Nonparametric Item Response Modeling for Identifying Differential Item Functioning in the Moderate-to-Small-Scale Testing Context

By

PETRONILLA MURLITA WITARSA

M. A. (Ed.) University of Victoria (1995)
Dokter (M. D.) University of Andalas (1982)

A Dissertation Submitted in Partial Fulfillment of
The Requirements for the Degree of

DOCTOR OF PHILOSOPHY

In the Department of Educational and Counselling
Psychology and Special Education
Program: Measurement, Evaluation, and Research Methodology

We Accept this Dissertation as
Conforming to the Required Standard

University of British Columbia
July 2003

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational and Counselling Psychology and Special Education

The University of British Columbia
Vancouver, Canada

Date July 25th 2003

ABSTRACT

Differential item functioning (DIF) can occur across age, gender, ethnic, and/or linguistic groups of examinee populations. Therefore, whenever there is more than one group of examinees involved in a test, a possibility of DIF exists. It is important to detect items with DIF with accurate and powerful statistical methods. While finding a proper DIF method is essential, until now most of the available methods have been dominated by applications to large scale testing contexts. Since the early 1990s, Ramsay has developed a nonparametric item response methodology and computer software, TestGraf (Ramsay, 2000). The nonparametric item response theory (IRT) method requires fewer examinees and items than other item response theory methods and was also designed to detect DIF. However, nonparametric IRT's Type I error rate for DIF detection had not been investigated.

The present study investigated the Type I error rate of the nonparametric IRT DIF detection method, when applied to moderate-to-small-scale testing context wherein there were 500 or fewer examinees in a group. In addition, the Mantel-Haenszel (MH) DIF detection method was included.

A three-parameter logistic item response model was used to generate data for the two population groups. Each population corresponded to a test of 40 items. Item statistics for the first 34 non-DIF items were randomly chosen from the mathematics test of the 1999 TIMSS (Third International Mathematics and Science Study) for grade eight, whereas item statistics for the last six studied items were adopted from the DIF items used in the study of Muñiz, Hambleton, and Xing (2001). These six items were the focus of this study.

The MH test maintained its Type I error rate at the nominal level. The investigation of the nonparametric IRT methodology resulted in: (a) inflated error rates for both a formal and informal test of DIF, and (b) a discovery of an error in the widely available nonparametric IRT software, TestGraf. As a result, new cut-off indices for the nonparametric IRT DIF test were determined for use in the moderate-to-small-scale testing context.

TABLE OF CONTENTS

Abstract ii

Table of contents iv

List of tables vii

List of figures x

Acknowledgments xi

CHAPTER I: INTRODUCTION 1

Setting the Stage for the Dissertation 1

Motivation for the Dissertation 2

Testing context 5

Systematic literature based survey 7

DIF detection methods 9

Problem Statement 12

CHAPTER II: LITERATURE REVIEW 14

Methods for DIF Detection for the Moderate-to-Small-Scale Testing Context 14

Mantel-Haenszel (MH) method 16

TestGraf DIF: Nonparametric regression to assess DIF 17

Findings from Simulation Studies 26

MH test 27

TestGraf 27

Research Questions 28

CHAPTER III: METHODOLOGY 31

Study Design 31

Distribution of the latent variable and sample sizes 32

Statistical characteristics of the studied test items 32

Computer simulation design and dependent variables 35

Procedure 35

Data Analysis of Simulation Results 36

Version of TestGraf Used in the Simulation 37

CHAPTER IV: RESULTS 38

Mantel-Haenszel 40

TestGraf 43

Beta of TestGraf 43

Type I error rate of TestGraf 50

Cut-Off Indices 52

CHAPTER V: DISCUSSION 62

Summary of the Findings 65

REFERENCES 67

APPENDIX A 83

APPENDIX B 87

B-1. Methodology 88

Description of DIF detection procedure 89

Variables in the study 90

Data generation 94

Procedure 96

Compute DIF statistics 97

Data analysis of simulation results 97

B-2. Results and Discussion 98

Summary 106

B-3. Conclusion 107

LIST OF TABLES

Table 1: Summary of the systematic based survey by category of research 9

Table 2: Item statistics for the 40 items 34

Table 3: Type I error rate of Mantel-Haenszel at nominal $\alpha = .05$ 42

Table 4: Type I error rate of Mantel-Haenszel at nominal $\alpha = .01$ 43

Table 5: Sampling distribution of beta of Item-35 44

Table 6: Sampling distribution of beta of Item-36 45

Table 7: Sampling distribution of beta of Item-37 46

Table 8: Sampling distribution of beta of Item-38 47

Table 9: Sampling distribution of beta of Item-39 48

Table 10: Sampling distribution of beta of Item-40 49

Table 11: Type I error rate of TestGraf at nominal $\alpha = .05$ 51

Table 12: Type I error rate of TestGraf at nominal $\alpha = .01$ 52

Table 13: Type I error of TestGraf DIF detection using the Roussos-Stout criterion across sample size combinations and item parameters 54

Table 14: Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels irrespective of the item characteristics..... 55

Table 15: Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a low discrimination level ($a = 0.50$) and low difficulty level ($b = -1.00$) 56

Table 16: Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a high discrimination level (a = 1.00) and low difficulty level (b = -1.00) 57

Table 17: Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a low discrimination level (a = 0.50) and medium difficulty level (b = 0.00) 58

Table 18: Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a high discrimination level (a = 1.00) and medium difficulty level (b = 0.00) 59

Table 19: Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a low discrimination level (a = 0.50) and high difficulty level (b = 1.00) 60

Table 20: Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a high discrimination level (a = 1.00) and high difficulty level (b = 1.00) 61

Table A1: Details of the systematic literature based survey by category of research 84

Table B1: Item statistics for the 40 items 93

Table B2: Item statistics for the six DIF items 94

Table B3: Type I error rate of TestGraf version pre-December 2002 at nominal α = .05 99

Table B4: Probability of rejecting the hypothesis for TestGraf version pre-December 2002 DIF detection in SMALL-DIF at α = .05 100

Table B5: Probability of rejecting the hypothesis for TestGraf version pre-December 2002 DIF detection in SMALL-DIF at $\alpha = .01$ 101

Table B6: Probability of rejecting the hypothesis for TestGraf version pre-December 2002 DIF detection in MEDIUM-DIF at $\alpha = .05$ 102

Table B7: Probability of rejecting the hypothesis for TestGraf version pre-December 2002 DIF detection in MEDIUM-DIF at $\alpha = .01$ 103

Table B8: Probability of rejecting the hypothesis for TestGraf version pre-December 2002 DIF detection in LARGE-DIF at $\alpha = .05$ 104

Table B9: Probability of rejecting the hypothesis for TestGraf version pre-December 2002 DIF detection in LARGE-DIF at $\alpha = .01$ 105

Table B10: Standard deviations and standard errors of TestGraf version pre-December 2002 DIF detection 106

LIST OF FIGURES

Figure 1: Item characteristic curve of the reference group with $N = 100$ 22

Figure 2: Item characteristic curve the focal group with $N = 50$ 22

Figure 3: Item characteristic curves for reference and focal groups with Ns = 100/50.

Curve 1 is for reference group, and curve 2 is for focal group 25

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation and thanks:

... to my supervisor, Dr. Bruno D. Zumbo, who has always been there whenever I needed his assistance, thoughts and suggestions.

... to my committee, Dr. Seong-Soo Lee and Dr. Anita Hubley; to my university examiners, Dr. Kimberly Schonert-Reichl and Dr. Lee Gunderson, to my external examiner, Dr. John O. Anderson, and to the Chair, Dr. Graham Johnson, of my final examination as well.

... to the former Department of Educational Psychology and Special Education Faculty and Staff; the Department of Educational and Counseling Psychology, and Special Education: the Head and Graduate Student Advisors, Faculty and Administrative Staff for the Research and Teaching Assistantship positions given to me, as well as for their kindness and assistance during my study.

... to the Universitas Terbuka in Jakarta, Indonesia, for allowing me to pursue and complete my study at the University of British Columbia.

... and to everyone who has been so kind and friendly to me during my stay in Vancouver, British Columbia, including some friends in Victoria.

CHAPTER I

INTRODUCTION

This chapter begins with the introduction of the dissertation problem and provides some general motivation and context for the dissertation. The second chapter deals with the literature review.

Setting the Stage for the Dissertation

Hattie, Jaeger, and Bond (1999, p.394) described educational test development and the educational test enterprise more generally as a cyclic process that involves the following tasks:

- Conceptual Models of Measurement: This involves alternative measurement models, classical test theory, item response theory, and cognitive processing based models.
- Test and Item Development: This involves selection of item formats, selection of scoring models, frameworks for test organization, test specifications, and test assembly.
- Test Administration: This involves classical group administration of tests, administering performance tests, accommodations in the standardized testing situation, and computer adaptive testing.
- Test Use: This includes using tests for school accountability purposes, ensuring the dependability of scoring, reporting test scores, setting standards for test performances, and linking scores from different tests.
- Test Evaluation: This includes estimation of the reliability of tests, generalizability of scores, reliability and performance assessment, estimation of

the standard error of measurement, estimating decision consistency, validity, dimensionality, and adverse impact and test bias.

- *The cycle continues at the top of this list.*

The above cycle is useful because it integrates the various activities and testing issues of measurement research and practice. It applies to all kinds of testing contexts from classroom testing to pilot studies and large-scale programs.

The Test Evaluation in the cycle of educational testing can be guided by the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999). In particular, Standard 7.3 stated:

> When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups (p. 81).

The Standards go on to state that differential item functioning (DIF) may happen when one or more groups of examinees with equal ability, on average, have different probabilities of getting a particular item correct. It further explained that the above circumstance could be considered as subject to DIF when the DIF results could be replicated, implying that the DIF results are found in another sample. The sample-to-sample variability is taken into account in the standard error of a statistic.

Motivation for the Dissertation

The motivation for this dissertation rests on the Test Evaluation part of Hattie et al.'s (1999) cycle, which is the task where item bias and validity are considered. Moreover, the precise focus of the present study was to investigate methods that statistically flag DIF items.

The statistical method involves testing a hypothesis of no-DIF; therefore, Type I error rate becomes very important because a statistical test must maintain its Type I error rate to be a valid test of its hypothesis. If the statistical null hypothesis of no-DIF were not rejected in a DIF analysis, a measurement analyst would conclude that the item contains no bias for that particular item. If it is rejected, it can be concluded that DIF exists and further evaluation will be needed to see if this DIF is attributable to item bias or to item impact reflecting relevant factors. What this means, then, is that a Type I error is quite important because a test user may be overlooking a potentially biasing item(s) and hence the test may be functioning differently and inappropriate decisions are made based on the test score. In the context of high-stakes testing this type of error may be of great concern because of the matter of test fairness.

DIF is a technical term to describe a situation where there is a difference in the item response performance between two or more groups of examinees, which possess equal underlying overall ability level required to respond correctly to the relevant item. It is worth noting that DIF only tells us that an item functions differently for a particular group or groups of overall equal ability level. Having found an item with DIF does not necessarily indicate bias, although the DIF term itself was derived from item bias research. Another source of DIF can be what is termed item impact. Therefore, DIF is a necessary but not a sufficient condition for item bias to exist.

Studies of the item bias were first introduced to educational measurement in the 1960s, when it was discovered that a significant discrepancy in test performance between Black and Hispanic students and White students existed on tests of cognitive ability (Angoff, 1982). The studies were designed to develop methods for exploring cultural differences. The more specific goal of those studies, however, was to identify any test items that were biased against minority

students and hence to remove the biased items from the tests. As many more studies report, it became clear that the use of the term "bias" referred to two different meanings, social and statistical (Angoff, 1993). This resulted in the expression 'differential item functioning' (DIF) for the statistical effect. DIF can be detected using statistical procedures or "item discrepancy methods" (Angoff, 1982), which will tell us whether or not a particular item functions differently in different groups that are equivalent on the latent variable being measured.

Therefore, a situation in which DIF occurs should be differentiated from the situations of either item impact or item bias. Item impact occurs when examinees of different groups possess true differences in the underlying overall ability being measured by the item, which results in differing probabilities of answering the relevant item correct. In this notion, the item should measure what it is purported to measure, but the differences in getting the item correct lie in the true differences of underlying ability needed to correctly answer the item. In other words, item impact should reflect the differences of the actual knowledge and skill of the construct being measured for all groups responding to the item.

In contrast, if the different probabilities of getting an item correct are due to some characteristics of the test item itself, or the testing situation, which is irrelevant to the purpose of the test, then there is item bias. Item bias can happen for any identified group of examinees however the most commonly investigated groups are formed based on gender, ethnic, culture, language, social, or/and economic differences. For item bias to happen, DIF is required but not sufficient. For example, put simply, item bias can occur, among several causes, when an item is worded in such a way that leads one or more groups having less probability of getting the item correct than the other group(s). As previously mentioned, the item bias exists only when all groups involved possess equal underlying overall ability which is required to respond correctly

to the relevant item, do not get the item correct due to some characteristics of the item itself, or the testing situation, which is irrelevant to the purpose of the test. It should be noted that only when a particular item persistently receives incorrect responses from one particular group over another, and the groups are equivalent in the measured latent variable, could bias be considered to have occurred.

As noted above, there is a link between item bias, item impact, and DIF. Such a connection is basically methodological in terms of a need for a statistical method for identifying item(s) with DIF. Once DIF is found, then further analysis or review will be necessary to determine whether DIF is attributable to item bias or is, in fact, attributable to item impact. A judgmental analysis or content review may be needed to determine the nature of the link.

There are two forms of DIF, uniform and non-uniform. Uniform DIF occurs when the difference in probability of success of getting a correct response between groups is consistent across all levels of ability. That is, there is no interaction effect between group membership and the ability level in getting an item correct. Non-uniform DIF can be found when there is an interaction effect between the ability of getting an item right and group membership. In non-uniform DIF, the difference in probability of success of responding correctly to the item between groups is not constant across ability levels. In the framework of item response modeling, uniform DIF occurs when the item characteristic curves (ICCs) of the groups studied are separated but parallel or not crossing, whereas non-uniform DIF implies that the ICCs are not parallel. ICC is a plot of probability of an item being answered correctly against ability.

Testing context.

At this point it is useful to distinguish between large-scale and smaller-scale testing because DIF has evolved out of large-scale testing's concern for litigation and capacity to

conduct formal analyses. The type of large-scale testing I am referring to is the type practiced by testing companies, such as ETS or ACT, who are concerned about ligitation from individuals who pay to take their tests. As stated by Hattie et al. (1999), the cycle of the tasks described earlier applies irrespective of the type of testing context – be it testing programs that are high volume wherein there is a testing session once or more times a year (e.g., provincial testing conducted by the British Columbia Ministry of Education, Test of English as a Foreign Language/TOEFL, or Graduate Record Examinations/GRE) or if it is smaller-scale testing contexts. The high volume testing context is often referred to as "large-scale" testing in contrast to what I will call "moderate-to-small-scale" testing. This moderate-to-small-scale testing involves those situations wherein one has a test developed for a specific testing or research purpose taken by 500 or fewer examinees on a single occasion. The tests used in moderate-to-small-scale testing contexts are often, but not always, much shorter than the ones produced in large-scale testing situations. Interestingly, as Hattie et al. point out, the educational measurement journals, such as the *Journal of Educational Measurement* and conference presentations and workshops (such as NCME), reflect a greater focus on large-scale testing than moderate-to-small-scale testing.

Two points are noteworthy at this point. First, the fact that there are many items and many examinees is somewhat intertwined. In the move toward using item response theory (IRT) in educational measurement, there is a need for many items so that one can accurately estimate the latent score (theta) or ability level for examinees. Likewise, an increase in the number of items, and IRT-based analysis, has led to the need for more examinees. In the end, large-scale testing emerged. Second, large-scale testing has evolved out of educational policy and accountability. That is, there is a keen interest on the part of policy makers and administration for

large-scale testing of students at various grades for the purposes of student accountability and educational program evaluation.

Systematic literature based survey.

As prominent as IRT has become, as Hattie et al. has observed, it is expected that not all measurement activity would involve IRT nor would it involve large number of items and examinees. What appears largely undocumented in the measurement literature are 'statistics' about the typical numbers of examinees and numbers of items used in a variety of educational testing contexts. To fill this void and help set the context for this dissertation, a systematic literature based survey was conducted.

