{k+1,q}}\right)\left[P^*_{kq}Q^*_{kq}(-Da)\right]^2\right\}A(X_q) \\ &= D^2a^2N\sum_{q=1}^{Q}\left\{\left[P^*_{kq}Q^*_{kq}\right]^2\left[\left(\frac{1}{P_{kq}Q_{kq}}\right) + \left(\frac{1}{P_{k+1,q}Q_{k+1,q}}\right)\right]\right\}A(X_q) \end{aligned}, \tag{58}$$

and

$$I_{b_k b_{k+1}} = N\sum_{q=1}^{Q}\left\{\left(\frac{1}{P_{k+1,q}Q_{k+1,q}}\right)\left[\left(P^*_{kq}Q^*_{kq}\right)\left(P^*_{k+1,q}Q^*_{k+1,q}\right)\left(-D^2a^2\right)\right]\right\}A(X_q)$$
$$= -D^2a^2N\sum_{q=1}^{Q}\left\{\left(\frac{1}{P_{k+1,q}Q_{k+1,q}}\right)\left(P^*_{kq}Q^*_{kq}\right)\left(P^*_{k+1,q}Q^*_{k+1,q}\right)\right\}A(X_q) \quad (59)$$

For non-adjacent categories (j not equal to 1), $I(b_k,b_{k+j}) = 0$. The covariance matrix is then found by taking the inverse of **I**.

DFIT analyses were then conducted using the DFIT7 (Raju, Oshima, & Wolach, 2005) program which provides NCDIF values for all items using Equation 27. Cutoff values for NCDIF were generated utilizing the IPR method. As previously described, the IPR method generates a large number of replications of item parameters. NCDIF values obtained from all replications are rank-ordered, and the 99th percentile rank score establishes the cutoff value for an alpha level of .01. Consistent with Oshima et al. (2006), the present study used 1000 replications. Items with initial NCDIF values larger than the cutoff value were identified as exhibiting DIF at the designated level of significance.

It should be noted the DFIT analysis is focused on detection of item-level DIF using the NCDIF index. This is consistent with the purpose of the IPR method. Further, this index offers the most direct comparison to the LR test.

LR Test. As described in the previous chapter, the LR test involves comparison of a series of nested models. To begin, the compact (or baseline) model was estimated in which the four anchor items were constrained to be equal and all other items were free to vary. Estimation of the compact model resulted in a baseline likelihood value of model

fit. Next, a series of augmented models were estimated. In each augmented model, item parameters for the studied item, in addition to the anchor items, were constrained to be equal. So, for each condition of the 40-item test, one compact model and 36 augmented models were estimated for each replication. This process produced a likelihood value of model fit for each augmented model. The likelihood ratio was then computed for each studied item and significance testing conducted to identify the items exhibiting DIF (see Equations 18 and 19). The.01 level of significance was again used.

Assessment of DIF Methods. Analyses were conducted to compare DIF-detection rates using IPR-based NCDIF cutoff values, two previously recommended NCDIF cutoff values, and the LR test. The .096 and .016 fixed NCDIF cutoff values proposed by Raju (1999b) and Flowers et al. (1999), respectively, were chosen for comparison purposes. Two statistics were examined: True positive (TP) rates and false positive (FP) rates.

A TP is defined as an item with true DIF that is correctly identified as exhibiting DIF. TP rates (or power) for each item were determined by computing the number of replications in which DIF was detected divided by the total number of replications (100). The mean of TP rates was then computed across items to arrive at overall indices of power for each condition.

A FP is defined as a non-DIF item that is incorrectly identified as exhibiting DIF. FP rates (or Type 1 error rates) for each replication were determined by computing the total number of non-DIF items falsely identified as DIF items divided by the total number of non-DIF items. The mean of FP rates was then computed across the 100 replications simulated for each condition to arrive at overall indices of Type 1 error.

Additional analyses were conducted to examine the degree of similarity between

methods across simulated conditions. Consistent with the second purpose of this study, this analysis focused specifically on agreement between the IPR-based NCDIF and LR tests. Agreement was first computed taking into consideration only those items with true-DIF. Specifically, the number of true-DIF items identified with significant DIF in common to the two methods was divided by the total number of true-DIF items identified by that pair (in common and not in common). Agreement was also computed across all items. Similar to the above computation, the number of items identified with significant DIF in common to the two methods was divided by the total number of items identified by that pair. Looking across all items, Cohen's Kappa was computed to determine how much better this result was than what would be expected by chance.

CHAPTER 4

RESULTS

Prior to examination of DIF detection, descriptive statistics were computed to better understand the magnitude of IPR-based NCDIF cutoff values relative to the previously recommended fixed cutoffs. Further, for descriptive purposes, estimated NCDIF values were compared to the true NCDIF associated with the generating item parameters. Finally, the accuracy of DIF detection using IPR-based NCDIF item cutoffs, previously recommended fixed cutoffs, and LR procedures was examined. The results of these analyses are presented in the sections to follow.

4.1 IPR-Based NCDIF Cutoff Values

Within the IPR method, a unique NCDIF cutoff value is generated for each item. For each condition, the average IPR-based NCDIF cutoff (α=.01) across all items was computed to provide a sense of the magnitude of cutoffs across conditions. These values are presented in Table 3. In addition, Table 3 presents the range of item cutoffs observed in each condition averaged across replications. As shown, there was variability in item cutoffs across conditions. Further analysis was conducted to better understand the factors driving this variability.

Looking across all conditions of DIF, a polynomial regression analysis was conducted regressing mean NCDIF item cutoffs on true *a*- and *b*-parameters and sample size. Item parameters and sample size accounted for 80% of the variance in cutoffs, F (5, 1114) = 889.28, $p < .001$. Incorporating impact in to the regression model explained an additional 6% of the variance, F (6, 1113) = 1125.76, $p < .001$. Sample size (Std. b = -.595, $p \leq .001$) and a curvilinear trend for the *a*-parameter (Std. b = 1.423, $p \leq .001$) had

the strongest influence on IPR-based NCDIF item cutoffs. Although still significant at $p \leq .001$, the *b*-parameter (Std. b = -.220) and impact (Std. b = .243) had relatively weaker, linear relationships with IPR-based NCDIF cutoffs.

Since the influence of item parameters on cutoffs can largely be attributed to the *a*-parameter, the curvilinear relationship between *a*- parameter values and mean NCDIF cutoffs for the null condition with N=1000 and no impact is presented for illustrative purposes (see Figure 6). As shown, cutoff values were notably larger for items with small *a*-parameters. Cutoffs decreased sharply up to a=1.36, where values then began to level off. Results were nearly identical to this illustration across all conditions studied. Further, results with regards to sample size indicate cutoffs tended to be higher (regardless of the condition of DIF or impact) when the sample size was 500 than when the sample size was 1000. Referring to the mean cutoffs presented in Table 3, the mean difference across sample sizes ranged from .012 to .016.

An unexpected result of this analysis is the significant effect found for impact. IPR-based NCDIF item cutoffs are derived from the variance and covariance of parameter estimates. The method used to compute the covariance matrix of parameter estimates relied only on focal group parameter estimates and sample size (see Equations 47-59); therefore, impact should have no effect on item cutoffs. Results, however, indicate cutoffs tended to be higher in conditions with impact. This may be explained in terms of the effect shifting the θ distribution has on parameter values. In conditions with impact, the focal group θ distribution is shifted downward such that the mean is equal to negative one. Item parameters are scaled relative to the θ distribution. When scaled relative to a lower θ value, there will be a corresponding increase in the *b*-parameters. In

other words, item difficulty is greater for those of lower ability. It is possible the observed differences in item cutoffs across conditions of impact are due to this parameter shift but in an unknown way. One possible explanation is that item parameters tend to be estimated most accurately when the θ distribution is centered near the item location parameter. Shifting the θ distribution of the focal group such that the mean is lowered by one standard deviation relative to the reference group may result in θ distributions that are less aligned with the item location, resulting in more estimation error.

In addition to examining IPR-based NCDIF item cutoffs across studied conditions, comparisons were made to the previously recommended fixed cutoff values of .096 and .016. Item cutoffs were consistently substantially lower than the .096 fixed cutoff. Across all studied conditions, the largest average item cutoff was equal to .075 (see Table 3). IPR-based NCDIF item cutoffs were more similar to the .016 fixed cutoff, though. Figure 7 illustrates item cutoffs for the null condition with no impact. When the sample size was 500, average IPR-based NCDIF item cutoffs were consistently greater than or approximately equal to .016. When the sample size was 1000, the average IPR-based NCDIF item cutoffs tended to be less than or approximately equal to .016. Similar results to those illustrated in Figure 7 were observed across conditions of DIF, as well as in conditions with impact.

4.2 Estimated NCDIF Values

For descriptive purposes, the mean NCDIF value across replications was computed for each true-DIF item in each condition to provide a sense of the similarity between estimated and true NCDIF values (see Tables 4 and 5). In the no impact conditions, estimated NCDIF values closely mirrored the true NCDIF values presented in

Table 2. Estimated NCDIF values in conditions with impact, however, were notably smaller than values reported in Table 2. This was true across sample sizes. In addition, the average NCDIF value across all non-DIF items (excluding the anchor items) in each condition was computed. As would be expected, these values were near zero across all conditions (range .004-.008).

It is not surprising that the estimated NCDIF values in the conditions with impact were smaller than the true NCDIF values presented in Table 2. The true NCDIF values were computed assuming a θ distribution with mean equal to zero and standard deviation equal to one. As described in the previous section, though, in conditions with impact the focal group θ distribution was shifted downward such that the mean was equal to negative one. Because the distribution of θ plays a role in the definition of NCDIF, the true NCDIF values in conditions with impact would actually be different than those presented in Table 2.

