2011

Item Purification in Differential Item Functioning Using Generalized Linear Mixed Models

Qian Liu

THE FLORIDA STATE UNIVERSITY

COLLEGE OF EDUCATION

ITEM PURIFICATION IN DIFFERENTIAL ITEM FUNCTIONING USING GENERALIZED LINEAR MIXED MODELS

By

QIAN LIU

A Dissertation submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Doctoral Philosophy

Degree Awarded:
Spring Semester, 2011

The members of the committee approve the dissertation of Qian Liu defended on 01/28/2011.

Betsy Jane Becker
Professor Co-Directing Dissertation

Akihito Kamata
Professor Co-Directing Dissertation

Xufeng Niu
University Representative

Yanyun Yang
Committee Member

Insu Paek
Committee Member

Approved:

Betsy Jane Becker, Chair, Department of Educational Psychology and Learning Systems

The Graduate School has verified and approved the above-named committee members.

For my dear husband, Shiling

and our beautiful daughters, Ellen and Emily

ACKNOWLEDGEMENTS

I would like to sincerely thank those who have helped me and inspired me during my doctoral study.

First and foremost, I want to express my deepest gratitude to my advisor, Dr. Aki Kamata, for his encouragement, patience and guidance to measurement and statistics field. I appreciate all his contributions of time and ideas to make my doctoral study productive and exciting, even after leaving Florida State University. Without his support, this dissertation would not have been possible.

I would like to extend my special thanks to Dr. Betsy Becker, who has been providing numerous advices on my dissertation and "adopt me" in the end to keep me on track. Her technical and editorial advice was essential to the completion of this dissertation and has taught me innumerable lessons and insights on the workings of academic research in general.

I would also like to thank my program buddies at Florida Department of Education who made it a pleasant place to work. Thanks for attending my defense practice and giving me critique, suggestions and encouragement.

I especially thank my husband Shiling Ding for his accompany and love during the whole PH.D study. His tolerance of my occasional vulgar moods is a testament in itself of his unyielding devotion and love. I also thank my two lovely daughters, Ellen and Emily. Their love is my motivation to work and study. I wish I could have spent more time with them. My deepest thanks go to my mother, Shoulan Qiao, for taking care of my little Emily to free me from daily house chores so that I could have more time to work on my dissertation. Her dedication and love was in the end what made this dissertation possible.

Finally, I wish to dedicate this dissertation to my father, Yongquan Liu. BABA, I know you were always proud of me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

For this dissertation, four item purification procedures were implemented onto the generalized linear mixed model for differential item functioning (DIF) analysis, and the performance of these item purification procedures was investigated through a series of simulations. Among the four procedures, forward and generalized linear mixed model (GLMM) iterative purification procedures attempt to remove the contamination of matching variables due to the inclusion of DIF items. The rank-based strategy and mean-DIF procedure are designed to locate DIF-free item(s) as anchor(s), and then the located DIF-free item(s) can be used to identify DIF items. Four variables were manipulated as simulation factors in this study: (1) sample size [500 examinees in each of the reference and focal groups (500R/500F), and 1,000 examinees in each group (1,000R/1,000F)]; (2) test length (20 and 50 items); (3) percentage of DIF items in the test (0%, 20% and 40%); and (4) DIF direction of DIF items (one-sided DIF, dominant DIF and balanced DIF). The type I error rate and power were calculated to evaluate the performance of the forward and GLMM iterative purification procedures. On the other hand, rank-based strategy and mean-DIF procedure were evaluated based on accuracy of DIF-free item identification. All four procedures were applied to simulated data and a real data set. The simulation results showed that the forward and iterative purification procedures were able to control type I error rates, and they were able to maintain a satisfactory power level when 20% of the test items were DIF items. When 40% of the items contained DIF, both procedures had good control over type I error rates and maintained adequate power under the dominant and balanced DIF conditions; however, both procedures lost control of type I error in one-sided DIF conditions. When larger amounts of DIF contaminations were in the tests, the iterative procedure performed better than the forward procedure by generating less error rates. Overall, the rank-based strategy and mean-DIF procedure were both promising for locating a set of up to four DIF-free items. By using a GLMM approach, this study compared the effectiveness of the four item purification approaches for the purpose of creating fairer tests. The comparisons provided practical knowledge that will benefit measurement professionals and enhance the psychometric literature.

CHAPTER 1

INTRODUCTION

Test fairness has always been a concern among test developers and users because important decisions may be made based on a person's test performance. For instance, school or college admittance, job promotions, graduation, certification, diagnosis of a special population, and other such decisions are often largely based on test scores. It is essential to avoid including biased items in a test since item bias may unfairly influence examinees' performance on a test (Hambleton & Rodgers, 1995). Item bias occurs when a test item unfairly gives advantage to a group of examinees taking the test, resulting in one group having a higher probability of answering the item correctly (Schumacker, 2005). As a result, the inclusion of biased items in a test may lead to inaccurate decisions and unfair treatment of examinees.

Since item bias is a potential threat to test fairness and validity, various studies have focused on how to detect biased items in a test. Differential item functioning (DIF) analysis is one convenient first step for studying item bias (Williams, 1997) and has become routine in psychometric bias analysis (Zumbo, 1999). DIF concerns situations in which examinees of equal ability from different groups have different probabilities of responding correctly to an item.

A variety of statistical procedures for detecting DIF have been developed. They include the standardization method (Dorans, 1989; Dorans & Kulick, 1986); the logistic regression method (Swaminathan & Rogers, 1990; Zumbo, 1999); the confirmatory factor analysis approach (Muthen, 1989); the simultaneous item bias test (SIBTEST: Shealy & Stout, 1993); the multiple indicators, multiple causes confirmatory factor analysis methods (MIMIC; Oort, 1998); IRT-based approaches, such as loglinear item response models (Kelderman, 1989); area measures (Raju, 1988; Raju, van der Linden, & Fleer, 1995); and the likelihood ratio method (IRT-LRT: Thissen, Steinberg, & Wainer, 1988). For a more detailed review of the DIF procedures, see Holland and Wainer (1993) and Zumbo (1999). Most DIF procedures are well developed and tested. Each technique has appropriate circumstances under which it demonstrates its strengths, and conversely, its weaknesses.

Even though the analysis of DIF seems quite routine and DIF detection procedures have matured since DIF analysis began, several issues still need to be addressed (Dorans & Holland, 1993; Zenisky et al., 2004)**.** For instance, scale indeterminacy of DIF items is a very important issue that requires further research. An important characteristic of DIF parameters is that they can only be scaled by providing a scaling reference point. That means the scale of the DIF for each item is relative to the other items. As a result, several empirical studies have reported different, even contradictory results for the various procedures for detecting DIF items (Ercikan, Gierl, McCreith, Puhan & Koh, 2004; Kim & Cohen, 1995; Wu, Adams & Wilson, 1997). The discrepancies among DIF estimates may come from different parameterization techniques used in the differing DIF detection procedures, as revealed in Cheong and Kamata's study (Cheong & Kamata, 2007). In order to meaningfully interpret DIF analyses, researchers must be informed about how the DIF parameters were scaled (Cheong & Kamata, 2007).

To deal with the scale indeterminacy of DIF parameters, one solution is to use all items in a test as anchor items, except the item that is being tested for DIF. This method is called the all-other anchor item method or all-other method for short (Wang, 2004). The all-other anchor method produces accurate DIF estimations if the studied item is the only DIF item or when there are no DIF items in a test (Wang, 2003; Cohen, Kim & Wollack, 1996; Kim & Cohen, 1998). In other words, if the anchor set contains DIF items, then the all-other method cannot function properly; therefore, as the number of DIF items increases in a test, the reliability of the performance of the all-others anchor method decreases (Wang, 2003). To diminish the effect of biased items in the anchor set, Lord (1980) proposed the idea of scale purification. Others have recommended, and sometimes examined, specific purification methods for IRT-based DIF analysis (Candell & Drasgow, 1988; Park & Lautenschlager, 1990) and for non-IRT based approaches, such as the Mantel-Haenzel method and the logistic regression IRT detection (Clauser & Hambleton, 1993; Miller & Oshima, 1992). Many researchers have demonstrated that DIF detection methods with scale purification are less affected by the inclusion of DIF items than DIF detection methods without scale purification (French & Maller, 2007; Wang, Shih & Yang, 2009).

Another possible solution to the scale indeterminacy of DIF parameters is constraining the DIF of an item (or a set of items) to pre-specified fixed values. When this is done, the item(s) can be used as reference item(s) to scale the estimates of DIF parameters of the other items. As a result, the DIF magnitude of the other items is the deviation from the DIF of the reference item(s). This scaling procedure is referred to as the constant anchor item method (Wang, 2006). Typically, an item or items that are believed to be free from DIF are chosen as reference items, and the DIF of the reference items is fixed to zero; therefore, when using the constant anchor item method, it is critical to choose a reference item correctly. When a DIF item is mistakenly chosen as a reference item, the DIF-free items in the test will be incorrectly detected as having DIF, that is almost the same magnitude of the anchor item's DIF (Wang, 2004).

Therefore, having DIF-free items as anchors is critical in DIF detection. When such an a priori, unbiased item is not available, the identification of DIF-free items is a necessary starting point in DIF analysis. In recognition of the importance of pure anchors, Meulders and Xie (2004) discussed the idea of using the generalized nonlinear mixed model (GLMM) for item purification, but they did not propose or specify specific procedures for investigation. Meanwhile, Wang, Shih, and Yang (2009) conducted a series of simulations to investigate the performance of the multiple indicators, multiple causes method with scale purification (or MIMIC-SP for short). They concluded that the MIMIC-SP procedure surpassed the MIMIC method without scale purification in controlling type I error rates if few DIF items present in a test. Moreover, with the purpose of finding DIF-free items as anchor items, Shih and Wang (2009) studied the MIMIC method, and Woods (2009) studied the rank-based method.

None of procedures described above has been studied using the framework of GLMM. Given the importance of pure anchor items for DIF analysis and the promising characteristics of GLMM, efforts should be made to verify the advantages and disadvantages of item purification procedures under the framework of GLMM. This study was conducted with the motivation of finding one or more efficient purification procedures, especially a purification procedure that can be used to locate DIF-free items as anchor items for subsequent DIF item detection. A GLMM DIF model has been applied to investigate the proposed procedures.

Research Purpose

The objective of this study is to investigate the performance of the proposed purification method and compare its performance with three other methods using GLMM. As previously stated, it is important to have pure anchor items available for DIF analysis. In addition, unlike other DIF models, the GLMM DIF model is fairly new and has not been as widely tested as other methods. Thus, a study is needed to investigate the performance of the purification procedures using the GLMM DIF model.

First, I am going to propose a forward item purification procedure (or forward procedure for short) based upon a GLMM DIF model (Meulders & Xie, 2004). The GLMM DIF model is constructed by adding a person-by-item interaction variable into the GLMM model to describe the DIF phenomenon. The interaction variable is derived as the product of the item indicator and a person predictor indicating group membership. The forward procedure consists of two steps: (1) the all-other anchor method is utilized to test each single item for DIF; and (2) all of the items with significant test statistics are included in the final model for DIF estimations.

Second, the performance of the proposed method is investigated along with three other methods (GLMM iterative purification procedure, the mean-DIF procedure and the rank-based strategy). The GLMM iterative purification procedure is an extension of the forward procedure. In the iterative purification, after a first round of item purification using the all-other anchor method, each item is tested for DIF. Unlike the forward procedure, the items with significant test result are not included in the final model. Rather, these possible DIF items are excluded from the anchor sets. The procedure continues with a second round of item purification with new anchor sets. This iterative process will continue until the same set of DIF items is identified on successive iterations.

The mean-DIF procedure and the rank-based strategy are used to identify pure anchors for subsequent DIF analysis. Using a MIMIC DIF method, Shih and Wang (2009) studied an iterative procedure to locate up to four DIF-free items. I call this procedure the mean DIF procedure since the mean DIF indices are used to identify DIF-free item in the last step. In the rank-based strategy, the DIF-free item is identified by

rank order of the items based on the magnitudes of the DIF test statistics after each item is tested for DIF using the all-other anchor method. This procedure was proposed by Woods (Woods, 2009). She assessed this procedure using the likelihood ratio test in IRT models (IRT-LR-DIF) and recommended to use this method with other DIF models.

This study applies the framework of the GLMM DIF model to these four purification procedures. A series of simulations is done to investigate their performances. Since the main purpose is to provide a set of general guidelines for the procedure that is determined to be advantageous, various factors, such as test length, sample size, the percentage of DIF items in a test, and DIF direction are manipulated. Following that, the four procedures are also applied to a real data set. This study is expected to produce a practical demonstration of item purification procedures using the GLMM DIF model.

Significance

In studying DIF, pure (i.e., DIF-free) anchor items are essential for many approaches in order to produce accurate parameter estimates and correct DIF detection (e.g., SIBTEST: Shealy & Stout, 1993a, b; MIMIC: Shih & Wang, 2009). When such a prior set of unbiased items is not available, purification procedures are highly recommended as the starting point in DIF detection.

Purification procedures have been widely implemented in many DIF methods in order to purify the contaminated matching variable when the all-other anchor method is applied. Many researchers have shown support for the use of purification with both IRT and non-IRT based DIF methods (Ackerman, 1992; Candell & Drasgow, 1988; Cohen & Kim, 1996; Holland & Thayer, 1988). Unlike other DIF detection methods (e.g., MIMIC purification, Wang, Shih & Yang, 2009; iterative Mantel-Haenszel method, Holland & Thayer, 1988; and iterative logistic regression method, French & Maller, 2007), few, if any, purification procedures have been adapted to the GLMM DIF method. Due to the flexibility and versatility of the GLMM models, it is essential to propose and study a purification procedure that can be implemented with the GLMM DIF method.

Purification procedures also can be utilized to identify DIF-free items as pure anchors for subsequent DIF item detection. Unlike the abundant research on the standard

purification procedure, there are relatively few studies on how to locate DIF-free items as anchor items. Among them, Shih and Wang (2009) adapted a scale purification procedure into a MIMIC DIF method in order to locate DIF-free items as pure anchors. Using a series of simulation studies, Shih and Wang suggested that there must be more DIF-free items than DIF items in a test in order to use this procedure appropriately. Also, Woods (2009) recently recommended a noniterative, rank-based strategy for selecting pure anchors. Following the test for DIF on all items, a DIF-free item is located if that item has the smallest magnitude of the likelihood ratio test statistic. However, when locating DIF-free items, it is unclear how sample size and DIF direction affect the performance of the rank-based procedure.

As Woods (2009) indicated, the challenge for purification studies is to find a method that is as accurate as possible in suboptimal situations. For example, very few purification procedures have been evaluated under conditions in which the percentage of DIF items is large. However, this is very likely to happen under special testing situations (e.g., test translation and adaptation, in which the percentage of DIF items might be large (Gierl, Gotzmann, & Boughton, 2004). For example, Gierl, Rogers, and Klinger (1999) investigated a Canadian Social Studies achievement test and noted that 26 out of 50 items (52%) were DIF items when the test was translated from English to French. Ercikan and McCreith (2002) found that 110 out of 139 items (79%) on the Third International Mathematics and Science Study displayed DIF when American and French examinees were compared. Rogers, Gotzmann, and Vanderberghe (2000) noted that 28 out of 50 items (56%) on the Hong Kong Certificate of Education Examination in Computer Studies contained DIF when English and Chinese examinees were compared. These studies raise concerns about the efficiency of the purification procedure when the proportion of DIF items in a test is large.

Moreover, the effectiveness of the existing purification procedures in identifying DIF-free items has not been documented when DIF items are pervasive and unbalanced. In fact, very few of the preferred purification methods have been evaluated in testing situations where the DIF items largely favor one group and/or the percentage of DIF items is large. Unfortunately, these conditions may occur when DIF analysis is conducted on test items that have not undergone rigorous content reviews for the focal group of

interest. For example, researchers who have studied the psychometric characteristic of tests for special populations have noted an important trend: the direction of DIF items is mostly one-sided or dominant (i.e., all or most DIF items provide advantages for or against the same group) (Cheong & Kamata, 2007).

This study has two features that distinguish it from other purification studies. First, DIF items were simulated in diverse conditions. Two test lengths were crossed with three levels of DIF contamination. The percentage of DIF items was manipulated so that it was 0%, 20%, and 40%, which represents a range from no DIF to a majority of DIF items in a given test. The large percentage of DIF items (40%) reflects a possible testing situation for a special population test or test adaption. The direction of DIF was also manipulated to reflect the general and the most extreme situations that can occur.

Second, this study is conducted using the GLMM DIF method. In general, GLMM DIF modeling provides several major strengths for DIF detection and research. Specifically, by including predictor variables in the GLMM DIF model, researchers can attempt to explain the reasons for DIF. Also, GLMM can be carried out using commonly available standard software, such as SAS Proc GLIMMIX or Proc NLMIXED and similar software for generalized linear and nonlinear models, which allows researchers to build models with more flexibility. Moreover, this approach allows for the simultaneous estimation of uniform DIF and nonuniform DIF. Therefore, this study is relevant to researchers and practitioners alike, especially those who study the special population testing or translated tests.