For this survey, four widely read educational research journals were selected. Two journals were devoted to educational measurement and two others to educational psychology. They include *Journal of Educational Measurement*, *Applied Measurement in Education*, *Journal of Educational Psychology*, and *British Journal of Educational Psychology*. To get a sense of the current state, the search was limited to issues from the years 1998 to 2002. Each research paper that reports the examinee/simulee/subject sample sizes, with or without reporting the number of items used was recorded. Articles were divided into four groups, achievement testing, simulation testing, psychological testing, and survey research. For the purposes of recording, some criteria were first established to achieve a consistency across the articles. The criteria and detailed data are provided in Appendix A and the summary of the information recorded in Table 1.

In summary, Table 1 shows that the 38 studies of achievement testing involved sample sizes ranging from 50 to nearly 88,000 examinees with numbers of items from 20 to 150 per test. There were 27 simulation studies reported to use sample sizes ranging from 25 simulees to

20,000 with numbers of items of 20 or more. Studies in the educational psychology used as small as 18 subjects up to 1,070 and from 10 to 56 items per questionnaire. Sample sizes in the survey category were from 3 to 2,000 respondents with numbers of items from 9 to 80 per survey.

In conclusion, the survey suggests that:

1. Achievement testing research uses larger sample sizes and larger number of items than psychological and survey research.
2. A majority of simulation research is similar both in terms of sample size and number of items to achievement testing (and not typical psychological research).
3. The median of the psychological and survey research reflect the use of sample sizes of moderate-to-small-scale context as defined by the present study.

Likewise, we will later see that studies investigating the statistical properties of DIF detection methods have focussed on large-scale testing. This focus on large-scale testing indicates that there is no widely studied DIF detection method for moderate-to-small-scale testing contexts, where there are fewer examinees and/or items than in the large-scale context.

Table 1.

Summary of the systematic literature based survey by category of research

	Achievement Tests		Simulation Studies		Psychological Test		Survey Studies	
	Examinees	Items	Simulees	Items	Subjects	Items	Subjects	Items
M	4,657.0	54.1	2,558.9	51	321.5	29.7	310.6	43.2
M	1,808.2*		1,962.6**					
Mdn	1,284	40	1,000	37.5	214.5	27	193.5	40
Min	50	20	25	20	18	10	3	9
Max	87,785	150	20,000	153	1,070	56	2,000	80
25^{th}%	371.5	31	250	30	148	19.5	33.5	26.3
75^{th}%	2,106.3	62	1,600	55	444.8	37.5	361.5	58.3
N	38	17	27	21	16	6	36	10

Note : * denotes mean without three studies that involved *N*s of 50, and over 25,000.
** denotes mean without two studies that involved *N*s of 25 and 20,000.

DIF detection methods.

To recapitulate what we have discussed so far, the presence of DIF may confound interpretations that one makes from test scores. However, the development of DIF methodology has been associated with three phenomena: (a) simulation studies involving thousands of examinees, (b) tests with many items, and (c) the widespread use of item response theory (IRT). These three phenomena are interconnected through the use of many items, many examinees, and use of IRT. Motivated by the fact that some testing contexts still exist wherein one has fewer numbers of items and examinees than found in the contexts where IRT is used, the purpose of this dissertation was to explore the operating characteristics of a psychometric DIF detection method useful in moderate-to-small-scale context.

Therefore, this section will briefly explore some possible pitfalls associated with DIF and statistical methods to detect DIF, as well as the uses of IRT. Then, the section will describe the existence of persistent issues relating to using moderate-to-small number of examinees for the same purposes of assessment and testing, and conclude by identifying the research problem that this dissertation is to address.

With the attention focused on the role of DIF in test development and adaptation, comes the concomitant need for statistical methods, to detect (i.e., flag) DIF items, which are accurate and useful in a variety of testing contexts. Methods for detecting items with DIF are normally used in the process of developing new psychological and/or educational measures, adapting existing measures, and/or validating test score inferences. DIF detection methods will only permit us to evaluate whether an item functions similarly for various examinee groups, or favours one or more groups over the other after conditioning on their underlying ability level, which is required to get the item correct (Zumbo & Hubley, 2003).

The majority of DIF methods have been developed with large sample size and large number of items in mind. This context of large testing organizations is mentioned because this large scale testing has been the place where a great deal of the contemporary measurement theory has evolved. This is particularly true in the almost meteoric rise of item response theory in educational measurement. One can easily see the prominence of IRT by scanning the National Council on Measurement in Education (NCME) conference program over the last 15 years wherein one sees many workshops, symposia, and conference papers on the topic of IRT. For a list of reasons for this increased visibility of IRT one can see Hambleton, Swaminathan, and Rogers's (1991) as well as Crocker and Algina's (1986) discussions of topics, such as invariance and sample-free estimates.

There are various methods found in the literature used to detect DIF items. Among those the Mantel-Haenszel (MH) DIF detection method has been investigated as a useful method for both large-scale and moderate-to-small-scale testing contexts. Several DIF studies have been conducted which referred to substantial small-scale testing, using this MH method. Some of those studies used sample sizes as small as 50 respondents per group in either reference or focal group (e.g., Muñiz, Hambleton, & Xing, 2001). Muñiz et al. investigated the DIF operating characteristics of Mantel-Haenszel using combinations of sample sizes ranging from 50 to 500 with a test of 40 items. Parshall and Miller (1995) analysed DIF with the MH method using sample sizes from 25 to 200 for each of the focal groups studied, each of which was combined with a reference group of 500. Although details will be provided later in the literature review, the upshot is that the MH method was not as successful as expected for moderate-to-small-scale studies in detecting DIF items. Because there was no method found to work consistently in the small-scale testing context, there was a need of a new method. The DIF detection method found in the software TestGraf may be a viable one for moderate-to-small-scale testing, because TestGraf was developed with moderate-to-small-scale testing in mind.

TestGraf is a relatively new software based method for evaluating items with a nonparametric item response modeling approach (Ramsay, 2000). TestGraf's DIF detection method was designed not only to detect the presence of DIF, but the magnitude of the DIF effect as well. For DIF to be practically useful, it is important to have a statistical method that can accurately detect it as well as measuring its magnitude. To date, TestGraf's DIF statistical test has not been investigated. In fact, the standard error of TestGraf's DIF statistic is reported, but it has not yet been used to construct a hypothesis test of DIF, similar to the hypothesis test found with the MH.

Although large-scale assessment and testing continues unabated, the need for moderate-to-small-scale testing has always persisted. Therefore, the present study was intended to highlight this new method of assessing DIF for moderate-to-small-scale testing (to contrast it with large-scale testing) wherein one has far fewer examinees and in combination with far fewer items or questions than the above defined large-scale testing context.

The present study is based on the pioneering work of Muñiz et al. (2001) by exploring smaller sample sizes and studying the TestGraf DIF detection method. Findings of this study will make a significant contribution to the field of educational and psychological measurement and evaluation, and primarily to the development of research methods for detecting items with DIF in the context of moderate-to-small-scale testing.

Problem Statement

During the last decades many assessment practitioners had been facing problems in assessing differential item functioning (DIF) when dealing with small number of examinees or/and small number of test items. Ramsay (2000) described a nonparametric item response modeling DIF detection method that required far fewer examinees and items than did the other DIF detection methods. This DIF detection method also provided a graphical method of evaluating multiple-choice test and questionnaire data (Zumbo & Hubley, 2003). However, Type I error rate of detecting any DIF has been unknown. The DIF detection method from TestGraf was new and had not been investigated in terms of its operating characteristics. This study was designed to specifically investigate the Type I error rate of the TestGraf DIF method in the context of moderate-to-small-scale testing as defined above.

As well, this study investigated the cut-off criteria for significant DIF provided by Roussos and Stout (1996). The formal statistical test is new, whereas the latter method (i.e., the cut-off approach) is known in the literature for some time.

CHAPTER II

LITERATURE REVIEW

In this chapter, the most frequently used method of DIF detection, the Mantel-Haenszel (MH) test, will be reviewed and the main research questions will be identified. Then, a description of the nonparametric item response modeling approach, which was the primary method of the study, is presented. Each of the methods to be presented will be accompanied by an example to demonstrate any critical differences in the methods to be implemented. Along the way, the Roussos-Stout cut-off criteria for significant DIF will be introduced.

One should recall that, for the purposes of the present study, large-scale testing is defined as a testing context involving more than 500 examinees and moderate-to-small-scale testing involves 500 or less examinees per group. It should be noted, however, that even though the above definitions do not involve the number of items, large-scale testing typically involves a large number of items.

Methods for DIF Detection for Moderate-to-Small-Scale Testing Contexts

The contingency table, MH, method will be reviewed below, followed by the nonparametric TestGraf item response modeling method. However, before delving into the details of the methodologies, a few overall remarks denoting the essential differences between MH and TestGraf (IRT) are appropriate at this point.

Zumbo and Hubley (2003) provide three frameworks for dealing with DIF: (1) modeling item responses via contingency tables and/or regression models, (2) item response theory (IRT), and (3) multidimensional models. The first framework for DIF includes two DIF detection

methods, logistic regression and MH. The MH method has been suggested for detecting DIF items in the context of moderate-to-small scale testing (e.g., Muñiz, Hambleton, & Xing, 2001). The second and third frameworks, IRT and multidimensional methods, are primarily targeted for large-scale testing.

There are two noteworthy points beyond the distinction in terms of the number of examinees between MH and IRT. First, although IRT is used in large-scale testing, the concept of item response modeling is applied in TestGraf's nonparametric regression method (Ramsay, 2000); TestGraf was designed with moderate-to-small scale testing in mind. Second, MH focuses on the observed differences in response proportions to each item whereas IRT focuses on the differences between the item response functions – the function that traces the relation between the latent variable score and the likely, or expected, item response.

A key feature of DIF is the "matching" to study the group differences. That is, the definition of DIF implies that groups of individuals, who are matched on the key variable being measured, should perform equally well on each item. A further distinction, however, between the MH and IRT (including TestGraf) DIF detection methods is that MH matches on the actual observed total scale score – a discretized total scale score – whereas the IRT methods integrate out the matching score. As Zumbo and Hubley (2003) write:

> In its essence, the IRT approach is focused on determining the area between the curves (or, equivalently, comparing the IRT parameters) of the two groups. It is noteworthy that, unlike the contingency table ... methods, the IRT approach does not match the groups by conditioning on the total score. That is, the question of "matching" only comes up if one computes the difference function between the groups conditionally (as in MH ...). Comparing the IRT parameter estimates or ICCs [item characteristic curves] is an unconditional analysis because it implicitly assumes that the ability distribution has been 'integrated out'. The mathematical expression 'integrated out' is commonly used in some DIF literature and is used in the sense that one computes the area between the ICCs across the distribution of the continuum of variation, theta. (p. 506-507).

Mantel-Haenszel (MH) method.

Developed by Mantel and Haenszel (1959; in Camilli & Shepard, 1994) for medical research this method was first applied to DIF research by Holland and Thayer (1988). The MH DIF detection method is considered to be one of the most widely used contingency table procedures. It uses the total test score for matching examinees of the reference and focal groups and compares the two groups in terms of their proportion of success on an item. MH discretizes the total scores into a number of category score bins. The MH method treats the DIF detection matter as one involving, in essence, a three-way contingency table (Zumbo & Hubley, 2003). The three-way contingency table consists of the correctness or incorrectness of a response, the group membership, and the total score category or score bin.

The MH procedure requires relatively fewer examinees than other DIF methods – with the exception of TestGraf. It is easy to understand and a relatively inexpensive in terms of computing time. The method yields a significance test distributed as a chi-square statistic with one degree of freedom for the null DIF hypothesis and an estimate of DIF effect size as the MH common odds ratio estimator. The null hypothesis chi-square statistic indicates that the likelihood of a correct response to a given item is equal for examinees of the reference and focal groups at equal ability level. If the chi-square is statistically significant, the item is considered to perform differentially for the compared groups. If the common odds ratio estimate is greater than one, it suggests that the item benefits the reference group and if it is less than one, the focal group.

Over many replications of a study, the statistical significance of chi-square statistics refers to the Type I error rate, which is expected to be near the nominal significance level alpha (α). With respect to this nominal, Bradley (1978) stated a liberal criterion of robustness. A test

conforms to Bradley's liberal criterion at $\alpha = .05$ if the Type I error rate is between 0.025 and 0.075.

Let me demonstrate the MH with a hypothetical example of the MH DIF method. The data set was simulated to be statistically non-significant (i.e., an investigation of Type I error rates) with a sample size combination of 100 and 50 for the reference and focal groups, respectively. The item parameters of this item for the two groups were:

$$a = 1.00, b = -1.00, c = 0.17;$$

where *a* refers to the item discrimination, *b* the item difficulty, and *c* the pseudo-guessing parameters in IRT. As expected, it produced not-significant *p*-value to suggest there was no-DIF, $\chi^2 (1) = 0.014, p = .905$.

TestGraf DIF: Nonparametric regression to assess DIF.

The present study used the computer program TESTGRAF (Ramsay, 2000; December 20th, 2002 version). TestGraf software was developed by Ramsay and was designed to aid the development, evaluation, and use of multiple choice examination, psychological scales, questionnaires, and similar types of data. As its manual describes, TestGraf requires minimal data, which are characterized by (1) a set of examinees or respondents, and (2) a set of choice situations, such as items or questions on an examination or a questionnaire. As Ramsay states, although it can be used with large-scale testing data, TestGraf was developed with the intention of aiding instructors and teachers, typically at the college or university level, with their test analysis.

TestGraf implemented nonparametric regression method. Nonparametric regression is the term used for a wide range of methods that directly estimate a functional relationship between an

independent variable X and a dependent variable Y. TestGraf applies what is considered the simplest and most convenient computational procedure, that is the normal kernel or the Gaussian kernel smoothing. This normal kernel is simply the standard normal density function,

$$K_N(z) = \frac{1}{\sqrt{2\Pi}} e^{-z^2/2}$$

where $K_N(z)$ refers to the normal kernel function (Fox, 2000).

The kernel smoothing is used to estimate an item characteristic curve (ICC) and an option characteristic curve for each item. The option characteristic curve is the ICC for each option for an item. The ICC is, in essence, the option characteristic curve for the correct option. The ICC curve displays the relationship between the probability that examinees choose the correct response and the proficiency variable or the expected score. Therefore, each ICC provides information about the probabilities of getting a correct response over the given range of proficiency levels or expected scores. The expected score refers to the expected or average number of correct items that an examinee at a particular proficiency level will achieve. Hence, it provides information over the range of the proficiency variable or the expected scores obtained.

TestGraf uses a kernel smoothing to estimate the probability of option m to item i, P_{im}. Ramsay (2000) described the proficiency value θ as the independent variable, and the dependent variable is the probabilities of an examinee, *a,* choosing option m for item i, which is the actual choice. The actual choices can be summarized numerically by defining an indicator variable of y_{ima}. In the binary situation, this is denoted as one if an examinee picks a correct option and zero if incorrect. Assigning θ_a to each examinee and then smoothing the relationship between the binary values of 0-1 and the examinee abilities will give an estimate of the probability function $P_{im}(\theta)$.

The estimation follows four sequential steps (Ramsay, 2000):

1. Estimate the rank of each examinee, r_a, by ranking their total scores on the exam.
2. Replace these ranks with the quantiles of the standard normal distribution. These quantiles will be used as the proficiency level values θ_a, $a = 1,\ldots, N$; and are calculated by dividing the area under the standard normal density function into $N + 1$ equal areas of size $1/(N + 1)$, where N denotes the total number of examinees.
3. Arrange the examinee response patterns $(X_{a1}, \ldots, X_{an})$ by the estimated proficiency rankings. That is, the a_{th} response pattern in the given test, $(X_{(a)1}, \ldots, X_{(a)n})$, is that of the examinee by the level of θ_a.
4. For the m^{th} option of the i^{th} item, estimate P_{im} by smoothing the relationship between the binary item-option indicator vector y_{ima} of length N and the proficiency vector $\theta_1, \ldots, \theta_N$ using the equation

$$P_{im}(\theta_q) = \sum_{a=1}^{N} W_{aq} Y_{ima}$$

where y_{ima} is the indicator variable as described earlier and w_{aq} is the weight applied at a particular point. The weight vector is computed from

$$W_{aq} = \frac{K[(\theta_a - \theta_q)/h]}{\sum_{b=1}^{N} K[(\theta_b - \theta_q)/h]},$$

where the h value used by TestGraf is set equal to $1.1N^{-1/5}$.