4.3 Detection of DIF

Consistent with the research hypotheses, examination of the methods of DIF detection was done in two stages. First, the IPR method was compared to previously recommended fixed cutoffs. Specifically, true positive (TP) and false positive (FP) rates were examined across IPR-based NCDIF cutoffs, Raju's (1999b) .096 fixed NCDIF cutoff, and Flowers et al's. (1999) .016 fixed NCDIF cutoff. Next, comparisons were then made between the DFIT and LR approaches.

Comparison to Fixed NCDIF Cutoffs. Tables 6 and 7 present overall TP and FP rates for each condition across DIF detection methods. Examination of TP rates across the three alternate NCDIF tests provides partial support for the first hypothesis. IPR-based

cutoffs were more likely to detect true DIF than the previously recommended .096 fixed cutoff value. The IPR-based and .016 fixed cutoffs, however, were comparable in their ability to detect DIF.

In the 5% DIF conditions (Conditions 1 and 4), the three DFIT approaches demonstrated perfect DIF detection across sample sizes and conditions of impact. IPR-based and .016 fixed cutoffs maintained TP rates equal to 1.0 in the 10% DIF conditions (Conditions 2 and 5), whereas TP rates for the .096 fixed cutoff dropped, ranging from .78 to .97 across sample sizes and conditions of impact. All three DFIT approaches had lower power to detect DIF in the 20% DIF conditions (Conditions 3 and 6). The .096 fixed cutoff again demonstrated the lowest power, with TP rates ranging from .39 to .68. TP rates for the .016 fixed cutoff ranged from .78-.88, and TP rates for the IPR-based cutoffs ranged from .61-.83. Within these conditions, the IPR approach demonstrated slightly less power at *N*=500. For the IPR-based cutoffs, TP rates were lowest in the *N*=500, impact conditions. Sample size had minimal impact on the fixed cutoff approaches.

Item-level TP rates were examined (see Tables 8-11) to gain better understanding of which items explained the observed differences in DIF detection across the three DFIT approaches. Results indicate each of the DFIT approaches had difficulty detecting items with non-uniform DIF embedded on just the *a*-parameter (Items 20 and 40 in Condition 3, and Items 5, 6, 15 and 16 in Condition 6). These items account for the decreased power observed in the 20% DIF conditions (Conditions 3 and 6). Looking at the .096 fixed cutoff, TP rates for these items were near zero across sample sizes and conditions of impact. TP rates of the IPR-based and .016 fixed cutoffs for Item 40 in Condition 3 were

near zero across conditions, but this item contained the smallest magnitude of DIF (true NCDIF=.001) of any simulated item. Power to detect the other items using IPR-based and .016 fixed cutoffs was better but still generally low. The IPR method was able to detect non-uniform DIF of greater magnitude, though, such as that observed in Items 15 and 35 of Condition 3. Power to detect these non-uniform DIF items decreased only in the condition with *N*=500 and impact.

Next, FP rates were examined. Referring again to Tables 6 and 7, results provide only modest support for the second hypothesis. IPR-based FP rates were close but consistently slightly higher than the nominal significance level (α=.01), ranging from .03 to .06. At *N*=1000, IPR-based FP rates were slightly higher than FP rates produced using the .016 fixed cutoff, which ranged from .01 to .04. At *N*=500, however, FP rates for the .016 fixed cutoff increased significantly, ranging from .10 to .14. FP rates for the IPR-based test, on the other hand, remained consistent across sample sizes. The .096 fixed cutoff maintained the tightest control of Type 1 errors, producing no FPs across items.

In summary, both IPR-based and .016 fixed cutoffs outperformed the .096 fixed cutoff recommended by Raju (1999b). The IPR method had TP rates comparable to the .016 fixed cutoff but maintained tighter control of Type 1 errors (FPs) in the smaller sample size conditions. Overall, therefore, the IPR method showed the best performance of the three DFIT approaches. As such, comparisons to the LR test focused solely on the IPR method.

Comparison to LR Test. The IPR-based NCDIF and LR tests were compared across three indices of DIF detection: TP rates, FP rates and agreement. Referring again to Tables 6 and 7, the LR test demonstrated strong power to detect DIF. TP rates across all

conditions studied were consistently high, ranging from .95 to 1.00. Looking at the 5 and 10% DIF conditions (Conditions 1, 2, 4, and 5), results were comparable to those observed under the IPR method. Notable differences between methods occurred only in the 20% DIF conditions (Conditions 3 and 6). Recall that in these conditions TP rates for the IPR-based NCDIF test ranged from .61 to .83. This is notably lower than the lowest value of .95 observed using the LR procedure. Examination of item-level TP rates again helps us to understand these results.

As can be seen in Tables 8-11, across many items there were minimal differences in TP rates between methods. The LR test, however, had substantially more power to detect items with small magnitudes of non-uniform DIF embedded on just the *a*-parameter (Items 20 and 40 in Condition 3, and Items 5, 6, 15, and 16 in Condition 6). TP rates for these items using the LR test ranged from .91 to 1.00 across conditions of sample size and impact, whereas TP rates using the IPR-based NCDIF test ranged from .00 to .99, with the vast majority of TP rates being less than .73. It is not clear why the IPR method had more power to detect Item 16 in Condition 6 in the large sample size, no impact condition (TP=.99; see Table 8) than it did to detect other items with similar types of DIF.

It should also be noted there were instances where the IPR method had markedly greater power to detect DIF than the LR test. Referring to Table 11, Item 10 in Conditions 2 and 3 had TP rates of .98 and .95, respectively, under the IPR method. TP rates using the LR test were .88 and .79, respectively. It is not clear why the LR test showed decreased power to detect this item relative to the IPR method in conditions with *N*=500 and impact. This does indicate, though, there will be circumstances where the

IPR method may demonstrate greater power than the LR test.

Examination of FP rates in Tables 6 and 7 suggests the LR test has slightly greater control of Type 1 errors than the IPR method. FP rates for the LR test were closer to the nominal significance level (α=.01), ranging from .02 to .05. Similar to the IPR method, the LR test maintained consistent control of Type 1 errors across sample sizes and conditions of impact.

To assess the similarity in items detected by the IPR method and LR test, agreement was first computed as the proportion of all items detected as DIF by either method that were detected as DIF by both methods. This was computed for each replication of each condition, and the average across replications provided an overall index of agreement for each condition. Due to the way agreement was defined, agreement could not be computed in replications where neither method identified any items as exhibiting DIF. This only occurred in the null conditions, and the percentage of such replications ranged from 17 to 35% (see Table 12). Agreement, in these instances, was computed based on all other replications. Agreement values in the null conditions ranged from .11 to .24, indicating the two methods generally did not falsely identify the same set of items.

Tables 13 and 14 present the average agreement between IPR and LR tests examined across all items for the studied conditions. When the sample size was 1000, the average agreement ranged from .63 to .80. When the sample size was 500, the average agreement ranged from .52 to .81. In the 20% DIF conditions, agreement was lower when N=500. Results across all other conditions were nearly identical for large and small sample sizes, though.

Tables 13 and 14 further present Cohen's Kappa computed and averaged across replications for each condition as an index of how much better agreement is than what would be expected by chance. Across all conditions, observed Kappa values ranged from .60 to .80. Landis and Koch (1977) suggest Kappa values ranging from .00 to .20 represent slight agreement, .21 to .40 represents fair agreement, .41 to .60 represents moderate agreement, .61 to .80 represents substantial agreement, and Kappa values above .81 represent almost perfect agreement. These results, therefore, demonstrate moderate to substantial agreement between the IPR method and LR test, supporting these results as better than what would be expected by chance.

A final measure of agreement was computed focusing only on true-DIF items. Specifically, the proportion of true-DIF items detected in common to the two procedures was computed and averaged across replications. These results are presented in Tables 15 and 16. When the sample size was 1000, the average agreement ranged from .70 to 1.00 across conditions. Agreement was lowest in the 20% DIF conditions, ranging from .70 to .82. Near perfect agreement was observed in all other conditions (range .98 to 1.00). Similar results were observed when N=500. In the smaller sample size, agreement within the 20% DIF conditions ranged from .59 to .76, and agreement across all other conditions ranged from .96 to 1.00. There was only one instance of perfect agreement at the smaller sample size. Cohen's Kappa could not be reported for agreement indices based on only true-DIF items. Cohen's Kappa is computed as $(HR - HR_e)/(1 - HR_e)$, where HR represents the observed hit rate (i.e., the number of correct identifications) and HR_e represents the expected hit rate. In replications where both methods demonstrated perfect DIF detection (i.e., TP=1.00), the expected hit rate will always equal one resulting in a

zero in the denominator. Given the large number of replications across conditions with TPs equal to 1.00, this statistic was not computed.

CHAPTER 5

DISCUSSION

This study examined the efficacy of the IPR method (Oshima et al., 2006) for determining cutoff values for polytomous items within the DFIT framework. It was hypothesized IPR-based NCDIF item cutoffs would be more likely to detect true-DIF than previously recommended fixed cutoffs and would have false positive rates close to the nominal significance level. In response to recent literature suggesting the DFIT framework to be overly conservative and less likely to detect DIF than the LR test (e.g., Bolt, 2002; Braddy et al., 2006; Meade & Lautenschlager, 2004), power and Type 1 error rates for IPR-based NCDIF and LR tests were also compared, and overall inter-method agreement was examined.

A Monte Carlo research design was used in which data for a 40-item test with 5 polytomous response categories was simulated under Samejima's (1969) GRM. Factors potentially influencing each statistic's ability to detect DIF were examined. Specifically, sample size, focal group ability distribution (or impact), proportion of test-wide DIF, and direction of DIF were manipulated.

Examination of IPR-based cutoffs supports the need to derive item-level indices. Results show that cutoffs varied across items both within and across conditions. The majority of the variance in item cutoffs was explained by the item parameters and sample size, with item discrimination having the largest overall effect. As found with dichotomous data (Oshima et al., 2006), the IPR method produced larger item cutoffs at the smaller sample size. This result makes sense when one considers how the IPR method works. Within the IPR method, the larger the standard errors of the item

parameter estimates, the larger the item cutoff will be (Oshima et al.). In essence, the IPR method takes into account unknown factors influencing the standard error of parameter estimates when generating a cutoff value. So, the increased cutoff values observed are likely due to potentially poorer parameter estimation at *N*=500.

