IMPACT OF DIFFERENTIAL ITEM FUNCTIONING ON STATISTICAL CONCLUSIONS

by

Zhen Li

B.Ed. Northeast Normal University, 1997
MA. Northeast Normal University, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUAIREMENTS FOR THE DEFREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Measurement, Evaluation & Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

November 2009

ABSTRACT

Differential item functioning (DIF), sometimes called item bias, has been widely studied in educational and psychological measurement; however, to date, research has focused on the definitions of, and the methods for, detecting DIF. It is well accepted that the presence of DIF may degrade the validity of a test. There is relatively little known, however, about the impact of DIF on later statistical decisions when one uses the observed test scores in data analyses and corresponding statistical hypothesis tests. This dissertation investigated the impact of DIF on later statistical decisions based on the observed total test (or scale) score. Very little is known in the literature about the impact of DIF on the Type I error rate and effect size of, for instance, the independent samples t-test on the observed total test scores. Five studies were conducted: studies one to three investigated the impact of unidirectional DIF (i.e., DIF amplification) on the Type I error rate and effect size of the independent samples t-test; studies four and five investigated the DIF cancellation effects on the Type I error rate and effect size of the independent samples t-test. The Type I error rate and effect size were defined in terms of latent population means rather than observed sample means. The results showed that the amplification and cancellation effects among uniform DIF items did transfer to the test level. Both the Type I error rate and effect size were inflated. The degree of inflation depends on the number of DIF items, magnitude of DIF, sample sizes, and interactions among these factors. These findings highlight the importance of screening DIF before conducting any further statistical analysis. It offers advice to practicing researchers about when and how much the presence of DIF will affect their statistical conclusions based on the total observed test scores.

Keywords: differential item functioning, DIF, impact, Type I error, effect size, t-test

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This dissertation could not have been done without my research supervisor Dr. Bruno Zumbo who not only guided me through my academic program but also encouraged and supported me in many ways. Without his guidance and support I would not have been able to go this far. I also gratefully thank my research committee members, Dr. Anita Hubley and Dr. Kimberly Schonert-Reichl, for their thoughtful comments on this Dissertation. In addition, I would like to thank my parents, my brother, my sister in-law, and my niece for their unconditional support through my life.

DEDICATION

To my parents

CO-AUTHORSHIP STATEMENT

The version of chapter two has been accepted for publication. Li, Z., & Zumbo, B.D. (in press). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica*.

This manuscript was co-written with my research supervisor Dr. Bruno Zumbo. My contributions to the dissertation are as follows:

- research design
- literature review
- data collection (simulation)
- data analyses
- writing and manuscript preparation

CHAPTER ONE: INTRODUCTION TO THE THEME OF THE DISSERTATON

General Introduction to the Research Problem

Differential item functioning (DIF) has been a focus of a great deal of attention in the educational and psychological measurement research literature. However, most of the previous research has primarily focused on the definitions of, and the methods for detecting DIF. Very few researchers have paid attention to the impact of DIF. Especially, no research has been found studying the effect of DIF on later statistical conclusions when one uses the observed test scores in data analyses and the properties of the corresponding statistical hypothesis tests. This is an important issue, however, because (1) educational and psychological researchers commonly conduct statistical analyses to answer their research questions based on observed test scores; (2) it is well accepted that DIF is pervasive and DIF items may be missed because DIF analyses have not been conducted, or DIF is present, unbeknownst to the researcher, as an artifact of DIF detection being a statistical decision method (e.g., making Type II error). Without knowing if DIF is present in the test, the statistical conclusions drawn based on the analyses could be a false statement because the effect found in the study could just be an artificial effect of DIF items rather than true effect. For example, when a researcher is investigating whether there are gender differences on a language proficiency test, the gender differences found in the observed score could just reflect the gender based DIF rather than true difference in performance between boys and girls. Therefore, the purposes of this dissertation are to (1) confirm that DIF does affect statistical conclusions based on observed test scores, and (2) investigate how DIF items affect the subsequent statistical results and conclusions, particularly, the Type I error rate and effect size of hypothesis tests from observed score test data. It should be noted that the term “test” is used hereafter in this dissertation to represent different measurement instruments (e.g., test, questionnaire, measure, scale) in different research areas.

To answer the question of how much DIF effects the eventual statistical conclusions two testing situations are particular of interest in this dissertation: (a) several DIF items consistently favor one group, and (b) some of the DIF items favor one group and some favor the other group. The first situation represents what I refer to as DIF amplification effects whereas the second situation represents DIF cancellation effects.

Five computer simulation studies were conducted. Studies one and two investigated the amplification effects of DIF on the Type I error rate and effect size of the independent samples t-test. The third simulation study investigated the impact of varying item parameter values on the Type I error rate of the subsequent t-test of the observed score means to confirm the generalizability of the results of study one. Studies four and five focused on the impact of DIF cancellation effects on the Type I error rate and effect size.

The population test data were simulated using item response theory (IRT) with varying degrees of DIF – that is, number of items exhibiting DIF and the magnitude of DIF. The observed (number correct) total scores were then computed for each simulated test taker and then these observed number correct scores were subjected to a t-test to test for the equality of sample means. Throughout this research I focused on the (Type I) error rates and effect sizes of the t-test under the null hypothesis of equal means.

Structure of the Dissertation

This dissertation includes three chapters: Chapter 1 (a) introduces the research questions and describes the structure of the dissertation, (b) provides the relevant statistical theories and introduces the related psychometric and statistical terms used in the following chapters, and (c) reviews the related literature with an eye toward setting the context for the simulation studies. Chapter 2 reports on the studies of the impact of DIF on the statistical properties of the independent samples t-test. This chapter includes five inter-related simulation studies as stated in the above section. Chapter 3, as a concluding chapter, generally discusses the study results as a whole and highlights for the reader the contributions and limitations of this dissertation to the research literature. In addition, this chapter discusses the implications of this dissertation for research and practice. This dissertation is written in a manuscript-based format (UBC, Faculty of Graduate Studies guideline, 2009, http://www.grad.ubc.ca/students/thesis/index.asp?menu=002,002,000,000) for the sake of gaining writing experience in a format used by researchers in education and psychology.

Methodological Review And Research Agenda

In order for the reader to clearly and sufficiently understand the content of this dissertation, this section provides information about the theories and terminologies relevant to IRT, DIF, simulation design, and hypothesis testing that are involved in this dissertation.

Item Response Theory (IRT) and Differential Item Functioning (DIF)

Introduction to IRT

Because of the properties of item and person parameters invariance (item and person parameter independence), item response models have been widely used to solve large scale assessment problems. In IRT, it is assumed that an examinee has an underlying ‘ability’, denoted θ, which determines his or her probability of giving a correct response to an item (Hambleton, Swaminathan, & Rogers, 1991). Conversely, the probability of an examinee endorsing an item can be predicted by his/her ability level conditioning on item properties (e.g., item difficulty and item discrimination). To model this relationship, IRT employs a nonlinear monotonically increasing function known as the item response function (IRF) or item characteristic curve (ICC) which relates ability to item performance. The ICC monotonically increases as θ increases, which indicates that, as the level of the ability increases, the probability of success on an item increases. For binary items, the probability of a correct response to an item *i* given ability (θ) is usually denoted as P_{i} (θ) whose values, because it is a probability value, range between 0 and 1. The ability (representing the construct of interest in a particular test) of the examinee is usually denoted by $\theta \in (-\infty, +\infty)$. The ability θ is conventionally considered a latent variable (trait, e.g., mathematics ability) whose values cannot be directly observed and must be inferred or estimated from the observed item responses of examinees (Baker & Kim, 2004). As θ cannot be measured directly, its scale is usually arbitrary defined on a z-score scale (μ=0 and σ=1) for its simplicity and usually ranges from -4 to +4. When only one construct of interest is measured by the test items, the test is said to be unidimensional.

One or more parameters are employed to describe an item in different models. The most popular unidimensional IRT model for binary response data is Birnbaum's (1968) 3-parameter (3PL) model

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp[-1.7a_i(\theta-b_i)]} \quad i = 1, 2, ..., \mathrm{k} \tag{1}$$

where

P_i (θ) is the probability of success on an item for an examinee with ability θ,

b_i represents item difficulty parameter (location parameter),

a_i represents item discrimination parameter (slope parameter),

c_i represents guessing parameter (lower bound of ICC),

k is the number of items in the test, and

1.7 is the scaling constant to adjust the difference between a logistic model and a normal ogive model.