In short, the TestGraf methodology is a nonparametric regression of the item response onto the latent score.

If there is a large number of examinees, tens or hundreds of thousands, estimation of the ICC at a display value θ_q is done by averaging the indicator values y_{ima} for the values of θ_a falling within the limits $(\theta_{q-1} + \theta_q)/2$ and $(\theta_q + \theta_{q+1})/2$; that is, between the centres of the adjacent

intervals, $[\theta_{q-1}, \theta_q]$ and $[\theta_q, \theta_{q+1}]$. The average of values that fall below the centre of the first interval denotes the smallest value of y_{ima} and that of above the centre of the last interval the largest value of y_{ima}. These Q averages are indicated by P_{imq}. The area under the standard normal curve between the two interval centres is computed as Q_q. This Q_q is smoothed by the equation

$$P_{im}(\theta_q) = \sum_{r=1}^{Q} W_{rq} p_{imr} .$$

The summation resulted in much smaller set of display values (by default is set to 51 by TestGraf) rather than over the potentially enormous number of examinee indices. The weight of the values in interval r for θ_q is computed by the equation

$$W_{rq} = \frac{Q_r K[(\theta_r - \theta_q)/h]}{\sum_{s=1}^{Q} K[(\theta_s - \theta_q)]}$$

where r is the interval of values that are close to q, Q_r is the area under the curve between the two intervals, θ_r is the value of the centre of the r interval, and θ_s is the index of the summed r intervals. As can be seen, the weight values do not depend on the item neither on the option; hence a matrix of order Q containing their values can be computed initially and then used for all curves.

Examples of TestGraf displays. Figures 1 and 2 provide some description of TestGraf graphical output. They present the ICCs of the same item used for the example of MH. Figure 1 represents a reference group with sample size of 100 and Figure 2 a focal group with sample size of 50. In these figures, the X-axis represents the expected scores, and the Y-axis is the likely correct item response at the corresponding expected scores. The vertical dashed lines across the entire plot indicating various quantiles of the standard normal distribution. The third vertical dashed line, from the right, indicates the median while 50% of the scores lie between the second

and fourth lines, and 5% beyond the first and fifth vertical dashed lines. Going from left to right, the vertical dashed lines indicate the 5th, 25th, 50th, 75th, and 95th percentiles, respectively.

For multiple-choice items or binary case, as mentioned earlier, the correct response has a weight of one and each of the other options weighs zero. The ICC displayed in Figure 1 and in Figure 2 is simply the ICC for a correct response. The vertical solid lines designate estimated 95% confidence limits for the value of the curve at specific expected score values. These are referred to as *point wise confidence limits*. The name was given for distinguishing them from confidence limits for the entire curve. As can be seen, these point wise confidence limits are broader for lower scores, because only few examinees fell into this range of expected scores. On the contrary, the point wise confidence limits are denser for higher scores indicating that those examinees of higher expected scores had more probability of getting correct answer than those of lower expected scores.

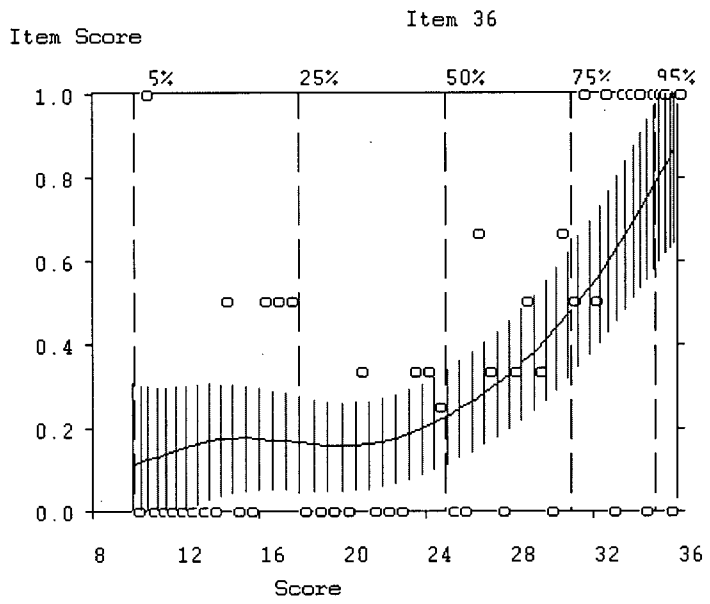

Figure 1. Item characteristic curve of the reference group with $N = 100$

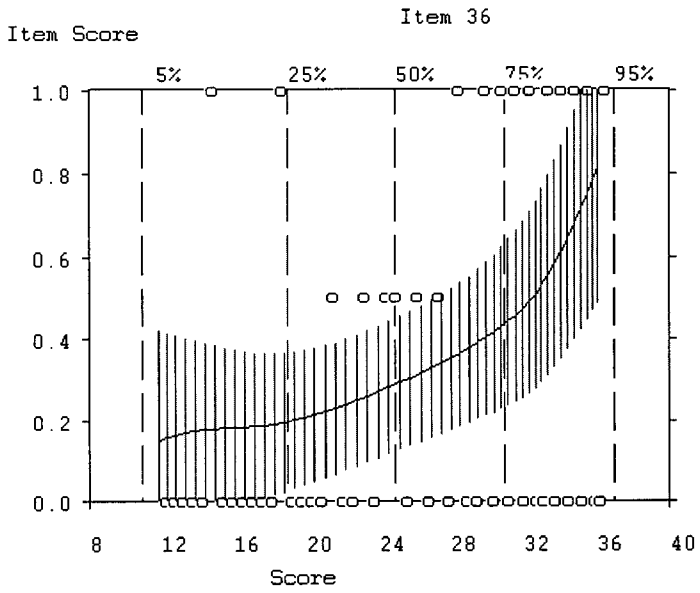

Figure 2. Item characteristic curve of the focal group with $N = 50$

TestGraf DIF detection method. As previously mentioned, TestGraf has also been developed to display the differences among two or more groups of examinees or respondents in terms of how they responded to items or questions. This indicated the capacity of TestGraf to assess DIF by showing whether there were systematic ways in which, for instance, females and males that possessed similar underlying ability level responded differently to a particular item or question, or whether different ethnic or language groups responded to particular items or questions in different ways. Because the TestGraf framework is a latent variable approach, it must be noted of the importance of a common metric which is often overlooked in the literature. If the ability distributions of the compared groups are not on the same metric, any DIF results are difficult to interpret.

When comparing two or more groups, in practice one of the groups is identified as the reference group and the remainder as focal group(s). The reference group is the standard of comparison, and considered the one being advantaged over the remaining group(s), while the focal group is being disadvantaged on an item or a test being studied. TestGraf assumed that the first file named during an analysis is the reference group. For the purposes of this study, however, neither group was identified as being disadvantaged nor advantaged. Instead, the main interest was only comparing two groups in assessing the operating characteristics of the nonparametric item response modeling method, i.e. its Type I error as previously mentioned.

Like other IRT methods (Zumbo & Hubley, 2003), TestGraf measures and displays DIF in the form of a designated area between the item characteristic curves. This area is denoted as beta, symbolized as β, which measures the weighted expected score discrepancy between the reference group curve and the focal group curve for examinees with the same ability on a particular item. This DIF summary index of β is computed in TestGraf as follows.

Let the proportion of the reference group having display variable value θ_q be indicated by p_{Rq}. And let $P_{im}^{(R)}(\theta)$ and $P_{im}^{(F)}(\theta)$ stand for the *option* characteristic curve values for the reference and focal groups, respectively. Then β is

$$\beta_{img}(\theta) = \sum_{q=1}^{Q} p_{Fq}\left[P_{im}^{(F)}(\theta) - P_{im}^{(R)}(\theta)\right].$$

In a situation where the focal group is found to be disadvantaged on average, this index β is negative.

As Ramsay (personal communication, December 19, 2002) stated, the variance of β, which is denoted as Var[β_{img}] is computed by the equation:

$$Var[\beta_{img}] = \sum_{q}^{Q} p_{q0}^2 \ \{Var[P_{img}(\eta_q)] + Var[P_{im0}(\eta_q)]\},$$

where the notation is the same as above with the addition of *g* to denote group. The standard error of β is the squared root of that variance as follows

$$SE_\beta = \sqrt{Var[\beta_{img}]}.$$

To illustrate how the TestGraf software can identify DIF across two groups of examinees, Figure 3 is given below. It brings together the ICCs displayed in Figures 1 and 2. As expected, given that the data was simulated with no DIF, it produced a negligible amount of DIF with β = 0.055 and the $SE = 0.033$.

Curve 1 indicates the scores obtained by the reference group, whereas curve 2 by the focal group. As previously described, the X-axis shows the expected scores to be gained by examinees, and the Y-axis the item score or the probability of getting the item correct given the examinees' expected score. The Y-axis ranges from zero to one, the X-axis indicates the range of scores from the expected minimum to maximum scores that obtained by the examinees of the two comparison groups overall.

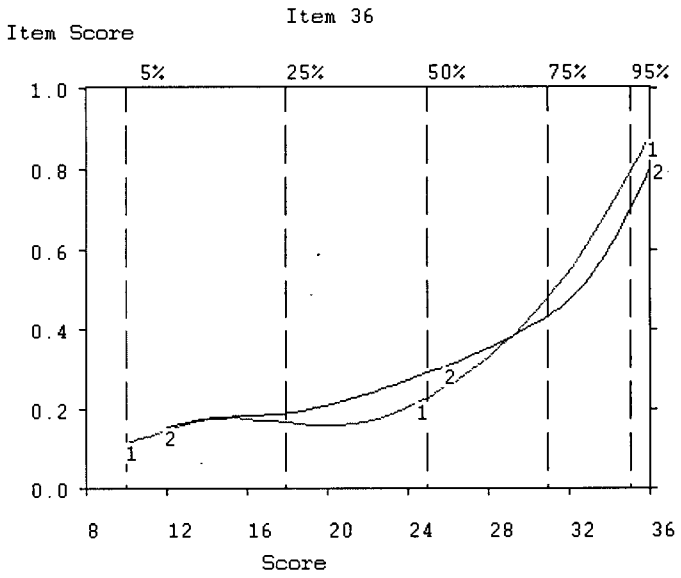

Figure 3. Item characteristic curves for reference and focal groups with each N=100 and 50. Curve 1 is for reference group, and curve 2 is for focal group.

The β and its corresponding standard error, which was provided by the TestGraf program could be used to obtain a confidence interval by dividing β by the standard error

$$\frac{\beta}{SE} \sim N(1,0)$$

According to Ramsay (personal communication, January 28, 2002), the distribution of this statistic should be a standard normal Z-distribution. The Z-value of the item illustrated in Figure 3 is Z = 0.055/0.033 = 1.667, which is not significant at nominal α = 0.05.

Testing for β DIF with TestGraf. Given the above-described test statistic for β, there are two approaches to testing the no-DIF hypothesis for β using the TestGraf DIF detection method:

1. Using the guideline proposed by Roussos and Stout (1996) for the SIBTEST procedure in identifying DIF items. They suggested to use the following cut-off indices: (a) negligible DIF if $|\beta| < .059$, (b) moderate DIF if $.059 \leq |\beta| < .088$, and (c) large DIF if $|\beta| \geq .088$.

These criteria were recently investigated by Gotzmann (2002). Applying the Roussos-Stout criterion, the example data displays negligible DIF.

2. Applying the statistical test described above by dividing β by its standard error and referring to a standard normal distribution. TestGraf produced both the β and the standard error of β. This is a new method that will be investigated for the first time in this dissertation. As Ramsay states:

> The program produces a composite index no matter whether one or two groups are involved. This index is always signed, and is given in equation (9) on page 65 of the latest version of the manual. It also produces a DIF index that is signed for each reference group (males in your case), and this is equation (8)on the previous page. Right, the standard errors can be used to give a confidence interval for the signed DIF value. (Ramsay, Personal Communication, January 28, 2002)

Of course, any decision rule for a hypothesis is a statistical test, of a sort. It should be noted, however, that of the two approaches, only the second is a conventional statistical hypothesis test as commonly used in statistics – i.e., a test statistic that refers to a particular sampling distribution to see if the result falls in the rejection region of the sampling distribution. Roussos & Stout's and Gotzmann's approach, although a test of DIF, serves as rather a heuristic device for helping interpret the magnitude of β and to aid the analysts in deciding if DIF is present. What is sought here in the present study is to develop the heuristic device toward to a decision rule for DIF and hence a statistical test of DIF.

Findings from Simulation Studies

This section presents relevant simulation studies of the statistical properties of the MH and TestGraf methods. Because the main interest of this dissertation was the context of moderate-to-small-scale testing, the literature review was limited to studies that investigated 500

or less examinees per group. In addition, only DIF studies that investigated the Type I error rate of DIF detection were reviewed. Given the hypothesis of no-DIF, Type I error rate in detecting DIF refers to declaring an item as DIF when it is not a DIF item (Null hypothesis). It is also called a false positive (Rogers & Swaminathan, 1993).

MH test.

In the study by Muñiz, Hambleton, and Xing (2001), they used various sample size combinations of 50/50, 100/50, 200/50, 100/100, and 500/500 for the reference and focal groups, respectively. Type I error rates were obtained based on the 34 non-DIF items in their simulation study. Findings of their study showed Type I error rates associated with the Mantel-Haenszel method under the combinations of sample sizes of 50/50, 100/50, 200/50 and 100/100, ranging from 0% to 6%. In contrast with the sample size combination of 500/500, Type I error rates ranged from 6% to 16%. The nominal Type I error rate (alpha) was set at 0.05.

Rogers and Swaminathan (1993) applied the MH method in detecting items with DIF. They used two levels of sample sizes of 250 per group and 500 per group. Three different test lengths were applied, 40, 60, and 80, each of which contained 10% uniform DIF items and 10% non-uniform DIF items, and the remaining 80% of non-DIF items were used to obtain the false positives or Type I error rates. Their findings showed that Type I error rates of the Mantel-Haenszel method consistently resulted in around 1% under all conditions of test lengths.

TestGraf.

From the above findings, it was evident that the formal hypothesis test of TestGraf had not been investigated in terms of its operating characteristics in detecting DIF items. Although it

involves 500 and more examinees per group, the Gotzmann (2002) study is reviewed because it is the only simulation study of the Roussos-Stout decision heuristic approach to DIF detection. Gotzmann conducted a simulation study that included the investigation of Type I error rate of TestGraf DIF detection method. The study used eight sample size combinations for the reference and focal groups: 500/500, 750/1000, 1000/1000, 750/1500, 1000/1500, 1500/1500, 1000/2000, and 2000/2000. The findings showed that Type I error rates at the nominal .05 level, with an exception of the 500/500 sample size condition, were considerably below .05. It is worth noting that Gotzmann applied, as previously mentioned, the guideline proposed by Roussos and Stout (1996).

Research Questions

It is clear from the literature review that:

- The MH test is, at best, inconsistent in terms of protecting its Type I error rate, sometimes being quite liberal in moderate-to-small-scale contexts.
- More research needs to be done to investigate the Type I error rate of (a) TestGraf's formal hypothesis test approach, and (b) TestGraf heuristic approach to detecting DIF with particular emphasis on moderate-to-small-scale contexts wherein the heuristic approach has not yet been investigated.

Hence, the primary purpose of this study was to investigate the operating characteristics of the nonparametric item response modeling method (both the formal test and heuristic rule) in the moderate-to-small-scale testing context. That is, do the statistical tests have a Type I error rate that was equal to or less than the nominal alpha set for the test statistics. The study of Type I

error rate is necessary before one can consider turning to investigating the probability of detecting DIF when DIF was actually present in the data – i.e. statistical power.