The finding of higher IPR-based cutoffs in conditions with impact was unexpected. This same result, however, was found using the 3-PL dichotomous IRT model (Oshima et al., 2006). The differences observed are likely due to increased estimation errors resulting from the change in item difficulty parameters that result from shifting the focal group ability distribution, but it is not clear why this same effect would not have been observed with the 1 and 2-PL dichotomous models (Oshima et al.). It would be interesting for future research to further examine this effect with both dichotomous and polytomous response data.

Comparison of approaches within the DFIT framework shows the IPR method outperformed previously recommended fixed cutoffs. Examination of TP and FP rates across the three DFIT approaches clearly indicates the .096 fixed cutoff is overly conservative and unlikely to detect smaller magnitudes of DIF. As previously concluded, this approach trades too much power for increased control of Type 1 errors (Meade et al., 2006). Both the IPR-based and .016 fixed cutoffs were effective in identifying DIF, with the exception of the 20% DIF conditions. Both approaches were less sensitive to the non-uniform DIF of extremely small magnitudes embedded within these conditions. These same conditions were problematic is past research as well (Flowers et al., 1999; Oshima et al., 2006).

Across conditions, both the IPR-based and .016 fixed cutoffs had FP rates slightly

higher than the nominal significance level. In interpreting FP results, though, one must determine what constitutes acceptable control of Type 1 errors. A FP rate equal to .05, for example, indicates that on average 5% of items are expected to be falsely identified as having DIF. So, in a 40-item test there will on average be 2 items incorrectly identified as having DIF. No attempts are made, however, to control for the Type 1 error inflation associated with conducting multiple tests of significance across items. The expected Type 1 error across multiple independent tests can be computed as $[1 - (1 - p)^n]$, where p represents the nominal significance level and n represents the number of independent tests. In the current study, the number of items tested ranged from 28 in the 20% DIF Conditions to 36 in the Null Conditions. So, at $\alpha = .01$, the expected Type 1 error rate in the 20% DIF conditions would equal $[1 - (1 - .01)^{28}]$ or .25. In the Null Conditions, the expected Type 1 error rate would equal $[1 - (1 - .01)^{36}]$ or .30. The largest observed FP rate for the IPR method was .06. Comparatively, this is much lower than what could be and so these results indicate that although the IPR method may falsely identify DIF more often than desired it still maintains reasonable control of Type 1 errors. FP rates for the .016 fixed cutoff were also consistently lower than these expected values, but the degree of control over Type 1 errors varied across sample sizes. In the larger sample size the .016 fixed cutoff had slightly better control of Type 1 errors than the IPR method, but in the smaller sample size it was overly sensitive and demonstrated FP rates well above those of the IPR method. As such, the IPR method is more generalizable and demonstrated the strongest overall performance within the DFIT framework.

It is not surprising that the IPR-based and .016 fixed cutoffs performed similarly when the sample size was 1000. The item parameters used to simulate data for the

current study were adapted from Flowers et al. (1999). As such, the characteristics of the data used in the current study were the same as the conditions under which .016 was empirically determined to be the optimal cutoff value. Flowers et al., however, did not include a sample size of 500 in their analysis. In conditions with *N*=500, IPR-based cutoffs increased and use of the .016 fixed cutoff resulted in high FP rates. This provides support for the idea that cutoffs are not generalizable across situations and again supports the need for a procedure to derive cutoffs that are tailored to a specific data set.

Because the IPR method outperforms prior fixed cutoffs, it stands to reason that current results change our understanding of the relative efficacy of the DFIT and LR approaches. Past research has generally concluded the LR test is more sensitive and provides a better measure of DIF than the DFIT framework (Bolt, 2002; Braddy et al., 2006; Meade & Lautenschlager, 2004). In the current study, however, IPR-based NCDIF and LR tests showed comparable TP and FP rates across a number of the conditions and items studied. With the exception of the 20% DIF conditions, there was high agreement between methods in detecting true-DIF items. The increased agreement observed across most conditions in the current study may be explained by how DIF was measured and manipulated in the past research.

First, some of the past research has used the .096 fixed NCDIF cutoff as the criterion for DIF detection (Braddy et al., 2002; Meade & Lautenschlager, 2004). As was found in the current study, this does indeed provide an overly conservative measure of DIF, and it is therefore not surprising DFIT was less likely to detect DIF in these studies. Research that has derived empirical cutoffs (Bolt, 2002), on the other hand, still supported the LR test but also noted DFIT may be preferable in some circumstances,

such as with larger sample sizes where the LR test may become too sensitive to model misfit. In addition, Braddy et al's. results were based on actual test data. So, while results suggested the LR test to be more sensitive, it cannot be known whether the items identified had true-DIF or whether they were false positives.

Further, Meade and Lautenschlager (2004) simulated differences on individual *b*-parameters. For example, in one condition of DIF studied, only the largest *b*-parameter differed between groups for each DIF item. This simulated a situation in which the most extreme response option was less likely to be used by one group than the other, while no differences occurred for the other response options. In the current investigation, however, all four *b*-parameters were consistently simulated to differ between groups. It has been acknowledged in the literature (Flowers et al., 1999) that polytomous DFIT allows for different patterns of DIF on the *b*-parameters to cancel each other out at the item level. This can occur because within the polytomous DFIT framework parameter values for each response category are combined to yield the expected score, or item true score function. Examination of DIF is then based on comparison of the item true score functions across groups, whereas the LR test compares the individual *b*-parameters. It therefore makes sense DIF simulated in this manner could be less likely to be detected by DFIT.

It is also worth noting that Meade and Lautenschlager (2004) acknowledged their study simulated relatively small amounts of DIF. Using empirical cutoffs, DFIT has demonstrated adequate power in subsequent research simulating moderate amounts of DIF on separate *b*-parameters (Meade et al., 2006). This suggests prior results may be more an artifact of the magnitudes of DIF simulated than of the patterns of DIF simulated.

Nonetheless, in the current study, the LR test performed well across conditions, outperforming the IPR method in the 20% DIF conditions. This can be attributed to the ability to detect non-uniform DIF. The LR test showed greater power to detect non-uniform DIF of small magnitudes. The IPR method, on the other hand, was less sensitive to DIF of this nature. The fact that DFIT is less sensitive to small magnitude non-uniform DIF is only a concern if these small amounts of DIF would be practically meaningful, and determination of what magnitude of DIF translates into practically meaningful outcomes is open for discussion. This poses a challenge for practitioners in deciding what method of studying DIF is most appropriate and suggests the need for a measure of effect size.

In the meantime, the optimal solution may be to examine DIF using multiple methods. If multiple DIF procedures arrive at the same conclusion regarding DIF, this may increase confidence that items detected contain non-trivial amounts of DIF. This approach may not be realistic, however, given practical time constraints. Alternatively, practitioners should choose a method of studying DIF taking into consideration the purpose of the test or questionnaire and consequences associated with failing to identify a true-DIF item (Braddy et al., 2006; Meade & Lautenschlager, 2004). In employment testing situations where the conclusions made based on test results have potential legal implications, practitioners may be wise to take a highly stringent approach and remove items with any magnitude of DIF. In this case a more powerful measure of DIF, such as the LR test, would always be preferred. With organizational attitude surveys, on the other hand, the consequences of including items with very small amounts of DIF are not as great, and practitioners may be willing to sacrifice power to detect small magnitudes of

DIF. In this case, either the DFIT framework or LR test would be an acceptable choice.

While the results of the current study tend to favor the LR test, it is important to remember the DFIT framework offers several practical advantages over the LR test. First, DFIT provides a means to study differential functioning at both the item and test levels. Whereas DFIT allows practitioners to statistically demonstrate equivalence at the test level, the LR test does not and practitioners must assume that removal of DIF items results in an unbiased measure. Further, DFIT offers two indices of DIF: NCDIF and CDIF. The NCDIF index is most comparable to the LR test, and both of these examine DIF for a single item at a time, assuming all other items to be DIF-free. Quite often, this assumption may not be true. The CDIF index, however, takes into account compensating bias across items and allows practitioners to understand the impact of each item on overall DTF. This is useful because practitioners often make decisions based on overall test or questionnaire results. If the DIF present in two items cancels out at the test level and only test level results will be interpreted, then the presence of DIF on these individual items is not of concern. Any item level group differences are lost in the aggregate. From a test or questionnaire development standpoint, this means the individual items can be retained in the measure, again assuming decisions will not be made based on item level results. Were the LR test to be used, on the other hand, all DIF items would automatically be discarded. In practice, the pool of items is often limited and so it is beneficial to have a means by which to be more selective in deciding which items will have the largest impact on reducing DTF.

It is also important to remember that this new method offers several advantages over the fixed cutoff DFIT approach (Oshima et al., 2006). First, the IPR method

provides cutoffs that are tailored to a specific data set. This overcomes concerns that have been raised regarding the generalizability of fixed cutoffs across testing situations. Although practitioners could derive empirical cutoffs to achieve the same result, the process of doing so is time consuming and often not realistic. The IPR method, on the other hand, is implemented within available computer software making it more accessible to use in practice. This makes the IPR method a valuable addition to the DFIT framework. Second, the IPR method produces a distinct cutoff for each item. This is advantageous because the standard errors of item parameter estimates vary across items and so the optimal criteria for detecting DIF varies as well.

In the current study, sample size, focal group distribution (i.e., impact), proportion of test-wide DIF, and direction of DIF had little effect on TP and FP rates across methods of DIF detection. This is an important point because it means that results were fairly consistent across a wide range of conditions. This suggests results are likely to generalize beyond the current study.