CHAPTER 2

LITERATURE REVIEW

To date, many DIF detection procedures have been proposed and evaluated. Most DIF detection methods are designed to identify individual items that show differential functioning between subgroups by implying a kind of anchoring in an implicit or explicit way (Clauser & Mazor, 1998; De Boeck, 2008). Anchoring is used to solve the scale indeterminacy of the DIF parameters in DIF study. In order to explain the anchoring issue and to prepare for describing the proposed solution, three anchor methods, based on Wang (2004), are described in this chapter (2004). The three anchor methods are termed as follows: all-other anchor item method (i.e., all-other method), equal-mean-difficulty anchor method (i.e., EMD method), and constant anchor item method (i.e., constant anchor method).

This chapter is divided into four sections. The first section introduces the basic concepts of DIF, item bias, and item impact. This knowledge is very helpful for understanding DIF terminology. The second section discusses some fundamental DIF terminologies, such as focal and reference groups, uniform and nonuniform DIF, and matching criteria. The third section describes the anchoring issue in DIF studies and introduces how each of the three anchor methods mentioned above are implemented in certain DIF detection models. Although some of these procedures have already been extended for polytomous data, all of the procedures discussed in this study are based on dichotomously scored items. Item purification and certain purification procedures are then demonstrated in the fourth section. Last, the idea that the one-parameter logistic (i.e., Rasch model) model can be conceptualized as the GLMM model is introduced and followed by a description of GLMM DIF modeling.

Item Bias, Item Impact, and DIF

In some literatures, the terms item bias and DIF are used interchangeably, but they are actually different ideas and should be treated as such. Angoff (1993) defined item bias by writing, "An item is biased if equally able (or proficient) individuals, but from different groups, do not have equal probabilities of answering the item correctly" (p. 4). Shepard et al. (1981) described bias as "a kind of invalidity that harms one group more than another" (p. 318). Based on the above information, the detection of biased items involves performance of the subjects, an evaluation of performance, and the judgment of unfair effects that result. The presence of bias has at least two different meanings: social and statistical (Angoff, 1993). DIF can be considered a statistical indication of bias. Whether statistical findings of DIF indicate item bias depends on later interpretation and judgment.

DIF occurs "when examinees from different groups have differing ability or likelihoods of success on an item, after they have been matched on the ability of interest" (Clauser & Mazor, 1998, p. 31). DIF analysis is one way to identify items that are functioning differently across different groups of individuals (e.g., male vs. female or native English speakers vs. English language learners). An item displays DIF if the proportion of individuals answering an item correctly is different for each subpopulation tested when all individuals have the same score on the test containing that item (Scheuneman, 1975). This definition of DIF is based on the observed score DIF approach. Lord (1980) provided another way, based on IRT methodology, of understanding DIF. From his point of view, an item shows DIF when that item has different item response functions for different groups of examinees after controlling for ability.

A DIF item is a *potentially* biased item. This means that an item showing DIF may not always be biased. If an item has DIF, this might simply indicate the multidimensionality of that item (Ackerman, 1992). From the multidimensionality perspective, the presence of DIF is indicative of the presence of a secondary factor related to group membership. Such secondary factors affect the item performance in a way that differs systematically for the reference group and focal group after conditioning on the target ability (Park & Lautenschlager, 1990; Penfield & Camilli, 2007). Most item

response models assume that the items are measuring a single latent ability (Hambleton & Murray, 1983; Lord, 1980). There are, however, some items that can measure multiple traits (Ansley & Forsyth, 1985). To answer such items correctly, the examinees need to master multiple abilities. For example, in order to get the correct response for a mathematical word problem, examinees are required to master both calculation and reading comprehension skills. If examinees are only matched on their ability to perform calculations and one group is less skillful than the other in reading comprehension, then a mathematical word problem will likely find between-group differences indicating the presence of DIF (Clauser & Mazor, 1998). However, one may not initially interpret that this item with DIF is unfair and thus biased. The relevance of reading comprehension ability to the testing purpose should determine whether or not such an item is biased. If reading comprehension ability is not the objective of the measurement, then this DIF item can be considered biased. The content expert establishes whether or not the observed DIF is the resulted of an unintended secondary factor. Since the content expert may determine that the presence of DIF is the result of intended factors (e.g., reading comprehension skills to solve mathematical word problems), the presence of DIF is an insufficient condition for claiming that an item is biased (Williams, 1997).

Item impact is another important concept in measurement theory. Item impact is defined as the group difference in the probability of getting an item correct due to the true group differences in the underlying ability measured by the items (Wainer, 1993; Zumbo, 1999). For example, on a typical reading test, Native English Speakers, as a group, usually score higher than English as Second Language (ESL) students. This difference in performance is not evidence of item bias; rather, it is evidence of item impact. Frequently, impact on any given item is consistent with impact on all other items of similar type (Dorans & Holland, 1993). For example, if Native English Speaker examinees perform better on one item in a typical reading test than ESL students, then it is probable that Native English Speaker students will perform better than ESL students on other items on that test.

Thus, if an item is declared as having no DIF, then there is no item bias. The presence of DIF on an item is a statistical indication of bias. A DIF item should not be considered as biased if the difference results from the groups' difference in ability to

respond to that item; rather, the performance difference between two groups is indicative of item impact (Perron, 2006).

DIF Terminology

Before introducing the specific DIF analysis procedures, it is important to explain the pertinent terminologies. First, a DIF study typically compares the performance of two manifest groups: a *focal group and a reference group*. The focal group is referred to as the group of interest in a particular DIF study. Usually, it is the group that is investigated to see whether or not it is disadvantaged by an individual item (De Ayada, 2008). The reference group is the group used for comparison purposes. In some literatures, the reference group and focal group are called the majority group and minority group, respectively. Typically, but not necessarily, the reference group consists of individuals that the test is expected to favor, while the focal group consists of individuals who are at risk of being disadvantaged by the test (e.g., minority students, students with learning disabilities). For example, when analyzing an English reading test, ESL students will probably be referred to as the focal group, and Native English Speaker students will probably be referred to as the reference group.

Second, DIF studies examine the "conditional" mean difference between subpopulations. An observed difference in performance between studied groups does not necessarily mean that DIF items exist in the studied test. In order to establish evidence for DIF, one needs to demonstrate the difference between the reference group and focal group by controlling for their latent trait levels. One way to control for the latent trait is to match examinees by relevant knowledge and skills. The relevant knowledge and skills can be referred to as the *matching criterion*. There are two types of matching criterion: internal and external. For example, an observed total raw score can be used to match examines in the Mantel-Haenszel approach (Mantel & Haenszel, 1959), and the estimate of latent ability can be used in the logistic regression method. This matching criterion is referred to as the internal matching criterion. Therefore, an internal matching variable is one that is derived from the test on which the DIF analysis is conducted. On the other hand, an external matching variable is obtained from an external source, such as a

different test or information other than the test on which DIF analysis is conducted (Longford, Holland, & Thayer, 1993). Regardless of whether the internal or external matching criterion used, it is preferable to obtain the matching variable from DIF-free items (Shealy & Stout, 1993). In order to obtain a "pure" matching variable, a purification process is recommended to purify the contaminated matching criterion due to the inclusion of DIF items.

Third, it is important to distinguish two forms of DIF: *uniform DIF and nonuniform DIF*. Graphically, DIF can be represented by the difference between two item response functions (IRFs) (Mellenberg, 1982) based on item response theory (IRT). Based upon the item parameter estimates, one IRF can be obtained for each of the studied groups. Two coinciding IRFs indicate the absence of DIF for that item after the abilities of the two groups are linked into the same metric. If the two IRFs do not match after the two groups are put on the same scale, then the item is detected as exhibiting DIF.

In the simplest case, the difference between the two IRFs has the same relative magnitude across the latent ability continuum, which means that the item consistently provides an advantage for one group regardless of the ability level. The item shown in Figure 2.1 displays an uniform DIF that favors the members in Group A over the members of Group B at all ability levels.

In contrast to uniform DIF, nonuniform DIF is present when the difference between IRFs changes in magnitude or in direction across the latent ability range. For example, an item may provide a small relative advantage to the reference group at low ability levels and a large advantage at high ability levels. A more complex case is crossing DIF. Crossing DIF exists when the item provides advantages to different groups depending on ability level, within the range of typically observable ability levels. Figure 2.2 shows an example of an item that exhibits crossing DIF. In summary, in the presence of nonuniform DIF, an item's IRFs for two groups differ in their slopes and potentially their locations; however, with uniform DIF, an item's IRFs for two groups only show different locations. Regardless of uniform or nonuniform DIF, the probability of success is different for examinees in different groups after the examinees have been matched on their underlying abilities.

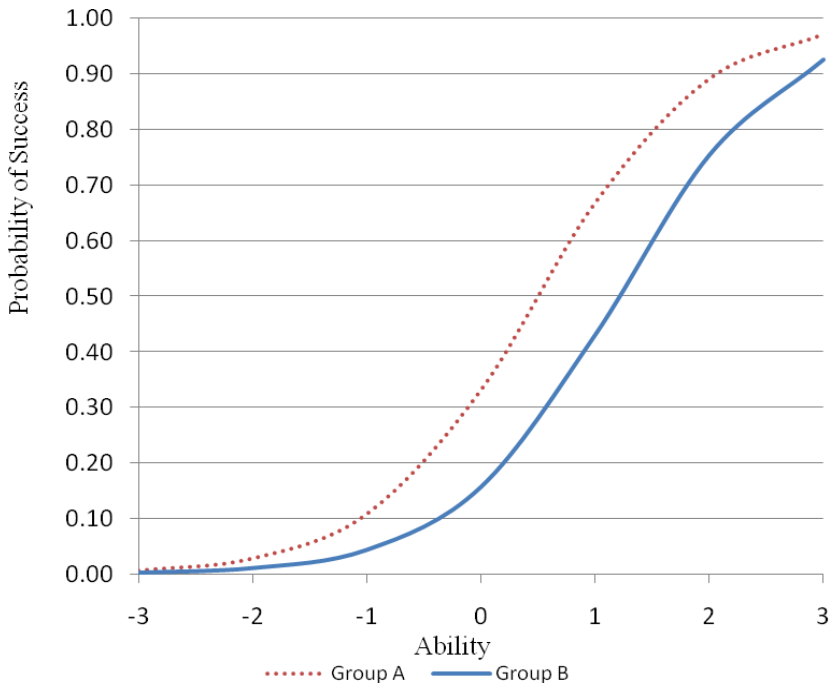

Figure 2.1 Example of Uniform DIF item

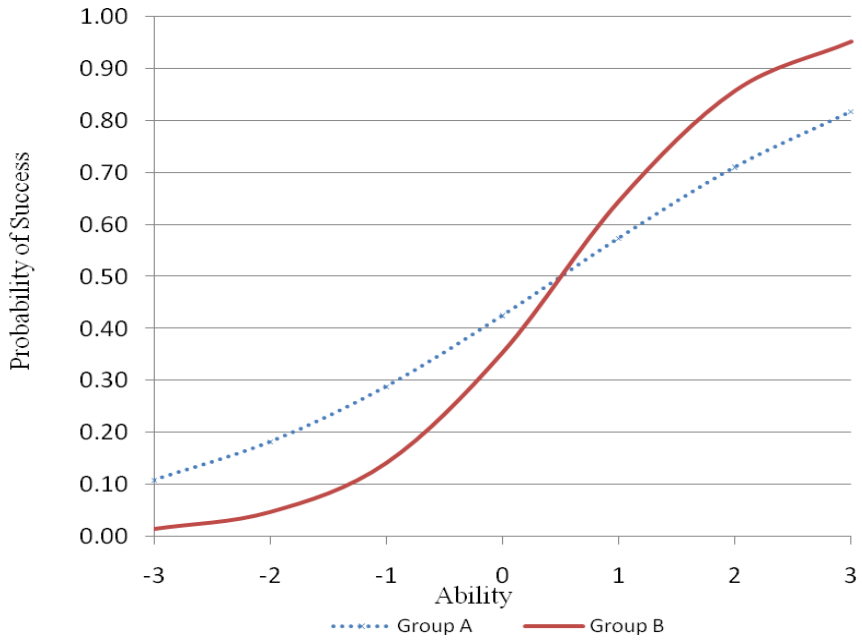

Figure 2.2 Example of Nonuniform DIF item

Anchor Methods in DIF Analysis

A variety of statistical methods have been developed for DIF analysis, including but not limited to the delta plot (Angoff & Ford, 1973), the Mantel-Haenszel (MH) approach (Holland & Thayer, 1988), the logistic regression procedure (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993), the confirmatory factor analysis approach (Muthen, 1989) and item response theory (IRT) methods (Candell & Drasgow, 1988; Lautenschlager, Flaherty & Park, 1994; Lord, 1980; Miller & Oshima, 1992). All of the methods available for identifying DIF have the issue of scale indeterminacy of the DIF parameters.

The scale indeterminacy of DIF results from the fact that DIF is a relative concept and DIF estimates depend on how the DIF parameters are scaled (De Boeck, 2008). For instance, according to the Rasch model, the presence of DIF is indicative of a difference in item difficulties between groups. However, in the IRT model, the origin of the scale is not fixed. Since Rasch is one of the IRT models, it has this limitation. In order to establish a difference in item difficulty, two groups under DIF analysis need to be linked to the same scales. As a consequence, the DIF estimate is a matter of choice for the scale linking. Or in other words, the anchoring method that is chosen has an effect on the DIF assessment (De Boeck, 2008). Various anchoring methods are available for DIF detection. In this section, the following three anchor methods based on Wang's (2004) descriptions are introduced: all-other anchor method, equal-mean-difficulty method and constant anchor method.

All-Other Anchor Method

All-other anchoring means that the anchor set consists of all items except the item studied for DIF. This anchoring method tests each item in turn for DIF using a different anchor set. In other words, if a test under DIF investigation contains 20 items, then 20 anchor sets are used and 20 DIF analyses are carried out using the all-other anchor method. This anchoring method has been applied in a variety of DIF detection procedures, either explicitly or implicitly.

For example, the all-other anchoring method has been adapted into an IRT-based DIF detection method using the likelihood ratio test (IRT-LR-DIF: Thissen, Steinberg, & Wainer, 1988, 1993) in the MULTILOG program (Thissen, 1991; Thissen, Chen, & Bock, 2002). In the IRT-LR DIF analysis, DIF is detected through three steps.

1) Estimate the compact IRT model, in which all items are constrained to have the same parameters in both groups, and calculate the likelihood deviance G_C^2 (= –2 × log-likelihood) of the maximum likelihood estimates.

2) Estimate the augmented model, in which all items but the studied one are constrained to have the same parameters in both groups and calculate the likelihood deviance G_A^2.

3) Obtain the difference of the two likelihood deviances $G^2 = G_C^2 - G_A^2$ and test its significance using the chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated in the compact and augmented models.

The studied item is identified as having DIF if G^2 is significant. This process should be repeated to detect DIF for each item in a test. For example, if a test contains 25 items, then 25 augmented models will be estimated and 25 G_A^2 values will be calculated. In each augmented model, each single item is tested for DIF by anchoring all other items. As demonstrated above, when applying this all-other anchor item method, the anchor item sets vary across studied items.

In addition to the IRT-based DIF model, the all-other anchoring method has been a part of standard procedure in the MH method (Holland & Thayer, 1988), the logistic regression method (Swaminathan & Rogers, 1990), and the standardization method (STD P-DIF: Dorans & Kulick, 1988). For instance, the MH method uses all but one item to calculate the test raw scores. Then the test raw scores serve as the internal matching criterion to establish a common scale over groups. Once a common scale is obtained, the performance of the reference and focal groups on each item can be examined for evidence of DIF. Therefore, the MH method can be viewed as an application of the all-other anchor method (Wang, 2004).

With the logistic regression method for DIF detection, the dependent variable is the response to an item (i.e., right or wrong), and predictors are group membership and the measure of the construct. Conceptually, the likelihood ratio (LR) strategy for DIF analysis is based on a series of nested model comparisons. The chi-square statistic is used to determine whether the full model differes significantly from the reduced model. Equation 2.1 represents the full LR-DIF model. The log-odds (or logit) of answering the item correctly (or endorsing the item) can be modeled as

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg \,, \qquad 2.1$$

where ln is the natural logarithm, x is a measure of an individual's trait level such as an observed total test score, g is the group membership indicator coded as 1 for the focal group and 0 for the reference group, and xg is the interaction term between trait levels and group membership (Swaminathan & Rogers, 1990). To detect DIF in an individual item, the interaction term β_3 is first tested for nonuniform DIF (β_3=0 or not). If there is no evidence of nonuniform DIF, the interaction term is dropped from the model. Next, the uniform DIF is assessed by testing whether β_2=0. Therefore, when the logistic regression procedure analyzes each item for DIF by controlling for latent ability, it is also an application of the all-other anchor method.

Several studies have demonstrated that the all-other anchor method performs properly when a test contains only one DIF item that is exactly the studied item or when there are no DIF items at all (Ankenmann, Witt, & Dunber, 1999; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Wang, 2004). However, if a test contains multiple DIF items, then the DIF detection procedure using the all-other anchor method produces inaccurate item parameter estimates and incorrectly identifies of DIF items (Clauser, Mazor, & Hambleton, 1993; Narayanan & Swaminathan, 1994, 1996; Wang, 2004).