In addition, Type I error rate of the MH DIF detection method was investigated for two reasons. First, to extend the measurement literature (i.e., the knowledge base) about the performance of the MH with smaller sample sizes, and at the same time allow a comparison of Type I error rate of the new TestGraf method with the more widely documented MH method. Second, given that the MH has been previously investigated, it was used as a validity check of the simulation data and the study design. A word of caution should be stated in comparing MH to TestGraf. Zumbo and Hubley (2003) explained that an item response modeling method such as TestGraf, integrates out the latent variable, while the MH statistically conditions (i.e., covaries) the empirical categorization of the total score. This distinction between hypotheses may be magnified in the case of moderate-to-small-scale testing because the conditioning bins for the MH will typically involve very wide bins (sometimes called thick matching).

The present study answered five specific research questions in the context of an educational testing and sample size. They were as follows:

I. What is the Type I error rate of the Mantel-Haenszel statistical test?

II. Is the shape of the sampling distribution of the beta TestGraf DIF statistic normal as assumed by the formal test of DIF from TestGraf?

III. Is the standard deviation of the sampling distribution of the beta TestGraf DIF statistic the same as the standard error computed in TestGraf?

IV. What is the Type I error rate of the formal statistical test of the TestGraf?

V. What is the Type I error rate of the statistical test of TestGraf using the cut-off index of $|\beta| < .059$ as proposed by Roussos and Stout (1996) for the SIBTEST procedure

when sample size combinations were 500/500 and smaller? If the Roussos-Stout cut-offs resulted in an inflated error rate, new cut-offs will be determined from the simulation data.

CHAPTER III

METHODOLOGY

The methodology and study design used in this simulation study was similar to the one used by Muñiz, Hambleton, and Xing (2001). Although the current study adopted the sample size combinations that were used in the Muñiz et al. study, three additional sample size combinations, to be described later, were included.

Study Design

To investigate the Type I error rate of the DIF detection methods, the simulation involves generating two populations of item responses. As is standard practice in the measurement literature, in each case, the item parameters were generated using the three-parameter item response theory model. The three item parameters include (a) item discrimination, (b) item difficulty, and (c) the lower asymptote (that is sometimes referred to as guessing or pseudo-guessing) parameters. In short, for each population, what the simulation process requires is generating a normal distribution with a specified mean and variance for each item and then using the three-parameter IRT model to compute a likely item response. In essence, the IRT model becomes a model of item responding for each item. In simulation studies, as reviewed in the previous chapter, one may manipulate any of the input values for the item parameters or mean and variance of the latent distribution. If the item parameters are different across the two populations, the simulation study depicts a DIF situation. If, on the other hand, the mean of the latent distribution in one population group is different from the other, the simulation study depicts a test impact situation. Once the populations are constructed, one would sample

randomly (with replacement) from the population, a type of bootstrapping, to generate samples of response simulees.

The DIF tests were conducted on each sample and a tally was recorded of the statistical decision for each test for each sample. If the item parameters and latent distribution characteristics were the same for the two populations, the tally is the Type I error rate.

Distribution of the latent variable and sample sizes.

In the present study, the two population groups being studied had equal ability distributions, normal distributions with $M = 0$ and $SD = 1.0$. The sample sizes of the reference and focal groups were varied in the study design. Building on the design used in Muñiz et al. (2001), the sample sizes used for the reference and focal groups were: 500/500, 200/100, 200/50, 100/100, 100/50, 50/50, 50/25, and 25/25 examinees in pairs, respectively. Five of the above combinations were the same with that used in the study by Muñiz et al. The additional sample size combinations, 200/100, 50/25, and 25/25 were included so that an intermediary between 500/500 and 200/50, and smaller sample size combinations were included. In addition, as Muñiz et al. suggested, these sample size combinations reflect the range of sample sizes seen in practice in, what I would refer to as, moderate-to-small-scale testing.

Statistical characteristics of the studied test items.

The item characteristics for the 40 item simulated test involved two parts. First, item statistics for the first 34 non-studied items were randomly chosen from the mathematics test of the 1999 TIMSS (Third International Mathematics and Science Study) for grade eight, whereas item statistics for the last six items studied were adopted from the DIF items used in the study of

Muñiz et al. (2001). The set of these six items were the focus of this study. The study design included two item discrimination levels of low and high, each of which consisted of three item difficulty levels low, medium, and high. Item statistics for the 40 items were presented in Table 2. The last six items were the items for which DIF was investigated – i.e., the studied DIF items. The *a* refers to the item discrimination parameter, *b* the item difficulty parameter, and *c* the pseudo-guessing parameter.

Both groups in the DIF analysis had the same population item characteristics; therefore, this was a study of the Type I error rate of the DIF detection methods. The current study used the three-parameter logistic item response-modeling program of MIRTGEN (Luecht, 1996) to generate data for the population of the simulation. Two population groups were generated each with 500,000 population simulees. One hundred data sets were randomly sampled from each population and analysed for each sample size combination. This made it possible to obtain Type I error rate in any combination of 100 replication pairs. Therefore, the data sets were reflective of actual test data.

Table 2

Item Statistics for the 40 Items

Item #	*a*	*b*	*c*
1	1.59	0.10	.19
2	0.60	-0.98	.20
3	0.75	-0.42	.06
4	1.08	0.43	.24
5	0.81	0.34	.32
6	0.66	-0.57	.38
7	0.81	-0.18	.20
8	0.43	-0.36	.30
9	0.94	0.45	.34
10	1.40	0.15	.07
11	0.98	-0.20	.18
12	1.28	-0.12	.23
13	1.18	0.18	.23
14	0.98	-0.63	.30
15	0.94	-0.14	.17
16	1.39	0.94	.43
17	0.78	0.25	.16
18	0.55	-0.82	.20
19	0.88	0.09	.27
20	1.10	0.14	.40

Item #	*a*	*b*	*c*
21	1.23	-0.43	.10
22	0.73	1.13	.27
23	0.54	-1.91	.23
24	0.71	-0.43	.31
25	0.66	-0.67	.16
26	1.14	0.59	.18
27	1.12	0.29	.26
28	0.96	-0.26	.23
29	0.95	0.13	.15
30	1.38	0.66	.16
31	1.38	1.11	.16
32	0.42	-0.02	.20
33	1.04	-0.01	.30
34	0.73	0.10	.13
35	0.50	-1.00	.17
36	1.00	-1.00	.17
37	0.50	0.00	.17
38	1.00	0.00	.17
39	0.50	1.00	.17
40	1.00	1.00	.17

Computer simulation design and dependent variables.

The computer simulation study was an 8 x 3 x 2 completely crossed design: sample size combinations, by item difficulty levels, and by item discrimination levels. The dependent variables recorded for each replication of the simulation were:

1. The beta (β) and the standard error of β for each item produced by TestGraf. From this information I was able to compute the DIF hypothesis test statistics (via a confidence interval) for each item for each replication. In addition, βs allowed me to apply such a heuristic test of DIF as based on the Roussos and Stout (1996) criterion.
2. The MH test was performed for each item and each replication.

Procedure

As an overview, the study was carried out following a modified procedure originally used in the study of Muñiz et al. (2001) as follows:

1. First, set a test of 40 items with the item parameters of the first 34 items were drawn from the TIMSS data, and those of the last six items were adopted from the studies of Muñiz et al. These item parameters were provided in Table 2.
2. Two populations of simulees were created with no DIF and no item impact – i.e., equal item parameters and equal ability means across populations.
3. One hundred replications from the reference and focal population groups were created for the eight sample size combinations: 500/500, 200/100, 200/50, 100/100, 100/50, 50/50, 50/25, and 25/25 examinees for reference and focal groups in pairs, respectively.
4. TestGraf was used to compute the nonparametric IRT procedure and SPSS was used to compute the MH test.

5. Repeated step 4 for the 100 replications of each of the 48 (8x3x2 cells) conditions in the study. Type I error rate and the other statistics were computed over the 100 replications.

Because of the statistical software limitations, I conducted all of the above steps manually. That is, a batch computer code could not be written to handle the sampling and computation. The computing time for the simulation studies including that reported in Appendix B, was approximately 8 months at 5 days a week and approximately six to eight hours a day.

Data Analysis of Simulation Results

For each of the 6 items studied and each of the sample size combinations, the Type I error rate of the DIF detection per item was obtained by dividing the number of times the null hypothesis was rejected by 100, the number of replications. The mean and standard deviation of the β values for the 100-replications of each sample size combination were computed. Distribution of β for each condition was examined for normality.

The following analyses were carried out to answer each of the research questions:

1. Tabulated for Type I error rate of the statistical test of the Mantel-Haenszel.
2. Conducted the Kolmogorov-Smirnov test on TestGraf β for normality. Because it was done for each item separately, the Type I error rate of the K-S test was set at .01.
3. Computed empirical standard error of β and compared to TestGraf standard error.
4. Tabulated for Type I error rate of the statistical test of TestGraf DIF detection method.
5. Computed Type I error rate of TestGraf DIF detection method using the Roussos-Stout criterion of $|\beta| < .059$.
6. Calculated the 90^{th}, 95^{th}, and 99^{th} percentiles of β values of each item across sample size combinations for the nominal alpha of .10, .05, and .01, respectively.

Version of TestGraf Used in the Simulation

The December 2002 version of TestGraf was used in this simulation. Appendix B reports on a study that used the pre-December 2002 version of TestGraf. The study in Appendix B resulted in a new version TestGraf because it was found via my simulation in Appendix B that TestGraf was incorrectly computing the standard error of β. Readers interested in the details of that simulation should see Appendix B.

CHAPTER IV

RESULTS

To answer each of the research questions, several analyses were conducted on the simulation data. The independent variables in the study design were sample size combinations of reference and focal groups, and the item discrimination and item difficulty parameters. There were eight sample combinations generated for the purpose of the study: 500/500, 200/100, 200/50, 100/100, 100/50, 50/50, 50/25, and 25/25. Item discrimination (a-parameter) consisted of two levels, low (a = 0.50) and high (a = 1.00). Item difficulty (b-parameter) consisted of three levels, low (b = -1.00), medium (b = 0.00), and high (b = 1.00).

As a reminder to the reader, and to help structure the reporting of the results, the Type I error rate of the Mantel-Haenszel DIF detection method was investigated to:

(i) extend the measurement literature (i.e., the knowledge base) about the performance of the MH with smaller sample sizes and, at the same time, allow a comparison of Type I error rate of TestGraf method with the more widely documented MH method, and concomitantly

(ii) provide supporting empirical evidence of the validity of the simulated data and study design.

The simulation outcomes of the TestGraf DIF detection method were examined to investigate:

(i) the purported normality of the sampling distribution of the TestGraf βs using the Kolmogorov-Smirnov test for one-sample on the TestGraf β for the six studied items and sample size combinations,

(ii) the mean of β over 100 replications (i.e., the mean and standard deviation of the sampling distribution of β) for each item at each sample size combination,

(iii) whether the standard deviation of β equals the standard error of β produced by TestGraf -- i.e., if the standard error produced by TestGraf is the expected standard error,

(iv) the Type I error rate of the formal hypothesis test of DIF based on the TestGraf method,

(v) the Type I error rate of the Roussos-Stout (1996) criterion of $|\beta| < .059$,

(vi) and finally, for the conditions for which this Type I error rate is inflated, new cut-offs by studying the percentiles of β values of each item across sample size combinations at 90%, 95%, and 99% for α of .10, .05, and .01, respectively.

All results with regard to the Type I error rates of MH, sampling distribution and statistics of TestGraf betas, Type I error rates of TestGraf, and the cut-off indices are presented in tables along with the explanation referring to the relevant tables. To determine if the observed Type I error rate is within an acceptable range of the nominal alpha, the confidence interval strategy from Zumbo and Coulombe (1997) was used. This strategy treats the empirical Type I error rate like a proportion computed from a particular sample size (i.e., the number of replications in the simulation design). A confidence interval was computed by the modified Wald method presented in Agresti and Coull (1998) to investigate if one has coverage of the nominal alpha. That is, if the confidence interval contains the nominal alpha, then the test is considered to be operating appropriately at that nominal alpha. In the following Tables that contain Type I error rates, a Type I error in bold font denotes an inflated Type I error. Note that if either of the

confidence bounds equals the nominal alpha, the test will be considered as operating appropriately.

A methodological question in analyzing the simulation results arises because statistical significance (and confidence intervals) is used in analyzing the simulation results. In short, the study design involves 48 cells (i.e., an 8x3x2 design) and a confidence interval (or hypothesis test for research question 2) is computed for each of these cells. This could result in inflation of the Type I errors of the methods used to examine the simulation results. As is typical of any research study, to address this matter, one needs to balance the inflation in Type I error rate with the potential loss of statistical power if one is conservative in correcting for this inflation. That is, if one corrects for every hypothesis test for each research question, the per comparison Type I error rate would be very small as would be the statistical power. The very liberal option is to ignore the concern of Type I error rate and simply conduct each analysis at an alpha of .05. To strike a balance, the following strategy was applied: (a) each research question was considered individually as was each Table in the results, and (b) for each Table of results the analysis error rate was set at .05/(the number of rows in the Table). Therefore, given that each of the studied Tables has eight rows, when a statistical test of the results is reported the alpha for the analysis of the simulation outcomes is conducted at an alpha of .006. Therefore, the Kolmogorov-Smirnov tests used the *p*-value threshold of .006 for significance and the confidence intervals were computed for 99% rather than 99.4%, because of software limitations that allowed only whole numbers for the percentage of coverage of the interval.

Mantel-Haenszel

Research Question 1: What is the Type I error rate of the Mantel-Haenszel statistical test?

The results of the Type I error rate for the MH at alpha of .05 and .01 are listed in Tables 3 and 4, respectively. The first column of these tables lists the sample size combinations, while the remaining columns are the six studied items. As an example of how to interpret these tables, the first row of the results in Table 3 is for a sample size combination of 500/500, and given that none of values are in bold font, all of the items have a Type I error rate within acceptable range of the nominal alpha.

Applying the confidence interval strategy for the Type I error rates in Tables 3 and 4, one can see that there are no inflated Type I errors at the nominal alpha of .05 nor .01*). In correspondence with Rogers and Swaminathan (1993), it was found that, as expected, the MH operates appropriately for the sample size of 500/500, hence adding empirical evidence as a design check on the simulation methodology used herein. Although the simulation methodology used in this dissertation is standard, I believe that it is always important to have an empirical crosscheck built into the simulation study.

*) Table 3 contains one result that rests on the borderline of admissible Type I error rate. For a sample size combination of 100/50, Item-40, the confidence interval is .049 to .219.

Table 3

Type I error rate of Mantel-Haenszel at nominal $\alpha = .05$

	Item Discrimination Level (*a*)					
	Low			High		
	Item Difficulty Level (*b*)					
	Low	Medium	High	Low	Medium	High
N_1 /N_2	Item35	Item37	Item39	Item36	Item38	Item40
500/500	.01	.06	.03	.05	.04	.01
200/100	.02	.06	.02	.03	.04	.07
200/50	.05	.04	.05	.03	.06	.01
100/100	0	.03	.01	.02	.03	.05
100/50	.02	.07	.05	.06	.02	.11
50/50	.02	.03	0	.03	.08	.01
50/25	.02	.05	0	.02	.05	.01
25/25	.03	.02	0	.02	.03	.06

Table 4

Type I error rate of Mantel-Haenszel at nominal $\alpha = .01$

	Item Discrimination Level (*a*)					
	Low			High		
	Item Difficulty Level (*b*)					
	Low	Medium	High	Low	Medium	High
N_1 /N_2	Item35	Item37	Item39	Item36	Item38	Item40
500/500	0	.01	.01	.02	.01	0
200/100	0	.02	0	.01	.03	.03
200/50	.01	.01	.01	0	.02	0
100/100	0	.01	0	0	.01	0
100/50	0	.01	.01	0	.01	.02
50/50	.01	.01	0	.01	.02	0
50/25	.02	.03	0	0	.03	.01
25/25	.01	0	0	.01	.01	.01

TestGraf

Beta of TestGraf.

Research Question 2: Is the shape of the sampling distribution of the beta TestGraf DIF statistic normal as assumed by the formal test of DIF from TestGraf?

The results of the one-sample Kolmogorov-Smirnov test are in Tables 5 to10. The first column of these tables refers to the various sample size combinations. The present research

question uses only Columns [2] and [3] (from the left). All of the sample combinations and items showed no statistical significance when tested for normality of the beta sampling distribution – the reader should recall that, as described above, the significance level for each of these K-S Z tests is .006.