The current study is not without limitations, however. First and foremost, the use of Monte Carlo research methods calls into question the generalizability of results. Data was simulated to represent a perfectly unidimensional test or questionnaire, as well as normally distributed test data. In practice, though, data seldom adheres to these conditions. It stands to reason that violations of these assumptions will impact DIF detection using IRT-based procedures. Future research should examine the robustness of IPR-based NCDIF and LR tests to violations of these statistical assumptions. Further, *b*-parameters were evenly spaced and DIF was always simulated on all four *b*-parameters. In practice, it is likely different patterns of DIF on the *b*-parameters, such as those

examined by Meade and Lautenschlager (2004), occur. Sensitivity to DIF of this nature using IPR-based cutoffs should be examined.

In addition, the GRM provided an exact fit to the simulated data used in the current study. This raises two issues. First, both the DFIT framework and LR test can be used with other polytomous IRT models. Future research should examine the efficacy of these approaches using data simulated according to different underlying response models, such as the generalized partial credit model (Muraki, 1990). Second, in practice, it is not known which IRT model best fits the data, and the model chosen to study DIF can only approximate the true underlying response process (Bolt, 2002). It is therefore prudent that one consider robustness to model misspecification in examining different DIF procedures. As noted before, Bolt found empirically derived DFIT cutoffs to be less affected by slight model misfit than the LR test. Sensitivity to model misfit increased at larger sample sizes, and therefore DFIT was suggested to be preferable to the LR test when working with large data sets. Future research should examine the impact of model misspecification using IPR-based cutoffs.

Further, the current study focused on sample sizes of 500 and 1000. IPR-based NCDIF tests demonstrated comparable TP and FP rates across both of these conditions. This "small" sample size, however, is still considerably larger than what can be commonly found in practice. Sample sizes in the current study were also equal for the reference and focal groups. In practice, this may often not be the case, and it is possible this could impact DIF detection. IPR-based NCDIF item cutoffs are based on the item parameter covariance matrix for the focal group and so are influenced only by the focal group sample size. A smaller or larger sample size for the reference group will affect the

sampling variance of the NCDIF statistic, though, and this will not be reflected in the cutoff. Oshima et al. (2006) studied the IPR method with dichotomous data under conditions of unequal sample sizes and still found support for the IPR-based cutoffs. Nonetheless, this should be taken into consideration with polytomous data as well.

The current study further used concurrent estimation procedures with four known non-DIF anchor items. The benefit of this approach is that the anchor items are known to be unbiased, and so there are no concerns regarding the potential impact of including a DIF-item in the anchor set on subsequent DIF analyses. However, this process has limited applicability to practice as it cannot be known which items are DIF-free in an empirical data set. Stark et al. (2006) described a procedure based on the LR test that could be used to identify a single anchor item in practice, but this requires multiple DIF analyses and comparison of results from different runs, making it very time consuming and labor intensive. Further, this does not extend to the DFIT framework.

It should again be noted that DFIT has typically not been implemented using concurrent estimation procedures either. This approach was felt to be advantageous in the current study because it offered a more even comparison to the LR test. In other words, it removed the possibility that any observed differences across methods could be due to the use of linking versus concurrent estimation procedures. Practitioners, however, are likely to continue using linking of reference and focal group parameter values. A logical next step would be to compare results within the DFIT framework across these techniques. If results are comparable, the above issue regarding selection of an appropriate anchor set becomes a moot point within DFIT because practitioners can continue to use iterative linking and achieve the same benefits observed in this study. This would be a notable

advantage over the LR method, in that selection of at least one anchor item will always be an issue using that approach.

Despite the advances of the IPR method, it does also introduce one unique challenge to using DFIT with polytomous data. Current polytomous IRT software packages do not provide estimates of item parameter covariances in the standard output. In the current study, alternative procedures to compute covariances from the item parameters had to be used (Li & Lissitz, 2004; Morris et al., 2007), and practitioners may not have the expertise to do this. This is not an issue with dichotomous data because the covariances needed are readily available in the output. Moving forward, it would be beneficial for vendors of IRT software to make this information more easily available for polytomous data as well.

In conclusion, this study supported the efficacy of the IPR method in detecting DIF, while also maintaining acceptable control of Type 1 error rates. Although questions regarding the generalizability of results are always an issue in research based on simulated data, the factors included in this study (sample size, group differences in ability, proportion of test-wide DIF, and direction of DIF) are felt to encompass conditions found in practice. These results are not meant to encompass all conditions found in practice, but rather serve as a starting point for understanding the value of the IPR method in working with polytomous data. Future research can build upon these results and incorporate other variables that may impact the efficacy of this procedure (e.g., IRT model misfit, smaller sample sizes), especially as it relates to other methods of DIF detection. Extension of this procedure to detection of DTF should also be addressed.

CHAPTER 6

TABLES

Table 1. Reference Group Item Parameters (Page 1 of 2)

Item	a	b_1	b_2	b_3	b_4
1	.55	-1.80	-.60	.60	1.80
2	.55	-1.80	-.60	.60	1.80
3	.73	-2.32	-1.12	.08	1.28
4	.73	-2.32	-1.12	.08	1.28
5[a]	.73	-1.80	-.60	.60	1.80
5[b]	.73	-2.30	-1.10	.10	1.30
5[c]	.73	-1.80	-.60	.60	1.80
6[a]	.73	-1.80	-.60	.60	1.80
6[b]	.73	-1.30	-.10	1.10	2.30
6[c]	1.23	-1.80	-.60	.60	1.80
7	.73	-1.80	-.60	.60	1.80
8	.73	-1.80	-.60	.60	1.80
9	.73	-1.28	-.08	1.12	2.32
10	.73	-1.28	-.08	1.12	2.32
11	1.00	-2.78	-1.58	-.38	.82
12	1.00	-2.78	-1.58	-.38	.82
13	1.00	-2.32	-1.12	.08	1.28
14	1.00	-2.32	-1.12	.08	1.28
15[a]	1.00	-2.32	-1.12	.08	1.28
15[b]	1.00	-2.57	-1.37	-.17	1.03
15[c]	1.00	-2.32	-1.12	.08	1.28
16[a]	1.00	-2.32	-1.12	.08	1.28
16[b]	1.00	-2.07	-.87	.33	1.53
16[c]	.50	-2.32	-1.12	.08	1.28
17	1.00	-1.80	-.60	.60	1.80
18	1.00	-1.80	-.60	.60	1.80
19	1.00	-1.80	-.60	.60	1.80
20	1.00	-1.80	-.60	.60	1.80
21	1.00	-1.80	-.60	.60	1.80
22	1.00	-1.80	-.60	.60	1.80
23	1.00	-1.80	-.60	.60	1.80
24	1.00	-1.80	-.60	.60	1.80

Table 1. Reference Group Item Parameters (Page 2 of 2)

Item	a	b_1	b_2	b_3	b_4
25[a]	1.00	-2.10	-.90	.30	1.50
25[b]	1.00	-1.28	-.08	1.12	2.32
25[c]	1.00	-2.10	-.90	.30	1.50
26[a]	1.00	-1.28	-.08	1.12	2.32
26[b]	1.00	-1.28	-.08	1.12	2.32
26[c]	1.00	-.78	.42	1.62	2.82
27	1.00	-1.28	-.08	1.12	2.32
28	1.00	-1.28	-.08	1.12	2.32
29	1.00	-1.90	-.70	.50	1.70
30[a]	1.00	-1.60	-.40	.80	2.00
30[b]	1.00	-.82	.38	1.58	2.78
30[c]	1.00	-.60	.60	1.80	3.00
31	1.36	-2.32	-1.12	.08	1.28
32	1.36	-2.32	-1.12	.08	1.28
33	1.36	-1.80	-.60	.60	1.80
34	1.36	-1.80	-.60	.60	1.80
35	1.36	-1.80	-.60	.60	1.80
36	1.36	-1.80	-.60	.60	1.80
37	1.36	-1.28	-.08	1.12	2.32
38	1.36	-1.28	-.08	1.12	2.32
39	1.80	-1.80	-.60	.60	1.80
40	1.80	-1.80	-.60	.60	1.80

[a]Parameters used in Conditions 1, 2, and 3.
[b]Parameters used in Conditions 4 and 5.
[c]Parameters used in Condition 6.

Table 2. Focal Group Item Parameters and True NCDIF Values by Condition

Item	a	b_1	b_2	b_3	b_4	Difference a	Difference b	True NCDIF
Condition 1								
5	.73	-.80	.40	1.60	2.80	----	+1.0	.47
10	.73	-.28	.92	2.12	3.32	----	+1.0	.42
Condition 2								
5	.73	-.80	.40	1.60	2.80	----	+1.0	.47
10	.73	-.78	.42	1.62	2.82	----	+.5	.11
15	1.00	-1.32	-.12	1.08	2.28	----	+1.0	.56
20	1.00	-1.30	-.10	1.10	2.30	----	+.5	.14
Condition 3								
5	.73	-.80	.40	1.60	2.80	----	+1.0	.47
10	.73	-.78	.42	1.62	2.82	----	+.5	.11
15	.50	-1.82	-.62	.58	1.78	-.5	+.5	.17
20	.50	-1.80	-.60	.60	1.80	-.5	----	.03
25	1.00	-1.10	.10	1.30	2.50	----	+1.0	.56
30	1.00	-1.10	.10	1.30	2.50	----	+.5	.14
35	.86	-1.30	-.10	1.10	2.30	-.5	+.5	.14
40	1.30	-1.80	-.60	.60	1.80	-.5	----	.00
Condition 4								
5	.73	-1.30	-.10	1.10	2.30	----	+1.0	.48
6	.73	-2.30	-1.10	.10	1.30	----	-1.0	.48
Condition 5								
5	.73	-1.30	-.10	1.10	2.30	----	+1.0	.48
6	.73	-2.30	-1.10	.10	1.30	----	-1.0	.48
15	1.00	-2.07	-.87	.33	1.53	----	+.5	.14
16	1.00	-2.57	-1.37	-.17	1.03	----	-.5	.14
Condition 6								
5	1.23	-1.80	-.60	.60	1.80	+.5	----	.01
6	.73	-1.80	-.60	.60	1.80	-.5	----	.01
15	.50	-2.32	-1.12	.08	1.28	-.5	----	.03
16	1.00	-2.32	-1.12	.08	1.28	+.5	----	.03
25	1.00	-1.10	.10	1.30	2.50	----	+1.0	.56
26	1.00	-1.78	-.58	.62	1.82	----	-1.0	.54
29	1.00	-1.40	-.20	1.00	2.20	----	+.5	.14
30	1.00	-1.10	.10	1.30	2.50	----	-.5	.13