Equal-Mean-Difficulty (EMD) Anchor Method

Another alternative to DIF parameter scaling is to constrain the mean item difficulty to a given common value. This anchor method is referred to as the equal mean difficulty (EMD) method (Wang, 2004). Usually, the common value is set to as zero. The EMD anchor method also could be accomplished by constraining the sum of DIF values

to zero. When the sum of DIF values is constrained to zero, the magnitude of DIF favoring one group is compensated for by the same magnitude of DIF against the same group. Therefore, the test as a whole is equally difficult for both groups under DIF investigation, resulting in the equal mean difficulty across groups. Based on the EMD constraint, DIF detection is then executed. In the EMD anchor method, the magnitude of the DIF parameter for each item is the deviation score from the mean DIF in the test.

Many DIF assessment procedures utilize the EMD anchor method to provide the scaling reference point for the DIF parameters. For example, Cheong (2006) utilized a hierarchical generalized linear modeling framework (HGLM) to detect DIF. In HGLM, DIF is conceptualized as an interaction effect between group membership and item difficulty. To detect DIF, a hierarchical, unconditional model is first set up and estimated. Then, a group membership predictor is entered into the model to test for DIF. For example, in a two-level HGLM DIF model, the log-odds of correctly answering an item on a test depend on the overall level of measured proficiency, group differences in the proficiency, an item effect for group membership variables, and group differences in the item effects plus individual student contributions to the performance on the measured proficiency. Equation 2.2 is an example of the combined model for two-level HGLM DIF, which is actually equivalent to the Rasch model with DIF parameters.

$$\eta_{ij} = \gamma_{00} + \gamma_{01} Focal_j + \sum_{q=1}^{k-1} (\gamma_{q0} + \gamma_{q1} Focal_j) X_{qij} + u_{0j}, \qquad 2.2$$

If there is no uniform DIF on a studied item, then the item effect for group membership should be null. As Cheong and Kamata (2007) demonstrated, in this HGLM DIF approach, DIF is parameterized such that the group differences in item difficulty sum up to zero and DIF is a zero-sum variable (Schulz, Perlman, Rice & Wright, 1996). As a result, the mean of item difficulty across groups is zero. Therefore, the EMD anchor method is applied to this HGLM DIF approach.

Although the EMD anchor method has been applied to several DIF detection procedures, this does not imply that the EMD method is always appropriate in DIF studies (Wang, 2004). One problem when using the EMD anchor method is that many applications of this anchor approach do not explicitly inform the assumption of equal mean difficulty. In other words, it is hard, if not impossible, to check whether this

constraint is applicable. As Wang stated, "if there is only one DIF item in a test, by definition the mean item difficulties of the reference and the focal group will never be equal" (Wang, 2004, p.223). Moreover, it is even more difficult to meet the EMD assumption in DIF analysis when there are multiple DIF items in a test because most DIF items favor reference groups in a real test (Budgell, Raju, & Quartetti, 1995; Walstad & Robson, 1997). Failing to meet EMD assumption may result in erroneous inferences about the DIF estimates as well as the mean ability difference (Cheong, Kamata, 2007; De Boeck, 2008; Wang, 2004). These biased estimates make the detection of DIF problematic. In sum, it is essential to have a valid scaling constraint for DIF parameters in DIF studies.

Constant Anchor Item Method

In the all-other anchor method, the anchor sets are different across studied items; however, the anchor items can be the same for all studied items. That is, all of the studied items are assessed for DIF with the same set of anchor items (Wang, 2004; Wang & Yeh, 2003). This anchoring method is referred to as a constant anchor item method. In a constant anchor item method, the anchor set may contain only one item (i.e., single-item anchoring) or more than one item (i.e., multiple-item anchoring) that does not show DIF (De Boeck, 2008). The pre-specified anchor item(s) can be selected based on theory, expert judgment, or earlier studies.

Several authors have proposed and examined the MIMIC method for DIF detection (Finch, 2005; Glockner-Rist & Hoitjink, 2003; MacIntosh & Hashim, 2003; Muthen, 1985, 1988; Oort, 1998; Wang, 2009). Among them, Wang (2009) proposed a MIMIC method with a pure short anchor (M-PA method) to assess DIF. The MIMIC method for DIF detection is an application of the confirmatory factor analysis model to item response data. The MIMIC model in the context of DIF can be written as:

$$y_i^* = \lambda_i \theta + \beta_i z + \varepsilon_i, \qquad 2.3$$

where y_i^* is a latent response variable and is transformed to a dichotomous item response variable y_i. The item response variable is measured by test items so that y_i=1 if $y_i^* > \tau_i$; otherwise 0. Here, τ_i is the threshold parameter on item i and is related to item difficulty

in IRT. θ is the latent ability, λ_i is the factor loading on item i and is equivalent to item discrimination in the IRT context. ε_i is random error with a mean of zero. z is an exogenous variable in the context of CFA and represents the group membership in the DIF context. Therefore, β_i represents the direct effect from z to response on item i. The uniform DIF is assessed by testing whether β_i is equal to zero. In Wang's MIMIC DIF approach, a single item (e.g., item 1) is selected and serves as an anchor in order to test DIF for all other items simultaneously; therefore, the anchor item is constant across all studied items. Wang (2004) also suggested that the optimal number of anchor items is 4.

Moreover, Kamata et al. (2001, 2006) adopted the constant anchor item method using Hierarchical Generalized Linear and Nonlinear Mixed Models. In Kamata's approach, the model formulation is the same as Cheong's HGLM framework DIF analysis. The two major differences are in their approaches for model identification. First, instead of assuming zero-sum DIF as in Cheong's model, Kamata's model assumes that the reference item is a DIF-free item and no centering is employed on item indicators. Moreover, Cheong's procedure constrains the mean item difficulty to zero, whereas Kamata's procedure constrains the mean ability to zero.

In both Wang's MIMIC DIF approach and Kamata's HGLM DIF method, valid anchors are essential to detect DIF appropriately. The constant anchor item method works well if the pure anchor item(s) assumption holds. For this reason, knowing how to locate DIF-free item(s) to serve as anchor item(s) for subsequent DIF detection is a critical issue in DIF detection.

Purification Procedures

The presence of DIF items in the anchor set may seriously affect methods used to detect DIF items in DIF studies. To remove, or at least reduce, the potential effects of the biased items on the DIF analysis, Marco (1977) and Lord (1980) suggested conducting purification first before the DIF detection is conducted. In other words, multistage DIF detection is recommended to assist with more accurate DIF detection. Thus, purification

can be defined as a process that intends to remove the effect of the DIF items in the anchor set so that the DIF items can be accurately detected (French & Maller, 2007).

Many researchers have recommended, and sometimes examined, specific purification methods for IRT-based DIF analysis (Candell & Drasgow, 1988; Park & Lautenschlager, 1990) and non-IRT approaches (Clauser & Hambleton, 1993; Miller & Oshima, 1992). As outlined by Marco (1977) and Lord (1980), the general purification strategy involves the following steps:

1) Initially, run the standard DIF analysis and estimate DIF;
2) Remove the DIF items found in step 1 and rerun the DIF analysis with the remaining items;
3) Estimate DIF using fixed latent ability obtained with the unbiased items.

For instance, in the two-stage MH approach, the DIF procedure is conducted as follows. First, use all items to derive the total raw score to be used as matching criterion for the initial DIF detection. Second, remove the DIF item identified in the first step and recalculate the total raw score. With a fixed total raw score, reestimate DIF for all items. The rationale for using the purification procedure is that the DIF items may contaminate the matching criterion, which in turn may adversely affect the DIF assessment (McCauley & Mendoza, 1985).

Some researchers have studied fully iterative procedures in the DIF analysis. In the iterative purification procedure, the purification process continues until the same set of items is found to be DIF items in two successive iterations. Plenty of iterative procedures have been proposed, including the iterative logistic regression method (French & Maller, 2007), the iterative linking IRT-based method (Candell & Drasgow, 1988; Park & Lautenschlager, 1990), the iterative logit method (Kok, Mellenbergh, & Van der Flier, 1985; Van der Flier, Mellenbergh, Ader, & Vijn, 1984), and MIMIC with iterative purification (Wang, Shih & Yang, 2009). For example, in the iterative logistic regression method (LR-DIF) (French & Maller, 2007), the DIF assessment is conducted as follows:

1) Use all items (K) in the test to obtain the total summed score as a matching variable. Assess item 1 for DIF based on the initial matching variable.
2) Repeat step 1 until the last item is tested for DIF.

3) Identify k DIF items based on the DIF identification criterion.
4) Use K-k items to obtain the updated matching variable and assess DIF for all items in the test.
5) Continue steps 3 and 4 until the same set of DIF items are identified in two successive analyses or no other items are identified as DIF items.

Empirical evidence showed that 1) the DIF assessment with purification can assist with more accurate parameter estimations and DIF detection; 2) the iterative purification procedure outperforms the noniterative purification procedure in regard to the DIF detection rates, especially when a large amount of DIF items appears in the test (Candell & Drasgow, 1988; Kok, Mellenbergh, & Wan Der Flier, 1985; Wang, Shih & Yang, 2009). On the other hand, the studies also indicated that the purification procedures begin to lose control over the type I error rate when a test contains more than 20% DIF items (Clauser, Mazor, & Hambleton, 1993; French & Maller, 2007; Holland & Thayer, 1988; Miller & Oshima, 1992; Wang & Su, 2004a, 2004b). This is especially true where most DIF items favor one group.

Another alternative method for identifying DIF items is to locate DIF-free items, then, use the identified DIF-free items as pure anchors to estimate DIF. This method is referred to as the DIF-free-then-DIF strategy (Shih & Wang, 2009). After a pure anchor set is established, the pure anchor set can either be used to obtain valid matching criteria or serve as a benchmark in the constant anchor method. Shih and Wang (2009) proposed and evaluated a MIMIC DIF method for locating DIF-free items. This procedure begins by testing all items except anchor items for DIF using a single anchor, with every item taking a turn as the anchor (Rensvold & Cheung, 2001; Wang, 2004). After each run, the DIF estimates are collected for all items except the anchor item. Thus, if the total number of items in a test is 10, then every item is tested for DIF nine times, resulting in nine DIF estimates for each item. In the last step, the mean absolute value of DIF indices is calculated for each item. Items with the smallest absolute value of the mean DIF would be the best candidates to serve as pure anchors for the subsequent DIF assessment. In another simulation study conducted by Shih and Wang (2004, 2008), they concluded that the optimal number of designated anchors is four in order to yield a very powerful level of DIF detection (Thissen et al., 1988; Wang, 2004, 2009; Wang & Wilson, 2005).

Conclusively, Wang's mean-DIF strategy yields nearly a perfect accuracy rate in DIF-free item identification, even when there are up to 40% DIF items. It would be useful to empirically evaluate Wang's procedure for DIF-testing methods other than the MIMIC approach.

In addition, Woods (2009) suggested identifying a single anchor based on a simple, noniterative procedure. In her study, a single item is identified as a DIF-free item if that item has the smallest value of likehood ratio test statistics following a test of all items using the all-other anchor method. The general idea of this strategy is that the magnitude of the LR statistic is positively related to the degree to which an item functions differently between groups. Therefore, a larger LR value is an indicator of the presence of DIF for the studied item (Woods, 2009). In Wood's study, she simulated item response data and DIF item responses using Samejima's graded response model (1969, 1997).The test length, percentage of DIF items, and DIF type were manipulated in the simulation study which began by testing a single item for DIF using the all-others anchor method. The simulation results provided evidence that supports the use of rank-based strategy for selecting pure anchors implemented in the IRT-LR-DIF method. Several issues remain to be clarified in Wood's strategy. In addition, it would be desirable to investigate this procedure in other IRT models.

Item Response Theory

Item response theory (IRT), a latent variable modeling approach, has been developed to model the clustered data in educational and psychological testing and other psychometric applications. In IRT models, the characteristics of items (item parameters) and characteristics of individuals (latent traits) are used as predictors of observed responses (e.g., the probability of the positive response) (De Ayala, 2008).

As in classical test theory, the ultimate goal of IRT is to estimate latent traits of individuals (i.e., motivation, ability, social anxiety, proficiency, and so on). Using an English test as example, the English proficiency would be the measured latent trait. To estimate English proficiency, individuals are administered an instrument containing test items. The latent trait can be conceptualized as a latent continuum (De Ayala, 2008). Each individual has a location on the continuum according to his/her English proficiency.

In IRT, persons and items are placed on the same continuum, so each item also can be located on the continuum. The item location, also called item difficulty (β), reflects the required English proficiency that individuals should possess in order to respond correctly to that item. On the continuum, the right end of the continuum represents greater English proficiency and the left side of the continuum represents lesser English proficiency. This means that an individual must possess greater proficiency to correctly answer the items located toward the right end of the continuum than the items located toward the left side of the continuum. For instance, if an individual is located at 0 on the continuum (i.e., θ_p =0) and three items are located at -2, 1, and 3, respectively (i.e., β_1=-2, β_2=1, and β_3 =3), then one can infer how an individual might respond to an item by comparing the locations of person and item on the continuum. Specifically, the individual located at 0 may have a higher probability of answering item 1 (β_1=-2) correctly, but a lower probability of correctly answering item 3 (β_3=3). Therefore, the main concern of this model is the difference between person and item locations (i.e., $\theta_p - \beta_i$). In fact, when only item difficulty parameter is involved to model the probability of correct response, the model is called the one-parameter logistic regression model, or Rasch model. The Rasch model in the logit format can be specified as

$$\eta_{pi} = \theta_p - \beta_i. \qquad 2.4$$

The item discrimination parameter (a) and the guessing parameter (c) are two additional item characteristic indexes. The item discrimination parameter allows items to discriminate differently among examinees located at different points along the continuum. The guessing parameter reflects the fact that individuals may answer item i correctly just by randomly guessing. Besides the Rasch model, numerous IRT models have been developed to model the probability of correct response to an item for an individual with specific latent ability. In addition, IRT models have been widely applied to model clustered data in social science. From a statistical point of view, IRT models represent some of the earliest generalized linear mixed models (Hedeker, 2008).

Generalized Linear Mixed Models

Generalized Linear Mixed Models

Generalized linear mixed models represent a class of mixed effects models for several types of dependent variables (i.e., continuous, dichotomous, and counts). With respect to the regular regression model, the basic difference is that GLMMs include random effects in addition to the fixed effects in the model. These models are well suited to analyze the clustered data and/or longitudinal data by adding the random cluster and/or subject effects into the classical regression models. Common GLMMs include the linear mixed model, mixed logistic regression model, and Poisson mixed model (Hedeker, 2005; Kachman, 2000; Raudenbush, & Bryk, 2002; Rijmen, et al., 2003). They also illustrated the structure and applications of the general GLMMs as follow.

All GLMMs are composed of three components. The first component is the random component. This component describes the distribution of outcomes Y_{ni} in terms of mean, μ_{ni}. Usually, given the random effects, fixed effects and covariates, the conditional distribution of the outcome can be realized from an exponential family distribution. In a GLMM, the expected value of the outcome can be expressed as:

$$\mu_{ni} = E\left[Y_{ni} \mid X_{ni}, Z_{ni}, \beta, \theta_n\right]. \qquad 2.5$$

Here, $\boldsymbol{X_{ni}}$ is a fixed effect covariate, $\boldsymbol{Z_{ni}}$ represents the random effect covariate; β is the fixed effect, and θ_n is the random effect covariate. For example, in educational measurement, the item response on a dichotomously scored item for an examinee is either zero or one, which is an independent Bernoulli distributed variable. In that case, $\mu_{pi} = \pi_{pi}$ in which π is the probability of success of a person p on a specific item i.

The second component is the linear predictor function. The linear predictor function defines η_{ni} (the expected value of the underlying continuous variable) as a linear function of predictors. The predictors include fixed effects Xs and random effects Zs. The general form of a linear predictor component can be written as:

$$\eta_{ni} = X'_{ni}\beta + Z'_{ni}\theta_n. \qquad 2.6$$

In equation 2.6, the random effects are assumed to be multivariate normal distributed with a mean vector of zero and covariance matrix $\sum$ such that $\theta_n \sim N(0, \Sigma)$.

The third component is the link function. The link component connects the linear function component and the expected value of the observed outcomes, denoted as $\eta_{ni} = f_{link}(\mu_{ni})$. The link function is typically selected in terms of the distribution of outcomes. For normally distributed data, the link function is the identity function $g(.)$ =1. For data with a Binomial distribution, either the logit or probit link function can be used.

According to the specification of each component, GLMM can generate various models. For example, if the data are normally distributed and the link function is identity, the resulting model is the linear mixed model. The mixed logistic regression model is another special case of GLMM for binary data. In this model, the observed outcomes are assumed to be independent Bernoulli distributed. If there is only one random effect predictor in the mixed logistic regression model, then the model will be the logistic random intercept model. The Rasch model (1-PL) is a logistic random intercept model, and is therefore a GLMM. Some IRT models can be conceptualized as nonlinear mixed models, but not GLMM. For example, the two parameters logistic (2-PL) model is a nonlinear mixed model in which the random effects are modeled nonlinearly.