Table 5

Sampling distribution of beta of Item-35

N_1 / N_2	Kolmogo-rov-Smirnov *Z*	Asymp. Sig. (two-tailed) of K-S *Z*	*M* of beta over 100 replica-tions	*SD* of beta over 100 replications (Empirical *SE*)	*M* of the TestGraf *SE* over 100 replica-tions	Compari-son ratio [6]/[5]
[1]	[2]	[3]	[4]	[5]	[6]	[7]
500/500	0.617	.841	-0.00037	0.02723	0.01089	.39993
200/100	0.582	.887	-0.00051	0.05401	0.02277	.42159
200/50	1.241	.092	0.00921	0.08276	0.03071	.37107
100/100	1.030	.239	0.00454	0.06020	0.02460	.40864
100/50	0.440	.990	0.00580	0.07900	0.03232	.40911
50/50	0.692	.724	-0.00060	0.08771	0.03547	.40440
50/25	0.707	.700	0.00962	0.13437	0.05292	.39384
25/25	0.581	.889	0.01526	0.13225	0.05623	.42518

Note: *N* = sample size, *M* = mean, *SD* = standard deviation, *SE* = standard error

Table 6

Sampling distribution of beta of Item-36

N_1 / N_2	Kolmogo-rov-Smirnov *Z*	Asymp. Sig. (two-tailed) of K-S *Z*	*M* of beta over 100 replica-tions	*SD* of beta over 100 replications (Empirical *SE*)	*M* of the TestGraf *SE* over 100 replica-tions	Compari-son ratio [6]/[5]
[1]	[2]	[3]	[4]	[5]	[6]	[7]
500/500	0.515	.953	0.00188	0.02968	0.01058	.35647
200/100	0.676	.750	-0.00603	0.06101	0.02260	.37043
200/50	0.596	.870	0.01422	0.07548	0.03043	.40315
100/100	0.386	.998	0.00485	0.05951	0.02434	.40901
100/50	0.713	.689	0.01351	0.08055	0.03223	.40012
50/50	0.555	.918	0.00111	0.09474	0.03549	.37460
50/25	1.198	.113	-0.00696	0.11153	0.05249	.47064
25/25	0.654	.786	-0.01110	0.12575	0.05574	.44326

Note: *N* = sample size, *M* = mean, *SD* = standard deviation, *SE* = standard error

Table 7

Sampling distribution of beta of Item-37

N_1 / N_2	Kolmogo-rov-Smirnov *Z*	Asymp. Sig. (two-tailed) of K-S *Z*	*M* of beta over 100 replica-tions	*SD* of beta over 100 replications (Empirical *SE*)	*M* of the TestGraf *SE* over 100 replica-tions	Compari-son ratio [6]/[5]
[1]	[2]	[3]	[4]	[5]	[6]	[7]
500/500	0.712	.691	-0.00348	0.02629	0.01079	.41042
200/100	0.533	.939	-0.00048	0.06237	0.02319	.37181
200/50	0.639	.808	0.00926	0.08198	0.03133	.38217
100/100	0.539	.934	0.00793	0.06046	0.02495	.41267
100/50	0.504	.961	0.01588	0.08446	0.03300	.39072
50/50	0.567	.905	-0.01508	0.09703	0.03621	.37318
50/25	0.732	.657	-0.00497	0.14431	0.05487	.38022
25/25	0.446	.989	-0.00402	0.13522	0.05733	.42398

Note: *N* = sample size, *M* = mean, *SD* = standard deviation, *SE* = standard error

Table 8

Sampling distribution of beta of Item-38

N_1 / N_2	Kolmogo-rov-Smirnov *Z*	Asymp. Sig. (two-tailed) of K-S *Z*	*M* of beta over 100 replica-tions	*SD* of beta over 100 replications (Empirical *SE*)	*M* of the TestGraf *SE* over 100 replica-tions	Compari-son ratio [6]/[5]
[1]	[2]	[3]	[4]	[5]	[6]	[7]
500/500	0.672	.758	-0.00104	0.02588	0.01009	.38988
200/100	0.631	.821	-0.00007	0.05884	0.02195	.37305
200/50	0.710	.695	-0.00097	0.07964	0.03005	.37732
100/100	0.433	.992	-0.00146	0.06377	0.02390	.37478
100/50	0.612	.848	-0.00675	0.08089	0.03181	.39325
50/50	1.373	.046	-0.00904	0.09651	0.03501	.36276
50/25	0.580	.890	-0.00118	0.10154	0.05316	.52354
25/25	0.878	.424	0.01751	0.12067	0.05629	.46648

Note: *N* = sample size, *M* = mean, *SD* = standard deviation, *SE* = standard error

Table 9

Sampling distribution of beta of Item-39

N_1 / N_2	Kolmogo-rov-Smirnov *Z*	Asymp. Sig. (two-tailed) of K-S *Z*	*M* of beta over 100 replica-tions	*SD* of beta over 100 replications (Empirical *SE*)	*M* of the TestGraf *SE* over 100 replica-tions	Compari-son ratio [6]/[5]
[1]	[2]	[3]	[4]	[5]	[6]	[7]
500/500	0.738	.648	0.00148	0.01864	0.00895	.48015
200/100	0.681	.743	-0.00681	0.04102	0.01849	.45076
200/50	0.498	.965	-0.00309	0.05953	0.02372	.39845
100/100	0.690	.727	-0.00421	0.04629	0.02054	.44372
100/50	1.026	.243	-0.00399	0.06256	0.02617	.41832
50/50	0.443	.990	-0.00844	0.06174	0.02999	.48575
50/25	0.605	.858	-0.01281	0.08551	0.04335	.50696
25/25	0.976	.296	-0.01372	0.08439	0.04759	.56393

Note: *N* = sample size, *M* = mean, *SD* = standard deviation, *SE* = standard error

Table 10

Sampling distribution of beta of Item-40

N_1 / N_2	Kolmogo-rov-Smirnov *Z*	Asymp. Sig. (two-tailed) of K-S *Z*	*M* of beta over 100 replica-tions	*SD* of beta over 100 replications (Empirical *SE*)	*M* of the TestGraf *SE* over 100 replica-tions	Compari-son ratio [6]/[5]
[1]	[2]	[3]	[4]	[5]	[6]	[7]
500/500	1.052	.219	0.00179	0.02121	0.00896	.42244
200/100	0.691	.727	-0.00779	0.04973	0.01889	.37985
200/50	1.151	.141	-0.00514	0.05693	0.02447	.42983
100/100	0.751	.625	-0.00038	0.05135	0.02104	.40974
100/50	0.796	.551	-0.00542	0.07530	0.02659	.35312
50/50	0.678	.748	0.01129	0.07078	0.03033	.42851
50/25	0.763	.605	-0.00325	0.08820	0.04501	.51032
25/25	1.019	.251	-0.00387	0.11184	0.04821	.43106

Note: *N* = sample size, *M* = mean, *SD* = standard deviation, *SE* = standard error

Research Question 3: Is the standard deviation of the sampling distribution of the beta TestGraf DIF statistic the same as the standard error computed in TestGraf?

First, it is important to note that in Tables 5 to 10, the fourth column from the left lists the mean (*M*) of the βs over the 100 replications for the various items and sample size combinations. A confidence interval of *M* over the 100 replications was computed for each item for each sample size using the same confidence level as in the K-S tests reported above (i.e., a 99.4% confidence interval). In all cases, the confidence around the mean β contained zero indicating

that the statistic is an unbiased estimate of the population β even at small sample sizes, -- i.e., the mean of the sampling distribution is the population value of zero because the item parameters are the same for the two groups. This is an important point that needed to be established before examining the standard errors.

Columns 5 and 6 from the left in Tables 5 to 10 are the standard deviation of β over the 100 replications (i.e., an empirical estimate of the standard error of β) and the average of the standard error of β produced by TestGraf, $\beta_{img}(\theta) = \sum_{q=1}^{Q} p_{Fq}\left[P_{im}^{(F)}(\theta) - P_{im}^{(R)}(\theta)\right]$, over the 100 replications. One can see that the empirical standard errors are larger than the standard error produced by TestGraf betas of each item for every sample combination. The difference between these two statistics appears to be effected by the sample size combination wherein, for sample sizes of 500/500 in Table 10, the standard error produced by TestGraf ([6] = .00896) is mostly less than half of what it should be (i.e., the empirical standard error in Column 5 of each table, which is [5] = .02121 in Table 10). This comparison ratio is listed in column 7 or the last column from the right.

Type I error rate of TestGraf.

Research Question 4: What is the Type I error rate of the formal statistical test of the TestGraf?

The Type I error rates for the formal statistical test of DIF produced by TestGraf are reported in Tables 11 and 12 for a nominal alpha of .05 and .01, respectively. These two tables have the same layout as Tables 3 and 4. By examining Tables 11 and 12 one can see that all of the Type I error rates are inflated. This is an expected result given that the standard error of β produced by TestGraf is too small in comparison to the empirical standard error.

Table 11

Type I error rate of TestGraf at nominal $\alpha = .05$

	Item Discrimination Level (a)					
	Low			High		
	Item Difficulty Level (b)					
	Low	Medium	High	Low	Medium	High
N_1 /N_2	Item35	Item37	Item39	Item36	Item38	Item40
500/500	**.44**	**.40**	**.39**	**.47**	**.50**	**.45**
200/100	**.33**	**.46**	**.38**	**.44**	**.50**	**.38**
200/50	**.45**	**.57**	**.46**	**.44**	**.42**	**.39**
100/100	**.46**	**.43**	**.43**	**.43**	**.46**	**.47**
100/50	**.45**	**.45**	**.34**	**.42**	**.46**	**.42**
50/50	**.42**	**.39**	**.30**	**.46**	**.41**	**.38**
50/25	**.46**	**.50**	**.42**	**.38**	**.35**	**.30**
25/25	**.42**	**.46**	**.24**	**.35**	**.35**	**.37**

Note: Bold denotes an inflated Type I error

Table 12

Type I error rate of TestGraf at nominal $\alpha = .01$

	Item Discrimination Level (a)					
	Low			High		
	Item Difficulty Level (b)					
	Low	Medium	High	Low	Medium	High
N_1 /N_2	Item35	Item37	Item39	Item36	Item38	Item40
500/500	**.30**	**.28**	**.18**	**.38**	**.36**	**.34**
200/100	**.26**	**.33**	**.22**	**.36**	**.41**	**.29**
200/50	**.33**	**.40**	**.35**	**.30**	**.32**	**.28**
100/100	**.36**	**.28**	**.29**	**.30**	**.29**	**.33**
100/50	**.33**	**.29**	**.24**	**.25**	**.31**	**.34**
50/50	**.29**	**.31**	**.26**	**.32**	**.30**	**.24**
50/25	**.34**	**.32**	**.26**	**.21**	**.15**	**.22**
25/25	**.26**	**.32**	**.17**	**.22**	**.25**	**.24**

Note: Bold denotes an inflated Type I error

Cut-Off Indices

Research Question 5: What is the Type I error rate of the statistical test of TestGraf using the cut-off index of $|\beta| < .059$ as proposed by Roussos and Stout (1996) for the SIBTEST procedure when sample size combinations were 500/500 and smaller? If the Roussos-Stout cut-offs result in an inflated error rate, new cut-offs will be determined from the simulation data.

To answer this research question, first the SIBTEST cut-off of $|\beta| < .059$ as suggested by Roussos and Stout (1996) for the SIBTEST procedure (Gotzmann, 2002) was applied to investigate the Type I error of TestGraf DIF detection method. The results can be seen in Table 13. As it was in Tables 3 and 4, the confidence interval strategy with the *p*-value of .006 was used. In Table 13 one can see that the test is operating appropriately for a sample combination of 500/500. However, for smaller sample sizes the criterion operates at best inconsistently (sample size combinations of 200/100, 200/50, 100/100) being quite inflated for most item parameter combinations and at worst always inflated for the smaller sample size combinations of 100/50, 50/50, 50/25, and 25/25. Clearly, a new cut-off is needed for cases of less than 500 per group.

By investigating the empirical sampling distribution of β, the simulation design in this study allowed me to compute new cut-off values for the sample sizes wherein the Roussos-Stout criterion resulted in inflated error rates. As a matter of completeness, the cut-off for the sample size combination of 500/500 will also be provided. To obtain the new cut-offs, the percentiles of β values of each item across sample size combinations were computed at 90%, 95%, and 99% for significance levels of .10, .05, and .01, respectively. These indices answered the fifth research question addressed in the present study and can be seen in Table 14 and Tables 15 to 20. Table 14 provides the criterion irrespective of the item characteristics, whereas Tables 15 to 20 provide the criterion for the variety of item characteristics in the present simulation study.

For example (Table 14), a researcher interested in investigating the gender DIF for their test conducts the nonparametric IRT analysis with TestGraf and computes the β for each of the items. Imagine further that this researcher has 100 male and 100 female examinees; she/he would compare each obtained beta to the cut-off of .0421 for a nominal alpha of .05. Any item with a β

greater than .0421 would mean that that item has been flagged as DIF. This allows the researcher to use the same cut-off value irrespective of the item's discrimination and difficulty.

Table 13

Type I error of TestGraf DIF detection using the Roussos-Stout criterion across sample size combinations and item parameters

	Item Discrimination Level (*a*)					
	Low			High		
	Item Difficulty Level (*b*)					
	Low	Medium	High	Low	Medium	High
N_1 /N_2	Item35	Item37	Item39	Item36	Item38	Item40
500/500	.01	.02	0	.03	0	0
200/100	.11	**.14**	.02	**.18**	**.17**	.08
200/50	**.28**	**.33**	**.15**	**.26**	**.21**	.09
100/100	**.17**	**.22**	.09	**.18**	**.17**	.10
100/50	**.24**	**.31**	**.15**	**.25**	**.22**	**.18**
50/50	**.29**	**.18**	**.14**	**.28**	**.22**	**.21**
50/25	**.39**	**.34**	**.20**	**.27**	**.31**	**.28**
25/25	**.37**	**.34**	**.17**	**.25**	**.33**	**.31**

Note: Bold denotes an inflated Type I error

Table 14

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels irrespective of the item characteristics

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0113	.0161	.0374
200/100	.0249	.0373	.0415
200/50	.0460	.0540	.0568
100/100	.0308	.0421	.0690
100/50	.0421	.0579	.0741
50/50	.0399	.0455	.0626
50/25	.0633	.0869	.1371
25/25	.0770	.0890	.1154

Table 15

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a low discrimination level (a = 0.50) and low difficulty level (b = –1.00)

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0369	.0440	.0639
200/100	.0687	.0917	.1519
200/50	.1039	.1100	.2006
100/100	.0839	.0998	.1179
100/50	.1140	.1406	.2560
50/50	.1213	.1386	.2353
50/25	.1680	.2657	.2940
25/25	.1999	.2314	.3774

Table 16

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a high discrimination level (a = 1.00) and low difficulty level (b = –1.00)

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0390	.0519	.0869
200/100	.0893	.0990	.1658
200/50	.1216	.1453	.1769
100/100	.0829	.0968	.1558
100/50	.1070	.1622	.2548
50/50	.1256	.1598	.2000
50/25	.1180	.1640	.1820
25/25	.1432	.1888	.4562

Table 17

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a low discrimination level (a = 0.50) and medium difficulty level (b = 0.00)

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0329	.0380	.0759
200/100	.0720	.0890	.1904
200/50	.1110	.1319	.1768
100/100	.0910	.1010	.1500
100/50	.1295	.1557	.2660
50/50	.1160	.1487	.2867
50/25	.1617	.2187	.3053
25/25	.1667	.2262	.2739

Table 18

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a high discrimination level (a = 1.00) and medium difficulty level (b = 0.00)

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0329	.0360	.0579
200/100	.0798	.0890	.1308
200/50	.1022	.1240	.2919
100/100	.0808	.1122	.1735
100/50	.1093	.1458	.1709
50/50	.1040	.1187	.2019
50/25	.1180	.1460	.2080
25/25	.1712	.2244	.3174

Table 19

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a low discrimination level (a = 0.50) and high difficulty level (b = 1.00)

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0249	.0290	.0410
200/100	.0428	.0520	.0780
200/50	.0754	.0926	.1516
100/100	.0565	.0710	.1059
100/50	.0906	.1120	.1350
50/50	.0638	.0954	.1490
50/25	.1029	.1309	.1700
25/25	.0937	.1756	.2296

Table 20

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels for an item that has a high discrimination level (a = 1.00) and high difficulty level (b = 1.00)

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0288	.0358	.0480
200/100	.0500	.0726	.1089
200/50	.0573	.0719	.1040
100/100	.0606	.0679	.1277
100/50	.0973	.1326	.1856
50/50	.0940	.1427	.1910
50/25	.0999	.1413	.1680
25/25	.1497	.1738	.3431

Chapter V

DISCUSSION

The purpose of this dissertation is to investigate psychometric methods for DIF detection in moderate-to-small-scale testing. In particular, the focus is on the test evaluation stage of testing wherein one may focus on differential item functioning (DIF) as a method to detect potentially biased items. In setting the stage for the investigation, I defined moderate-to-small-scale testing contexts as those involving 500 or fewer examinees. The literature-based survey I reported in the beginning of this dissertation shows that psychological studies and surveys involve far fewer participants than achievement testing research. The survey also highlights the observation that simulation research on psychometric methods tends to focus on large-scale testing contexts. Both findings empirically confirm what are fairly widely held beliefs about achievement testing and simulation research.