Table 3. Average IPR-based NCDIF Cutoffs (α=.01)

		N=1000		*N*=500	
Impact	Condition	Mean Cutoff	Range	Mean Cutoff	Range
No Impact					
	Null Condition	.012	.006-.024	.025	.011-.048
	Unidirectional DIF Conditions				
	Condition 1	.012	.006-.025	.025	.011-.048
	Condition 2	.012	.006-.025	.025	.011-.049
	Condition 3	.013	.006-.025	.026	.012-.048
	Balanced Bidirectional DIF Conditions				
	Condition 4	.012	.006-.025	.025	.011-.049
	Condition 5	.012	.006-.025	.025	.011-.048
	Condition 6	.013	.006-.025	.025	.011-.048
Impact					
	Null Condition	.017	.007-.034	.032	.015-.064
	Unidirectional DIF Conditions				
	Condition 1	.016	.007-.035	.031	.014-.065
	Condition 2	.016	.007-.035	.032	.015-.066
	Condition 3	.017	.007-.034	.034	.015-.067
	Balanced Bidirectional DIF Conditions				
	Condition 4	.017	.007-.034	.032	.015-.067
	Condition 5	.017	.007-.034	.032	.015-.065
	Condition 6	.017	.007-.039	.033	.015-.075

Table 4. Mean NCDIF Values for *N*=1000

Condition	DIF Item	Mean NCDIF No Impact	Mean NCDIF Impact
Unidirectional DIF Conditions			
Condition 1	Item 5	.477	.370
	Item 10	.424	.269
Condition 2	Item 5	.475	.369
	Item 10	.119	.084
	Item 15	.582	.523
	Item 20	.149	.125
Condition 3	Item 5	.489	.380
	Item 10	.118	.084
	Item 15	.175	.082
	Item 20	.027	.042
	Item 25	.593	.470
	Item 30	.142	.115
	Item 35	.143	.095
	Item 40	.003	.004
Balanced-Bidirectional DIF Conditions			
Condition 4	Item 5	.485	.449
	Item 6	.498	.418
Condition 5	Item 5	.499	.445
	Item 6	.498	.431
	Item 15	.143	.161
	Item 16	.142	.134
Condition 6	Item 5	.010	.018
	Item 6	.011	.014
	Item 15	.034	.029
	Item 16	.035	.035
	Item 25	.569	.503
	Item 26	.558	.405
	Item 29	.153	.137
	Item 30	.124	.081

Table 5. Mean NCDIF Values for *N*=500

Condition	DIF Item	Mean NCDIF No Impact	Mean NCDIF Impact
Unidirectional DIF Conditions			
Condition 1	Item 5	.463	.370
	Item 10	.411	.275
Condition 2	Item 5	.504	.358
	Item 10	.128	.091
	Item 15	.585	.526
	Item 20	.152	.129
Condition 3	Item 5	.468	.349
	Item 10	.119	.083
	Item 15	.171	.081
	Item 20	.027	.046
	Item 25	.583	.477
	Item 30	.141	.112
	Item 35	.143	.096
	Item 40	.006	.007
Balanced-Bidirectional DIF Conditions			
Condition 4	Item 5	.497	.434
	Item 6	.515	.424
Condition 5	Item 5	.485	.451
	Item 6	.498	.415
	Item 15	.153	.149
	Item 16	.151	.148
Condition 6	Item 5	.011	.023
	Item 6	.013	.017
	Item 15	.039	.037
	Item 16	.034	.037
	Item 25	.582	.484
	Item 26	.554	.401
	Item 29	.156	.130
	Item 30	.144	.083

Table 6. Average True Positive and False Positive Rates for N=1000 (α = .01)

Impact	Condition	No. of DIF Items	IPR-Based NCDIF		.016 NCDIF Cutoff		.096 NCDIF Cutoff		LR Test	
			TP	FP	TP	FP	TP	FP	TP	FP
No Impact										
	Null Condition	0	----	.05	----	.02	----	.00	----	.02
	Unidirectional DIF Conditions									
	Condition 1	2	1.00	.05	1.00	.01	1.00	.00	1.00	.03
	Condition 2	4	1.00	.06	1.00	.02	.92	.00	.99	.04
	Condition 3	8	.82	.04	.84	.02	.68	.00	1.00	.02
	Balanced-Bidirectional DIF Conditions									
	Condition 4	2	1.00	.05	1.00	.02	1.00	.00	1.00	.03
	Condition 5	4	1.00	.04	1.00	.02	.97	.00	1.00	.03
	Condition 6	8	.81	.04	.77	.01	.47	.00	1.00	.02
Impact										
	Null Condition	0	----	.04	----	.03	----	.00	----	.03
	Unidirectional DIF Conditions									
	Condition 1	2	1.00	.04	1.00	.03	1.00	.00	.99	.03
	Condition 2	4	1.00	.03	1.00	.03	.79	.00	.98	.03
	Condition 3	8	.83	.04	.87	.03	.47	.00	.98	.03
	Balanced-Bidirectional DIF Conditions									
	Condition 4	2	1.00	.03	1.00	.03	1.00	.00	1.00	.03
	Condition 5	4	1.00	.05	1.00	.04	.93	.00	1.00	.04
	Condition 6	8	.70	.05	.79	.03	.40	.00	1.00	.04

Table 7. Average True Positive and False Positive Rates for $N = 500$ ($\alpha = .01$)

Impact	Condition	No. of DIF Items	IPR-Based NCDIF		.016 NCDIF Cutoff		.096 NCDIF Cutoff		LR Test	
			TP	FP	TP	FP	TP	FP	TP	FP
No Impact										
	Null Condition	0	----	.06	----	.12	----	.00	----	.03
	Unidirectional DIF Conditions									
	Condition 1	2	1.00	.04	1.00	.11	1.00	.00	1.00	.03
	Condition 2	4	1.00	.06	1.00	.13	.89	.00	.97	.03
	Condition 3	8	.77	.04	.84	.11	.65	.00	.97	.03
	Balanced-Bidirectional DIF Conditions									
	Condition 4	2	1.00	.04	1.00	.10	1.00	.00	.99	.03
	Condition 5	4	1.00	.06	1.00	.12	.94	.00	.98	.03
	Condition 6	8	.67	.06	.78	.11	.47	.00	.00	.04
Impact										
	Null Condition	0	----	.04	----	.13	----	.00	----	.02
	Unidirectional DIF Conditions									
	Condition 1	2	1.00	.03	1.00	.12	1.00	.00	.99	.03
	Condition 2	4	1.00	.03	1.00	.12	.78	.00	.96	.03
	Condition 3	8	.73	.03	.88	.14	.47	.00	.95	.03
	Balanced-Bidirectional DIF Conditions									
	Condition 4	2	1.00	.04	1.00	.13	1.00	.00	.99	.03
	Condition 5	4	1.00	.04	1.00	.13	.93	.00	.97	.03
	Condition 6	8	.61	.04	.82	.13	.39	.00	.99	.05

Table 8. Item Level True Positive Rates for *N*=1000, No Impact Conditions (α = .01)

Condition	Item	IPR-Based NCDIF	.016 NCDIF Cutoff	.096 NCDIF Cutoff	LR Test
Unidirectional DIF Conditions					
Condition 1	5	1.00	1.00	1.00	1.00
	10	1.00	1.00	1.00	1.00
Condition 2	5	1.00	1.00	1.00	1.00
	10	1.00	1.00	.70	.99
	15	1.00	1.00	1.00	1.00
	20	1.00	1.00	.97	.98
Condition 3	5	1.00	1.00	1.00	1.00
	10	1.00	1.00	.68	1.00
	15	1.00	1.00	.97	1.00
	20	.53	.72	.00	1.00
	25	1.00	1.00	1.00	1.00
	30	1.00	1.00	.90	1.00
	35	1.00	1.00	.89	1.00
	40	.03	.00	.00	1.00
Balanced-Bidirectional DIF Conditions					
Condition 4	5	1.00	1.00	1.00	1.00
	6	1.00	1.00	1.00	1.00
Condition 5	5	1.00	1.00	1.00	1.00
	6	1.00	1.00	1.00	1.00
	15	1.00	1.00	.94	1.00
	16	1.00	1.00	.93	1.00
Condition 6	5	.59	.18	.00	.99
	6	.16	.26	.00	.99
	15	.70	.86	.00	1.00
	16	.99	.89	.00	1.00
	25	1.00	1.00	1.00	1.00
	26	1.00	1.00	1.00	1.00
	29	1.00	1.00	.97	1.00
	30	1.00	1.00	.80	.98

Table 9. Item Level True Positive Rates for *N*=1000, Impact Conditions (α = .01)