In the last few years, several scholars have discussed and demonstrated the connection between IRT models and GLMM. (e.g. Cheong, 2006; Kamata, 2001; Kamata & Cheong, 2006; Mellenbergh, 1994; Moustaki & Knott, 2000; Rijmen et al., 2003). Specifically, in the GLMM formulation of the Rasch model, the latent ability level for person p, θ_p, is treated as the random parameter. Item difficulty parameters, β_i s, are regarded as fixed parameters (De Boeck, 2008; Kamata, 2001; Kamata & Cheong, 2006). X_{ik} is the item indicator with a value of 1 when $i = k$, otherwise 0. This indicator variable represents the response for person p to item i and serves as the fixed effect covariate in the model. Thus, the probability of success for person p on item i is modeled as a function of the latent trait and item difficulty. The GLMM representation of the Rasch model can be expressed as follows:

$$\log it(\pi_{pi}) = \log\left(\frac{\pi_{pi}}{1-\pi_{pi}}\right) = \theta_p - \beta_i X_{ik}. \qquad 2.7$$

In this equation, the logit link function is used to connect the linear function to non-normal distributed outcomes. Therefore, in this GLMM approach, the probability of answering an item correctly depends on the random effect and item-level fixed effect.

Many researchers have applied GLMM to educational measurement (Kamata, 2001; Meulders & Xie, 2004; Van Den Noortgate, De Boeck & Meulder, 2003; Van Den Noortgate & Paek, 2004). Among them, Meulders and Xie (2004) proposed that DIF can be detected for each item by including a person-by-item covariate into the GLMM. The following section details this GLMM DIF modeling.

GLMM DIF Model

Equation 2.7 represents the GLMM model for valid items (i.e., the items without DIF). To describe DIF phenomena, a person-by-item covariate is introduced into the GLMM model. For instance, The GLMM DIF model that only involves item difficulty can be written as:

$$\eta_{pi} = \theta_p - (\beta_i X_{ik} + \delta_i^{\beta} W_{pik}). \qquad 2.8$$

In equation 2.8, the logit of the success probabilities for the item suspected of having DIF depends on the latent trait, item difficulty, and DIF parameter. The interaction variable W_{pik} actually is derived from the product of the group indicator Z_p and item indicator X_{ik}. Usually, the group indicator Z is coded as 1 for the focal group and 0 for the reference group. For a person from the reference group (Z_p=0), the estimate of the interaction effect δ_i^{β} is zero, and the equation 2.8 is reduced to equation 2.7. For a person from the focal group (Z_p=1), the effect of δ_i^{β} is estimated and the item difficulty becomes $\beta_i + \delta_i^{\beta}$. Therefore, the interaction effect actually represents the difference in item difficulty between the reference group and focal group. Therefore, δ_i^{β} represents the DIF on the item difficulty parameter. As a result, DIF in a single item can be detected with a Wald test for the null hypothesis that the interaction parameter equals zero.

The DIF in equation 2.8 only involves the item difficulty parameter; therefore, it is uniform DIF. This model can easily be expanded to a nonlinear mixed model (NLMM) for nonuniform DIF, by inserting a parameter α_i as a weight of θ_p and a person-by-item interaction effect δ^a. The non-uniform DIF is the DIF that involves $\delta^a \neq 0$ and possibly also the item difficulty parameter (e.g., the item may show a difference in location among different groups or it may not); however, this extension was not implemented here. In this study, only uniform DIF is modeled.

In DIF studies, it is common to find mean ability differences between the focal and reference groups. Thus, the random effect in GLMM DIF is assumed to be normally distributed with group-specific mean and variance (Meulders & Xie, 2004). Note that in order to make the model be identified, the mean of one group is constrained to be zero.

GLMMs have advantages over regular IRT models, including the capacity to explain person effects and item difficulties and to explore DIF sources (Rijmen, Tuerlinckx, De Boeck and Kuppens, 2003). In addition, GLMMs provide the measurement context with a broad selection of statistical models. A researcher can extend the model to fit many situations. Also, researchers can use more flexible and efficient software related to broad statistical models. As presented in this literature review chapter, many DIF detection models and purification procedures have been proposed. However, few, if any of those studies, were conducted on the GLMM DIF detection with purification. This current research is concentrated on the GLMM DIF analysis with purification.

CHAPTER 3

METHODOLOGY

This chapter has two sections. First, it introduces the four purification procedures that are investigated in this study. Second, the design of the simulation study is presented. The evaluation criterion for these procedures is also briefly described. In this simulation study, the manipulated variables are sample size, percentage of DIF items, test length, and DIF direction. As a result, 28 simulation conditions are created by crossing the four variables.

3.1 Purification Procedures

The forward approach and iterative procedure apply an all-other anchor method to purify the anchor set so that the DIF items can be detected accurately. The rank-based strategy and mean-DIF procedure are two applications of DIF-free-then-DIF strategy in the DIF analysis. The DIF-free-then-DIF strategy is a two-step purification procedure:

Step 1: Execute a procedure to locate a set of DIF-free items.

Step 2: Use the identified DIF-free items as the anchor and assess the other items for DIF simultaneously. In this study, only step 1 of the DIF-free-then-DIF strategy is examined. In other words, the rank-based strategy and mean-DIF procedure are performed with the purpose of identifying a set of DIF-free items to serve as pure anchors for the subsequent DIF assessment.

Procedure 1: Forward Approach

In most DIF detection procedures, a prior set of pure anchors is needed to align the scales of both reference and focal groups. Meulders and Xie (2004) suggest using an explanatory approach to detect DIF when such pre-identified DIF-free items are not available. In particular, one can follow a forward approach to assess DIF. In a forward approach, every single item is tested for DIF using an all-other anchor method, and then the items flagged as having DIF (i.e., items with significant interactions) are included in the final model and assessed for DIF simultaneously; however, it is not yet known whether the forward approach method functions appropriately in the DIF assessment.

Here, based on the suggestion by Meulders and Xie (2004), I have developed and implemented a "forward procedure" with the GLMM method to detect DIF items in a test. This procedure follows four steps, which can be conducted as follows:

Step 1: Test item 1 in the GLMM DIF model by setting all other items in the test as anchor items.

Step 2: Test item 2 in the GLMM DIF model by setting all other items in the test as anchor items.

Step 3: Repeat step 2 for every item in the test until the last item has been tested for DIF.

Step 4: Include items that displayed significant DIF estimates (δ^{β}) into the final model and test for DIF for these items simultaneously.

The 0.05 significance level (two-tailed) is used to test the null hypothesis that the DIF parameter is statistically different from zero. We can use a 10-item test as an example. In the forward procedure, DIF is tested on item 1 first by estimating the group-by-item interaction effect of δ_1^{β}. Items 2 through 10 are constrained to have no interaction effect in the GLMM DIF model. Accordingly, the structural model part of the GLMM DIF model for testing DIF on item 1 is $\eta_{p1} = \theta_p - \left(\beta_1 X_1 + \delta_1^{\beta} X_1\right)$. If the Wald test outcome for δ_1^{β} is statistically significant based on the α-level of .05, then item 1 will be considered as having non-zero DIF. Likewise, when testing DIF on item 2, item 1 and items 3 through 20 are constrained to have no DIF parameters tested in the model; whereas item 2 is allowed to have a DIF parameter, which is δ_2^{β}. The DIF model is thus $\eta_{p2} = \theta_p - \left(\beta_2 X_2 + \delta_2^{(\beta)} X_2\right)$. If the statistical test for δ_2^{β} is significantly different from zero, then item 2 will be deemed as having non-zero DIF. The same strategy is applied to all other items. After all items have been tested for DIF, the K items with significant δ^{β} values are then included simultaneously in one model for the final DIF assessment. The final model can be expressed as $\eta_{pi} = \theta_p - \sum_{k=I} (\beta_i X_{ik} + \delta_i^{(\beta)} W_{pik}); i = 1, \ldots 10$. However, for the purpose of model identification, at the most only i -1 items can be tested for DIF in one model.

Procedure 2: Iterative Procedure

Originally, Marco (1977) and Lord (1980) brought up the idea of "purification" for DIF analysis. Many scholars have subsequently tested purification procedures implemented with IRT-based DIF detection methods (Candell & Drasgow, 1988; Lord, 1980; Park & Lautenschlager, 1990) and non-IRT-based methods (French & Maller, 2007; Holland & Thayer, 1988; Wang, Shih & Yang, 2009). In this study, the iterative purification procedure was implemented with the GLMM for DIF detection. In the GLMM iterative purification, the first three steps are the same as the first three steps in the aforementioned forward approach. The difference is that if an item is detected as having DIF, then it will not be included in the anchor set for the second iteration of DIF detection. The iterative purification procedure is actually a two-stage purification process, which was implemented as follows: (1) conduct initial DIF screening on all items, and then obtain the "purified anchor" set by removing potential DIF items identified at the first step; (2) test DIF for all items with the purified anchor; and (3) enter all items having significant DIF in one model and estimate DIF.

The GLMM with iterative purification uses the following five steps:

Step 1: Test item 1 in the GLMM DIF model by setting all other items in the test as anchor items.

Step 2: Test item 2 in the GLMM DIF model by setting all other items in the test as anchor items.

Step 3: Repeat step 2 for every item in the test until the last item is tested for DIF.

Step 4: Remove the DIF items that were identified in the previous DIF assessment from the anchor set. Test DIF for all items using the new anchor set by applying the all-other anchor method.

Step 5: Enter all items having significant DIF in one model and estimate DIF.

Steps 1 to 3 serve as the initial DIF screening. For example, assume the test contains 10 items and items 1 and 2 are detected as having significant DIF in the initial DIF screening. In step 4, items 1 and 2 are then removed from the anchor set so that the new anchor set now consists of items 3 through 10. The second round of DIF assessment

is conducted with this "purified" anchor set. During this process, DIF items are tested together with the new anchor set. Specifically, when detecting item 1 for DIF, item 1 is allowed to have an interaction effect by estimating δ_1^{β}. If δ_1^{β} is significantly different from zero, then item 1 will be classified as a DIF item. Likewise, if δ_2^{β} is significantly different from zero, then item 2 will be classified as a DIF item. For non-DIF items, each item is tested separately using the new anchor set with the exception of the item of interest. Specifically, when analyzing item *i* for DIF (*i* = 3, …, 10), items 3 through 10 (excluding item *i*) are constrained to have no person-by-item interaction effect and serve as anchors, while item *i* is allowed to have an interaction effect by estimating δ^{β}. If δ^{β} is significantly different from zero, then item *i* is deemed as having DIF. Assume this time items 3 and 5 are detected as having significant DIF. Then in step 5, items 1, 2, 3, and 5 are entered into one model and tested for DIF.

To eliminate the impact of DIF contamination on DIF assessment, Shih and Wang (2009) adopted a DIF-free-then-DIF strategy in the MIMIC method for DIF detection. The DIF-free-then-DIF strategy identifies DIF-free item(s) first, and then uses identified DIF-free items to serve as anchor item(s) for subsequent DIF item classifications. In addition, Woods (2009) developed a rank-based strategy for locating DIF-free items. In this study, these two procedures are applied to the GLMM method. The details for these two procedures in GLMM are presented below.

Procedure 3: Mean-DIF Procedure

Shih and Wang (2009) examined an iterative MIMIC procedure to locate a set of up to four DIF-free items as pure anchors so that the MIMIC method could be performed properly. In this procedure, DIF-free items are identified based on the mean absolute values of DIF indices over all iterations. For this reason, it will be called the mean-DIF procedure in this study. This iterative procedure contains the following steps:

Step 1: Set item 1 as the anchor item, test DIF on all other items with constant anchor item method, and obtain a DIF index for each tested item.

Step2: Set item 2 as the anchor item, test DIF on all of the other items with the constant anchor item method, and obtain a DIF index for each tested item.

Step 3: Repeat step 2 until the last item is set as the anchor.

Step 4: Compute the mean absolute values of DIF indices for each item over all iterations and locate the item that has the smallest mean absolute value. The located items are most likely the DIF-free items and can serve in the anchor set.

The purpose of the procedure is to locate a set of DIF-free items to serve as anchors. The underlying logic of this iterative procedure was given by Shih and Wang (2009) in the study of the MIMIC iterative procedure. Specifically, in step 1, when item 1 (i.e., the chosen anchor item) is indeed a DIF-free item, all other items will be classified correctly. That means the DIF parameter (δ^{β}) is significantly different from zero for DIF items, while δ^{β} is not significantly different from zero for DIF-free items; however, if item 1 is actually a DIF item but served as an anchor, then all of the other items will be classified incorrectly. The result is that actual DIF-free items may be misclassified as DIF items with the DIF magnitude of item 1, while the DIF items may be misclassified as DIF-free items. All items take turns serving as anchors, and the same logic is applied to all items until the last item functions as anchor. Under the assumption that there are more DIF-free items than DIF items in the test, the DIF-free items would be classified as displaying DIF less frequently than the DIF items after all iterations. Therefore, the mean absolute value of DIF indices for actual DIF-free items will be smaller than those for items having DIF. Based on this logic, we can choose the items with the smallest mean absolute DIF indices as the anchor items. Obviously, to apply the mean-DIF strategy appropriately, one assumption has to be made: there are more DIF-free items in the test than DIF items.

Procedure 4: Rank-based Strategy

Many purification procedures are complicated and time consuming. To address this concern, Woods (2009), among others, suggested using a noniterative, rank-based strategy to identify a set of DIF-free items. In her simulation study, a graded model for items with ordinal responses was used to generate data, and the rank-based strategy was

investigated in the context of likelihood-ratio (LR) comparisons between nested two-group IRT models. The procedure starts by testing all items for DIF using the all-other anchor method. Then the items should be ranked based on the ratios of LR statistics to the number of free parameters. The items with the smallest ratio would be the best candidates to serve as anchors. The magnitude of the LR statistic is positively related to the degree to which an item functions differently between two groups, and this is why the rank-based strategy is a good method for locating DIF-free items (Woods, 2009). In the current study, the rank-based strategy is implemented and investigated in the GLMM model. In the GLMM DIF method, the DIF-free item is identified in such a way that the Wald test statistic reflects the degree to which the item functions differently between two groups, with larger values indicating greater DIF. To apply Wood's idea to the GLMM DIF model with the purpose of identifying DIF-free items, the following steps are implemented:

Step 1: Test DIF on all items by using the all-other anchor method. Collect the Wald statistic for each item.

Step 2: Rank the items in order based on the value of Wald statistics.

Step 3: Locate the desired number of items that have the smallest values of the Wald statistic. The located items are the best candidates to serve as the anchor set.

No conclusive decision has been made on the optimal number of anchors in DIF detection. A longer anchor set can yield higher power in DIF detection than a shorter anchor set (Wang & Yeh, 2003). However, each additional anchor increases the chance of contamination (Woods, 2009). In terms of IRT-based DIF methods with pure anchor sets, it has been demonstrated that one anchor can yield an acceptable DIF detection rate and four anchors is long enough to produce stable results in parameter estimation and DIF detection (Wang, 2004; Wang & Wilson; Wang & Yeh, 2003). One of the major purposes of this study is to evaluate the frequency of accurately identifying a set of DIF-free items using the mean DIF procedure and the rank-based strategy. The two procedures are evaluated on the identification of 1, 2, 3, and 4 DIF-free items.

In order to examine the performance of the above four purification procedures in the GLMM model, a simulation study was conducted. The data generation procedure, simulation conditions, and evaluation criteria are described below.

3.2 Simulation Study

Data Generation

Data generation was accomplished using SAS/IML® software (See Appendix 1 for syntax). I adapted existing SAS/IML® code (Kromrey et al., 1999) to generate examinees' latent ability and item responses (zeros and ones).

In GLMM DIF models, responses are explained on the basis of subject effect, item effects, and person-by-item interaction effects. The latent abilities were generated from a normal distribution with group-specific means and variances. For model identification purposes, the mean of the reference group was fixed to be 0. The mean ability difference between the studied groups was set to be 0.2.

The Rasch model was selected for the data generation. All of the α parameters were set at unity. The item difficulties were arbitrarily set as -1, -0.5, 0, 0.5, and 1. This pattern was repeated four times for a 20-item test and 10 times for a 50-item test. The item difficulty distribution thus had a mean of 0 and standard deviation of 0.73 in the 20-item test, and it had a mean of 0 and standard deviation of 0.71 in the 50-item test. The equivalently distributed item difficulty parameters between the 20-item and 50-item conditions were accomplished in the data generation. Ultimately, the goal of using equivalently distributed parameters was to avoid certain situations, such as not knowing from where an effect is derived. For example, the DIF effect may come from the simulated DIF or item difficulties. According to the simulation design in this study, the percentage of DIF items and DIF directions should both be considered to evenly distribute item difficulties in relation to the distribution of DIF items. Table 1 lists the item parameters for the reference group in the 50-item test. The first 20 items in the 50-item test are those selected for the 20-item test.