The survey findings therefore set the stage by demonstrating the need for more research into moderate-to-small-scale methods that can be used in educational psychology studies, surveys, pilot studies or classroom testing. The sort of classroom testing I have in mind is the same sort envisioned by Ramsay in the development of his TestGraf non-parametric item response theory model and software: tests used in classes the size of those found in post-secondary institutions sometimes include as few as 25 students per comparison group, to as many as 500 per group. This is in a stark contrast to the numbers found in large-scale testing involving thousands of examinees.

Doing psychometric research on classroom tests may appear out of step with contemporary thinking in classroom assessment, which de-emphasizes the psychometric properties of classroom tests. However, if one is to take the goal of fair classroom assessment

seriously, then investigating test properties for potentially biased items is a natural practice. I am not suggesting that every instructor assess all items in all contexts for potential bias, but rather that a post-secondary instructor *may* want to maintain an item bank for their tests and hence may want to investigate matters such as potential language or cultural bias on, for example, their mid-term and final examinations. Furthermore, textbook writers and textbook test-bank developers may want to investigate these potential group biases in developing their item banks. As is common in measurement practice, once one has flagged an item as DIF it does not mean that the item is necessarily biased nor is it necessarily removed from the test. It is a common practice, for example, that an item be "put on ice" (i.e., set aside) for further investigation and revision rather than simply removed from a test bank. As Zumbo and Hubley (2003) remind us, in the end, DIF research is about asking questions of whether a test is fair and about providing empirical evidence of the validity of the inferences made from test scores.

In addition, DIF research is important in educational psychology research because DIF may threaten the correct interpretation of group differences on psychological scales. For example, when one finds cultural differences in response to items on a self-esteem scale, one needs to rule out the inferential validity threat that these differences may arise from different interpretations of items on the scale. In short, one needs to rule out that the differences are an artefact of the measurement process rather than actual group difference.

The above two examples of the use of DIF methods in moderate-to-small-scale testing highlight the fact that appropriately functioning DIF statistical techniques are needed for sample sizes less than or equal to 500 respondents or examinees. The psychometric literature offers three alternatives: two methods based on non-parametric item response theory and the Mantel-Haenszel (MH) test. The non-parametric item response modelling method is incorporated in the

software TestGraf uses either (i) a heuristic decision rule, or (ii) a formal hypothesis test of DIF based on the TestGraf results. The TestGraf methods, which were developed with moderate-to-small-scale testing in mind, are becoming more widely used for two reasons: (a) the methodology is graphically based and hence easier to use by individuals who do not have specialized psychometric knowledge of item response theory, and (b) the software is made freely available by its developer, Jim Ramsay of McGill University, via the internet. On the other hand, the MH test is an extension of widely used methods for the analysis of contingency tables and is available in commonly used software such as SPSS or SAS.

The purpose of the present study is to investigate the operating characteristics of these three DIF methods in the context of moderate-to-small-scale testing. The DIF methodologies (either the formal hypothesis test or the heuristic cut-off values established by Roussos and Stout) in TestGraf have not been investigated in moderate-to-small-scale testing. In fact, the formal hypothesis testing in detecting DIF is, to my knowledge, first introduced in this dissertation.

The findings in this dissertation lead to recommending the following practice for moderate-to-small scale testing:

- The MH DIF test maintains its Type I error rate and is recommended for use.
- TestGraf can be used as a method of DIF detection; however, neither the formal hypothesis test nor Roussos-Stout's heuristic criteria should be used because of the inflated error rates. Instead, a practical decision rule based on the cut-off indices for TestGraf's beta statistic reported in Tables 14 to 20 should be used. One should note that Tables 15 to 20 provide cut-offs for particular item characteristics, whereas Table 14 provides cut-off indices irrespective of item properties but specific to sample sizes. All

of these tables provide values for eight combinations of sample sizes so, until further research, the practitioner may need to interpolate cut-offs for intermediate sample sizes.

- The reader should note that computing the 90th, 95th, and 99th percentiles of the null sampling distribution of the beta statistic, for the various sample size and item characteristics in the simulation, resulted in the cut-off values. This is based on the statistical principle that creating a statistical hypothesis tests, in essence, requires one to divide the sampling distribution of a statistic into a region of acceptance and a region of rejection. In the case reported in this dissertation for the new cut-off values, because the statistic beta did not follow its proposed sampling distribution, this division was done empirically by computing the percentiles from the empirical sampling distribution.

Summary of the Findings

This dissertation has found that:

- The MH DIF test maintains its acceptable Type I error rate.
- The original version of TestGraf, i.e. the version prior December 20, 2002, produced the incorrect standard error that was too large and hence, as expected, resulted in Type I error rate that was very small nearly zero. Ironically, with the computing error corrected and a revised version of TestGraf released, the new TestGraf program produced Type I error rate that was too large.
- Regardless of the error in the standard error produced by TestGraf, the beta statistic is shown to be an unbiased and hence accurate estimate of DIF magnitude.

- Likewise, the shape of the sampling distribution of beta is normal, but the standard error of beta produced by TestGraf is often too small resulting in too large of a test statistic value and hence an inflated Type I error rate.
- Building on the beta statistic's unbiasedness, it has been shown that the Roussos-Stout criterion does not work for moderate-to-small-scale sample sizes. However, new criteria are provided based on the simulation data.

Future research needs to further explore the new cut-off criteria proposed in this dissertation with a variety of sample size combinations. Furthermore, this dissertation focused on Type I error rates, which are necessary for further study of statistical power in detecting DIF. The Type I error rate for DIF tests is not only important for statistical reasons (i.e., power is not formally defined unless the test protects the Type I error rate) but also useful for the type of decision being made with DIF tests. That is, as I described in the Introduction of this dissertation, a Type I error in a DIF analysis means that an analyst will conclude that there is no DIF when in fact DIF is present. This type of error has obviously serious implications for research and practice that the Type I error rate needs be controlled in DIF. Once one finds a DIF test that protects the Type I error rate, then the needs turns to statistical power. Future research should investigate the statistical power of the cut-off scores for beta presented in Tables 14 to 20 and compare this power to that of the MH DIF method for the same data.

REFERENCES

References marked with an asterisk (*) indicate studies included in the small systematic literature-based study.

*Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. Applied Measurement in Education, 14, 219-234.

Agresti, A. & Coull, B. A. (1998). Approximate is better than "Exact" for interval estimation of binomial proportions. The American Statistician, 52, 119-126.

*Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. Journal of Educational Measurement, 36, 185-198.

*Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of Scholastic Aptitude Tests. Journal of Educational Measurement, 35, 31-47.

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

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: John Hopkins University.

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.

*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.

*Ban, J. C., Hansen, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. Journal of Educational Measurement, 38, 191-212.

*Ban, J. C., Hansen, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. Journal of Educational Measurement, 39, 207-218.

*Bennett, R. E., Morley, M., Quardt, D., & Rock, D. A. (2000). Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. Applied Measurement in Education, 13, 303-322.

*Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. Journal of Educational Measurement, 38, 51-77.

*Biggs, J., Kember, D., & Leung, D. Y. P. (2001). The revised two-factor Study Process Questionnaire: R-SPQ-2F. British Journal of Educational Psychology, 71, 133-149.

*Bishop, N. S., & Frisbie, D. A. (1999). The effect of test item familiarization on achievement test scores. Applied Measurement in Education, 12, 327-341.

*Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. Applied Measurement in Education, 12, 383-407.

*Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement, 39, 331-348.

Bradley, J. V. (1978). Robustness? The British Journal of Mathematical and Statistical Psychology, 31, 144-152.

*Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and bookmark standard setting methods. Journal of Educational Measurement, 39, 253-263.

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

*Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. Applied Measurement in Education, 12, 151-165.

*Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. Journal of Educational Measurement, 37, 245-261.

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

*Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. Journal of Educational Measurement, 37, 163-178.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: CBS College Publishing.

*Cross, L. H. & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. Applied Measurement in Education, 12, 53-72.

*De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. Journal of Educational Measurement, 38, 213-234.

*De Champlain, A. & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. Applied Measurement in Education, 11, 231-253.

*De Mars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. Applied Measurement in Education, 11, 279-299.

*De Mars, C. E. (2000). Test stakes and item format interactions. Applied Measurement in Education, 13, 55-77.

*Engelhard, G., Jr., Anderson, D. W. (1998). A binomial trials model for examining the ratings of standard-setting judges. Applied Measurement in Education, 11, 209-230.

*Engelhard, G., Jr., Davis, M., & Hansche, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. Applied Measurement in Education, 12, 199-210.

*Enright, M. K., Rock, D. A., & Bennett, R. E. (1998). Improving measurement for graduate admissions. Journal of Educational Measurement, 35, 250-267.

*Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. Journal of Educational Measurement, 35, 137-154.

*Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. Applied Measurement in Education, 11, 159-177.

*Feltz, D. L., Chase, M. A., Moritz, S. E., & Sullivan, P. J. (1999). A conceptual model of coaching efficacy: Preliminary investigation and instrument development. Journal of Educational Psychology, 91, 765-776.

*Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item cluster in a large scale hands-on science performance assessment. Journal of Educational Measurement, 36, 119-140.

*Fitzpatrick, A. R., Lee, G., & Gao, F. (2001). Assessing the comparability of school scores across test forms that are not parallel. Applied Measurement in Education, 14, 285-306.

*Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. Applied Measurement in Education, 14, 31-57.

*Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. Applied Measurement in Education, 11, 195-208.

Fox, J. (2000). Nonparametric simple regression: Smoothing scatterplots. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-130. Thousand Oaks, CA: Sage.

*Fox, R. A., McManus, I.C., Winder, B. C. (2001). The shortened Study Process Questionnaire: An investigation of its structure and longitudinal stability using confirmatory factor analysis. British Journal of Educational Psychology, 71, 511-530.

*Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., Katzaroff, M. (2000). The importance of providing background information on the structure and scoring of performance assessments. Applied Measurement in Education, 13, 1-34.

*Gallagher, A., Bridgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. Journal of Educational Measurement, 39, 133-147.

*Gao, X., Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. Applied Measurement in Education, 14, 191-203.

*Garner, M., & Engelhard, D., Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. Applied Measurement in Education, 12, 29-51.

*Ghuman, P. A. S. (2000). Acculturation of South Asian adolescents in Australia. British Journal of Educational Psychology, 70, 305-316.

*Gierl, M. J., Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. Journal of Educational Measurement, 38, 164-187.

*Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline examinees. Applied Measurement in Education, 12, 13-28.

Gotzmann, A. J. (2002). The effect of large ability differences on Type I error and power rates using SIBTEST and TESTGRAF DIF detection procedures. Paper prepared at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002). (ERIC Document Reproduction Service No. ED 464 108)

*Hall, B. W., & Hewitt-Gervais, C. M. (2000). The application of student portfolios in primary-intermediate and self-contained-multiage team classroom environments: Implications for instruction, learning, and assessment. Applied Measurement in Education, 13, 209-228.

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

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

*Hanson, K., Brown, B., Levine, R., & Garcia, T. (2001). Should standard calculators be provided in testing situations? An investigation of performance and preference differences. Applied Measurement in Education, 14, 59-72.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. In A. Iran-Nejad & P. D. Pearson (Eds.), Review of Research in Education (Vol. 24, pp. 393-446). Washington, DC: The American Educational Research Association.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

*Hong, S., Roznowski. (2001). An investigation of the influence of internal test bias on regression slope. Applied Measurement in Education, 14, 351-368.

*Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. Journal of Educational Measurement, 35, 69-81.

*Janosz, M., Le Blanc, M., Boulerice, B., & Tremblay, R. E. (2000). Predicting different types of school dropouts: A typological approach with two longitudinal samples. Journal of Educational Psychology, 92, 171-190.

*Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied Measurement in Education, 14, 329-349.

*Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. Journal of Educational Measurement, 37, 39-57.

*King, N. J., Ollendick, T. H., Murphy, G. C., & Molloy, G. N. (1998). Utility of relaxation training with children in school settings: A plea for realistic goal setting and evaluation. British Journal of Educational Psychology, 68, 53-66.

*Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. Applied Measurement in Education, 11, 121-137.

*Lee, G. (2000). Estimating conditional standard errors of measurement for test composed of testlets. Applied Measurement in Education, 13, 161-180.

*Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. Journal of Educational Measurement, 39, 149-164.

*Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. Applied Measurement in Education, 12, 237-255.

*Lee, W. C., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. Journal of Educational Measurement, 37, 1-20.

*Lingbiao, G. & Watkins, D. (2001). Identifying and assessing the conceptions of teaching of secondary school physics teachers in China. British Journal of Educational Psychology, 71, 443-469.

*Ma, X. (2001). Stability of school academic performance across subject areas. Journal of Educational Measurement, 38, 1-18.

*McBee, M. M., & Barnes, L. L. B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. Applied Measurement in Education, 11, 179-194.

*Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. Journal of Educational Measurement, 39, 219-233.

*Mokhtari, K. & Reichard, C. A. (2002). Assessing students' metacognitive awareness of reading strategies. Journal of Educational Psychology, 94, 249-259.

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

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

*Nichols, P., & Kuehl, B. J. (1999). Prophesying the reliability of cognitively complex assessments. Applied Measurement in Education, 12, 73-94.

*O'Neil, H. F., Jr., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. Applied Measurement in Education, 11, 331-351.

*Oosterheert, I. E., Vermunt, J. D., & Denessen, E. (2002). Assessing orientations to learning to teach. British Journal of Educational Psychology, 72, 41-64.

*Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. Applied Measurement in Education, 11, 353-369.

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

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

*Pommerich, M., Nicewander, W. A., & Hanson B. A. (1999). Estimating average domain scores. Journal of Educational Measurement, 36, 199-216.

*Pomplun, M., & Omar, M. H. (2001). Do reading passages about war provide factorially invariant scores for men and women? Applied Measurement in Education, 14, 171-189.

*Pomplun, M., & Omar, M. H. (2001). The factorial invariance of a test of a reading comprehension across groups of limited English proficient students. Applied Measurement in Education, 14, 261-283.

*Pomplun, M., & Sundbye, N. (1999). Gender differences in constructed response reading items. Applied Measurement in Education, 12, 95-109.

*Ponsoda, V., Olea, J., Rodriguez, M. S., & Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. Applied Measurement in Education, 12, 167-184.

*Powers, D. E., & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. Applied Measurement in Education, 12, 257-279.

*Powers, D. E., & Fowles, M. E. (1998). Effects of preexamination disclosure of essay topics. Applied Measurement in Education, 11, 139-157.

*Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. Journal of Educational Measurement, 36, 93-118.

*Rae, G. & Hyland, P. (2001). Generalisability and classical test theory analyses of Koppitz's Scoring System for human figure drawings. British Journal of Educational Psychology, 71, 369-382.

Ramsay, J. O. (2000). TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data. Montreal, Quebec, Canada: McGill University.

*Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. Applied Measurement in Education, 14, 369-415.

*Rest, J. R., Narvaez, D., Thoma, S. J., & Bebeau, M. J. (1999). DIT2: Devising and testing a revised instrument of moral judgment. Journal of Educational Psychology, 91, 644-659.