Condition	Item	IPR-Based NCDIF	.016 NCDIF Cutoff	.096 NCDIF Cutoff	LR Test
Unidirectional DIF Conditions					
Condition 1	5	1.00	1.00	1.00	.99
	10	1.00	1.00	1.00	.99
Condition 2	5	1.00	1.00	1.00	1.00
	10	1.00	1.00	.35	.94
	15	1.00	1.00	1.00	1.00
	20	1.00	1.00	.79	.97
Condition 3	5	1.00	1.00	1.00	1.00
	10	1.00	1.00	.30	.96
	15	.99	1.00	.30	1.00
	20	.58	.93	.02	1.00
	25	1.00	1.00	1.00	1.00
	30	1.00	1.00	.71	.96
	35	1.00	1.00	.44	.98
	40	.03	.02	.00	.95
Balanced-Bidirectional DIF Conditions					
Condition 4	5	1.00	1.00	1.00	1.00
	6	1.00	1.00	1.00	1.00
Condition 5	5	1.00	1.00	1.00	1.00
	6	1.00	1.00	1.00	1.00
	15	1.00	1.00	.90	.98
	16	1.00	1.00	.82	1.00
Condition 6	5	.58	.45	.00	1.00
	6	.09	.31	.00	1.00
	15	.23	.78	.01	1.00
	16	.73	.80	.01	1.00
	25	1.00	1.00	1.00	1.00
	26	1.00	1.00	1.00	1.00
	29	1.00	1.00	.88	.97
	30	1.00	1.00	.29	1.00

Table 10. Item Level True Positive Rates for *N*=500, No Impact Conditions (α = .01)

Condition	Item	IPR-Based NCDIF	.016 NCDIF Cutoff	.096 NCDIF Cutoff	LR Test
Unidirectional DIF Conditions					
Condition 1	5	1.00	1.00	1.00	0.99
	10	1.00	1.00	1.00	1.00
Condition 2	5	1.00	1.00	1.00	1.00
	10	1.00	1.00	0.71	0.91
	15	1.00	1.00	1.00	1.00
	20	1.00	1.00	0.85	0.96
Condition 3	5	1.00	1.00	1.00	.98
	10	1.00	1.00	.66	.92
	15	1.00	1.00	.92	.99
	20	.00	.65	.00	.98
	25	1.00	1.00	1.00	1.00
	30	1.00	1.00	.85	.97
	35	1.00	1.00	.80	.97
	40	.00	.05	.00	.92
Balanced-Bidirectional DIF Conditions					
Condition 4	5	1.00	1.00	1.00	.99
	6	1.00	1.00	1.00	.99
Condition 5	5	1.00	1.00	1.00	.98
	6	1.00	1.00	1.00	1.00
	15	1.00	1.00	.86	.96
	16	1.00	1.00	.89	.98
Condition 6	5	.27	.24	.00	1.00
	6	.03	.32	.00	1.00
	15	.36	.86	.02	1.00
	16	.66	.78	.02	1.00
	25	1.00	1.00	1.00	1.00
	26	1.00	1.00	1.00	1.00
	29	1.00	1.00	.86	.99
	30	1.00	1.00	.83	.98

Table 11. Item Level True Positive Rates for *N*=500, Impact Conditions (α = .01)

Condition	Item	IPR-Based NCDIF	.016 NCDIF Cutoff	.096 NCDIF Cutoff	LR Test
Unidirectional DIF Conditions					
Condition 1	5	1.00	1.00	1.00	.98
	10	1.00	1.00	1.00	.99
Condition 2	5	1.00	1.00	1.00	1.00
	10	.98	.99	.43	.88
	15	1.00	1.00	1.00	.99
	20	1.00	1.00	.68	.96
Condition 3	5	1.00	1.00	1.00	.99
	10	.95	1.00	.34	.79
	15	.62	1.00	.30	1.00
	20	.23	.90	.05	.99
	25	1.00	1.00	1.00	.99
	30	1.00	1.00	.61	.94
	35	.98	1.00	.44	.99
	40	.04	.10	.00	.91
Balanced-Bidirectional DIF Conditions					
Condition 4	5	1.00	1.00	1.00	.99
	6	1.00	1.00	1.00	.99
Condition 5	5	1.00	1.00	1.00	.99
	6	1.00	1.00	1.00	.99
	15	1.00	1.00	.88	.93
	16	.99	1.00	.82	.97
Condition 6	5	.34	.53	.00	.99
	6	.01	.45	.00	.99
	15	.08	.82	.00	1.00
	16	.42	.77	.03	1.00
	25	1.00	1.00	1.00	1.00
	26	1.00	1.00	1.00	1.00
	29	1.00	1.00	.74	.96
	30	1.00	1.00	.35	.94

Table 12. Average Agreement Between IPR-based NCDIF and LR Tests for the Null Conditions

Null Condition	Percentage of Replications with No False Positives	Agreement*
$N = 1000$, No Impact	17%	.18
$N = 1000$, Impact	28%	.14
$N = 500$, No Impact	23%	.24
$N = 500$, Impact	35%	.11

*Computed across replications with at least one false positive across methods

Table 13. Average Agreement Between IPR-based NCDIF and LR Tests across All Items for *N*=1000

Impact	Condition	Agreement	Cohen's Kappa
No Impact			
	Unidirectional DIF Conditions		
	Condition 1	.70	.78
	Condition 2	.79	.85
	Condition 3	.74	.80
	Balanced-Bidirectional DIF Conditions		
	Condition 4	.69	.77
	Condition 5	.80	.85
	Condition 6	.72	.78
Impact			
	Unidirectional DIF Conditions		
	Condition 1	.71	.78
	Condition 2	.80	.86
	Condition 3	.74	.80
	Balanced-Bidirectional DIF Conditions		
	Condition 4	.74	.81
	Condition 5	.78	.83
	Condition 6	.63	.70

Table 14. Average Agreement Between IPR-based NCDIF and LR Tests across All Items for *N*=500

Impact	Condition	Agreement	Cohen's Kappa
No Impact			
	Unidirectional DIF Conditions		
	Condition 1	.71	.78
	Condition 2	.74	.81
	Condition 3	.67	.74
	Balanced-Bidirectional DIF Conditions		
	Condition 4	.69	.77
	Condition 5	.77	.82
	Condition 6	.58	.65
Impact			
	Unidirectional DIF Conditions		
	Condition 1	.69	.77
	Condition 2	.81	.86
	Condition 3	.65	.72
	Balanced-Bidirectional DIF Conditions		
	Condition 4	.72	.80
	Condition 5	.79	.85
	Condition 6	.52	.60

Table 15. Average Agreement Between IPR-based NCDIF and LR Tests across True DIF Items for *N*=1000

Impact	Condition	No. of DIF Items	Agreement
No Impact			
	Unidirectional DIF Conditions		
	Condition 1	2	1.00
	Condition 2	4	.99
	Condition 3	8	.82
	Balanced-Bidirectional DIF Conditions		
	Condition 4	2	1.00
	Condition 5	4	1.00
	Condition 6	8	.80
Impact			
	Unidirectional DIF Conditions		
	Condition 1	2	.99
	Condition 2	4	.98
	Condition 3	8	.82
	Balanced-Bidirectional DIF Conditions		
	Condition 4	2	1.00
	Condition 5	4	1.00
	Condition 6	8	.70

Table 16. Average Agreement Between IPR-based NCDIF and LR Tests across True DIF Items for *N*=500

Impact	Condition	No. of DIF Items	Agreement
No Impact			
	Unidirectional DIF Conditions		
	Condition 1	2	1.00
	Condition 2	4	.97
	Condition 3	8	.76
	Balanced-Bidirectional DIF Conditions		
	Condition 4	2	.99
	Condition 5	4	.98
	Condition 6	8	.66
Impact			
	Unidirectional DIF Conditions		
	Condition 1	2	.99
	Condition 2	4	.96
	Condition 3	8	.71
	Balanced-Bidirectional DIF Conditions		
	Condition 4	2	.99
	Condition 5	4	.97
	Condition 6	8	.59