The item difficulty parameter (b_i parameter) is measured on the same scale as ability, with values theoretically ranging from -4 to +4. In practice, the item difficulty parameters are usually found to range from -2.5 to +2.5. The item difficulty parameter is also called the location parameter, which indicates the position of ICC in relation to the ability scale -- which may be located approximately along the θ scale at the point at which $P_i(\theta)$ is $(1+c_i)/2$. The larger the value of b_i, the harder the item and the greater the ability required to answer the item correct at 50% chance. In essence, IRT is a threshold model, with b_i serving the role of a threshold.

The a_i parameter represents the discrimination ability of an item. The a_i parameter is not on the same scale as θ and its values range usually from 0 to 2. The a_i parameter is proportional to the slope of the ICC at the point b_i on the ability scale. Therefore, the discrimination parameter reflects the steepness of the ICC and measures the ability of an item to distinguish an examinee with high ability from those with low ability. The larger the value of $a_{i\,,}$ is the steeper the slope and the more capable an item is in distinguishing the examinees with high ability from those of low ability. An item's discrimination is best at about $b_{i.}$

The c_i parameter, which is usually called the *guessing* parameter, represents the probability that an examinee with extremely low ability will succeed on an item. The c_i is also called the lower asymptote because it represents the nonzero asymptote of the ICC. Its values fall between 0 and 1, with higher values indicating higher probability that an examinee with extremely low ability can succeed in an item by guessing.

When the assumption of no guessing behavior holds in the data (i.e., guessing is negligible), the 3-parameter logistic model (3PL) may be reduced to a 2-parameter (2PL) model. In this case, the lower asymptote of the ICC is zero.

$$P_i(\theta) = \frac{1}{1 + \exp[-1.7a_i(\theta - b_i)]} \qquad (2)$$

When the assumptions of all items are equally discriminating and the guessing is negligible, the 3-parameter model can be reduced to a 1-parameter (1PL) model (includes only b-parameter) or Rasch model (includes b-parameter and a fixed a-parameter).

$$P_i(\theta) = \frac{1}{1 + \exp[-1.7\bar{a}(\theta - b_i)]} \qquad (3)$$

For more information about IRT, please refer to Baker and Kim (2004), Hambleton, Swaminathan and Rogers (1991), Lord (1980), and van der Linden and Hambleton (1997).

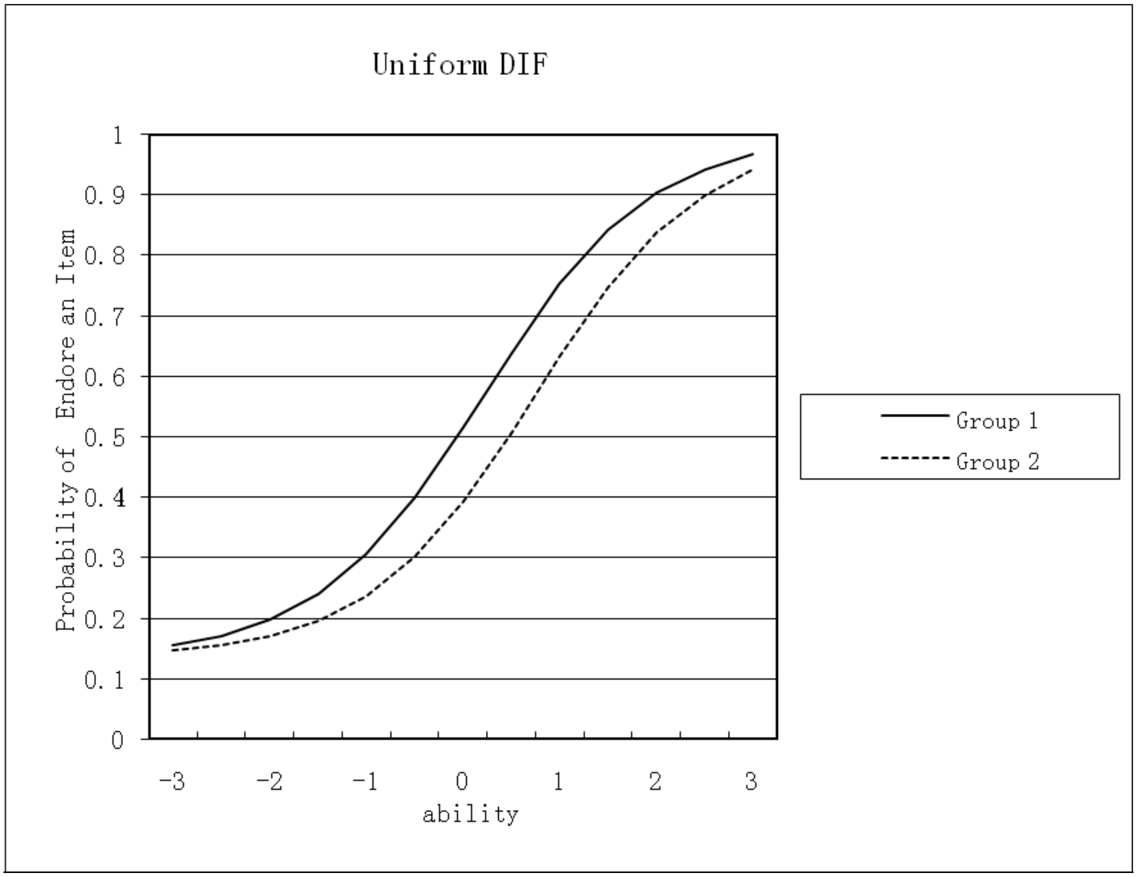
DIF and IRT

Differential item functioning (DIF) occurs when examinees with equal ability but from different groups (e.g., gender, age, cultural) have different probability endorsing an item. The purpose of DIF investigations is to find out whether the observed group mean differences are real (i.e., an accurate reflection of true performance differences owing to factors measured by the test) or artifactual (caused by the measuring process itself) (Camilli & Shepard, 1993). Since the 1970s, many DIF detection methods have been invented to identify DIF items and analyze the sources of DIF. The most popular DIF methods are the Mantel-Haenszel (MH), methods based on item response theory (IRT), and logistic regression. For more information about DIF and DIF investigation, please refer to Camilli (2006), Camilli and Shepard (1993), Holland and Wainer (1993), and Zumbo (2007).

The introduction of IRT into educational and psychological measurement provided a new way to define and detect DIF. As an ICC is uniquely characterized by an examines' ability (θ), a, b, and c parameters, DIF and the magnitude of DIF can be mathematically studied and measured by investigating differences in item a-, b-, and c- parameters---and the combination of them (Lord, 1977, 1980). Theoretically, examinees with the same level of ability should always have identical ICCs when answering the same item. However, non-identical ICCs can be obtained when poorly constructed items are used or when items are used in non-standard conditions. These conditions artificially change the item properties which, in turn, change the ICCs accordingly. Under this condition, DIF occurs. Non-negligible difference between the ICCs of two groups indicates the presence of DIF (Hambleton, Swaminathan, & Rogers, 1991).