The item response data were simulated for a unidimensional model, and item bias was introduced by adding the magnitude of DIF to the item difficulty parameter in the focal group (Miller, 1992; Shepard, Camilli, & Williams, 1985). The magnitude of DIF

present in the target item was 0.6 on the log-odds-ratio scale. This value was comparable to those used in previous research, including 0.4 (Wang & Yeh, 2003); 0.75 (Navas-Ara & Gomez-Benito, 2002); and 0.48 and 0.64 (Swaminathan & Rogers, 1990). Moreover, according to the ETS criterion (Zieky, 1993), 0.6 represents a moderate or large magnitude of DIF effect, which will raise a concern.

For instance, considering evenly distributed item difficulties, if the 20-item test had 20% DIF items, the DIF items were items 4, 7, 14, and 17 with item difficulties of 0.5, -0.5, 0.5, and -0.5, respectively. In addition, under one-sided DIF direction, each DIF item was simulated to favor the reference group; therefore, for each one-sided DIF item, the *b* parameter for the focal group was 1.1, 0.1, 1.1, and 0.1, respectively.

Table 3.1 Item Parameter Values for the Reference Group

Item	*b*
1	-1
2	-0.5
3	0
4	0.5
5	1
6	-1
7	-0.5
8	0
9	0.5
10	1
11	-1
12	-0.5
13	0
14	0.5
15	1
16	-1
17	-0.5
18	0
19	0.5
20	1

Table 3.1 (continued)

Item	*b*
21	-1
22	-0.5
23	0
24	0.5
25	1
26	-1
27	-0.5
28	0
29	0.5
30	1
31	-1
32	-0.5
33	0
34	0.5
35	1
36	-1
37	-0.5
38	0
39	0.5
40	1
41	-1
42	-0.5
43	0
44	0.5
45	1
46	-1
47	-0.5
48	0
49	0.5
50	1

First, to generate item responses (0/1 values), the logit of obtaining the correct response on item i for examinee p was calculated using equation 3.1. All of the variables were defined in chapter 2. The logit was

$$\eta_{pi} = \theta_p - \sum_{k=I_a+1}^{I} \left(\beta_k X_{ik} + \delta_k^{\beta} W_{pik} \right). \qquad 3.1$$

Second, this logit was transformed into a probability through the following logit link:

$$P(Y_{pi} = 1) = \frac{\exp(\eta_{pi})}{1 + \exp(\eta_{pi})}. \qquad 3.2$$

The probabilities of examinees obtaining the correct response on each item (P_i) were compared to a uniform random number (U) on the range from 0 to 1. For each item, if $P > U$, then the examinee would receive a score of 1 for that item (i.e., get the correct response). However, if $P < U$ or $P = U$, then the examinee would be assigned a score of 0 (i.e., an incorrect response). As a result, a vector of 1s and 0s was generated in SAS to represent the item responses. The difference between the reference and the focal group in the data generation was the presence of the group-by-item interaction effect. For the reference group, the interaction effect was zero. The reference and focal groups had all other effects in common. The presence of the group-by-item interaction effect indicated that examinees from different groups had differing probabilities of answering an item correctly after being matched on the ability of interest.

The data for this study was generated using SAS software, Version 9.2 of the SAS System for PC (Copyright © 2010 SAS Institute, Inc). The Proc GLIMMIX procedure by default estimates parameters in generalized linear mixed models by applying pseudo-likelihood techniques as in Wolfinger and O'Connell (1993) and Breslow and Clayton (1993). The tests of hypotheses for the DIF effect were based on the Wald-type test and the estimated variance-covariance matrix (SAS Institute, Inc. 2010). To classify DIF, the 0.05 significance level was used to justify a claim of the DIF parameter being statistically different from zero.

Simulation Design

Four variables were considered in the simulation study: (1) sample size--the number of examinees in a test [1,000 total with 500 for the reference group and 500 for the focal group and 2,000 total with 1,000 for the reference group and 1,000 for the focal group (1000R/1000F)]; (2) test length--the number of items in a test (20 items, 50 items); however, the test length was fixed at 20 items to evaluate the forward and iterative purification procedures; (3) percentage of DIF items in a test (0%, 20%, 40%); and (4) DIF direction (one-sided, dominant, and balanced). A total of 28 simulation conditions were generated. Four baseline samples of 1,000 and 2,000 examinees were generated for 20-item and 50-item tests. In the baseline models, no DIF items were simulated. Thirty-six biased conditions were simulated to examine the purification procedures. The biased conditions were based on a completely crossed design with $2 \times 2 \times 2 \times 3=24$ combinations based on the number of examinees (1,000 and 2,000) × the test length (20 and 50) × the percentage of biased items (20%, 40%) × DIF direction (one-sided, dominant, and balanced). Each simulation condition was replicated 100 times in order to obtain reliable results.

Simulation Conditions

Group size. The number of subjects was varied to examine the effects of group size on item purification, that is, to examine effects when both groups were equally large ($N_{refer} = N_{focal}$=1,000) versus equally small ($N_{refer} = N_{focal}$=500). Group size affects the power of the GLMM procedure through its effect on parameter estimation and test statistics. In the GLMM DIF model, the test statistics may not be valid indicators for the presence of DIF in a small sample. Many studies have indicated that sample size affects the performance of DIF detection models. For example, Finch (2005) compared the MIMIC model for DIF detection with MH, the simultaneous item bias test (SIBTEST, Shealy & Stout, 1993b), and IRT likelihood ratio test (LR) (IRT-LR, Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988). The results indicated that the small sample size in the focal group resulted in a deflated true-positive rate. In the study by Shin and Wang where the MIMIC method was used to identify DIF-free items, samples were simulated to have from 500 to 1,500 subjects per group (Shin & Wang, 2009). Their simulation study found that the smaller sample size yielded lower rates of

accuracy; therefore, it would be useful to know how different GLMM DIF methods would perform with varying number of subjects.

There are two reasons to simulate 500 and 1,000 subjects per group in the current study. First, large-scale assessments commonly test this quantity of students with special needs. With the institution of the Individuals with Disabilities Education Act (IDEA) and the Americans with Disabilities Act (ADA), states were encouraged to include individuals with disabilities in various levels of assessment (Kamata & Vaughn, 2004; Pitoniak & Royer, 2001). For example, the tested population for the 2010 Grade 3 Florida Comprehensive Assessment Test® (FCAT) in Reading included 1,533 students with emotional or behavioral disabilities and 588 students with intellectual disabilities (Florida Department of Education, 2010). For equity purposes, DIF assessment with purification is highly recommended when testing students with special needs. Knowing whether the GLMM DIF purification is robust to a sample size as small as 500 subjects per group would be useful for researchers and test developers. Second, such a selection of sample sizes makes it possible to compare the current study with other DIF simulation studies that used similar numbers of subjects. Many DIF simulation studies have manipulated sample sizes ranging from 500 to 2,000 per group (De Boeck & Leuven, 2008; Park, Lautenschlager & Georgia, 1990; Shih & Wang, 2009; Wang & Su, 2004a, 2004b; Wood, 2009).

Test length. Test length was also manipulated in this study. Many DIF studies have shown increased statistical power and decreased type I error rates with longer tests in DIF analysis with purification (Clauser et al., 1993; Narayanan & Swaminathan, 1996). Moreover, it was found that shorter tests and higher percentages of DIF resulted in decreased efficiency of test purification. (Clauser, Mazor, & Hambleton, 1993; Donoghue, Holland, & Thayer, 1993; Fidalgo et al., 2000; Wang & Su, 2004). A variety of DIF simulation studies have manipulated test length with a range of 20 to 80 items (e.g., Clauser et al., 1993; Cohen, Kim & Wollack 1996; Finch, 2005; Narayanan & Swaminathan, 1994, 1996; Rogers & Swaminathan, 1993; Shih & Wang, 2009). Therefore, a 20-item test and a 50-item test were included to represent a short and a long test in order to investigate the impact of the test length on the performance of GLMM test purification. In practice, it is common to have 20 to 50 items in large-scale assessments.

For example, the numbers of multiple-choice items in FCAT tests across all subjects and all grades range from 28 to 45. The selection of the test length, therefore, reflects theoretical and practical considerations.

DIF direction. Depending on which group (i.e., the reference group or the focal group) DIF favors, Wang and Su (2004a) categorized DIF into three categories: one-sided, dominant, and balanced. One-sided DIF refers to the case where all of the DIF items favor the same group (in most cases, the reference group). Dominant DIF represents the situation in which most of the DIF items favor one group while the rest of the DIF items favor another group. Balanced DIF is a special DIF pattern where some DIF items favor the reference group and the others favor the focal group, with equal magnitudes of DIF. In this case, the DIF magnitudes for the reference and focal groups cancell out as a whole; therefore, neither group is favored.

Some researchers have concluded that the DIF direction also plays an important role in the performance of DIF detection. For example, Wang and Su (2004a) indicated that for balanced DIF, the MH procedure works well, even when there are more than 20% of DIF items in the test. However, for one-sided DIF, even when 10% of the test items are DIF items, the MH procedure lose control of type I error. In this study, bias was introduced into the item difficulty parameters in three directions: one-sided, dominant and balanced. This simulation of DIF covers all possible DIF directions so that the effect of DIF direction on test purification can be investigated. Under the one-sided DIF direction, all DIF items were generated to favor the reference group. To generate the dominate DIF direction, three of fourth DIF items were simulated to favor the reference group, whereas one of fourth were simulated to favor the focal group. For example, if 40% of items in 20-item test are simulated to be DIF items, 6 of 8 DIF items were simulated to favor the reference group whereas 2 DIF items were simulated to favor the focal group. For the balanced DIF condition, 4 items were simulated to favor the focal group whereas the rest of the 4 DIF items were favoring the focal group.

Percentage of DIF items in the test. One conclusion that has been made across existing purification studies is that the degree of DIF contamination in the test is a critical factor that affects the performance of purification (Candell & Drasgow, 1988; Shin & Wang, 2009; Wang & Su, 2004a, 2004b). In general, the more the DIF contamination, the

worse the purification would be. Accordingly, the purification procedures are useful in controlling type I error rate and in maintaining high power when tests only contain certain amounts of DIF items (Lautenschlager, Flaherty, & Park, 1994; Miller & Oshima, 1992; Park & Lautenschlager, 1990; Holland & Thayer, 1988). For example, Wang, Shih, and Yang (2009) observed increased type I error rates and decreased power in the MIMIC purification when there were 20% one-sided DIF items in the test. Candell and Drasgow (1988) reported a dramatic impact of the number of biased items on bias classification using an iterative linking procedure. The percentage of DIF items in the test, combined with the DIF direction, made the purification more complicated. In the current simulation, three DIF directions (i.e., one-sided, dominant, and balanced) are crossed with two levels of DIF contamination (20% and 40% DIF items), which represent medium and large levels of DIF contaminations, respectively. It is desirable to evaluate how the purification procedures described in this chapter perform under various DIF contamination conditions.

3.3 Evaluation Criteria

The effectiveness of the forward procedure and iterative purification are evaluated using two criteria: type I error and power. Type I error is the proportion of times a DIF-free item is incorrectly identified as having significant DIF by the procedure. In this study, the type I error is calculated by taking the total number of simulated DIF-free items incorrectly identified over 100 replications divided by the total number of simulated DIF-free items. On the other hand, power is the proportion of times a DIF item is indeed identified as having significant DIF by the procedure. Power is calculated by taking the total number of simulated DIF items correctly identified over 100 replications divided by the total number of items simulated to have DIF.

For the mean-DIF procedure and rank-based strategy, the purpose is to locate a set of DIF-free items as the anchor set; thus, the evaluation criterion is the accuracy of the identification of a set of DIF-free items. In this study, the accuracy rates were calculated for the identification of 1, 2, 3 and 4 DIF-free items. The proportion of times a located item was indeed DIF-free over 100 replications provided the estimate of the accuracy. This proportion was also compared with a baseline criterion, which is the chance level of

random selection of an anchor item in a test with a certain number of DIF items. Shih and Wang (2009) recommended using equation 3.3 to calculate the chance level for random selection of DIF-free items. They used

$$\binom{T\times(1-D\%)}{t}\Big/\binom{T}{t} \qquad 3.3$$

where T is the total number of items in a test, $D\%$ is the proportion of DIF items in the test, and t is the number of DIF-free items that the procedure will locate. For example, in a 10-item test containing 20% DIF items, the chance level of locating a set of four DIF-free items will be $\binom{10\times(1-20\%)}{4}\Big/\binom{10}{4}$ = 70/210=0.33. A useful procedure is expected to produce a higher rate of accuracy than the chance level.

CHAPTER 4

SIMULATION RESULTS

In this chapter, simulation study results are presented to evaluate the forward procedure, iterative purification procedure, rank-based strategy, and mean-DIF procedures in the GLMM DIF model. Type I error and power were examined to evaluate the performance of the forward procedure and iterative purification. On the other hand, the mean-DIF procedure and rank-based strategy were evaluated in terms of the rate of accuracy in locating a set of up to four DIF-free items.

Forward Procedure and Iterative Purification Procedure

To evaluate the effectiveness of the forward procedure and iterative purification procedure, three variables were manipulated in this simulation study: sample size, percentage of DIF items in the test, and DIF direction. The test length was fixed at 20 items.

Tables 4.1 to 4.4 present the type I error rates and power across all conditions for the forward iterative purification procedures. For the type I error, the numbers reported were averaged across all DIF-free items. For the power, the numbers reported were averaged across all DIF items.

Type I Error Rates

Table 4.1 lists the type I error rates for the forward procedure and iterative purification procedure when there were no DIF items in the test. The type I error rates for the no-DIF conditions served as a baseline, which can help the evaluation of the effectiveness of the purification procedures in DIF conditions (French & Maller, 2007). It appears that both methods yielded a type I error around its expected value of 0.05 under the no-DIF conditions. Specifically, for the forward procedure, the range of type I error rates were from 0.037 to 0.043 with a mean of 0.040. For the iterative purification procedure, the range of type I error rates were from 0.041 to 0.045 with a mean of 0.043. These values are all slightly below the nominal α-level of 0.05. In addition, the confidence interval for these type I error rates was calculated as: $.05 \pm \sqrt{.05(.95)/n_{rep}}$,

which is [.028, .072]. Thus, it appears that both methods were able to yield good control over type I errors when tests did not contain any DIF items.

Table 4.1 Type I Error When No DIF Items are in the Test

% DIF Items	Sample Size	Purification Method	Type I Error Rate
No DIF	500F/500R	Forward	0.037
		Iterative	0.041
	1000F/1000R	Forward	0.043
		Iterative	0.045

The Type I error rates for the one-sided DIF conditions are summarized in Table 4.2. When 20% of the test items were DIF items, it was observed that both procedures yielded similar type I error rates in the condition of 500F/500R. The error for the forward procedure was 0.044, while the error was 0.043 for the iterative purification procedure. As the sample size increased to 1000F/1000R, the type I error rates also increased in both procedures. The increase of error in the forward procedure was larger than that in iterative purification procedure. Overall, both procedures yielded good control over type I errors under the 20% one-sided DIF conditions. The only exception was that the forward procedure yielded a slightly inflated type I error rate in the condition of 1000F/1000R. When the percentage of DIF items reached 40%, both procedures began to yield inflated type I error rates. Specifically, the forward procedure yielded a seriously inflated type I error rate, whereas the iterative purification procedure yielded a marginally inflated error rate. The type I error rates ranged from 0.111 to 0.213 for the forward procedure, and from 0.059 to 0.118 for the iterative purification procedure. The inflation in error rate for the iterative purification procedure was smaller than that for the forward procedure.

Under the dominant DIF conditions, both procedures continued to maintain acceptable type I error rates, even with up to 40% DIF items in the test. The only exception was that the forward procedure produced a type I error above 0.05 in the condition of 40% DIF and 1000F/1000R. As shown in Table 4.3, the type I error rates ranged from 0.036 to 0.053 for the forward procedure and from 0.036 to 0.043 for the

iterative purification procedure. The forward procedure produced slightly lower error rates than the iterative purification procedure under the 20% dominant DIF conditions. Conversely, when there were 40% dominant DIF items in the test, the iterative purification procedure was slightly superior to the forward procedure in controlling type I errors. Besides, under the 40% DIF items condition, the type I error increased with the increase of the sample size for both procedures.

Table 4.4 shows the type I errors in the balanced conditions. It was found that both procedures yielded a type I error rate under 0.05 in all simulation conditions. Thus, both procedures yielded well-controlled error rates, even when there were up to 40% DIF items in the tests. Specifically, the type I error rates ranged from 0.031 to 0.043 with the mean of 0.037 for the forward procedure, and from 0.040 to 0.049 with the mean of 0.045 for the iterative procedure. The percentage of DIF items appeared to have little effect on the performance of forward and iterative procedures under the balanced conditions. Also, it was found that the forward procedure outperformed the iterative purification procedure in controlling type I error rates under the balanced conditions.

Overall, both procedures were able to yield a satisfactory type I error rate when 20% of the test items were DIF items. Both procedures lost control of type I error when 40% of the test items were one-sided DIF items. However, under the dominant and balanced DIF conditions, both procedures continued to have good control on type I errors, even when the percentage of DIF reached 40%. Thus, the percentage of DIF items was not a critical factor for a type I error; rather, the overall DIF contamination in the whole test affected the forward and iterative purification procedures. Wang and Su (2004) proposed to use the mean item difficulty difference (MIDD) to measure the DIF contamination. In the balanced DIF conditions, the MIDD would be zero. Under the one-sided DIF and dominant DIF, the larger the percentage of DIF items, the higher the probability of DIF contamination. In general, with the same amount of DIF items in the test, the one-sided DIF conditions contain the greatest DIF contamination, followed by dominant DIF conditions, with the least contamination for balanced DIF conditions. The iterative procedure performed better than the forward procedure in generating lower error rates when a larger amount of DIF contamination was involved in the test.