CHAPTER 7

FIGURES

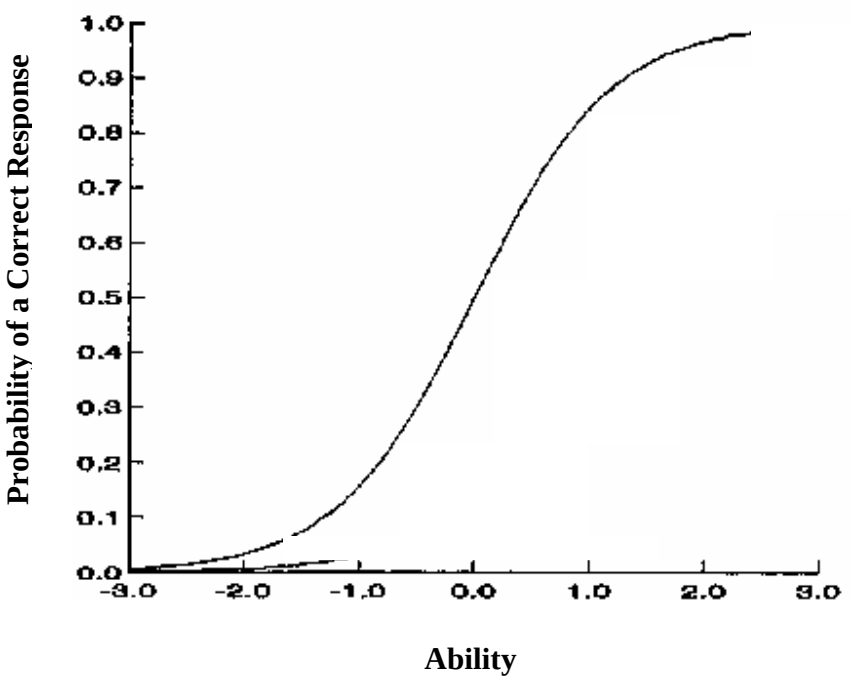

Figure 1. Example Item Response Function

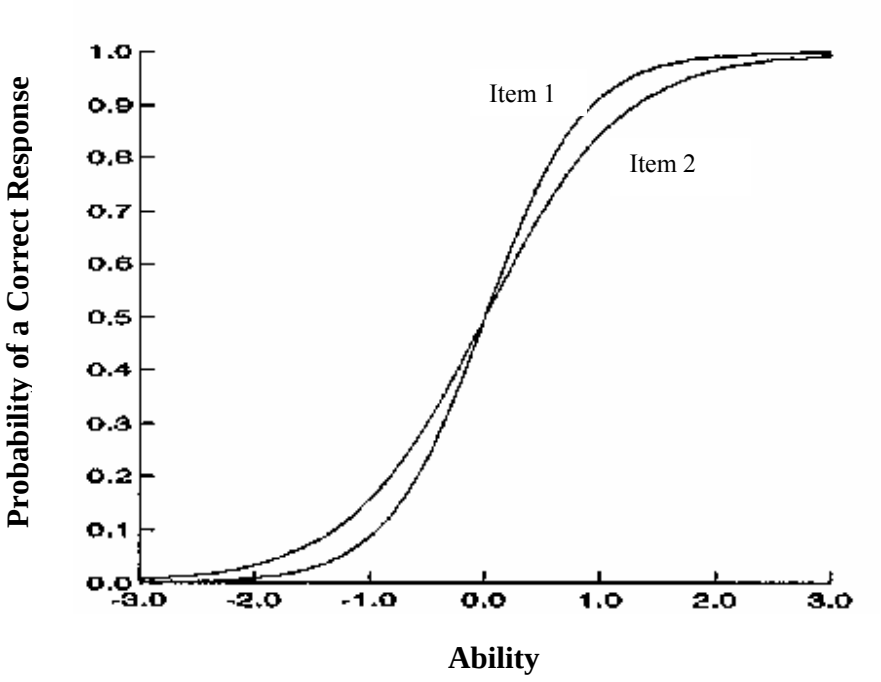

Figure 2. Example IRFs for Two Items under the 2-PL Model

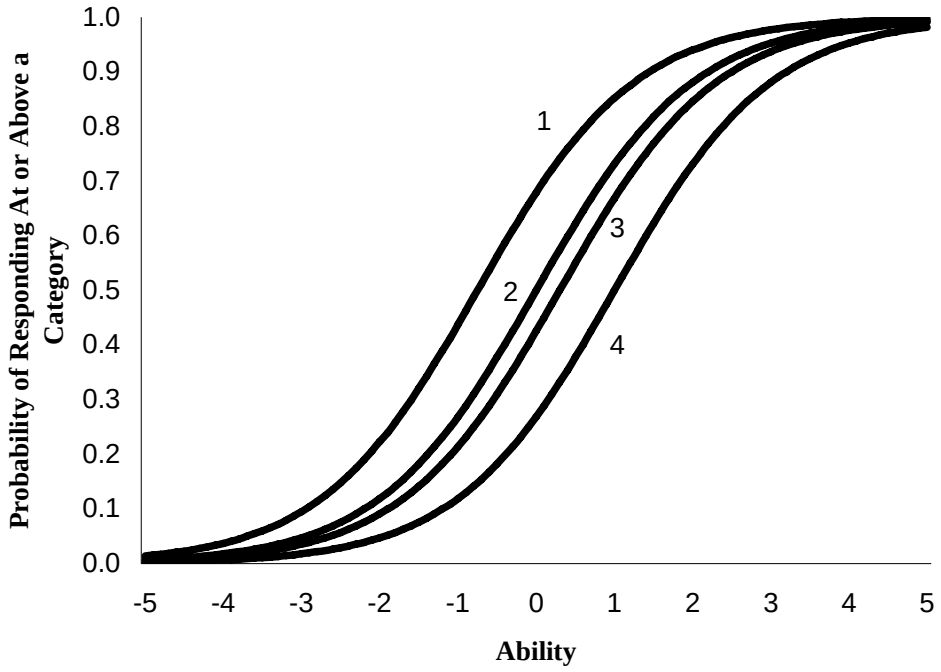

Figure 3. Examples of Boundary Response Functions for a 5-Category Item

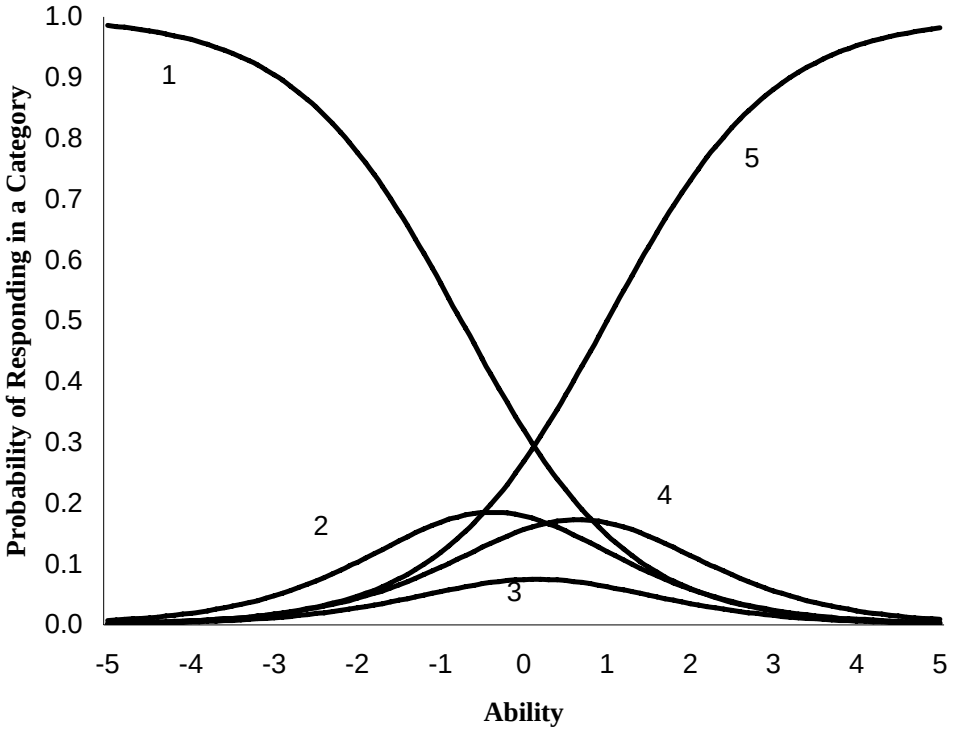

Figure 4. Examples of Category Response Functions for a 5-Category Item

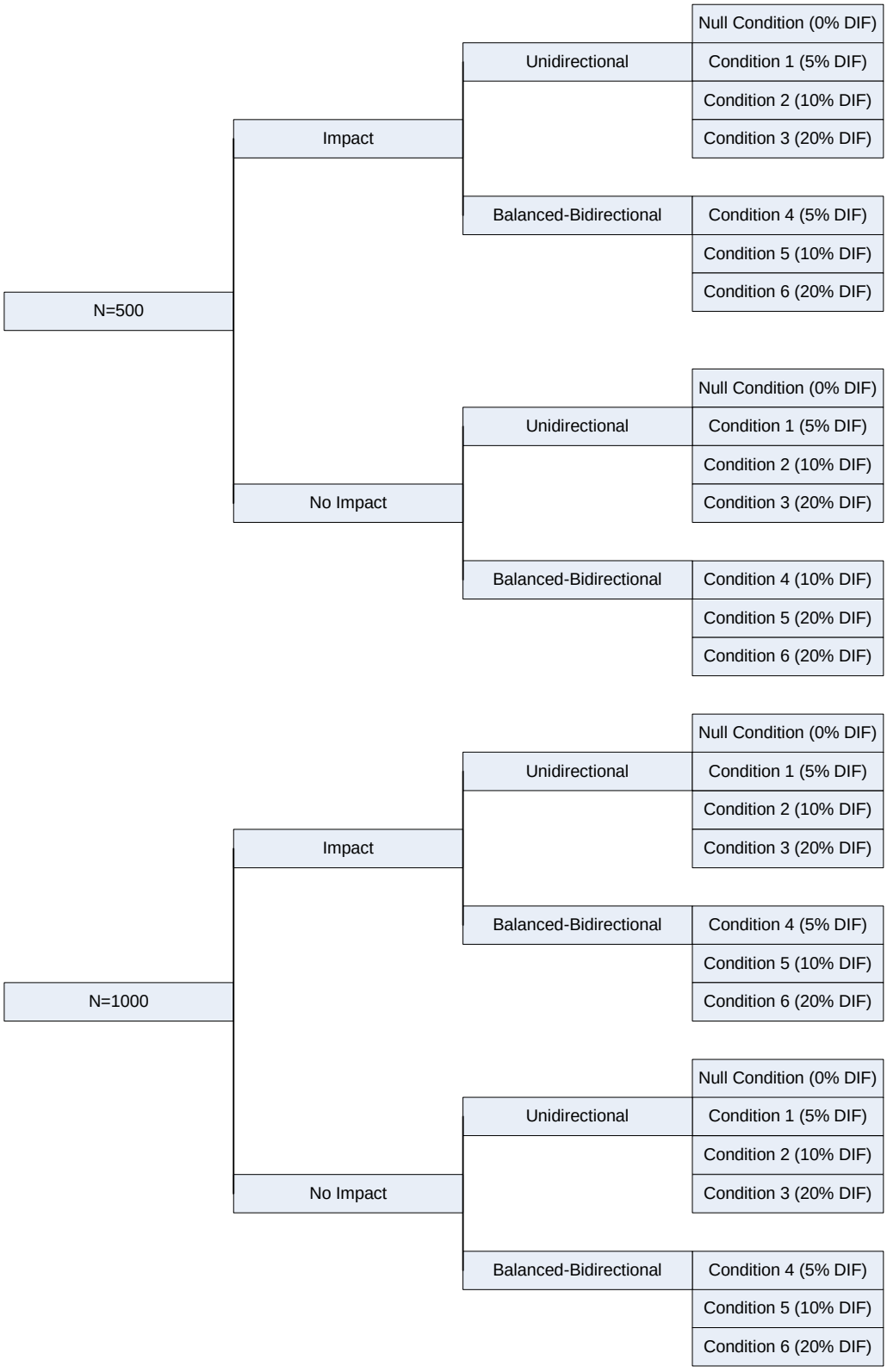

Figure 5. Simulation Design

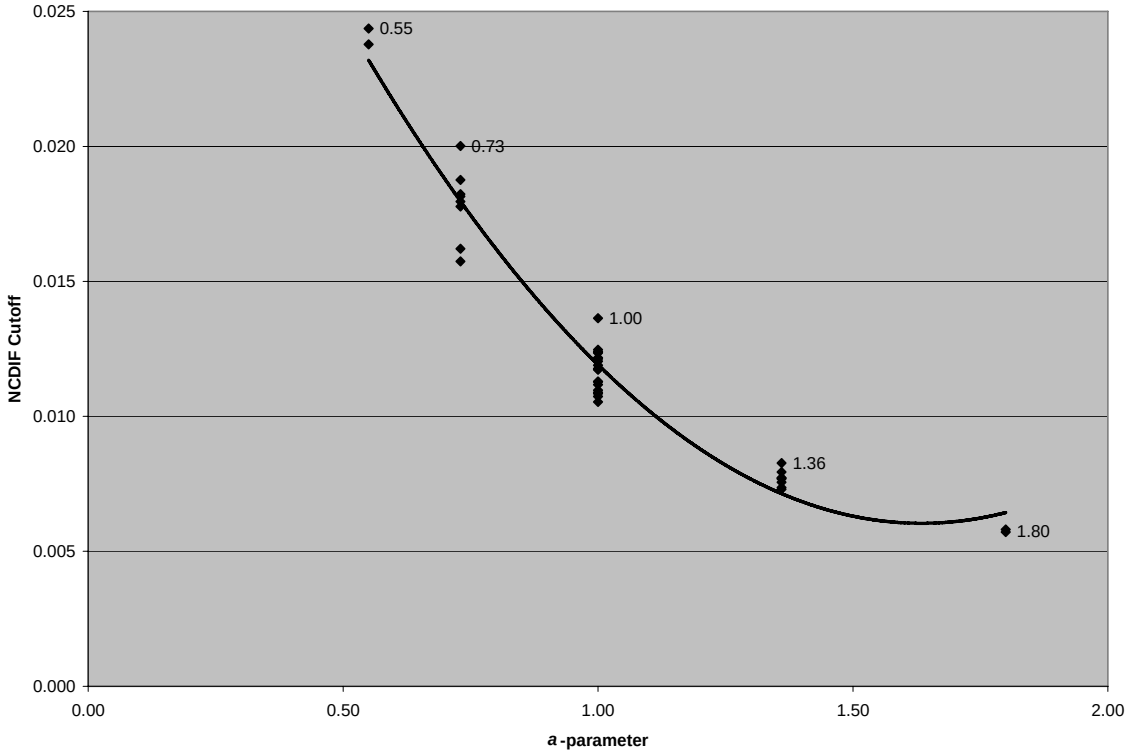

Figure 6. Mean NCDIF Cutoffs (α=.01) as a Function of *a*-Parameter Values

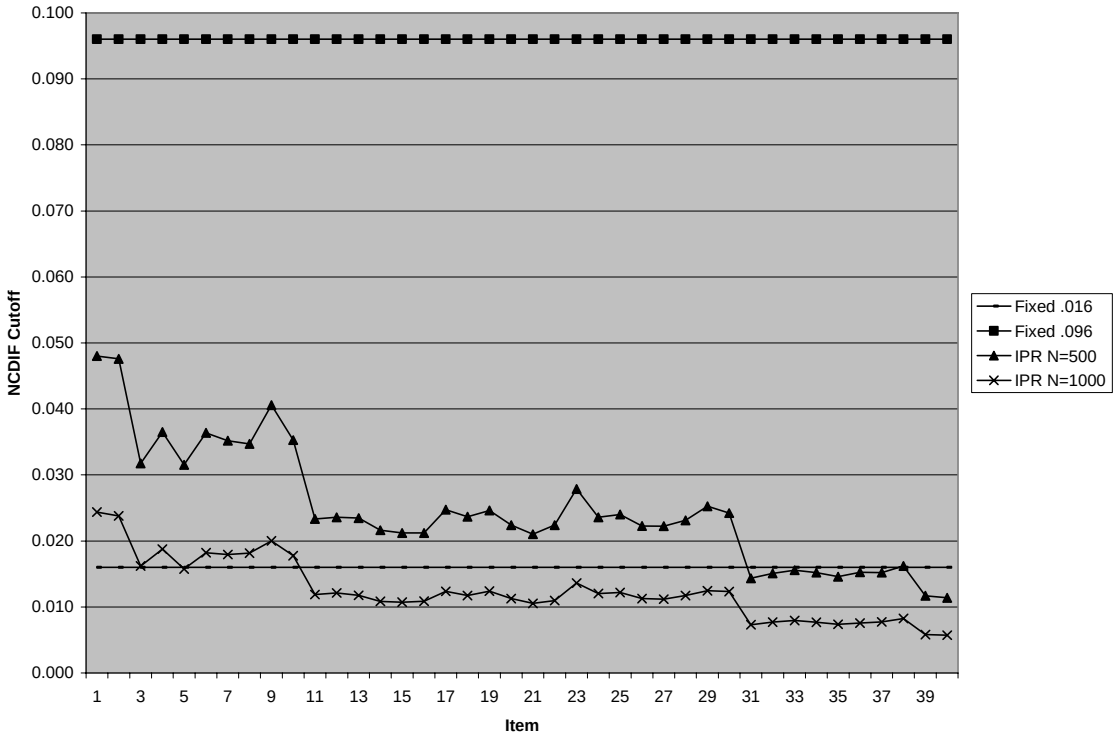

Figure 7. Mean NCDIF Item Cutoffs (α=.01) for the No Impact Null Condition

BIBLIOGRPAHY

Ankenmann, R. D., & Stone, C. A. (1992, April). More results on parameter recovery in the graded model using MULTILOG. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

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

Braddy, P. W., Meade, A. W., & Johnson, E. C. (2006, April). *Practical implications of using different tests of measurement invariance for polytomous measures.* Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Chamblee, M. C. (1998). *A monte carlo investigation of conditions that impact type 1 error rates of differential functioning of items and tests.* Unpublished doctoral dissertation, Georgia State University.

Cohen, A. S., Kim, A. –H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17* (4), 335-350.

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

Drasgow, F. & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70* (4), 662-680.

Fleer, P. F. (1993). *A monte carlo assessment of a new measure of item and test bias.* Unpublished doctoral dissertation, Illinois Institute of Technology.

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

Fortmann, K. A., Raju, N. S., Oshima, T. C., & Morris, S. B. (2006, May). *The item*

parameter replication method for detecting differential functioning in the DFIT framework. Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Graybill, F. A. (1969). *Introduction to matrices with applications in statistics*. Belmont, CA: Wadsworth Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals *of item response theory*. Thousand Oaks, CA: Sage.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26* (1), 3-24.

International Mathematical and Statistical Library. (1984). User's manual: IMSL library, problem-solving system for mathematical and statistical FORTRAN programming (Vol. 3, Ed. 9.2). Houston, TX.

Kim, S. –H., & Cohen, A. S. (1998a). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22* (2), 131-143.

Kim, S. –H., & Cohen, A. S. (1998b). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22* (4), 345-355.

Kim, S. –H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26* (1), 25-41.

Kim, S. –H., Cohen, A. S., & Park, T. –H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32* (3), 261-276.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Li, Y. H., & Lissitz, R. W. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement, 41*, 85-117.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47* (2), 149-174.