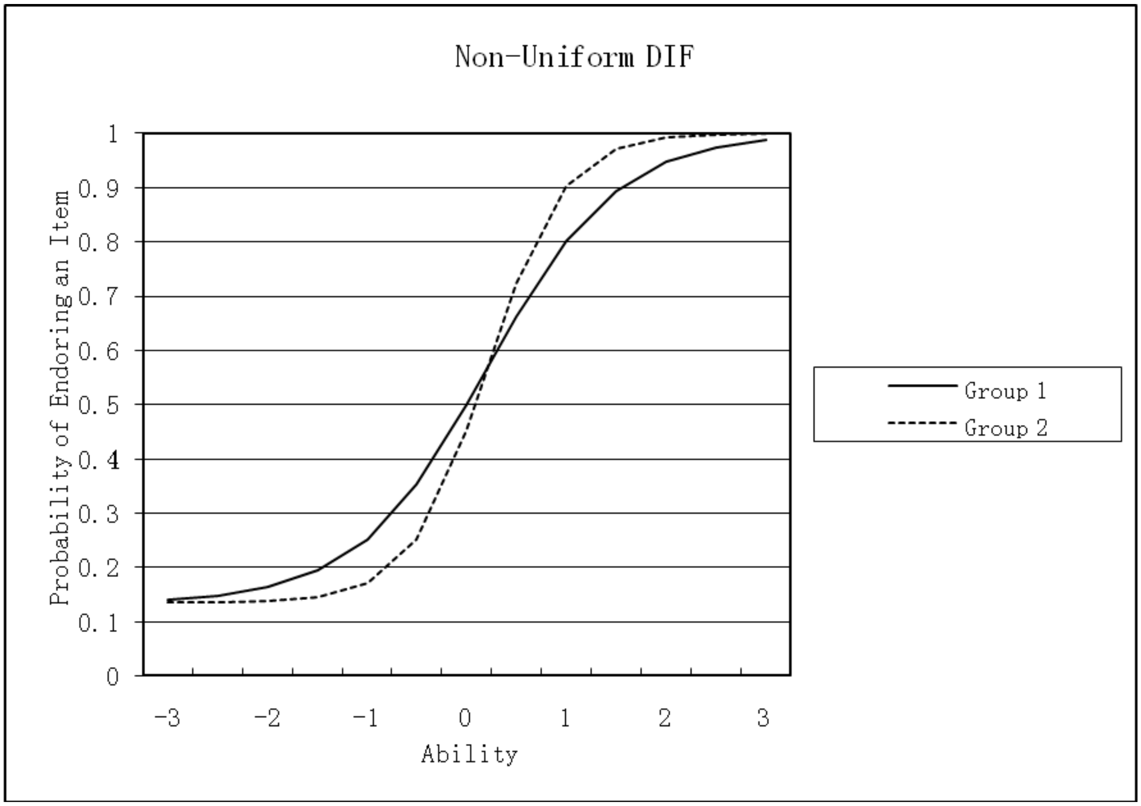
DIF can be classified into two broad categories: uniform DIF and nonuniform DIF (Mellenbergh, 1982). Figure 1 demonstrates uniform DIF and Figure 2 demonstrates non-uniform DIF. Uniform DIF occurs when there is no interaction between ability level and group membership; using IRT terminology, this happens when the ICCs for two groups are different but do not cross (when a- and c-parameters are equal but the b-parameter of the ICC of one group shifting to the right or left). In this case, one group has a relative advantage over the entire ability range. Nonuniform DIF occurs when there is an interaction between ability level and group membership; in IRT, this happens when the ICCs for two groups are both different and cross or are, at least, non-parallel (when a-parameters are different across groups). In this case, DIF favors one group at certain ability level but favors another group at other level. Under this condition, for a certain group, the DIF effect of favoring and against can balance or cancel each other to some degree.

Figure 1. Uniform DIF.



Under uniform DIF, the ICCs of two groups do not across (there is no interaction between proficiency level and group membership, Mellenbergh, 1982).

Figure 2. Non-Uniform DIF.



Under non-uniform DIF, ICCs from two groups across at certain point of ability scale (there is interaction between proficiency level and group membership, Mellenbergh, 1982).

DIF is caused by systematic errors introduced by the item and examinee parameters that are estimated (Hambleton, Swaminathan, & Rogers, 1991). This can also be conceptualized as multidimensionality of the test. No test is perfectly unidimensional in a strict sense, even though, it may be well designed. There are always secondary dimensions (abilities) involved to endorse an item. These untargeted secondary dimension(s) bring systematic error(s) into data which distort group differences on the target dimension. Therefore, if one applies a unidimensional ICC to item scores that require two different abilities, the knowingly or unknowingly untargeted abilities may nevertheless affect the ICC of a group (Camilli & Shepard, 1993). In other words, DIF occurs when examinees from different group having the same level of ability on the dimension(s) of interest but different level of abilities on the secondary untargeted dimension(s). However, multidimensionality does not necessarily cause DIF. It is the unequal ability on the secondary dimensions that causes the presence of DIF. It should be noted that these secondary dimensions are also sometimes called ‘minor secondary dimensions’ in the research literature.

IRT methods of DIF analysis

Based on IRT, DIF may be investigated by comparing the item response functions (IRT or ICC) of different groups by examining the differences in item parameters between groups employing a statistical test (Lord, 1977, 1980) and by measuring the areas between the ICCs of the groups (e.g., Raju, 1988). In the first method, no DIF occurs if there are no significant differences in the item parameters. In the second method, no DIF occurs if the area between two ICCs is zero where DIF is present if the area between two ICCs is nonzero. Raju (1988) proposed general equations for computing the exact area between the ICCs for the one-, two-, and three-parameter IRT models. As the area measures are effect size measures which can be used to quantify the magnitude of DIF for the study items -- equation (1) for uniform DIF and equation (2) for non-uniform DIF -- Raju's area method was often used in simulation studies as a criterion to quantify DIF (e.g., French & Maller, 2007; Narayanon & Swaminathan, 1996; Rogers & Swaminathan, 1993). For uniform DIF, wherein the *a*-parameters are equal between study groups, the Raju's area is expressed as the absolute difference between the *b*-values for the two groups (as shown in equation 4).

$$Area = (1-c)\left|b_2 - b_1\right| \tag{4}$$

$$Area = (1-c)\left|\frac{2(a_2 - a_1)}{Da_1a_2}\ln 2\right| \tag{5}$$

Educational Testing Service, ETS, also established DIF quantification rules to quantify the magnitude of DIF for use with the Mantel-Haenszel (MH) DIF detection method (Zwick & Ercikan, 1989). This rule was also used to quantify uniform DIF in an IRT context (Zumbo, 2003):

1. A-level DIF, negligible DIF,
2. B-level DIF, moderate amount of uniform DIF: a difference in difficulty -- i.e., b-parameter -- of 0.50, and
3. C-level DIF, large amount of uniform DIF: a difference in difficulty -- i.e., b-parameter -- of 1.0.

Simulation Design

A computer simulation is needed in this dissertation because the purpose of this study is to find out the robustness of the independent samples t-test in the presence of DIF without knowing the true parameter values. To identify under what condition and to what extent the Type I error rate is inflated, the study focuses on the effect of DIF. In other words, I did not consider other conditions in which the Type I error rate is inflated by violating the assumptions of the t-test. The factors that were usually considered in simulation studies investigating DIF were: sample properties (e.g., sample size, sample distribution), test properties (e.g., test length, item characteristic, such as dichotomous and polytomous), DIF related factors: type of DIF (uniform and non-uniform), proportion of DIF items, and magnitude of DIF (e.g., Swaminathan & Rogers, 1990; Zwick, Donoghue, & Grima, 1993; Zwick, Thayer, & Wingersky, 1994). The current study is the first study of its kind therefore the simulation situation was intended to be idealized to set up the baseline for future follow-up studies. That is, for example, I did not examine the effect of test-length, mixed magnitude of DIF (e.g., some items have large DIF while others have moderate or small DIF), non-normal ability distributions, non-uniform DIF, and polytomous DIF.

Sample Properties

Sample size.

Sample size directly affects the power of the statistical methods. Larger sample size results in greater statistical power to the statistical model to detect small effects. Increasing sample size will increase the accuracy of the parameter estimation, especially in IRT parameter estimation which often used in DIF analysis. However, it is not always true that the larger the sample size the better. With large enough sample size, the statistical models will have enough power to detect even a trivial, practically non-significant effect (Cohen, 1988). Thus, seeking an appropriate sample size that gives enough power while maintaining the generalibility of results, is one of the main issues in statistical and psychometric research. In the DIF simulation literature, study sample sizes are usually generated ranging from 20 to 3000 per group (e.g., Rogers & Swaminathan, 1993; Shealy & Stout, 1993; Swaminathan & Rogers, 1990; Zwick, Thayer, & Wingersky, 1994, Zumbo, 2003; Zumbo & Koh, 2005). In addition, the effect of imbalanced samples (e.g., unequal sample size, missing cells) is also a main issue in both statistical significance tests and psychometric research (e.g., 900:100 for the two groups, Zwick, Thayer, & Wingersky, 1994) as it is well known that the violation of the assumption of equal variances when sample sizes are unequal inflate the Type I error rate substantially (e.g., Zimmerman, 2004).

Sampling distribution.

Sampling distribution (e.g., distribution shape, mean and variance differences) also has an effect on hypothesis testing and DIF detection techniques because, as mentioned above, unequal variance may inflate the Type I error rate. These simulation conditions are usually achieved by varying the mean and standard deviation (and skewness and kurtosis, etc.) of one study group. The distribution of the reference group is usually simulated as a standard normal *N* (0, 1). The focal group distribution was usually simulated aberrant from normal with varying degree, such as, *N*(0.5, l), *N*(0, l) or *N*(-1, l), wherein *N* denotes the normal distribution, with the first number in the parentheses being the mean and the second the variance of that normal distribution.

Test Properties

Generally, longer tests have greater reliability and provide more accurate and stable parameter estimates in IRT applications (Baker, 2001; Fitzpatrick & Yen, 2001; Hambleton & Cook, 1983; Hambleton, 1989; Lord, 1980; Swaminathan & Gifford, 1983). In addition, for DIF detection methods whose matching criterion is based on observed total score, test length will also have an effect (Swaminathan & Rogers, 1990; Zwick, Donoghue, & Grima, 1993). In simulation studies, test length was usually generated from 2 to 80 items. Both binary and polytemous items were widely studied.