Power

In terms of power, it appeared that both procedures were able to maintain a satisfactory power level with one exception, which was the forward procedure in the conditions of one-sided DIF and 500F/500R. Tables 4.2 to 4.4 include the power for forward and iterative purification procedures across all simulation conditions.

Several conclusions can be drawn based on the simulated data sets. First, it appeared that the larger the sample size, the higher the power for both procedures. Sample size was a critical factor that determined the correct identification of DIF items. This finding was clearly demonstrated when there were large amounts of DIF contamination in the tests. For example, in the condition of 40% one-sided DIF, when the sample size increased from 500R/500F to 1000R/1000F, the power was increased from 0.705 to 0.880 for the forward procedure and from 0.813 to 0.974 for the iterative purification procedure.

Table 4.2 Type I Error and Power Under One-Sided DIF

% DIF Items	Sample Size	Purification Method	Type I Error Rate	Power
20%	500F/500R	Forward	0.044	0.935
		Iterative	0.043	0.955
	1000F/1000R	Forward	0.058	0.980
		Iterative	0.048	1.000
40%	500F/500R	Forward	0.213	0.705
		Iterative	0.118	0.813
	1000F/1000R	Forward	0.216	0.880
		Iterative	0.059	0.974

Second, when there were larger amounts of DIF contamination in the test, the iterative purification procedure was superior to the forward procedure in maintaining adequate power. The largest power improvement was found when 40% of the test items

were one-sided DIF items. Specifically, in the condition with 40% one-sided DIF and 500R/500F, the difference of the power in the forward procedure and iterative procedure is .108. However, for the 20% dominant DIF and balanced DIF, the forward procedure and iterative purification procedure yielded a similar level of power.

Third, the large DIF contamination caused not only inflation in incorrect identification but also a deflation in the identification of DIF items. DIF contamination is the overall magnitude of DIF within the items (Wang & Su, 2004a). Both of the number of DIF items and DIF direction affect the DIF contamination. For the balanced DIF, the DIF contamination is zero even there are up to 40% of DIF items in the test since the DIF are cancelled out between groups. With the same percentage of DIF items in the test, the one-sided DIF contains largest DIF contamination followed by the dominant DIF, then balanced DIF. Therefore, it is not surprising to find that the balance DIF conditions have the largest power compared with one-sided and dominant conditions holding the percentage of DIF items in the test constant. For example, in the condition with 40% DIF items, 500R/500F, and one-sided DIF, the forward and iterative procedures had a power of 0.705 and 0.813, respectively. However, in the condition with 40% DIF items, 500R/500F, and balanced DIF, the forward and iterative procedures had a power of 0.978 and 0.976, respectively.

Table 4.3 Type I Error and Power Under Dominant DIF

% DIF Items	Sample Size	Purification Method	Type I Error Rate	Power
20%	500F/500R	Forward	0.036	0.975
		Iterative	0.043	0.968
	1000F/1000R	Forward	0.032	1.000
		Iterative	0.038	0.940
40%	500F/500R	Forward	0.040	0.940
		Iterative	0.036	0.961
	1000F/1000R	Forward	0.053	1.000
		Iterative	0.042	1.000

Table 4.4 Type I Error and Power Under Balanced DIF

% DIF Items	Sample Size	Purification Method	Type I Error Rate	Power
20%	500F/500R	Forward	0.031	0.958
		Iterative	0.040	0.980
	1,000F/1,000R	Forward	0.036	0.990
		Iterative	0.041	1.000
40%	500F/500R	Forward	0.038	0.978
		Iterative	0.043	0.976
	1,000F/1,000R	Forward	0.043	0.990
		Iterative	0.049	0.980

Rank-Based Strategy and Mean-DIF Procedure

The purpose of the rank-based strategy and mean-DIF procedure is to identify a set of DIF-free items as anchors. To evaluate each method's performance, the rates of accuracy in locating the DIF-free items were calculated over 100 replications. The rate of accuracy is the percentage of actual DIF-free items located by the procedure. To assess whether a procedure is useful in locating DIF-free items, the rate of accuracy was compared with a baseline criterion. This baseline criterion is the chance of correctly selecting DIF-free items at random, which is also called the chance level. It is expected that a useful procedure can yield a rate of accuracy that is higher than its corresponding chance level (Shih & Wang, 2009). In this study, the rank-based strategy and mean-DIF procedure were evaluated for locating one, two, three, and four DIF-free items. Tables 4.6 to 4.13 list the rates of accuracy in locating up to four DIF-free items and the corresponding chance levels for the rank-based strategy and mean-DIF procedures. In general, five conclusions have been drawn from these results.

Rate of Accuracy

First, both procedures appeared to work very well in selecting DIF-free items within the GLMM DIF context. All accuracy rates were higher than their corresponding chance levels, even when multiple DIF-free items were identified. Both procedures yielded perfect or close-to-perfect rates of accuracy in most conditions. For example,

when one DIF-free item was identified in the 20-item test, the rank-based strategy yielded perfect rates of accuracy in 9 out of 12 data sets, and the mean-DIF procedure yielded perfect rates in 10 out of 12 data sets. Even when four DIF-free items were identified, the perfect rates were produced in 8 out of 12 data sets for the rank-based strategy and 9 out of 12 data sets for the mean-DIF procedure. This pattern also can be observed in a 50-item test in Tables 4.9 to 4.12. In the 50-item test, using rank-based strategy, the perfect accuracy rates were observed in 9 out of 12 data sets in locating one DIF-free item whereas 7 out of 12 data sets in locating two DIF-free item. Using mean-DIF procedure, the accuracy rates appeared in 10 out of 12 data sets in locating one DIF-free item whereas 8 out of 12 data sets in locating two DIF-free items.

Second, both procedures yielded a higher rate of accuracy when the sample size was larger. Specifically, for the rank-based strategy, the accuracy rate in locating one DIF-free item under the condition of 500R/500F was 86%, and the rate was increased to 97% when the sample size increased to 1000R/1000F holding other conditions constant. The same outcome was observed with the mean-DIF procedure: the accuracy rate in locating one DIF-free item increased from 95 to 100 when the sample size increased from 500R/500F to 1000R/1000F. The rates of accuracy in locating two, three and four DIF-free items were also higher for 1000F/1000R than 500F/500R. The same conclusion can be drawn in the 50-item test.

Third, DIF direction affected the efficiency of both procedures in locating DIF-free items. Specifically, the lowest rates of accuracy were found when 40% of the items were one-sided DIF items. However, with the same percentage of DIF items in the test, both procedures almost always could correctly locate a set of up to four DIF-free items if all DIF items favored one group. Similarly, with the same percentage of DIF items in the test, both procedures performed better in dominant DIF conditions than in one-sided DIF conditions. Therefore, DIF contamination, determined by the percentage of DIF items and DIF directions, was a critical factor that affected the accuracy rate of identifying DIF-free items.

Table 4.5 Rates of Accuracy and Chance Levels in Locating One DIF-Free Item with Rank-Based Strategy and Mean-DIF Procedure in a 20-item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	80	Rank	100	100
			Mean-DIF	99	100
	40%	60	Rank	86	97
			Mean-DIF	95	100
Dominant	20%	80	Rank	100	100
			Mean-DIF	100	100
	40%	60	Rank	99	100
			Mean-DIF	100	100
Balanced	20%	80	Rank	100	100
			Mean-DIF	100	100
	40%	60	Rank	100	100
			Mean-DIF	100	100

Table 4.6 Rates of Accuracy and Chance Levels in Locating Two DIF-Free Items with Rank-Based Strategy and Mean-DIF Procedure in a 20-item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	63	Rank	100	100
			Mean-DIF	99	100
	40%	35	Rank	76	91
			Mean-DIF	90	100
Dominant	20%	63	Rank	100	100
			Mean-DIF	100	100
	40%	35	Rank	98	100
			Mean-DIF	100	100
Balanced	20%	63	Rank	100	100
			Mean-DIF	100	100
	40%	35	Rank	99	100
			Mean-DIF	100	100

Table 4.7 Rates of Accuracy and Chance Levels in Locating Three DIF-Free Items with Rank-Based Strategy and Mean-DIF Procedure in a 20 item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	49	Rank	100	100
			Mean-DIF	99	100
	40%	19	Rank	68	82
			Mean-DIF	87	100
Dominant	20%	49	Rank	100	100
			Mean-DIF	100	100
	40%	19	Rank	98	100
			Mean-DIF	99	100
Balanced	20%	49	Rank	100	100
			Mean-DIF	100	100
	40%	19	Rank	99	100
			Mean-DIF	100	100

Table 4.8 Rates of Accuracy and Chance Levels in Locating Four DIF-Free Items with Rank-Based Strategy and Mean-DIF Procedure in a 20 item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	38	Rank	100	100
			Mean-DIF	99	100
	40%	10	Rank	49	71
			Mean-DIF	84	100
Dominant	20%	38	Rank	100	100
			Mean-DIF	100	100
	40%	10	Rank	97	100
			Mean-DIF	99	100
Balanced	20%	38	Rank	100	100
			Mean-DIF	100	100
	40%	10	Rank	99	100
			Mean-DIF	100	100

Fourth, with every additional DIF-free item identified, the accuracy rates dropped. In general, more DIF contamination correlated to lower accuracy rates. For example, under the condition of 20 items and sample sizes of 500F/500R, the rank-based strategy yielded rates of accuracy of 99, 98, 98, and 90 for 40% dominant DIF items in identifying one, two, three and four DIF-free items, respectively. However, it was observed that with all other conditions remaining the same, the rank-based strategy yielded the accuracy of 86, 75, 68, and 49 for 40% one-sided DIF items in the identification of one, two, three and four DIF-free items, respectively. The accuracy rates were much lower in the one-sided DIF situations than in the dominant DIF situations. Basically, the same conclusions for the mean-DIF procedure can be drawn here. With more DIF-free items identified, the accuracy rates of DIF-free item identification were generally reduced for the mean-DIF procedure. This was clearly demonstrated in 40% one-sided DIF conditions. In addition, the deflation rate for accuracy was much more serious in the rank-based strategy than in the mean-DIF procedure. Specifically, under the condition of 20 items and 500F/500R, the mean DIF procedure yielded the rates of accuracy of 95, 90, 87, and 84 for 40% of the one-sided items in identifying one, two, three and four DIF-free items, respectively. These accuracy rates were higher than those for rank-based strategy under the same conditions. Moreover, the lower accuracy rates generally improved with a larger sample size.

Fifth, test length appeared to have little effect on the performance of both procedures in locating DIF-free items. Tables 4.9 to 4.12 list the rates of accuracy and corresponding chance levels in locating up to four DIF items for both procedures in the 50-item test. Basically, all of the conclusions made based on the 20-item test can be observed in the 50-item test. These conclusions include that the mean-DIF procedure and rank-based procedure worked well in selecting DIF-free items; both of item purification procedures perform better with large sample size; DIF direction affects the efficiency of both procedures in locating DIF-free items and the accuracy rates dropped with the increase of the number of DIF-free items identified.

Table 4.9 Rates of Accuracy and Chance Levels in Locating One DIF-Free Item with Rank-Based Strategy and Mean-DIF Procedure in a 50-item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	80	Rank	100	100
			Mean-DIF	100	100
	40%	60	Rank	88	96
			Mean-DIF	97	100
Dominant	20%	80	Rank	100	100
			Mean-DIF	100	100
	40%	60	Rank	99	100
			Mean-DIF	100	100
Balanced	20%	80	Rank	100	100
			Mean-DIF	100	100
	40%	60	Rank	100	100
			Mean-DIF	100	100

Table 4.10 Rates of Accuracy and Chance Levels in Locating Two DIF-Free Items with Rank-Based Strategy and Mean-DIF Procedure in a 50-item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	64	Rank	100	100
			Mean-DIF	100	100
	40%	36	Rank	98	100
			Mean-DIF	74	93
Dominant	20%	64	Rank	87	100
			Mean-DIF	100	100
	40%	36	Rank	100	100
			Mean-DIF	96	100
Balanced	20%	64	Rank	99	100
			Mean-DIF	100	100
	40%	36	Rank	98	100
			Mean-DIF	99	100

Table 4.11 Rates of Accuracy and Chance Levels in Locating Three DIF-Free Items with Rank-Based Strategy and Mean-DIF Procedure in a 50-item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	50	Rank	100	100
			Mean-DIF	99	100
	40%	20	Rank	62	80
			Mean-DIF	85	100
Dominant	20%	50	Rank	96	100
			Mean-DIF	98	100
	40%	20	Rank	98	100
			Mean-DIF	99	100
Balanced	20%	50	Rank	100	100
			Mean-DIF	100	100
	40%	20	Rank	99	100
			Mean-DIF	100	100

Table 4.12 Rates of Accuracy and Chance Levels in Locating Four DIF-Free Items with Rank-Based Strategy and Mean-DIF Procedure in a 50-item Test (Multiplied by 100)

DIF Direction	% of DIF Items	Chance Level	Purification Method	Accuracy Rate	
				500R/500F	1,000R/1,000F
NO DIF		100		100	100
One-sided	20%	40	Rank	100	100
			Mean-DIF	99	100
	40%	12	Rank	52	76
			Mean-DIF	83	98
Dominant	20%	40	Rank	100	100
			Mean-DIF	100	100
	40%	12	Rank	96	100
			Mean-DIF	99	100
Balanced	20%	40	Rank	100	100
			Mean-DIF	100	100
	40%	12	Rank	98	100
			Mean-DIF	100	100

Conclusively, a larger sample size and lower DIF contamination produced a higher rate of accuracy. Moreover, it appeared that it was much easier to locate a single DIF-free item than multiple DIF-free items. Every additional item seemed to represent an opportunity for incorrect identification. Considering the first four findings, it was not a surprise to find that the lowest rate of identification happened when the sample size was smaller (i.e., 500F/500R), the percentage of DIF items was larger (i.e., 40%), the DIF contamination was larger (i.e., one-sided DIF), and more DIF-free items were to be located (i.e., four). Overall, the rank-based strategy and the mean-DIF procedure both seemed promising for locating a set of up to four DIF-free items.

CHAPTER 5

REAL DATA ANALYSIS

In addition to the simulated data, the forward procedure, GLMM iterative purification, rank-based strategy, and the mean-DIF procedure were applied to real data. In this chapter, the data information, DIF detection results, and conclusions are provided.

Data

The data were obtained from the Florida Comprehensive Achievement Test® (FCAT), which is administered to Florida public school students in grades 3 through 11. In Florida, FCAT assessments are designed to measure student achievement of the Sunshine State Standards in reading, mathematics, writing, and science (FDOE, 1996). In this study, data were sampled from the 2009 Grade 3 FCAT Reading administration. The Grade 3 FCAT Reading assessment contains 45 operational multiple-choice (MC) items. All items are scored dichotomously (correct or incorrect). The items are designed to measure four reporting categories (subscales): Words, Main Idea, Comparisons, and Reference. Once the parameters are estimated, the process of equating is used to determine students' scores on the FCAT score scale, which ranges from 100 to 500.

All of the FCAT test items receive intensive review by expert panels to detect any possible gender or ethnicity bias (FDOE, May 2001). In this real data analysis, DIF analysis was conducted to compare Standard Curriculum (SC) students and Learning Disabled (LD) students. Students in the SC group were treated as the reference group, while students in the LG group were the focal group. In Florida, the LD students obtain a regular diploma by learning the Sunshine State Standards and taking the FCAT. In addition, the LD students may have testing accommodations in an effort to ensure equal access. It is important to make sure that the FCAT tests are equally difficult for all students so that fair conclusions can be made and, therefore, treatment can be made.

In 2009, a total of 206,920 students took the Grade 3 FCAT Reading test. Among these, 205,274 examinees received their FCAT performance reports. Among these 205,274 examinees, 178,427 examinees were students in the standard curriculum program and 12,672 examinees were students with specific learning disabilities. In this study, 1,000 SC group examinees and 1,000 LD group examinees were randomly

sampled for DIF analysis; therefore, a total of 2,000 grade 3 examinees were included in this DIF study. Table 5.1 displays the descriptive statistics for the sample.

Table 5.1 Summary Statistics of Scale Scores in the Sample Data

	n	Mean Scale Score	SD of Scale Score
LD Students	1,000	252.10	59.52
SC Students	1,000	320.90	55.72

In this real data analysis, the DIF detection analysis was conducted with each of four purification procedures using the GLMM DIF method. A rejection of the null hypothesis at the significance level of 0.05 indicates the presence of DIF for the studied item.