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.

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

*Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. Journal of Educational Measurement, 35, 1-30.

*Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. Applied Measurement in Education, 14, 73-90.

*Sachs, J. & Gao, L. (2000). Item-level and subscale-level factoring of Biggs' Learning Process Questionnaire (LPQ) in a Mainland Chinese sample. British Journal of Educational Psychology, 70, 405-418.

*Schwarz, R. D. (1998). Trace lines for classification decisions. Applied Measurement in Education, 11, 311-330.

*Scrams, D. J., & McLeod, L. D. (2000). An expected response function approach to graphical differential item functioning. Journal of Educational Measurement, 37, 263-280.

*Sireci, S. G., & Berberoğln, G. (2000). Using bilingual respondents to evaluate translated-adapted items. Applied Measurement in Education, 13, 229-248.

*Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. Applied Measurement in Education, 12, 301-325.

*Smith, M., Duda, J., Allen, J., & Hall, H. (2002). Contemporary measures of approach and avoidance goal orientations: Similarities and differences. British Journal of Educational Psychology, 72, 155-190.

*Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. Journal of Educational Measurement, 39, 187-206.

*Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-Index for detecting answer copying. Journal of Educational Measurement, 39, 115-132.

*Stecher, B. M., Klein, S. P., Solamo-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., et al. (2000). The effects of content, format, and inquiry level on science performance assessment scores. Applied Measurement in Education, 13, 139-160.

*Stocking, M. L., Lawrence, I., Feigenbaum, M., Jirele, T., Lewis, C., & Van Essen, T. (2002). An empirical investigation of impact moderation in test construction. Journal of Educational Measurement, 39, 235-252.

*Stocking, M. L., Ward, W. C., & Potenza, M. T. (1998). Simulating the use of disclosed items in computerized adaptive testing. Journal of Educational Measurement, 35, 48-68.

*Stricker, L. J., Rock, D. A., & Bennett, R. E. (2001). Sex and ethnic-group differences on accomplishments measures. Applied Measurement in Education, 14, 205-218.

*Stricker, L. J., & Emmerich, W. (1999). Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. Journal of Educational Measurement, 36, 347-366.

*Stuart, M., Dixon, M., Masterson, J., & Quinlan, P. (1998). Learning to read at home and at school. British Journal of Educational Psychology, 68, 3-14.

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

*Sykes, R. C., & Yen, W. M. The scaling of mixed-item-format tests with the one-parameter and two-parameter partial credit models. Journal of Educational Measurement, 37, 221-244.

*Tsai, T. H., Hansen, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item non-equivalent groups design. Applied Measurement in Education, 14, 17-30.

*Van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. Applied Measurement in Education, 13, 35-53.

*Vispoel, W. P., Claigh, S. J., Bleiler, T., Hendrickson, A. B., & Ihrig, D. (2002). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? Journal of Educational Measurement, 39, 311-330.

*Vispoel, W. P., & Fast, E. E. F. (2000). Response biases and their relation to sex differences in multiple domains of self-concept. Applied Measurement in Education, 13, 79-97.

*Vispoel, W. P., Rocklin, T. R., Wang, T., & Bleiler, T. (1999). Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test? Journal of Educational Measurement, 36, 141-157.

*Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. Journal of Educational Measurement, 35, 155-167.

*Walker, C. M., Beretvas, S. N., & Ackerman, T. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. Applied Measurement in Education, 14, 3-16.

*Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. Journal of Educational Measurement, 37, 141-162.

*Waugh, R. F. (1999). Approaches to studying for students in higher education: A Rasch measurement model analysis. British Journal of Educational Psychology, 69, 63-79.

*Waugh, R. F. & Addison, P. A. (1998). A Rasch measurement model analysis of the revised approaches to studying inventory. British Journal of Educational Psychology, 68, 95-112.

*Webb, N. M., Schlackman,J., & Surged, B. (2000). The dependability and interchangeability of assessment methods in science. Applied Measurement in Education, 13, 277-301.

*Wightman, L. F. (1998). An examination of sex differences in LSAT scores from the perspective of social consequences. Applied Measurement in Education, 11, 255-277.

*Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. Journal of Educational Measurement, 35, 93-107.

*Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. Journal of Educational Measurement, 39, 1-37.

*Wise, S. L., Finney, S. J., Enders, C. K., Freeman, S. A., Severance, D. D. (1999). Examinee judgments on changes in item difficulty: Implications for item review in computerized adaptive testing. Applied Measurement in Education, 12, 185-198.

*Wolfe, E. W., & Gitomer, D. H. (2001). The influence of changes in assessment design on the psychometric quality of scores. Applied Measurement in Education, 14, 91-107.

*Yu, F., & Nandakumar, R. (2001). Poly-Detect for quantifying the degree of multidimensionality of item response data. Journal of Educational Measurement, 38, 99-120.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-formal populations with unequal variances: The case of reaction time. Canadian Journal of Experimental Psychology, 51, 139-150.

Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In Rocio Fernandez-Ballesteros (Ed.), Encyclopaedia of psychological assessment, (pp. 505-509). Sage Press, Thousand Oaks, CA.

*Zuriff, G. E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. Applied Measurement in Education, 13, 99-117.

APPENDIX A

Criteria for reporting sample sizes of the articles recorded the small systematic literature based study:

1. All articles that reported sample sizes found in Journal of Educational Psychology and British Journal of Educational Psychology were categorized as educational psychology research. Wherever applicable, rater or interviewer sample sizes were labelled as survey.
2. Articles in Journal of Educational Measurement and Applied Measurement in Education were classified as achievement testing, simulation testing, or survey. Sample sizes of raters, interviewers, judges, were grouped in to survey. Test scores used in a study if they were obtained from a testing database of real testing were classified as achievement testing unless otherwise it denoted as a simulation testing.
3. For articles with one study, if it used more than one-sample sizes, the smallest sample size used – either subjects or items - was chosen to be reported in Table A1.
4. For articles with multiple studies, sample size of each study was coded in Table A1 with the application of point 3 above as well.
5. Concerning achievement testing, number of complex items was not coded in Table 1, because when such item format involved, the relevant study would use only one or two items. This can lead to a wrong conclusion on the number of items involved.
6. DIF studies with various combinations of sample sizes in subjects were coded for the smallest sample size used in any groups.
7. The above criteria resulted in some studies to fall into more than one categories coded, for instance, Rae and Hyland (2001; coded as an educational psychology and a survey).

Table A1

Details of the systematic literature based survey by category of research

Achievement Tests		Simulation Studies		Psychological Test		Survey Studies	
Examinees	Items	Simulees	Items	Subjects	Items	Subjects	Items
163	40	1,000	36	1,070	18	293	80
2,080	70	12,000	30	75	10	900	20
1,392	48	7,300	25	791	30	26	25
448	100	100	25	174	56	16	40
1,401	125	500	30	18	24	12	53
1,351	20	1,000	40	450	40	50	9
8,454	62	100	200	85	-	186	30
717	31	250	55	229		2,000	75
1,485	35	250	36	475		3	60
2,595	40	1,700	30	443		36	40
198	49	1,500	56	169		200	-
4,637	20	300	24	369		16	
388	30	6,515	30	189		390	
633	150	3,000	60	200		15	
366	40	100	40	346		314	
6,883	34	500	20	61		44	
4,712	26	1,000	100	-		335	
2,115	-	200	43			662	
55		250	100			243	

Table A1 (continued)

Achievement Tests		Simulation Studies		Psychological Test		Survey Studies	
Examinees	Items	Simulees	Items	Subjects	Items	Subjects	Items
8,285		250	39			115	
1,217		250	30			728	
200		1,000	-			307	
1,493		1,000				4	
1,065		8,000				20	
2,002		1,000				187	
73		25*				77	
600		20,000*				39	
182		-				810	
57						782	
1,493						98	
629						300	
1,862						658	
107						26	
1,029						603	
2,919						352	
50*						335	
25,844*						-	
87,785*							

Table A1 (continued)

	Achievement Tests		Simulation Studies		Psychological Test		Survey Studies	
	Examinees	Items	Simulees	Items	Subjects	Items	Subjects	Items
M	4,657.0	54.1	2,558.9	51	321.5	29.7	310.6	43.2
M	1,808.2**		1,962.6**					
Mdn	1,284	40	1,000	37.5	214.5	27	193.5	40
Min	50	20	25	20	18	10	3	9
Max	87,785	150	20,000	153	1,070	56	2,000	80
25^{th}%	371.5	31	250	30	148	19.5	33.5	26.3
75^{th}%	2,106.3	62	1,600	55	444.8	37.5	361.5	58.3
N	38	17	27	21	16	6	36	10

Note: * Denotes studies that were excluded from *M* with **

APPENDIX B

An Initial Study of the Operating Characteristics of TestGraf's DIF Detection Statistic: Discovering an Error in TestGraf98's Calculation of the Standard Error of Beta[1]

As a consequence of sharing the results reported in this Appendix and getting clarification from Ramsay on the computation of the standard error of beta, a revised version of TestGraf98 was released by Ramsay on December 20, 2002. The purpose of this Appendix is to provide a record, for historical and archival reasons, of the Type I error rate and statistical power of TestGraf98's pre-December 20, 2002 version of the statistical test of DIF. Applied measurement researchers who used the pre-December 20th 2002 version of TestGraf98 for DIF will have had the Type I error rate and power reported in this Appendix B. Anyone using the Roussos-Stout criterion are not affected by the change in TestGraf versions.

That is, the version of TestGraf98 that was used in this Appendix is the version dated July 31, 2001. Upon completion of this simulation study and working through the results of the Type I error rates and statistical power patterns, I corresponded with Ramsay to follow up on the results I found. An outcome of this correspondence about this simulation study was that Ramsay found a computational error in TestGraf's computer code for the standard error of the Beta statistic and he released a revised version of TestGraf98 on December 20th 2002 that had the computation of the standard error of Beta corrected. The revised version of the software was used in the study reported in the main body of this dissertation.

This Appendix is written, as a freestanding study with more detail than would be typically found in an Appendix because some readers may be interested only in this study;

---

[1] Author note: I would like to thank Professor Jim Ramsay for his encouragement in this project and for so promptly providing the corrected version of TestGraf98.

therefore, some of the information is repetitive of the information found in the main body of this dissertation. It was decided to report this study in an Appendix because the conclusion was basically to point out an error in a widely used software package. Although I consider this an important finding I have not included it as part of the main body of my dissertation because I would like to build on the error I found in TestGraf to go back to initial purpose of studying the operating characteristics of TestGraf's DIF detection test. As Ramsay noted, "Actually, it's an excellent project to research the *SE* of the DIF index as TestGraf actually produces it, and it sounds like your work has revealed what needed to be revealed, namely that there's something wrong." (personal communication, December 20, 2002).

The study was originally designed to answer four broad research questions in the context of an educational testing and sample size. They were as follows:

VI. What was the Type I error rate of this statistical test of TestGraf?

VII. What was the power of this statistical test of TestGraf in detecting DIF?

VIII. Was the standard deviation of the sampling distribution of the beta TestGraf DIF statistic the same as the standard error computed in TestGraf?

B-1. Methodology

The methodology in the study by Muñiz, Hambleton, and Xing (2001) is used by many DIF researchers. Although the present study adopted the sample size combinations that were used in their study, there were differences in generating the reference and focal sample groups, for the investigation of Type I error.

It is important to make a methodological note about the significant difference between the present study and the previous study by Muñiz et al. (2001). As discussed in the literature

review, to compare two or more DIF detection methods in assessing their statistical properties – that are their Type I error rate and power of DIF detection -- the item parameters used will have to be obtained from the same item(s). In their study, however, Muñiz et al. computed Type I error rate from different items than the items used for assessing power of detecting DIF. By computing Type I error rate of different items to those for computing power, hence different item parameter values are used to compare the statistical properties of a DIF detection method in assessing the studied method. With this, they associated Type I error rate of items that were different than items for power of detecting DIF. Therefore, there may be confounding in interpreting the results for evaluation purpose.

For example, imagine we have a test of 40 items. Thirty-four items are used to compute Type I error rate, and the remaining six items are used to compute power of detecting DIF. Although this strategy is computationally efficient in terms of computing time, the principles of experimental design applied to simulation experiment would suggest that the experimental manipulation (Type I error and power) should be applied to the same item that the simulation is run for. For this to be an appropriate experiment, we should look at the same items that are simulated to be in either no-DIF or DIF condition.

Unlike Muñiz et al., the present study was carried out as a complete factorial design in which Type I error rate was explored from the same items that were used to assess power of detecting DIF. More detail of the methodology of this study is presented below.

Description of DIF detection procedure.

The Ramsay TestGraf non-parametric DIF detection method with the following factors was applied in detecting potential DIF items that were set for the purposes of this study. The

method displayed the findings graphically. The two population groups that were studied and compared to run the simulation were set so as to have equal ability and normal distribution with $M = 0$ and SD equal to 1.0. The manipulated factors were:

1. The sample sizes of the reference and focal groups. There were five categories of sample sizes used for the reference and focal groups: 50/50, 100/50, 200/50, 100/100, and 500/500.
2. The discrimination and difficulty parameters of the items in which DIF was found. Two item discrimination levels were used, low and high; each of which contained three different difficulty levels of easy, medium, and hard.
3. The amount of DIF in the studied items. The amount of DIF in the items will produce four categories of no-DIF, small DIF, medium DIF, and large DIF.

Variables in the study.

Sample sizes. As previously mentioned, the study used the same sample size combinations that were used in the small-scale study by Muñiz et al. (2001). Therefore, combinations of 50/50, 100/50 and 200/50 for reference and focal groups representatively were analyzed. Furthermore, considering a sample size of 100 per group was more realistic in practice, the study also looked at a combination of 100/100 in which each of the two groups contained 100 examinees. In order to mediate between the small-scale of this study and the large-scale testing context of over a thousand of examinees, the study also used a sample of 500 per group. As a result, the following combinations of sample size variables were analyzed:

Reference Group	Focal Group
50	50
100	50
200	50
100	100
500	500

Statistical characteristics of the studied test items with DIF. Previous research (Clauser, Mazor, & Hambleton, 1994) suggested that item statistics were related to power of DIF detection and therefore, to study this point further, DIF was simulated in easy, medium difficulty, and hard items each of which with either low or high discriminating power. If the *a* value indicates the proportional to the slope of the curve at the point of inflection and corresponding to the discriminating power, the *b* value for an item refers to the proficiency level (θ) at the point of inflection representing item difficulty parameter. When a *b* value scored 0-1 is reported on the ability scale, it indicates an examinee has a 60% chance of giving a correct answer to a five-option multiple-choice item. Higher *b* values indicate higher ability level required to get an item correct. In other words, the higher the *b* is, the harder the item will be.

Whereas for the *a*, values between zero and two are common when ability scores are scaled to an *M* of zero and an *SD* of one. The higher the *a* values, the steeper the item characteristic curves are and the more discriminating the items will be. Two *a* values of the studied DIF items were applied: 0.5 and 1.0. The third item characteristic involved in the three-parameter item response model is symbolized with a *c*. The *c* depicts the lower asymptote corresponding to the probability of a person with no ability at all in getting the item correct will provide right answer. Therefore, this parameter is called as the pseudo-guessing parameter. The *c*

value for the six DIF items was set at .17, which is fairly typical for a multiple-choice item with four or five options. In each test of 40 items, DIF was built into the last six items. Item statistics for the 40 items were presented in Table B1 and that for the six studied items in Table B2. It should be noted that all four reference population groups were generated from the test without DIF.

Amount of DIF. The amount of DIF was introduced through *b* value differences in the DIF items between the two test sets for generating the reference and focal population groups (See Table B2). Three levels of *b* value differences were applied: 0.5 for small DIF, 1.0 for medium DIF, and 1.5 for large DIF. Based on the amount of DIF as the variables that were investigated (dependent variables), there were four conditions generated: no-DIF, small DIF with 0.5 *b* value difference, medium DIF with 1.0 *b* value difference, and large DIF with 1.5 *b* value difference.