DIF item generation

DIF item generation was usually conducted using a two or three parameter IRT model with varying item parameters to reflect uniform and non-uniform DIF of different magnitudes (Hambleton & Rovinelli, 1986; Swaminathan & Rogers, 1990). Uniform DIF is usually simulated by shifting the b-parameter, with the a-parameter remaining the same while non-uniform DIF is usually simulated by changing a-parameter values between groups while keeping the b-parameter values the same across groups (Swaminathan & Rogers, 1990). The magnitude of DIF is usually quantified by either Raju's area (1988), as shown in equation (4) for uniform DIF and equation (5) for non-uniform DIF, or ETS's DIF classification rules (Zwick & Ercikan, 1989). Using Raju's area, the studied magnitudes of DIF vary from trivial non-zero DIF to the magnitude measured by Raju's area of 0.2, 0.4, 0.6, and 0.8 (Rogers & Swaminathan, 1993). In simulating the dichotomous item parameters, a-parameters were randomly sampled from a uniform distribution on the interval [.40 - 2.00] (or with natural log transformation); b-parameters were usually drawn from uniform distribution on the interval [-2.5 - +2.5]; and c-parameters were set to a fixed value (Hambleton & Swaminathan, 1985; Hambleton & Rovinelli, 1986; Rogers & Swaminathan, 1993; Zwick, Thayer, & Wingersky, 1994). The proportion of DIF items usually range from 0 % to 20% (French & Maller, 2007; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Zwick, Donoghue, & Grima, 1993).

Generation of item response data

The widely used conventional method for generating item responses was used in the present studies (Hambleton & Rovinelli, 1986). That is, to generate the item response data, the item and θ parameters were substituted into the appropriate IRT model to obtain a probability of correct response on each item for each examinee. These probabilities were converted to item responses by comparing each probability with a random number from a uniform distribution on the interval [0, 1]. The item was scored correct (i.e., 1) if the probability exceeded the random number and 0 otherwise (French & Maller, 2007; Hambleton & Rovinelli, 1986; Rogers & Swaminathan, 1993). Fifty to100 replications were usually used in simulation studies (French & Maller, 2007; Rogers & Swaminathan, 1993; Shealy & Stout, 1993; Zwick, Donoghue, & Grima, 1993).

Hypothesis Testing

Studies on hypothesis tests usually focused on investigating the robustness of these statistical models under various violations of statistical assumptions such as normality, equal variances and presence of outliers. However, the performance of hypothesis tests of equal observed score means when DIF is present has not been studied. This dissertation takes the independent samples t-test as a case in point to examine the impact of DIF on the subsequent conclusions of hypothesis testing. Particularly, this dissertation investigates the properties of the independent samples t-test under two situations: amplification DIF effect and cancellation DIF effect on the Type I error rate and effect size under different DIF and sample size combinations.

To help readers better contextualize the content of this study by putting everyone on the same foundational footing, this section provides a brief description of Type I error rate, statistical power, and effect size. More detailed information can be found in any introductory statistical textbooks (e.g., Gravetter & Wallnau, 2002).

Hypothesis testing is the most widely applied statistical technique for analyzing data in a variety of different disciplines. By testing competing alternative hypothesis, hypothesis testing draws inference about population parameters from sample data. This goal is achieved by comparing the results under certain situations of interest (e.g., intervention) with the results under the null situation (completely random, happen by chance, random fluctuation expected by chance) through assessing statistical significance of the findings. Hypothesis testing consists of a series of tests established for different purposes and situations. Among them, the independent samples t-test is probably the simplest, most useful, and most widely used one.

The Independent Samples t-test applies to the situations where the mean difference between populations is of interest. The independent samples test requires three assumptions:

1. The observations within each sample must be independent.
2. The two populations from which the samples are selected must be normal.
3. The two populations from which the samples are selected must have equal variances.

Type I and Type II Error

Two types of errors can be made while applying hypothesis testing: Type I error and Type II error. Type I error happens when one rejects a null hypothesis when it is actually true. In other words, one makes a Type I error when he/she concludes that there is effect/difference when in fact there is no effect. This can happen when one uses a sample which does not represent a population or simply by chance. On the contrary, Type II error happens when one fails to reject a false null hypothesis which means he/she fails to capture the real effect/differences. This can happen when the sample size is not large enough to capture the small true effect, or again, simply by chance.

Statistical power

Statistical power evaluates the ability of a statistical test correctly reject a false null hypothesis. The higher the power the more capable of a statistical test is in capturing the true effect. Formally, one can state that power equals one minus the probability of a Type II error.

Effect size

Though useful, a statistical test is not able to find the size of the true effect. Thus, effect size is usually required to be reported. The simplest and most often used methods for measuring effect size is Cohen’s d (Cohen, 1988), as indicated in equation (6).

$$\text{Cohen's d}=\frac{\text{mean difference}}{\text{standard deviation}}\ . \qquad (6)$$

After studying a large number of empirical results in published studies, the magnitude of Effect Size quantified by Cohen (1988) is as follows:

Small effect: $0 < d < 0.2$;

Medium effect: $0.2 < d < 0.8$; and

Large effect: $d > 0.8$.

Robustness

Maronna, Maritin, and Yohai (2006) indicated all statistical methods rely explicitly or implicitly on a number of assumptions and it is generally understood that the resulting formal models are simplifications of reality and that their validity is, at best, approximate. It often happens in practice that these assumptions (e.g., normality, independence, and equal variances) hold only approximately in the data. Bradley (1978) called for more attention to the robustness of tests and gave the quantification standard/criterion on robustness. Bradley indicated that: When one or more of a test’s assumptions are violated and the null hypothesis is true, the true probability of a Type I error tends to differ from the nominal significance level α. He defined three different levels of Type I error rate robustness which he terms as fairly stringent, moderate, and very liberal. Thus, using his equation, for a nominal Type I error rate of .05, the fairly stringent criterion for robustness requires that the empirical Type I error rate lie between .045 and .055, for the moderate criterion between .040 and .060, and finally for the very liberal criterion between .025 and .075.

Literature Review of Closest Related Studies

After an extensive literature search (e.g., EBSCO between 1950 to 2008, and google scholar), no studies were found directly investigating the impact of DIF on subsequent statistical tests based on the observed test score data. There are, however, a series of tangentially related studies that investigated the impact of DIF on subsequent factor analysis results, IRT scoring and the test characteristic curves, as well as predictive validity and overall measurement quality (e.g., Drasgow, 1987; Pae, 2004; Pae & Park, 2006; Roznowski, 1987; Roznowski & Reith, 1999; Shealy & Stout, 1991, 1993; Zumbo, 2003; Zumbo & Koh, 2005).

The Impact of DIF on Subsequent Factor Analysis Results

Several studies (Pae & Park, 2006; Zumbo, 2003; Zumbo & Koh, 2005) investigated the relationship between DIF and subsequent analyses of the factorial invariance of the test data. Zumbo’s (2003) simulation study found that the item level DIF did not manifest itself in the test-level results using the conventional widely recommended factor analysis techniques of that time: factor analyses of Pearson correlation matrices. Zumbo and Koh (2005) concluded from their simulation studies that item level bias can be present even when a confirmatory factor analysis (based on Pearson correlation matrices) of the two different groups reveals an equivalent factorial structure of rating scale items. Based on the analysis of real test data (non-simulated data), Pae and Park (2006) found flagging item level DIF does impact subsequent test level factorial invariance. Over and above the research methods, the disparity between Zumbo (2003) and Pae and Park (2006) studies are mainly on the model fit criterion chosen at test level analysis. Pae and Park (2006) used chi-square difference tests as well as practical fit indices as the criterion whereas Zumbo (2003) used practical fit indices to assess the factorial invariance.

Impact of DIF on Subsequent IRT Scoring

Several studies (e.g., Rupp & Zumbo, 2003 (mathematical results), 2006; Stahl, Bergstrom, & Shneyderman, 2002 (simulation findings); Takala & Kaftandjieva, 2000 (empirical findings); Witt, Stahl, Bergstrom, & Muckle, 2003 (simulation findings)) investigated the influence of DIF (or item parameter drift, IPD) on IRT examinee parameter (theta score) estimates. These studies found that, despite the presence of DIF items (or IPD) in favor of different groups, there was only a small effect on ability parameter estimates.