Meade, A. W., & Lautenschlager, G. J. (2004, April). *Same question, different answers:*

CFA and IRT approaches to measurement invariance. Symposium presented at the 19th annual conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Meade, A. W., Lautenschlager, G. J., & Johnson, E. C. (2006, April). *Alternate cutoff values and DFIT tests of measurement invariance.* Paper presented at the 21st annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

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

Morris, S. B., Fortmann, K. A., & Oshima, T. C. (2007, April). *An evaluation of the item parameter replication method for DFIT analysis of polytomous items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59-71.

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 (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43,* 1-17.

Ostini, R. & Nering, M. L. (2006). *Polytomous item response theory models.* Thousand Oaks, CA: Sage.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C: The art of scientific computing.* New York, NY: Cambridge University Press.

Raju, N. S. (1999a). *DFIT framework: Note 1. How CDIF and NCDIF indices are used in deleting items to make DTF non-significant.* Unpublished manuscript.

Raju, N. S. (1999b). *DFIT framework: Note 2. Suggested cut-offs for NCDIF*. Unpublished manuscript.

Raju, N. S. (1999c). *DFIT framework: Note 3. Cut-offs for DTF*. Unpublished manuscript.

Raju, N. S., Burke, M. J., & Normand, J. (1990). A new approach for utility analysis.

Journal of Applied Psychology, 75 (1), 3-12.

Raju, N. S. & Ellis, B. B. (2002). Differential item and test functioning. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 156-188). San Francisco: Jossey-Bass Inc.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factory analysis and item response theory. *Journal of Applied Psychology, 87* (3), 517-529

Raju, N. S., Oshima, T. C., Fortmann, K. A., Nering, M., & Wonsuk, K. (2006, February). *The new significance test for Raju's polytomous DFIT.* Paper presented at the New Directions in Psychological Measurement with Model-Based Approaches conference, Atlanta, GA.

Raju, N. S., Oshima, T. C., & Wolach, A. (2005). *Differential functioning of items and tests (DFIT): Dichotomous and polytomous* [Computer program]. Chicago: Illinois Institute of Technology.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19,* 353-368.

Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27* (2), 133-144.

Samejima, F. (1969). Estimation *of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa City, IA: Psychometric Society.

Stark, S., Chernyshenko, O., S., & Drasgow, F. (2006). *Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy*. Journal of Applied Psychology, 91 (6), 1292-1306.

Thissen, D. (1991). *MULTILOG User's guide* (Version 6.0). Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99* (1), 118-128.

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

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

Measurement, 27 (6), 479-498.

Zickar, M. J. (2002). Modeling data with polytomous item response theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 123-155). San Francisco: Jossey-Bass Inc.