Table B1

Item statistics for the 40 items

Item	*a*	*b*	*c*
1	1.59	0.10	.19
2	0.60	-0.98	.20
3	0.75	-0.42	.06
4	1.08	0.43	.24
5	0.81	0.34	.32
6	0.66	-0.57	.38
7	0.81	-0.18	.20
8	0.43	-0.36	.30
9	0.94	0.45	.34
10	1.40	0.15	.07
11	0.98	-0.20	.18
12	1.28	-0.12	.23
13	1.18	0.18	.23
14	0.98	-0.63	.30
15	0.94	-0.14	.17
16	1.39	0.94	.43
17	0.78	0.25	.16
18	0.55	-0.82	.20
19	0.88	0.09	.27
20	1.10	0.14	.40

Item	*a*	*b*	*c*
21	1.23	-0.43	.10
22	0.73	1.13	.27
23	0.54	-1.91	.23
24	0.71	-0.43	.31
25	0.66	-0.67	.16
26	1.14	0.59	.18
27	1.12	0.29	.26
28	0.96	-0.26	.23
29	0.95	0.13	.15
30	1.38	0.66	.16
31	1.38	1.11	.16
32	0.42	-0.02	.20
33	1.04	-0.01	.30
34	0.73	0.10	.13
35	0.50	-1.00	.17
36	1.00	-1.00	.17
37	0.50	0.00	.17
38	1.00	0.00	.17
39	0.50	1.00	.17
40	1.00	1.00	.17

Table B2

Item statistics for the six DIF items

Item	Item difficulty level (*b*)				
	Reference group	Focal group			
		No-DIF	Small-DIF	Medium-DIF	Large-DIF
35	-1.00	-1.00	-0.50	0.00	0.50
36	-1.00	-1.00	-0.50	0.00	0.50
37	0.00	0.00	0.50	1.00	1.50
38	0.00	0.00	0.50	1.00	1.50
39	1.00	1.00	1.50	2.00	2.50
40	1.00	1.00	1.50	2.00	2.50

Data generation.

The present study used the three-parameter logistic item response-modeling program of MIRTGEN (Luecht, 1996) to generate data for the population of the simulation. It should be noted that although an item-response theory framework was used in generating the data, the study was not about item-response theory. The use of the three-parameter logistic item-response model was because the test items being studied contained differences in the difficulty and discrimination parameters, as well as taken into consideration the pseudo-guessing parameter. Therefore, the model provided a convenient way for generating examinee item-response data for analysis.

Two population groups for each of the four conditions in regard to the amount of DIF were generated. Each of both reference and focal populations contained 500,000 examinees. Each population corresponded to a test of 40 items. Item statistics for the first 34 non-DIF items were randomly chosen from the mathematics test of the 1999 TIMSS (Third International Mathematics and Science Study) grade eight, and item statistics for the last six items that were set as containing DIF, were adopted from the DIF items used in the study by Muñiz et al. (2001). These six items were the focus of this study.

For the focal population, the amount of DIF across the six DIF items, which are Items 35 to 40, was fixed similarly to the replicating study except for the non-DIF. Under no-DIF for assessing Type I error rate, the item characteristics were set the same for both reference and focal groups. Focal populations were considered being disadvantaged. Therefore, for the focal populations with DIF, the item difficulty parameter values were set higher or harder than the difficulty levels of the test for the reference population. There were small DIF with a *b* or item difficulty value difference equals 0.5, medium with a *b* or item difficulty value difference of 1.0, and large with a *b* or item difficulty value difference equals 1.5. On the other hand, in the reference populations, all of the item difficulty parameters remained the same across all categories, the no-DIF, small DIF, medium DIF, and large DIF groups. The item difficulty parameter values of the six DIF items can be seen in Table B2.

For each combination of variables, 100 data sets were generated and analysed. This made it possible to obtain Type I error rate and power of DIF detection in any combination of variables of the 100 replication pairs. Therefore, the data sets were reflective of actual test data, except for the insertion of the six DIF items.

Procedure.

The study was carried out following a modified procedure originally used in the study of Muñiz et al. (2001) as follows:

1. First, set a test of 40 items with the item parameters of the first 34 items were drawn from the TIMSS data, and those of the last six items were adopted from the studies of Muñiz et al. In each given data set, the amount of DIF across the six DIF items were set at no-DIF (*b*-value for both reference and focal populations was the same), small (*b*-value difference = 0.5), medium (*b*-value difference = 1.0), or large (*b*-value difference = 1.5). These item parameters are provided in Tables B1 and B2.
2. Then set a normal and equal ability distribution, $N(0,1)$.
3. Next, sample sizes for the reference and focal groups as described earlier were generated in five combinations. These five combinations contained 50/50, 100/50, 200/50, 100/100 and 500/500 examinees for reference and focal groups in pairs, respectively.
4. Generated population groups of 500,000 examinees/simulees for the reference and the focal groups of each DIF conditions.
5. Generated item-response data for the reference and focal groups for each condition, no-DIF, small DIF, medium DIF, and large DIF. (Steps 1 to 4 produced 20 different combinations of variables.) Each of these combinations was repeated 100 times.
6. Applied TestGraf procedure to the examinee item-response data and compared each pair of the reference and focal groups.
7. Repeated step 6 for 100 replications of each of the 20 combinations of variables in the study (5 sample combinations x 4 conditions of DIF), and determined Type I error rate and power of DIF detection.

Because of the graphical interface and the need to check each replication for convergence as well as the appropriate bin width, the seven steps of the simulation study were conducted without the use of batch computing – instead I conducted each DIF test and TestGraf run for each replication individually. Each of the 20 cells of the simulation design (with 100 replications per cell) required roughly 20 hours of computational time resulting in 400 hours of computation.

Compute DIF statistics.

When two or more groups were compared in the Compare step, TestGraf produced a file containing statistics summarizing the differences between the studied groups for each item. This file could be identified by its extension of .cmp (Ramsay, 2000, pp. 55-56). The file provided the amount of DIF or beta (β) and the standard error of beta for each item resulted in from comparing multiple groups.

From this output, the values of sampling distribution of β divided by the standard error of β was computed. The hypothesis was that the $\frac{\beta}{SE_\beta}$ distribution would look like a standard normal *Z*-distribution. Type I error rate and power of DIF detection at the nominal alpha of .05 ($Z = 1.96$) and .01 ($Z = 2.576$) were computed. Any significant result indicated the presence of DIF items, either as Type I error rate or false positive under no-DIF, or power of DIF detection under DIF condition.

Data analysis of simulation results.

For each of the 6 items studied and each of the sample size combinations, the Type I error rate of the DIF detection per item was obtained by dividing the number of times H_0

hypothesis is rejected by 100. The mean and standard deviation of the β values for the 100-replications of each sample size combination were computed.

The following analyses were done to answer each of the three research questions:

1. Tabulated Type I error rate.
2. Tabulated Power.
3. Computed empirical standard error of β and compared to TestGraf standard error.

The design was a 2 (levels of item discrimination) x 3 (levels of item difficulty) x 5 (sample size combinations).

B-2. Results and discussion

This section will present the results of the above simulation study to answer the four research questions. A brief conclusion will conclude this Appendix.

Research Question 1: What was the Type I error rate of this statistical test of TestGraf?

Table B3 presents Type I error rate resulted in from TestGraf DIF detection version pre-December 20, 2002. As can be seen, at nominal alpha of.05 the old TestGraf produced almost all zero Type I error rate, except for Item-36 and Item-37 with sample size combination of 50/50, Type I error rate was .01. The study found Type I error rate was all zero at nominal alpha of .01 across all sample size combinations.

Table B3

Type I error rate of Old-TestGraf at nominal $\alpha = .05$

	Item discrimination level (a)					
	Low			High		
	Item difficulty level (b)					
	Low	Medium	High	Low	Medium	High
N_1 / N_2	Item35	Item37	Item39	Item 36	Item38	Item40
50/50	0	.01	0	.01	0	0
100/50	0	0	0	0	0	0
200/50	0	0	0	0	0	0
100/100	0	0	0	0	0	0
500/500	0	0	0	0	0	0

Research Question 2: What was the power of this statistical test of TestGraf in detecting DIF?

The probability of rejecting the hypothesis for the old TestGraf DIF detection, i.e. the power of detecting DIF items given the small amount of DIF represented in the 0.50 difference between the item difficulty levels, was ranged from zero to .07 with a combination of sample sizes 50/50, 0 to .09 for 100/50, 0 to .09 for 200/50, 0 to .07 for 100/100, and 0 to .58 for 500/500. All of the above values were at nominal alpha of .05 as presented in Table B4. At nominal alpha of.01 the DIF detection power values range from zero to .01, 0 to .02, 0 to .02, all 0, and 0 to .16, for the designed sample size combination, respectively (Table B5).

Table B4

Probability of rejecting the hypothesis for the Old-TestGraf DIF detection in SMALL-DIF at $\alpha = .05$

	Item discrimination level (*a*)					
	Low			High		
	Item difficulty level (*b*)					
	Low	Medium	High	Low	Medium	High
N_1 / N_2	Item35	Item37	Item39	Item 36	Item38	Item40
50/50	.01	.07	.02	0	.01	0
100/50	.02	.09	.02	0	.01	0
200/50	.03	.09	.01	0	.02	0
100/100	.01	.07	.01	0	0	0
500/500	.28	.58	.07	.08	.12	0

Table B5

Probability of rejecting the Hypothesis for the Old-TestGraf DIF detection in SMALL-DIF at $\alpha = .01$

	Item discrimination level (*a*)					
	Low			High		
	Item difficulty level (*b*)					
	Low	Medium	High	Low	Medium	High
N_1 / N_2	Item35	Item37	Item39	Item 36	Item38	Item40
50/50	.01	0	0	0	.01	0
100/50	.01	.02	0	0	0	0
200/50	.01	.02	.01	0	0	0
100/100	0	0	0	0	0	0
500/500	0	.16	0	.01	0	0

When the difference of item difficulty level was set at 1.00, which was set to represent a medium amount of DIF, the probability of rejecting the hypothesis or power of DIF detection ranged from 0 to .38, .01 to .46, .07 to .49, 0 to .56, and .57 to 1.00, for the sample size combinations of 50/50, 100/50, 200/50, 100/100, and 500/500, respectively, at nominal alpha of .05. These findings can be seen in Table B6 below. At nominal alpha of .01 the power of detecting DIF items was found ranging from 0 to .08, 0 to .14, 0 to .13, 0 to .24, and .08 to 1.00, respectively. The values are presented in Table B7.

Table B6

Probability of rejecting the hypothesis for the Old-TestGraf DIF detection in MEDIUM-DIF at $\alpha = .05$

	Item discrimination level (*a*)					
	Low			High		
	Item difficulty level (*b*)					
	Low	Medium	High	Low	Medium	High
N_1 / N_2	Item35	Item37	Item39	Item 36	Item38	Item40
50/50	.15	.38	0	.06	.16	.01
100/50	.33	.46	.06	.12	.16	.01
200/50	.39	.49	.07	.16	.24	.07
100/100	.56	.52	.05	.22	.23	0
500/500	1.00	.96	.73	.99	.97	.57

Table B7

Probability of rejecting the hypothesis for the Old-TestGraf DIF detection in MEDIUM-DIF at $\alpha = .01$

	Item discrimination level (a)					
	Low			High		
	Item difficulty level (b)					
	Low	Medium	High	Low	Medium	High
N_1 / N_2	Item35	Item37	Item39	Item 36	Item38	Item40
50/50	.02	.08	0	.01	.06	0
100/50	.05	.14	.01	.03	.01	0
200/50	.13	.02	0	.04	.06	0
100/100	.16	.24	0	.01	.03	0
500/500	1.00	.96	.22	.86	.78	.08

Tables B8 and B9 present the probability of rejecting hypothesis when the difference in the item difficulty level was set at 1.50 to represent a large amount of DIF. At nominal alpha of .05 it was found that the probability ranged from .01 to .67 for the combination of sample sizes of 50/50. For the sample size combinations of 100/50, 200/50, 100/100, and 500/500, the power of DIF detection ranged from .06 to .93, .11 to .93, .01 to .98, and .94 to 1.00, respectively. At nominal alpha of.01 it was found that power of detecting large amount of DIF ranged from 0 to .35, 0 to .64, .01 to .71, .01 to .78, and .67 to 1.00, under each of the five sample conditions, respectively.

Table B8

Probability of rejecting the hypothesis for the Old-TestGraf DIF detection in LARGE-DIF at $\alpha = .05$

	Item discrimination level (a)					
	Low			High		
	Item difficulty level (b)					
	Low	Medium	High	Low	Medium	High
N_1 / N_2	Item35	Item37	Item39	Item 36	Item38	Item40
50/50	.67	.64	.05	.44	.36	.01
100/50	.79	.93	.06	.56	.63	.07
200/50	.87	.93	.11	.72	.69	.18
100/100	.98	.96	.01	.75	.73	.14
500/500	1.00	1.00	.94	1.00	1.00	.95

Table B9

Probability of rejecting the hypothesis for the Old-TestGraf DIF detection in LARGE-DIF at $\alpha = .01$

	Item discrimination level (*a*)					
	Low			High		
	Item difficulty level (*b*)					
	Low	Medium	High	Low	Medium	High
N_1 / N_2	Item35	Item37	Item39	Item 36	Item38	Item40
50/50	.34	.35	0	.17	.16	0
100/50	.43	.64	0	.25	.26	0
200/50	.59	.71	.01	.32	.32	.01
100/100	.78	.08	.01	.42	.39	.02
500/500	1.00	1.00	.67	1.00	1.00	.71

Research Question 3: Was the standard deviation of the sampling distribution of the beta TestGraf DIF statistic the same as the standard error computed in TestGraf?

Table B10 presents the *SD* values of DIF β and the averages of the TestGraf *SE* values over 100 replications for each of the six studied items across the five sample size combinations. While the two scores should refer to the same statistical characteristics of TestGraf DIF detection, the study found the TestGraf *SE* scores were substantially higher than their counterpart scores. As a consequence of the higher *SE* produced by TestGraf, Type I error rate and power of TestGraf in detecting DIF items resulted in lower values than it was expected.

Table B10

Standard deviations and standard errors of the Old-TestGraf DIF detection

Item	Over 100 replications	N_1 / N_2				
		50/50	100/50	200/50	100/100	500/500
35	*SD* of beta (Empirical *SE*)	0.088	0.079	0.083	0.060	0.027
	M of the TestGraf *SE*	0.171	0.158	0.152	0.128	0.067
36	*SD* of beta (Empirical *SE*)	0.095	0.081	0.076	0.060	0.030
	M of the TestGraf *SE*	0.171	0.157	0.149	0.126	0.064
37	*SD* of beta (Empirical *SE*)	0.097	0.085	0.082	0.061	0.026
	M of the TestGraf *SE*	0.175	0.161	0.154	0.129	0.066
38	*SD* of beta (Empirical *SE*)	0.097	0.081	0.080	0.064	0.026
	M of the TestGraf *SE*	0.168	0.153	0.146	0.123	0.060
39	*SD* of beta (Empirical *SE*)	0.062	0.063	0.060	0.046	0.019
	M of the TestGraf *SE*	0.147	0.130	0.120	0.109	0.057
40	*SD* of beta (Empirical *SE*)	0.071	0.075	0.057	0.051	0.021
	M of the TestGraf *SE*	0.146	0.132	0.124	0.111	0.057

Note: *N* = sample size, *M* = mean, *SD* = standard deviation, *SE* = standard error

Summary.

Type I error rate was nearly all zero at nominal alpha of .05 and all zero at nominal alpha of .01 across all sample size combinations. Although it is possible with 100 replications, all zero of Type I error rate was inappropriate. But carried it on with the power of TestGraf DIF detection, the average of the standard error over 100 replications was far off than the empirical

standard error. Test statistics resulted in from TestGraf are too liberal. As expected as sample size goes up, the power goes up as well. The operating characteristics of TestGraf DIF detection suggested that the TestGraf DIF detection method was a very conservative test. Furthermore, the old-version of TestGraf DIF detection prior to December 20, 2002 produced larger *SE* values in averages than the empirical *SE* of the sampling distribution, i.e. *SD* of DIF β values.

B-3. Conclusion

Anyone who has computed a hypothesis test from TestGraf prior to December 20, 2002 will have Type I error rate that was very low, and the power that was very low as well. The former is not a problem unless it comes with the latter. It should be noted that Type I error rate was inflated everywhere in the main body of this dissertation.