Impact of DIF on Test Characteristic Curves

Drasgow (1987), Pae (2004), Rupp and Zumbo (2003, 2006) (mathematical results), and Wells, Subkoviak, and Serlin (2002) (simulation results) investigated whether identifiable differences in item characteristic curves (ICCs) between groups of comparable proficiencies translate to test characteristic curves (TCCs). Results of these studies indicate that DIF/IPD has little effect on TCCs. The results also shown that, when the direction of DIF was inconsistent across items (non-unidirectional DIF, DIF items were found favoring both groups), their effects on total test scores tended to cancel out.

Impact of DIF Amplification or Cancellation

Drasgow (1987), Humphreys (1970), Maller (2001), Nandakumar (1993), Roznowski (1987), Roznowski and Reith (1999), and Shealy and Stout (1991, 1993) mathematically and empirically studied amplification and cancellation effects on item level bias. Nandakumar (1993) and Shealy and Stout (1991, 1993) investigated amplification and cancellation effects among nontrait variances in a multidimensional framework. These studies highlighted that if several DIF items act in the same direction, their effects, even though trivial individually, can have a cumulative effect at the test score level to produce differential test functioning (DTF).

Alternatively, items acting in different directions or influenced by different nuisance determinants can cancel out each other and result in little or no DIF at the test level.

Impact of DIF on predictive validity and measurement quality

Roznowski (1987) and Roznowski and Reith (1999) investigated the effect of DIF items on the predictive validity and measurement quality of the test empirically. These studies showed that, somewhat paradoxically, retaining DIF items in the test can increase predictive validity and measurement quality if the sources of DIF are diverse and multiply determined. Humphreys (1970) has argued against homogenous tests and recommended retaining diverse nontrait determinants in test items in order to avoid influence of any particular nontrait variance overpowering the influence of dominant trait variance of interest at the test score level.

These previous related studies explored the influences of DIF (or, equivalently, IPD) on test level statistics in different psychometric contexts. The information conveyed from above literature review about the influences of retaining DIF in the test is mixed. However, given that DIF does, in some contexts play a role, one may conjecture from these previous studies that if DIF items are multidirectional, or multiply determined, the overall DIF effect at test level tends to cancel out resulting in little effect at test level. On the other hand, if DIF items function in the same direction, they may have a significant effect at the test level. However, none of these studies directly study the question of impact of DIF on subsequent statistical conclusions and hence I will now turn my attention to the first study of its type to directly address this research question. Particularly, I adopt Shealy and Stout (1991, 1993)'s perspective defining DIF in two ways: unidirectional and multidirectional. Thus, this dissertation will investigate DIF amplification effect and cancellation effect on the performance of hypothesis testing.

References

Baker, F. B. (2001). *The basics of item response theory (2nd ed).* http://echo.edres.org:8080/irt/baker/final.pdf.

Baker, F. B., & Kim, S-H. (2004). *Item response theory: Parameter estimation techniques (2nd ed.).* New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397 - 424). Reading, MA: Addison-Wesley.

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

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

Camilli, G., & Shepard, L. A. (1993). *Methods for identifying biased test items, vol. (4).* Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for behavioral sciences (2nd ed.)*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72,* 19-29.

Faculty of Graduate Studies, University of British Columbia (UBC). (2009). Master's and Doctoral Thesis Preparation and Submission. http://www.grad.ubc.ca/students/thesis/index.asp?menu=000,000,000,000

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.

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

Gravetter, F. J. & Wallnau, L. B. (2002). *Statistics for the behavioral science*. Belmont, CA: Wadsworth.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (pp. 147–200). New York: American Council on Education and Macmillan.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 31–49). New York: Academic.

Hambleton, R. K. & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring professor Paul Horst* (pp. 23-32). Seattle: University of Washington.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets & Zeitlinger.

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

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

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods.* The Atrium, England: John Wiley & Sons.

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

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

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

Pae, T.-I. (2004). DIF for examinees with different academic background. *Language Testing*, *21,* 53-73.

Pae, T.-I., & Park, G-P. (2006). Examining the relationship between differential item functioning and different test functioning. *Language Testing, 23,* 475-496.

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.

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

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72,* 480-483.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59,* 248-269.

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49,* 264-276.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*, 63-84.

Shealy, R., & Stout, W. (1991). *An item response theory model for test bias* (Office of Naval Research Tech. Rep. No. 4421-548). Urbana: University of Illinois, Department of Statistics.

Shealy, R., & Stout, W. (1993). An item response theory model for test and differential item functioning. In H. Wainer & P. Holland (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum.

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

Stahl, J. A., Bergstrom, B. A., & Shneyderman, O. (2002, April). *Impact of item drift on test-taker measurement.* A paper presented at American Educational Research Association annual meeting, New Orleans, LA.

Swaminathan, J., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 13–30). New York: Academic.

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

Takala, S. & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17,* 323-340.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response Theory.* New York: Springer.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77 - 87.

Witt, E. A., Stahl, J. A. Bergstrom, B. A. & Muckle, T. (2003, April). *Impact of item drift with non-normal distributions*. A paper presented at American Educational Research Association annual meeting, Chicago, IL.

Zimmerman, D. W. (2004). Inflation of Type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation test. *Psicol ógica, 25,* 103-113,

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20,* 136-147.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Zumbo, B. D., & Koh, H. K. (2005). Manifestation of differences in item-level characteristics in scale-level measurement invariance tests on multi-group confirmatory factor analysis. *Journal of Modern Applied Statistical Methods, 4*, 275-285.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55-66.

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

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

CHAPTER TWO[1] : IMPACT OF DIFFERENTIAL ITEM FUNCTIONING ON SUBSEQUENT STATISTICAL CONCLUSIONS BASED ON OBSERVED TEST SCORE DATA

Differential item functioning (DIF) has been widely studied in educational and psychological measurement. For recent reviews, please see Camilli (2006) and Zumbo (2007). Previous research has primarily focused on the definitions of, and the methods for, detecting DIF. It is well accepted that the presence of DIF might degrade the validity of a test. There is relatively little known, however, about the impact of DIF on later statistical decisions when one uses the observed test scores in data analyses and corresponding statistical hypothesis tests. For example, let us imagine that a researcher is investigating whether there are gender differences on a language proficiency test. What is the impact of gender-based differential item functioning on the eventual statistical decision of whether the group means (male versus female) of the observed scores on the language proficiency test are equal? There is remarkably little research to help one directly answer this question.

DIF may be present in a test because either (a) DIF analyses have not been used as part of the item analyses, (b) it is there unbeknownst to the researcher, as an artifact of DIF detection being a statistical decision method, and hence true DIF items may be missed, or (c) as a result of the practice of leaving items flagged as DIF in a test. Irrespective of how the DIF items got there,

[1] A version of this chapter has been published. Li, Z., & Zumbo, B.D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30, 343-370.*

it is still unknown how such DIF items affect the subsequent statistical results and conclusions, particularly, the Type I error rate and effect size of hypothesis tests from observed score test data.

In order to directly answer this research question of the effect of DIF items on the eventual statistical conclusions from the test total scores, I conducted five interrelated simulation studies wherein I simulated population test data using item response theory (IRT) with varying degrees of DIF -- i.e., number of items exhibiting DIF and the magnitude of DIF. In order to answer the hypothetical researcher's research question, the observed (number correct) scores were then subjected to a t-test to test for the equality of sample means. Throughout this research I focus on the (Type I) error rates and effect sizes of the independent samples t-test under the null hypothesis of equal means.

It is important to note that the Type I error rate herein, in essence, was the probability of rejecting the null hypothesis when the latent means (rather than the observed test score means) were equal across groups. That is, using IRT one notes that an item response is a function of item parameters and examinee ability. By definition, when DIF items were retained in a test, these DIF items might result in differences in item responses of different group of examinees of comparable abilities. Accordingly, the research question more formally can be stated as: What is the probability of rejecting the null hypothesis of equal observed test score means when the latent means are equal but DIF is present in the test? Likewise the effect size reflects those settings in which the latent variable population means are also equal.

Based on tangentially related research that investigates the impact of DIF on person parameter estimates (i.e., the latent variable score) from IRT, scale scores, and predictive validity (e.g., Drasgow, 1987; Maller, 2001; Roznowski, 1987; Roznowski & Reith, 1999; Rupp & Zumbo, 2003, 2006; Shealy & Stout, 1991, 1993; Wells, Subkoviak, & Serlin, 2002) I predict that the Type I error rate and effect sizes will be inflated, however, the extent and under what conditions it will be inflated are unknown. To answer the question of how much DIF effects the eventual statistical conclusions I am interested in two testing situations: (a) several DIF items consistently favor one group, and hence, of course are against the other one; (b) some of the DIF items favor one group and some favor the other group. The first situation represents what I refer to as DIF amplification effects (which were the focuses of studies one, two, and three) whereas the second situation as DIF cancellation effects (which were the focuses of studies four and five). Of course, other test data situations may arise but given that this is the first study of its kind I wanted to address two fairly straightforward, but of course plausible, testing situations.