Results

Forward Purification

The real data were fitted using the GLMM DIF model with the forward purification procedure. In the first step of the forward purification procedure, the following 25 out of 45 items were detected as having significant DIF: 2, 5, 6, 8, 9, 11, 12, 15, 18, 19, 21, 27, 28, 29, 30, 31, 34, 37, 38, 40, 41, 42, 43, 44, and 45; therefore, these 25 items were included in one model for the final DIF assessment. Ultimately, 22 items were detected as having significant DIF. The results of the DIF analysis with the GLMM forward procedure are in Table 5.2.

In the GLMM DIF model, negative signs for the DIF estimates indicate that chances of answering correctly are lower for students in the LD group (focal group). According to the ETS DIF classification system (Dorans & Holland, 1993), the item is classified as a “B item” if the absolute value of DIF magnitude is larger than 0.43 but smaller than 0.6 logits. The item is classified as a “C item” when the absolute value of DIF is larger or equal to 0.6 logits. In addition to using the effect size, ETS criteria also evaluate hypothesis test results. For a B item, it has to be significantly different from 0 in the logit scale. For C item, the DIF index has to be significantly different from 0.43 on the logit scale. If the DIF size is smaller than 0.43 on the logits, DIF is negligible. When considering the values of the DIF indices, the following 11 items were between 0.40 and

0.80 in absolute magnitude: 6, 11, 15, 21, 27, 28, 34, 38, 43, 40, and 44. Items with positive values (items 15, 28, and 34) indicate DIF with lower performance by students in the SC group, while items with negative DIF (items 6, 11, 21, 27, 38, 40, 43, and 44) indicate DIF disadvantaging students in the LD group. The mean ability difference between the LD group and the SC group was 1.22 in the theta scale.

GLMM DIF Iterative Purification

The FCAT Reading data also were fitted into GLMM with iterative purification. At this point, 25 out of 45 items were detected as having significant DIF at the initial DIF screening; therefore, theoe items were removed from the anchor set. The second iteration of purification starts with the new anchor set containing items 1, 3, 4, 7, 10, 13, 14, 16, 17, 20, 22, 23, 24, 25, 26, 32, 33, 35, 36, and 39. The GLMM procedure was used to reanalyze all 45 items in the new anchor set. All DIF items were tested using the new anchor set. For non-DIF items, the anchor set excluded the studied item and the DIF items identified at the previous step.

Since no additional items were identified as having significant DIF, the purification process was stopped. The results after purification are included in Table 5.2. The conclusion is that the iterative purification produced exactly the same DIF assessment results as the forward procedure.

Mean DIF Procedure

The DIF-free-then-DIF strategy was then applied in the GLMM analysis with the mean-DIF procedure. In the first step of the mean-DIF procedure, each item in turn is considered a pure anchor, while all other items are investigated for DIF. The items with smallest mean absolute value of DIF indices are selected as DIF-free items for subsequent DIF assessment. In this study, items 23, 19, 1, and 6 had the smallest absolute value of DIF (0, 0.02, 0.04, and 0.04, respectively). Thus, item 23 was located as a pure anchor item for subsequent DIF estimation. Table 5.2 provides the DIF assessment results using item 23 as the anchor item.

Using the mean-DIF procedure, 21 of 45 items were detected as showing significant DIF. Considering the ETS DIF classification criterion, 10 items had slight-to-moderate DIF effects, while 5 items had moderate-to-large DIF effects. Items 2, 12, 15, 19, 28, 29, and 45 favored the LD. Items 6, 11, 21, 38, 40, and 43 favored the students in the SC

group. The difference in mean ability between the SC group and the LD group was 1.36. Compared with the previous two procedures, the mean-DIF procedure identified one less significant DIF item; and the DIF items disadvantaging the SC group have increased DIF estimates while the DIF items favoring the SC group have lower DIF estimates.

Rank-based Strategy

The rank-based strategy is also an application of the DIF-free-then-DIF strategy. Two steps were involved in this procedure. In step 1, each item was investigated separately using the all-other anchor method. Once all items were tested, a set of items more likely to display DIF were identified. In the rank-based strategy, the items that had the smallest magnitudes of Wald test statistics were the best DIF-free candidates. In step 2, these identified DIF-free items acted as the anchor set, and the other items were simultaneously assessed for DIF in one model. In this study, the optimal number of anchors was not clear for the GLMM DIF method; therefore, only one DIF-free item was located for subsequent DIF assessment using the rank-based strategy.

After running step 1 in the FCAT data, item 1 was identified as the DIF-free item with the smallest Wald test statistics; therefore, items 2 through 44 were tested for DIF simultaneously with item 1 serving as the anchor item. The DIF assessment results for the GLMM with the rank-based strategy are presented in Table 5.2. Among 45 tested items, the following 19 items were detected as having significant DIF: 2, 5, 6, 9, 11, 12, 15, 19, 21, 27, 28, 29, 34, 38, 40, 41, 43, 44, and 45.

Adopting the ETS DIF classification as in the forward procedure, four items were "B items," which are the items with slight-to-moderate magnitude of statistically significant DIF, while seven items were "C items" with moderate-to-large DIF effects. Items 6, 11, 21, 27, 38, 43, and 44 disadvantaged students in the LD group, and items 15 and 28 disadvantaged students in the SC group. The average ability difference between the SC group and the LD group was 1.245. Compared with the forward and GLMM iterative procedures, the rank-based strategy identified 3 fewer items with significant DIF. In regards to the DIF estimates, the rank-based strategy produced larger DIF value for the items that disadvantage the SC group, but smaller DIF value for the items that favor SD students. The rank-based strategy and mean-DIF procedures identified different pure anchors. Using the different pure anchors, the two procedures identified DIF items

in a similar manner. But the magnitudes of DIF for the positive DIF that was estimated in the rank-based strategy were smaller than those estimated in the mean-DIF procedure. Conversely, the magnitudes of DIF for the negative DIF estimated with the rank-based strategy were larger than those estimated using the mean-DIF procedure.

Table 5.2 DIF Assessment in GLMM with Four Different Purification Procedures

Item	DIF Estimates		
	Forward and Iterative Purification	Rank-Based Strategy	Mean-DIF Procedure
1	0	0	0.11
2	0.29**	0.31*	0.43**
3	0	-0.03	0.08
4	0	0.07	0.19
5	-0.33**	-0.31*	-0.19
6	-0.81**	-0.79**	-0.68**
7	0	-0.1	0.01
8	0.18	0.21	0.32*
9	-0.40**	-0.38*	-0.26
10	0	-0.13	-0.02
11	-0.56**	-0.54**	-0.43*
12	0.33**	0.35*	0.46**
13	0	-0.14	-0.02
14	0	-0.07	0.04
15	0.51**	0.53**	0.65**
16	0	0.19	0.30*
17	0	0.11	0.22
18	0.19	0.21	0.32*
19	0.39**	0.41**	0.52**
20	0	0.13	0.24
21	-0.90**	-0.88**	-0.77**
22	0	0.08	0.19
23	0	-0.11	0
24	0	0.06	0.17
25	0	0.07	0.18
26	0	-0.06	0.06
27	-0.49**	-0.47**	-0.35*
28	0.59**	0.62**	0.73**
29	0.38**	0.40**	0.51**
30	-0.28*	-0.26	-0.15

Table 5.2 (continued)

Item	DIF Estimates		
	Forward and Iterative Purification	Rank-Based Strategy	Mean-DIF Procedure
31	0.23*	0.25	0.37*
32	0	0.18	0.29
33	0	0.08	0.19
34	0.64**	0.66**	0.77**
35	0	-0.16	-0.05
36	0	0.14	0.25
37	0.19	0.21	0.33*
38	-0.60**	-0.58**	-0.46**
39	0	0.16	0.27
40	-0.56**	-0.54**	-0.43*
41	-0.40**	-0.37*	-0.26
42	-0.24*	-0.22	-0.11
43	-0.59**	-0.57**	-0.46**
44	-0.67**	-0.64**	-0.53**
45	0.34**	0.36*	0.48**
Mean Ability Difference	1.22	1.25	1.36

* $p < .05$; ** $p < .01$

Summary

DIF analysis was conducted on the 2009 Grade 3 FCAT Reading assessment using GLMM with four different purification procedures. The studied groups were the Standard Curricula (SC) group as the reference group and the Learning Disabled (LD) group as the focal group. Since no additional DIF items were identified in the iterative purification procedure after the initial DIF detections, the iterative purification procedure and the forward procedure produced exactly the same DIF assessment results as reported in Table 5.2. Three conclusions may be drawn from these results.

First, when four purification methods were applied to the FCAT Reading data set, the DIF identifications were quite consistent across the four purification procedures. One exception was that four additional DIF items were identified by the mean DIF

procedure. The forward procedure/iterative purification and rank-based strategy identified eight items that had slight-to-moderate DIF and three items with moderate-to-large DIF. Among these DIF items, eight DIF items showed bias against the students in the LD group, while three items showed bias against the students in the SC group.

Second, the mean-DIF procedure flagged more items as DIF than the rank-based strategy when only one anchor item was used. A total of 15 items were detected as having slight-to-large DIF by the mean-DIF procedure. In addition, the positive DIF values were larger and the negative DIF was smaller than those estimated using the rank-based strategy and forward procedure.

Third, when applying the DIF-free-then-DIF strategy in DIF assessment, multiple purification procedures are recommended to test the soundness of DIF-free item identification. In this study, item 1 seemed to be a good pure anchor candidate. More DIF-free items could have been identified and included in the anchor set to yield higher power of DIF detection. However, as discussed in the simulation study, with the increase of the anchor item identified in the item purification process, the accuracy rate of the anchor identification dropped. Besides, the appropriate number of anchor item is unknown when apply the item purification in GLMM model. Therefore, only one DIF-free item was identified and used as anchor for DIF-item identification in this study.

CHAPTER 6

CONCLUSIONS

The present study examined the process known as item purification, a process used to identify a set of DIF-free items from the instrument under investigation to be used as anchor items in DIF detection. The use of item purification can improve the detection error in DIF analysis; thus, creating fairer tests to all examinees because it eliminates, or at worst reduces, the prevalence of DIF items in the anchor set. The evaluation of the item purification methods is therefore required. By using a GLMM approach, this study compared the effectiveness of four approaches to item purification for the purpose of creating fairer tests.

The four item purification procedures examined in this study are: the forward procedure, GLMM iterative purification procedure, mean-DIF procedure, and rank-based strategy. To evaluate the performance of these four purification procedures, a simulation study was conducted followed by a real data analysis. The simulation results indicated that these four purification procedures work well in certain conditions.

Four independent variables, sample size, DIF direction, percentage of DIF items and test length, were included in the simulation design to investigate the impact of the four factors to the performance of the four item purification procedures. Specifically, the type I error rates and power rates generated by the forward procedure and GLMM iterative procedure were evaluated and compared. It appeared that both methods yielded a type I error around its expected value of .05 under the no-DIF conditions (with one exception in iterative procedure with 40% DIF item and 500F/500R). The confidence interval of all type I error rates was [.028, .072]. Type I error rates systematically increased as the sample size increased in most conditions. I concluded that the sample size affected the performance of the two procedures.

In addition, type I error rates, either acceptable or not, for the iterative procedure were smaller than those for the forward procedure across all one-sided conditions and dominant DIF with 40% DIF item conditions. I concluded that the iterative procedure was superior to the forward procedure when there was a large amount of the DIF contamination in the test. However, the forward procedure was more efficient than

iterative procedure when the DIF effect was cancelled out between groups or the DIF contamination was little. The advantage of purifying the anchor set was most apparent where the level of DIF contamination was large. However, if 40% or more items were one-sided DIF items, iterative purification yielded an inflated type I error rate and deflated power.

The power rates for both of the forward and iterative procedures were acceptable, which were greater than .80 (except the power in forward procedure with 40% one-sided DIF condition is .705). Sample size was an important factor that positively affected the power rates for both of the procedures. The power rates for forward procedure systematically increased as the sample size increased across all of the conditions. Similar results were found for iterative procedure with one exception. Moreover, the power rates for the iterative procedure were greater than those for the forward procedure when there was a large amount of DIF contamination in the test. In contrast, the power rates for the iterative procedure were smaller than those for the forward procedure when the overall DIF contamination was zero or very little. This conclusion was consistent with the conclusion made for the type I error rates when comparing the forward and iterative procedures.

To evaluate the performance of mean-DIF procedure and rank-based procedure, the accuracy rates of the DIF-free items identification were reported and compared. Both of procedures yielded perfect or close-to-perfect rates of accuracy in most conditions. The accuracy rates systematically increased as the sample size increased in all conditions for both procedures indicating that sample size is an important factor affecting the accuracy rates of the DIF-free item identification. In addition, the lowest rates of the accuracy were found under the condition of the 40% one-sided DIF items with up to 4 DIF-free items identified. This was true for both of the procedures. I concluded that the DIF contamination affected the locating of the DIF-free items. And, with additional DIF-free items identified, the accuracy rates dropped. In contrast, no systematical variation based on different test length could be found across the two procedures. I concluded that the test length did not affect the performance of two procedures.

This study also compared the accuracy rates of both procedures to their corresponding chance level. All of the accuracy rates were higher than their corresponding chance levels.

In addition, the accuracy rates for the mean DIF procedure were greater than or equal to the rank based strategy across all conditions. However, it was difficult for the mean-DIF procedure to select pure anchor sets when there were more DIF items than DIF-free items in the test.

These four item purification procedures also applied to a real data set analysis. In the real data analysis, DIF detection with item purification was conducted between Standard Curriculum (SC) students and Learning Disabled (LD) students using 2009 Grade 3 FCAT reading data. The students in SC groups were treated as the reference group while the students with disabilities were selected as the focal group in DIF analysis. After the DIF assessment with purification, the potential DIF items were identified for further review. The DIF identifications were quite consistent across the four item purification procedures with slightly differences.

Practical Implication for Education

As stated in the first chapter, DIF analysis is conducted with the need to fairly and equitably assess examines without bias (Standard for Educational and Psychological Testing, 1999). For the DIF model assuming pure anchor, it is essential to have pure anchor available for the accurate identification of the DIF items. If the DIF status of all items is initially unknown, using item purification in DIF detection is perhaps the only reasonable way to reduce the type I error of the DIF item detection. In addition, for the DIF detection model assuming zero-sum DIF, if the DIF is pervasive and when every item is bias in favor or disadvantage the same group, the DIF detection technique assuming zero-sum DIF may fail to detect any substantial DIF in any of the DIF items. Therefore, it is very important to have pure anchor.

In this study, I proposed an item purification procedure called forward procedure and investigated its performance with other three item purification procedures. The comparisons provided practical knowledge that will benefit measurement professionals and enhance the literature. By comparing these four item purifications procedures, I

illustrated how to use these procedures step by step and generalized a series of rule to use these purification methods. As user can easily identify the possible DIF-free item(s) using the generalized linear mixed model framework, one could first start with the item purification, pick DIF-free items(s) suggested by the analysis, submit it to a model assuming the pure anchor and check for the DIF items.

The findings also have implications for practitioners, especially for those in the area of test construction and validation. Identification of pure anchor items is not an easy task, especially when the instrument is intended for special populations or when the instrument was originally constructed in one language and later translated to other languages. In such cases, practitioners usually need some guidance as to what modeling techniques and DIF procedures to use and which would better detect errors in DIF analysis. This study, along with other studies in DIF analysis, would provide some general guidelines for modeling techniques and procedures. The study results suggest that the proposed procedure, forward procedure generated results that are comparable to those from other procedures when there is a small amount of DIF items in the test. When a large amount of DIF items are suspected in the test, the iterative procedure is recommended. In addition, the sample size is very important to the efficiency of the item purification. To have accurate results, large sample size needs to be provided. Regarding the DIF-free-then-DIF strategy, the mean-DIF strategy is recommended over the rank-based strategy. However, there should be more DIF-free items in the test than the DIF items.

This study compared four item purification procedures implemented in the GLMM. Generalized linear mixed models provide a flexible way to model the situation in which the assumption of the linear mixed model is not valid. This flexibility allows researcher to build model in a broader statistical frameworks as opposed to measurement models. Depending upon research questions and the information collected, other types of analysis could be specified using the GLMM model. For example, in this study, the person-by-item covariates were included into the GLMM to determine whether DIF exists for a specific item. In addition, person or item covariates could be added in a subsequent attempt to find the source of DIF.

Limitations and Future Studies

The results from this study provide researchers and practitioners with some insights into the implementations of the item purification into the DIF detection analysis. However, the majority of the conclusions were made based on the simulated data. The item parameters used for data generation were systematically manipulated in the simulation study. In real testing situations, the DIF conditions may be more complex. Therefore, the results obtained from this simulation study should be generalized to realistic testing situations with caution. Specifically, the limitations of this study include:

First, in this study, an equal number of examinees in both of reference group and focal group were simulated to examine the effect of group size on item purification (e.g., $N_{focal}=N_{refer}=500$ and $N_{focal}=N_{refer}=1{,}000$); however, it is common to find fewer examinees in the focal group than in the reference group because DIF is commonly present for minority populations, such as special populations or minority groups. Therefore, it is desirable to examine the effects of group size under the condition in which one group (e.g., minorities) was substantially smaller (i.e., $N_{focal}=300$ and $N_{refer=}1,\ 000$) than another group.