Five inter-related computer simulation studies were conducted. The first study focused on the amplification effects of DIF on the Type I error rate of the hypothesis test of equality of means of the observed test scores. The second simulation study focused on the amplification effects of DIF on the effect size. The third simulation study investigated the impact of varying the test item parameter values on the Type I error rate of the subsequent t-test of the observed score means. Note that in studies one and two the items used to generate DIF were sampled in a fixed manner. Influences of the different values of the item parameters on the Type I error rate were not considered. Therefore, study three was added to confirm the generalizability of the results of this study. Study four focused on the impact of DIF cancellation effects on Type I error rate, and finally study five focused on the impact of DIF cancellation effects on the effect size. In order to organize the findings and convey them in a clear manner, I organized the five simulation studies into two sections: section one being the amplification effects and section two being the cancellation effects. Each section will have a brief discussion section and then a general discussion will be reserved for the end.

I focused this research on the widely used two independent samples case of testing the equality of observed score group means; that is, the conventional (pooled variances) independent samples t-test. This scenario reflected the all too widely observed case wherein researchers investigate mean differences on their test data (a) without having first studied whether DIF exists, or (b) when one decides to retain the items even though DIF is found.

It is important to note that the DIF was aligned with the hypothesis test of mean differences itself (i.e., there were potential gender DIF items when one was investigating gender differences on the observed test scores). Without loss of generality to other assessment and measurement approaches (such as psychological or health measurement), I will use educational achievement testing and gender DIF as example to contextualize this research. Of course, the DIF could be due to test translation or adaptation, or any other situation that results in a lack of measurement invariance (Zumbo, 2007).

Section One: Impact of Differential Item Functioning (DIF) on Statistical Conclusions, I: Amplification Effect

Study One: Type I Error Rates

The purpose of this first simulation study was to document the effect of DIF items on the eventual Type I error rates of the t-test on the observed (number correct) total test score data. In this study I focused on the amplification effect of DIF item. That is the situation where DIF items favor a group consistently.

Methods

Simulation factors.

The simulation factors manipulated in this study were magnitude of DIF, number of DIF items, and sample size. There are three levels of magnitude of DIF -- small, moderate, and large as defined by Raju's 1988 area statistic of .4, .6., and .8, and four levels of number of DIF items (1, 4, 8, and 16 items out of 38 items in the test), and four levels of sample size (25, 50, 125, and 250 examinees per group). In addition, for comparison purposes, I investigated the no DIF condition as a baseline for the four sample sizes – it is expected, of course, that in this condition the observed Type I error rate would be at the nominal level. Therefore, for the Type I error rate simulation, the resultant simulation experiment was a 4x4x3 completely crossed factorial design. In addition, there are four no-DIF conditions (for the four sample sizes) resulting in a total of 52 cells in the simulation design.

I focused the investigation on binary items. The data were generated using item response theory (IRT). In order to examine the amplification effect of DIF items, I focused on unidirectional uniform DIF. Unidirectional DIF (Shealy & Stout, 1991, 1993) occurs when the DIF items are against the same group for all levels of ability (θ). Thus, in this study, DIF items were simulated consistently favoring the reference group. In addition, I adopted Zumbo's (2003) simulation design and therefore I did not vary the test length, and I used real item parameters based on the TOEFL test to generate the item data for a 38 item test.

The first factor of interest in this study was the magnitude of DIF. Theoretically, I expected larger magnitude of DIF would enlarge the differences in item responses between groups and hence the combined DIF effect across items might result in greater Type I error rate. Previous studies (French & Maller, 2007; Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993) also simulated DIF in this manner. Following these earlier studies, the uniform DIF items were simulated by shifting the b-parameter in the focal group to manipulate the area between two item response functions. In the situations wherein there was more than one DIF item, all the items in that situation had the same magnitude of DIF. That is, for the ease of interpretation, I did not investigate the effect of having a mixed magnitude of DIF – e.g., when one studies the effect of small DIF, for the situation in which more than one item had DIF, all of the items were simulated having, for instance, a .40 DIF (small DIF) effect. Likewise, when one studies the moderate DIF effect, all DIF items were simulated having a .60 DIF effect.

Similarly, I expected that the Type I error rate might be affected by the proportion of DIF items in the test. In the unidirectional uniform DIF case, the hypothesis was that the more DIF items were retained in the test, the larger differences would be resulted in observed response data across groups, then the more likely that the Type I error rate would be affected. Note that following Zumbo (2003) I did not vary the total of number of items in the test – I investigated 1, 4, 8, and 16 DIF items, out of a total 38 items.

Sample size was another factor that might affect the Type I error rate in terms of latent means as the larger the sample size the more likely one is to reject the null hypothesis. Sample size was set equal in both comparison groups.

Simulation procedures.

Following Zumbo (2003), the parameters for the 38 items for the Structure and Written Expression section of Test of English as a Foreign Language (TOEFL) were used to simulate the data in the various conditions. The means and standard deviations of item parameters in the reference and focal groups were presented in Table 1 and Table 2.

Table 1

Means (M) and Standard Deviations (SD) of Item Parameters in Reference Group

Item Parameter	M	SD	Minimum	Maximum
a-parameter	0.986	0.304	0.535	1.890
b-parameter	-0.133	0.861	-2.494	1.537
c-parameter	0.231	0.106	0.029	0.448

Table 2

Means (M) and Standard Deviations (SD) of b-parameters under Different Simulation Conditions in Focal Group

Number of DIF items	Small DIF (Raju's area = 0.4)		Moderate DIF (Raju's area = 0.6)		Large DIF (Raju's area = 0.8)	
	M	SD	M	SD	M	SD
1	-0.121	0.83	-0.115	0.816	-0.109	0.804
4	-0.081	0.875	-0.054	0.892	-0.028	0.915
8	-0.028	0.886	0.025	0.916	0.078	0.957
16	0.083	0.884	0.19	0.923	0.298	0.978

Examinee response data were generated using a three-parameter unidimensional logistic item response model (Birnbaum, 1968) as shown in equation (1),

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp[-1.7a_i(\theta-b_i)]}, \qquad (1)$$

where a_i, b_i, and c_i are the item discrimination, difficulty, and guessing parameters of item *i*, respectively. The latent variable is denoted as θ, whereas $P_i(\theta)$ denotes the probability of answering item *i* correctly with ability θ. Five thousand replications were conducted in each cell (i.e., study condition) of the simulation design; therefore, each Type I error rate value was based on 5000 replications of the simulation experiment.

Three steps were conducted to generate the item response data for Type I error rate studies.

Step #1: In the first step I generated the reference population data. In this step, the ability values, θ, for the reference group were generated from a standard normal distribution (*M* = 0, *SD* = 1). The probabilities, $P_i(\theta)$, were calculated using equation (1) and the values of a, b, c, and the generated θ. Then, uniformly distributed random numbers with interval [0, 1] were generated. To obtain the binary item response, the item response probabilities, $P_i(\theta)$, were converted to *1*s when the probability was larger than the corresponding random number, whereas the probabilities were converted to *0*s otherwise (Hambleton & Rovinelli, 1986). Next, the observed total test scores (number correct) were computed. And finally, samples with a particular sample size were randomly sampled from the reference population.

Step #2: In step two I generated the focal population data using exactly the same procedures except that some of the item parameter values were changed to reflect DIF on selected items and depending on the cell of the simulation design.

Step #3: In step three, the two generated populations were merged into one file and the independent sample t-tests were conducted, and the Type I error rates were computed as the number of rejections of the null hypothesis out of the 5000 replications. Our nominal significance level was 0.05 throughout this study. Therefore, empirically, the Type I error is defined as the proportion of times that a true null-hypothesis was falsely rejected at the 0.05 level.

Analysis of the Type I error rate simulation results.

I used the Bradley (1978) approach to documenting the inflation in Type I error rate. Bradley defined three different levels of Type I error rate robustness which he terms as fairly stringent, moderate, and very liberal. Thus, for a Type I error rate of .05, the fairly stringent criterion for robustness requires the empirical Type I error rate lie between .045 and .055. The moderate criterion requires the empirical Type I error rate lies between .040 and .060. And the very liberal criterion requires the empirical Type I error rate lies between .025 and .075. Please recall from the definition above that these proportions of rejected t-tests were the Type I error rates because the population means for the latent variable, θ, were equal.

In addition to Bradley's descriptive method, I also used regression modeling to investigate the effect of the simulation factors, treating the design and analysis as a type of response surface modeling (Zumbo & Harwell, 1999). The dependent variable is a proportion (i.e., the empirical Type I error rate based on 5000 replications), therefore, the logit transformation was applied (Cohen, Cohen, West, & Aiken, 2003, p. 240). The regression modeling was conducted in two steps. In the first step, a model was fit with main effects, and then an assessment was made whether the interactions statistically contributed to the model. In the second step, graphical methods were used to describe the main effects and/or interactions.

Results and Conclusions

Table 3 lists the simulation results and the description based on Bradley’s criteria. The Type I error rate for different sample sizes were computed for the no DIF conditions to establish baselines for comparisons with the conditions wherein different DIF conditions were manipulated. Under the no DIF condition, as shown in second column of Table 3, the Type I error rates range, as expected, from 0.052 to 0.053 for sample size from 25 to 250 per group (the column labeled ‘No DIF’ in Table 3). This also serves as a check on the simulation methodology.

Table 3

Type I Error Rates of t-test under Amplification Effect for Different Sample Sizes, Number of DIF, and Magnitude of DIF Combinations

	Magnitude of DIF												
	No DIF	Raju's area 0.4				Raju's area 0.6				Raju's area 0.8			
Number of DIF	0	1	4	8	16	1	4	8	16	1	4	8	16
N(per group)													
25	0.053	0.045	0.058	0.060↑	0.104↑↑	0.057	0.051	0.079↑↑	0.172↑↑	0.047	0.059	0.095↑↑	0.265↑↑
50	0.052	0.047	0.052	0.074↑	0.165↑↑	0.045	0.060↑	0.105↑↑	0.308↑↑	0.048	0.068↑	0.131↑↑	0.488↑↑
125	0.052	0.054	0.056	0.120↑↑	0.341↑↑	0.055	0.077↑↑	0.181↑↑	0.628↑↑	0.054	0.092↑↑	0.278↑↑	0.853↑↑
250	0.053	0.056	0.069↑	0.175↑↑	0.601↑↑	0.057	0.094↑↑	0.319↑↑	0.898↑↑	0.063↑	0.126↑↑	0.505↑↑	0.990↑↑

Note:

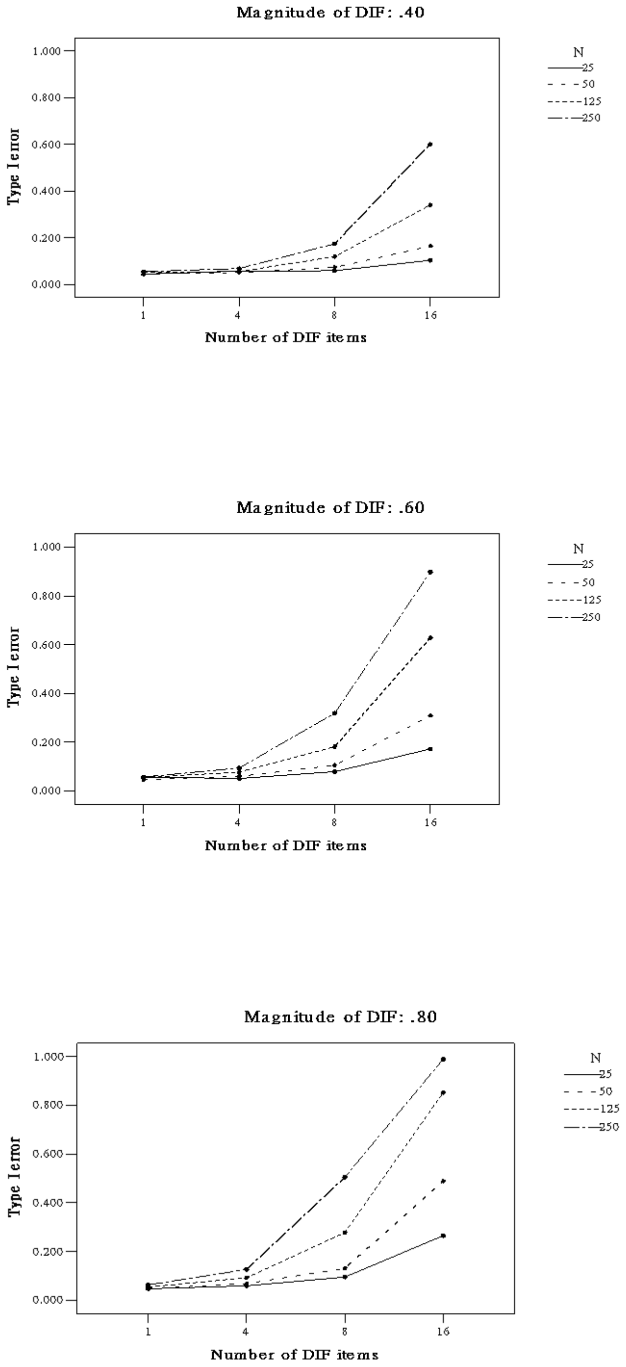
Type I error rate	Bradley (1978) criterion
$\alpha < 0.055$	Meet the stringent criterion
$0.055 \leq \alpha < 0.060$	Meet the moderate criterion
$0.060 \leq \alpha < 0.075$ ↑	Violates the moderate but meets the liberal criterion
$\alpha \geq 0.075$ ↑↑	Violates the liberal criterion, therefore inflated

Table 3 also displays the results of the Type I error rates for the case of the DIF amplification effect for the (a) different magnitudes of DIF, (b) number of DIF items, and (c) sample size combinations. Please recall that DIF items in this study were all simulated favoring the reference group. The far left column of Table 3 lists the sample sizes per group and next to it is the baseline Type I error rates. The remaining nine columns were divided into three magnitudes of DIF (i.e., Raju's area of .40, .60, and .80). Within each magnitude of DIF, the four columns represent the cases wherein there are 1, 4, 8, and 16 DIF items. For example, focusing on Raju's area of .40, with one uniform DIF retained in the test (column 2), the Type I error rates were between 0.45 and 0.56 for the sample sizes of 25 to 250 per group. In this situation, none of the Type I error rates were inflated for all studied sample sizes using Bradley's (1978) moderate criterion. As the sample size increased to 250, the Type I error rate inflated with large DIF compared against the moderate criterion. In terms of categorizing the resultant Type I error rates:

- When only one of the 38 items had DIF, the Type I error rate met the moderate criterion, except for Raju's area of .80 with 250 examinees per group wherein the Type I error rate only met the liberal degree of robustness.
- Irrespective of the magnitude of DIF and sample size, with 16 out of 38 of the items having DIF the Type I error rate was inflated. Likewise, for Raju's area of .60 and .80 with 8 out of 38 of the items having DIF the Type I error rate was inflated, and
- For 4 or 8 out of 38 items having DIF, the classification of the Type I error rates were dependent on the sample size and the magnitude of DIF – ranging from moderate to liberal inflated Type I error rates.

These classifications of DIF are informative in terms of deciding whether one should treat the Type I error rate as too large and hence invalidating the t-test of the hypothesis of equal population observed score means, but these classifications do not clearly provide a description of how the simulation factors, as main effects or interactions, effect the Type I error rate of the t-test. To address this latter question, I conducted the regression analysis of the simulation results treating each factor in the simulation (Number of DIF items, Magnitude of DIF, Sample size) and the interactions among them as explanatory variables in the multiple regression analysis (Zumbo & Harwell, 1999). In the first step of the modeling, the main effects were entered into the model with a resultant R-squared of 0.782 (F (3, 44) = 52.53, p < .0001), then the three two-way interactions were entered into the model for a resulting R-squared of 0.973 (R-squared change was statistically significant, F (3, 41) = 96.24, p < .0001), and finally the three-way interaction was entered into the model resulting in an eventual model R-squared of 0.985 (R-squared change was statistically significant, F (1,40) = 33.67, p < .0001). Please note that, because of the use of interaction terms, all of the explanatory variables were first centered before product terms were computed for the interactions. Clearly, the three-way interaction was statistically significant; therefore, Figure 3 was used to graphically interpret the three-way interaction. Figure 3 depicts the three-way interaction by plotting the two-way interaction of number of DIF items and sample size for each magnitude of DIF. For each magnitude of DIF, it can be clearly seen that the inflation of the Type I error rate increases as the number of DIF items and the sample size increase.

Figure 3. A plot of the effect of number of DIF items and sample size, for each magnitude of DIF, on Type I error rate.