Second, since the intention of this study was to first explore the purification procedures with a rather simple and robust model, the Rasch model was selected for the data generation and analysis. Therefore, the present study concentrated on the purification procedure on uniform DIF. Non-uniform DIF and uniform DIF in a more complex model are other important topics that were not investigated in this study. The current study would serve as an initial step to further explore non-uniform DIF in more complex models.

For the rank-based strategy and mean-DIF procedure, the main purpose is to investigate their performance in the locating of DIF-free items as anchor items; however, the optimal number of anchor items in the GLMM DIF model was not investigated. In future studies, it would be helpful to examine how the number of anchors affects the parameter estimations. Previous studies revealed that a four-item anchor set would be enough to yield a high power of DIF detection in IRT-based DIF methods with pure

anchors (Thissen et al., 1988; Wang, 2004; Wang & Yeh, 2003; Wang, 2009). Further studies can be done to investigate these issues.

Therefore, more research is needed to evaluate the forward procedure, GLMM iterative procedure, mean-DIF methods and rank-based procedure using item parameters derived from real tests. This expansion will, in turn, make the study more generalizable to real testing situations.

APPENDIX A

COMPUTER SYNTAX

Syntax for the Data Generation of 20-items and 1000 Examinees

```
%macro datagen_20 (sim=&sim);

Options nodate nonumber pagesize=30000;
Proc printto print='F:\Para_20\sim_31.txt';
PROC IMPORT OUT= WORK.&sim
            DATAFILE= "H:\para_20\param_20"
            DBMS=EXCEL REPLACE;
     SHEET="'&sim$'";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

proc iml;


start irtscore (nitms, simulees, gender, idn2, idn3, rrv, popa, popb,
popd, score);
eta=simulees*popa-popb-gender*popd;
expeta=exp(eta);
pi=(expeta/(1+expeta))`;

score = pi>rrv;
finish;

Use &sim;
Read all var {a} into popa;
Read all var {b} into popb;
Read all var {d} into popd;

nitms=nrow(popb);

DO REP=1 to 100;    /*Replicate 100 times*/
 Do I=1 to 1000; /*1000 examinees*/
seed1=round(100000000*ranuni(0));
idn2=i;
idn3=rep;

gender = RAND('BERNOULLI',0.5); sim1=gender*(-0.2)+(1)*rannor(seed1);
simulees =sim1;

rrv= J(1,nitms,0);
do k=1 to nitms;
rrv[1,k] =RANUNI(seed1);
end;
```

```
run irtscore (nitms, simulees, gender, idn2, idn3, rrv, popa,
popb,popd, score);
replica =idn3[1,1];
idnum =idn2[1,1];
thet=simulees[1,1];
male=gender[1,1];
itm1=score[1,1];
itm2=score[1,2];
itm3=score[1,3];
itm4=score[1,4];
itm5=score[1,5];
itm6=score[1,6];
itm7=score[1,7];
itm8=score[1,8];
itm9=score[1,9];
itm10=score[1,10];
itm11=score[1,11];
itm12=score[1,12];
itm13=score[1,13];
itm14=score[1,14];
itm15=score[1,15];
itm16=score[1,16];
itm17=score[1,17];
itm18=score[1,18];
itm19=score[1,19];
itm20=score[1,20];

file print;

put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm1 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm2 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm3 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm4 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm5 1.;
put
   @1 replica 4.
   @8 idnum 4.
```

```
   @16 male 1.
   @20 itm6 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm7 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm8 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm9 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm10 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm11 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm12 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm13 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm14 1.;


put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm15 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm16 1.;
put
   @1 replica 4.
   @8 idnum 4.
```

```
   @16 male 1.
   @20 itm17 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm18 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm19 1.;
put
   @1 replica 4.
   @8 idnum 4.
   @16 male 1.
   @20 itm20 1.;
end;
end;
quit;

%mend datagen_20;
%datagen_20(sim=sim_29);
%datagen_20(sim=sim_30);
```

APPENDIX B

DATA REQUEST LETTER

10/20/2010
Sharon Koon, Assistant Deputy Commissioner
Florida Department of Education
Turlington Building, Suite 414
325 West Gaines Street
Tallahassee, Florida 32399

Dear Dr. Koon:

I am a doctoral student in the program of Measurement and Statistics at Florida State University. I am writing this letter to request FCAT data for use in my academic research. I am writing my dissertation titled „Item Purification in Differential Item Functioning Using Generalized Linear Mixed Models', under the supervision of Dr. Akihito Kamata. In my dissertation, I have proposed an item purification procedure and evaluated its performance in DIF item identification with other three purification procedures.

I have compared the performances of four purification procedures. It is expected that the four purification procedures (forward procedure, GLMM iterative purification, rank-based strategy and mean-DIF procedure) will yield more accurate DIF item detection than the DIF analysis without purification. In other words, the purification procedures studied are efficient in producing reliable and valid test to all test takers.

I have investigated how these four purification procedure work using a series of simulated data sets and would like to extend my investigation to a real data set. Thus, I would like to request to have access to 2009 Grade 3 Reading data with the variable indicating whether student is standard curriculum student or learning disable or not. The student identification information will not appear in the data set. In addition, I pledge to provide results of my data analysis after I successfully defend my dissertation.

If you have any questions regarding my research, please contact me by phone at [redacted] or via email ([redacted]), or my dissertation supervisor Dr. Akihito Kamata by phone at ([redacted]) or via email ([redacted]).

Sincerely,

Qian Liu
[redacted]

APPENDIX C

RESPONSE TO DATA REQUEST

FLORIDA DEPARTMENT OF EDUCATION

Dr. Eric J. Smith
Commissioner of Education

Just Read, Florida!

November 03, 2010

Qian Liu
2534 Lagrange Trail
Tallahassee, Florida 32312

Dear Ms. Liu:

This letter is in response to your request for permission to use 2009 Grade 3 FCAT Reading item-level data for the full population, including student-level characteristics. As you have identified, the data requested will be used as a part of your dissertation in the study of item purification of differential item functioning analysis using generalized linear mixed model.

I have reviewed your request and authorize the release of the Grade 3 FCAT Reading item-level test data for 2009, without student identifications.

I look forward to receiving a copy of your research results. If you have further questions concerning this matter, please feel free to let me know.

Sincerely,

[redacted]

Sharon Koon

SHARON KOON, PH.D.
ASSISTANT DEPUTY COMMISSIONER
ACCOUNTABILITY, RESEARCH, AND MEASUREMENT
OFFICE OF ASSESSMENT

325 W. GAINES STREET • SUITE 414 • TALLAHASSEE, FL 32399-0400 • (850) 245-0513 • www.fldoe.org

APPENDIX D

HUMAN SUBJECTS APPROVAL MEMORANDUM

Office of the Vice President For Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 3/7/2011

To: Qian Liu

Address: 2534 [redacted]
Dept.: EDUCATIONAL PSYCHOLOGY AND LEARNING SYSTEMS

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research
Item Purification in Differential Item Functioning Using Generalized Linear Mixed Models

The application that you submitted to this office in regard to the use of human subjects in the research proposal referenced above has been reviewed by the Human Subjects Committee at its meeting on 03/02/2011. Your project was approved by the Committee.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 2/29/2012 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the Chair of your department and/or your major professor

is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

Cc: Besty Jane Becker, Advisor
HSC No. 2011.5775

REFERENCES

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

Angoff, W. H. & Ford, S. E. (1973). Item race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95-105.

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

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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*, 277-300.

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

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

Vaughn, Brandon K. (2006). A Hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian Multilevel Approach. Unpublished doctoral dissertation, Florida State University, College of Education.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist-Revised. *Psychological Assessment*, *16*, 155-158.

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

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P.W. Holland

H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253-260.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, *6*, 57-79.

Cheong, Y. F., & Kamata, A (2007). A comparison of DIF detection procedures using hierarchical generalized linear and nonlinear mixed models. Under review by *Journal of Applied Measurement*.

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

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

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

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

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement*: pp. 201-219, Phoenix, AZ: Oryx Press.

De Boeck, P., & Wilson, M. (Ed.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*, New York: Springer.

De Ayala, R. J. (Eds) (2008). Differential Item Functioning. *The theory and practice of item response theory.* pp324. New Your, NY: Guilford Press.

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

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*(3), 217-233.

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

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W.Holland & H. Wainer (Eds.), *Differential item functioning* (pp.171-196). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Ercikan, K. Gierl, M. J., McCreith, T., Puhan, G., & Koh, L. (2004). Comparability of bilingual versions of assessments: sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, *17(3),* 301-321.

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

Florida Department of Education (2010). FCAT 2010 reading, mathematics, and science test demographic reports. Tallahassee, FL.

Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items. In D. Robitaille & A. Beaton (Eds.), *Secondary analysis of the TIMSS results: A synthesis of current research*. Dordrecht, the Netherlands: Kluwer.

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

Gierl. M. J., Cheng, L., Rogers, W. T., Gotzmann, A., & Vanderberghe, C. (2000). *Translation differential item functioning on the Hong Kong Certificate of Education Examination in six content areas* (CRAME Research Rep. No. RR-00-01). Edmonton, Alberta, Canada: University of Alberta, Centre for Research in Applied Measurement and Evaluation.

Gierl. M. J., Gotzmann, A., & Boughton K. (2004). Performance of SIBTEST When the Percentage of DIF items is Large, *Applied Measurement in Education*, *17 (3),* 241-264.

Gierl. M. J., Rogers, W. T., & Klinger, D. (1999). Using statistical and substantive reviews to identify and interpret translation DIF. *Alberta Journal of Educational Research*, *45*, 353-376.

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

Goldstein, H. (2003). Multilevel statistical models (3rd.). London: Arnold.

Green, J. L., Camilli, G., Elmore, P. B. & Skukauskaite, A. (2006). Handbook of complementary methods in education research, American Educational Research Association, Elizabeth Grace.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 71-94). Vancouver BC: Educational Research Institute of British Columbia.

Hambleton, Ronald & Rodgers, Jane (1995). Item bias review. *Practical Assessment, Research & Evaluation*, *4(6).*

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *V.8*, 35 – 41.

Hedeker, D. (2005). Generalized linear mixed models. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Wiley, New York.

Hedeker, D. (2008). A mixed-effects multinomial logistic regression model, *Statistics in Medicine, 22*, 1433-1446.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the 27th Annual Conference of the Military Testing Association,* (Vol. 1, pp. 282-287). San Diego.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-45). Hillsdale, NJ: Lawrence Erlbaum.

Kachman, S. D. (2000). An Introduction to Generalized Linear Mixed Models. In: Proc. of a Symposium at the Organizational Meeting for a NCR Coordinating Committee on "Implementation Strategies for National Beef Cattle Evaluation," Oct 20-21, Athens, GA. 59-73.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

Kamata, A., & Cheong, Y. F. (2006). Multilevel Rasch models. In M. von Davier & C. H. Carstensen (Eds.) Multivariate and mixture distribution Rasch models: Extension and applications (pp.217-232) New York: Springer.

Kamata, A., & Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disability: A Contemporary Journal, 2*(2), 49-69.

Kelderman, H. (1989). Item bias detection using loglinear IRT. Psychometrika, *54*, 681-697.

Kim, W. (2003). *Development of a differential item functioning (DIF) procedure using the hierarchical generalized model: a comparison study with logistic regression procedure.* Unpublished doctoral dissertation. Pennsylvania State University.

Kim, S. –H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 551-566.

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measure, and the likelihood ratio test on detection of differential item functioning. *Applied Measuremnet in Educaiton, 8, 291-312.*

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

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

Kromrey, J. D., Parshall, C. G., Chason, W. M., & Yi, Q (1999). Generating item response based on multidimensional item response theory. *Proceeding of the 24th Annual SASs User's Group International Conference*, Poster. University of South Florida.

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

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

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

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement*, *27,* 372-379.

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

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139-160.

McCullagh, R., & Nelder, J. A. (1989). *Generalized linear models* (2nd 4.). London: Chap-man and Hall.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7(2),* 105-118.

Mellenbergh, G. J. (1994). Generalized Linear Item Response Theory. *Psychological Bulletin*, 1994, Vol. *115*, No. 2, 300-307.

Meulders, M. & Xie, Y. (2004). Person-by-item predictors. In: *Explanatory Item Response Model: A generalized linear and nonlinear approach*, P. De Boeck and M. Wilson, eds, New York, Springer: 214-40.

Muthen, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, *10*, 121-132.

Muthen, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Erlbaum.

Muthen, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

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

Moustaki, I., Knott, M. (2000). Generalized latent trait models. *Psychometrika, 65*, 391-441.

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

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

Navas-Ara, M. J.,& Gomez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment*, *18(1),* 9-15.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107-124.

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

Penfield, R. D. & Camilli, G. (2007). Differential Item Functioning and Item Bias. *Handbook of Statistics*, *Vol. 26*.

Perrone, M. (2006). Differential Item Functioning and Item Bias: Critical Consideration in Test Fairness. *TESOL & Applied Linguistics*, 2006, *Vol. 6*, No. 2.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research*, *71*, 53-104.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *54*, 495-502.

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

Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Model in Social and Behavioral Research: Applications and Data-Analysis Methods*, 2nd Edition, Sage Publications, Thousand Oaks.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structure equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Reider (Eds.), *Equivalence in measurement:Research in management* (pp. 25-50). Greenwich, CT: Information Age.

Rijmen, F., Tuerlinckx, F., Paul De Boeck, & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory, *Psychological Methods*, *8*,185-205.

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

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3–32.

SAS Institute Inc., SAS 9.1.3 Help and Documentation, Cary, NC: SAS Institute Inc., 2000-2004.

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. In Hancock, G.R., & Samuelsen, K. M. (Eds.), *Advances in latent variable mixture model* (pp. 177-199). Information Age Publishing, Inc.

Scheuneman, J. D. (1975). *A new method for assessing bias in test items.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 106-359).

Scheuneman, J. D. (1981a). A new look at bias in aptitude test. In P. Merrifield, (Ed.*), New Directions for Testing and Measurement: Measuring human abilities* (Vol. 12, pp. 3-33). San Francisco: Jossey-Bass.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing differential item functioning. In M. Wilson & G. Engelhard (Eds.), *Objective Measurement: Theory into practice* (Vol. 3, pp. 65-82). Norwood, NJ: Albex Publishing Corporation.

Schumacker, R. (2005). *Test bias and differential item functioning*. Retrieved November 18, 2006, from http://www.appliedmeasurementassociates.com/White%20Papers /TEST%20BIAS%20AND%20DIFFERENTIAL%20ITEM%20FUNCTIONING. pdf.

Shealy, R. & Stout, W. (1993a). An item response theory model for test bias and differential item functioning. In P. W. Holland & W. Howard (Eds), Differential Item Functioning (pp.197-239). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Shepard, L.A. (1982). Definitions of bias. In R. A. Berk (Ed.). *Handbook of methods for detecting test bias* (pp. 9-30). Baltimore: Johns Hopkins University press.

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

Shih, C.-L. & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*, 184-199.

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

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

Soper, J. C., & W. B. Walstad. (1987). Test of economic literacy: Examiner's manual. 2nd. New York: Joint Council on Economic Education (now the National Council on Economic Education.

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

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0) [Computer program]. Chicago: Scientific Software.

Thissen, D., Chen, W.-H., & Bock, R. D. (2002). MULTILOG [Computer program]. Lincolnwood, IL: Scientific Software International.

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

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Uiterwijk, H. & Vallen, T. (2003). Test bias and differential item functioning: A study of the suitability of the cito primary education final test for second generation

immigrant students in the Netherlands. *Studies in Educational Evaluation, 29* (2003) 129-143.

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

Walstad, W., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *Journal of Economic Education, 28*, 155 - 171.

Wang, W.-C., & Wilson, M. R. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*, 549-576.

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

Wang, W. –C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72 (3),* 221-261.

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

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

Wang, W.-C., Shih C.-L, & Yang, C. –C. (2009). The MIMIC method scale purification for detecting DIF. *Educational and Psychological Measurement Online First*.

Wang, W.-C, & Wilson, M. R. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*, 549-576.

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

Wainer, H. (1993). Measuring differential impact of items. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp.3-23). Hillsdale, NJ: Erlbaum.

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

Williams, V. S. L. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10,* 253-267.

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

Wolfinger, R. & O'Connell, M. (1993). Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation*, *4*, 233–243.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Generalized item response modeling software*. Melbourne, Australia: ACER.

Zenisky, A. L., Hambleton, R. K. & Robin, F. (2004). DIF detection and interpretation in large-scale science assessment: informing item writing practices. *Educational Assessment*, *9 (1 & 2),* 61-78.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), Differential item functioning (pp. 337-348). Hillsdale, NJ: Erlbaum.

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

.

BIOGRAPHICAL SKETCH

Qian Liu was born in September 1977 in China. She earned her Bachelor's degree in Early Childhood Education in 2001 at Shandong Normal University, China. In 2002, Qian came to United States of America to study in Educational Measurement and Statistics at Florida State University. While she pursued her doctoral degree in Measurement and Statistics at Florida State University, she worked in the Psychometrics and Research Services, Office of Assessment at the Florida Department of Education (FDOE) as an intern psychometrician. Currently, she is working in the Scoring and Reporting, Office of Assessment at the FDOE. Her research interests are differential item functioning, large scale assessment, scaling and equating.