Study Two: Effect Size

A second computer simulation study was designed to investigate the effect of DIF on the effect size of the independent samples t-test when DIF items were retained in the tests. Cohen's d is the appropriate effect size measure to use in the context of a t-test of independent means; d is defined as the difference between two means divided by the pooled standard deviation for those means. I computed this for both the observed total scores and latent variables; which allowed me to index the impact of DIF on the effect size. For the observed score d, the means and standard deviations were computed from the observed total test scores, whereas for the latent variable d the mean and standard deviations were computed from the latent variable scores.

As in study one, the observed score effect size was computed when the latent means (rather than the observed group test score means) were equal across groups. Therefore, the research question can be stated as: What is the effect size for the observed test score means when the latent means are equal but DIF is present in the test?

Methods

The simulation factors manipulated in this study, as well as the simulation methodology, were the same as those in study one except for one experimental factor, sample size. Like Zumbo (2003) I were not interested in the sample-to-sample variation in effect size estimates but instead focused on (the population analogue of) the bias in effect size. I manipulated number of DIF items and the magnitude of DIF. As in study one, there were four levels of number of DIF items (1, 4, 8, and 16 items out of 38 items in the test), and three levels of magnitude of DIF -- small, moderate, and large as defined by Raju’s area statistic of .4, .6, and .8 (Raju, 1988). In addition, I investigated the no DIF condition as a baseline for the four sample sizes for comparison purposes. This resulted in a 4x3 completely crossed factorial design and an additional no-DIF condition resulting in a total of 13 cells in the simulation design.

I simulated 10,000 examinees in each cell of the simulation design for the pseudo-populations. For each cell, I computed the effect size for the observed total test score mean difference and for the latent mean difference (and their corresponding standard deviations). Because both the observed score and latent variable effect size values are on the same metric (both being standardized), I was able to compute the difference between them as an index of how much the DIF biases the effect size.

Results and Conclusions

As was noted above, because the effect sizes are on the same metric (i.e., both are standardized), Table 4 lists differences between the effect sizes of the observed and the latent variable score for the three magnitudes of DIF and the four different number of DIF items (1, 4, 8, 16). One can see that, when no DIF exists, the effect sizes of the latent mean and observed mean are, as expected, equal (to the third decimal point and hence within sampling). Again, this serves as a check of the simulation methodology. However, when DIF (unidirectional uniform DIF) appeared in the test, the effect size differences increased. The more DIF items one has in their test and the larger the DIF, the greater the effect size differences with the observed mean differences being spuriously inflated by the presence of DIF.

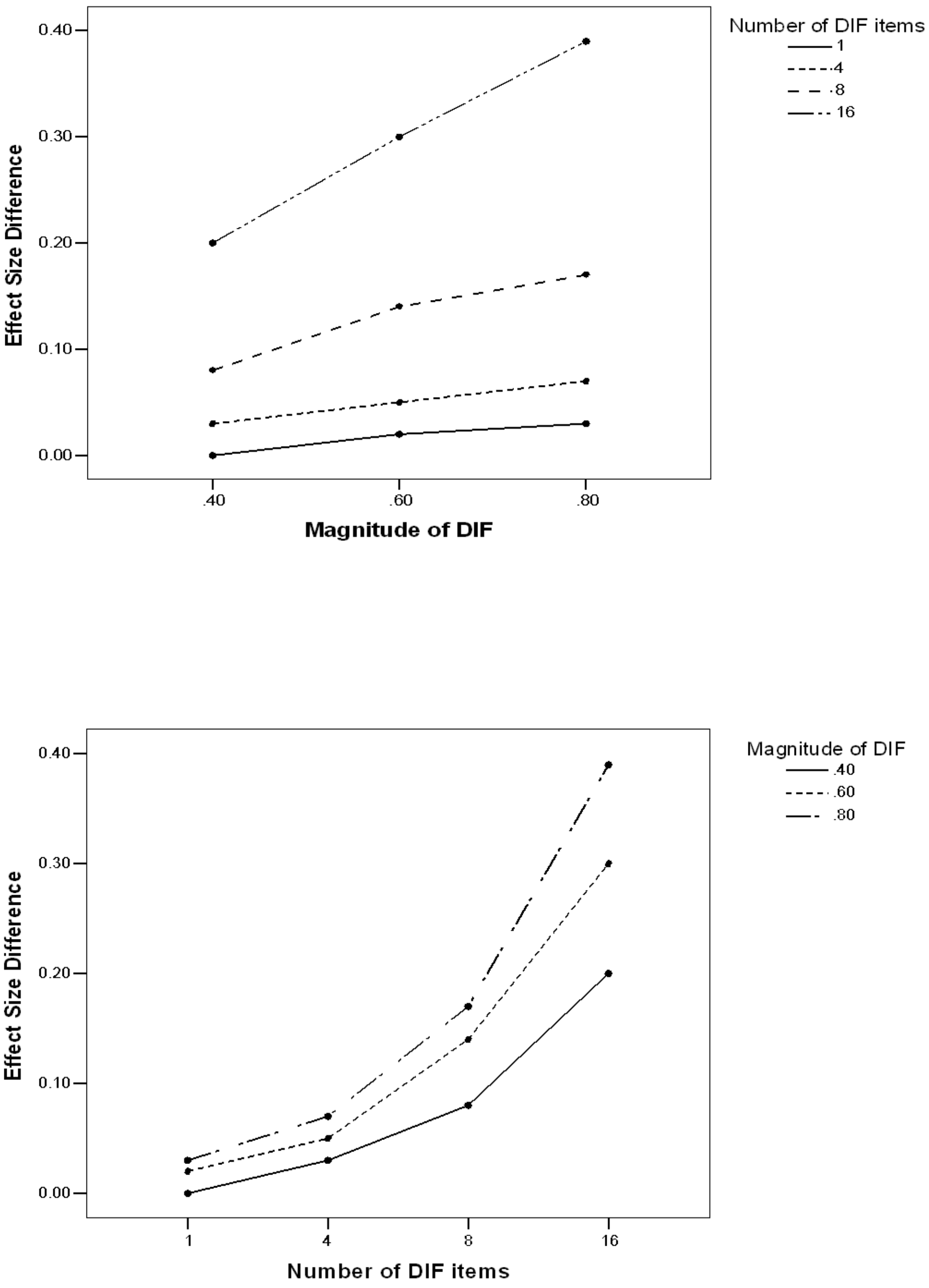
Table 4

Differences Between Latent and Observed Effect Size

Number of DIF Items	ES(Observed Score) – ES(Latent Variable Score)
0	-0.001
	Small DIF (Raju's area = 0.4)
1	0.003
4	0.029
8	0.088
16	0.196
	Moderate DIF (Raju's area = 0.6)
1	0.021
4	0.049
8	0.136
16	0.297
	Large DIF (Raju's area = 0.8)
1	0.033
4	0.063
8	0.167
16	0.396

Using the same analysis methodology as study one, the simulation results were analyzed using regression analysis with effect size differences as the dependent variable and magnitude of DIF, number of DIF items, and their interaction as independent variables. The model was statistically significant (F (3, 8) = 287.9, p < .0001) with an R-squared of 0.991 and an adjusted R-squared of .987. All of the predictors were statistically significant, including the interaction term. Figure 4 was used to graphically interpret the two-way interaction. Upon close inspection, it can be clearly seen that the effect size differences increase as the number of DIF items and the magnitude of DIF increase.

Figure 4. A plot of the effect of number of DIF items and magnitude of DIF on effect size differences.



A research question naturally arises from the findings to this point. Given that in studies one and two I treated the item parameter values as fixed values, I do not know the impact of varying item difficulty, discrimination and guessing on the Type I error rate.

Study Three: Impact of Item Parameter Values on Type I Error Rates

Studies one and two investigated the impact of DIF on the Type I error rate and effect size; however, the items used to generate DIF were sampled from 38 items in a fixed form of a test. Influences of the values of the item parameters were, therefore, not considered in either of the first two studies. Study three focuses on the impact of varying item parameter values on the Type I error rate. In essence, study three is an empirical check as to whether the findings in study one are generalizable to item parameter values other than just the ones under investigation therein – in essence, an investigation into the external validity of the findings in study one.

In this study I investigated the impact of item properties (values of a-, b-, and c-parameters), magnitude of DIF (quantified by Raju's area) and Δb (*b*-parameter differences between groups) and sample size on Type I error rate. This study focused on the case of one DIF item (i.e., the case in which the Type I error rates are protected in study one). I did not investigate the case of more than one DIF item because, in those cases, the Type I error rate is already inflated and hence of little practical value to investigate how the item parameter values may further inflate the error rates.

This study is different in purpose and design than the typical computer simulation study in psychometric research. The typical psychometric simulation study, such as studies one and two, have, in experimental design terms, fixed experimental factors. Therefore, as is well known in experimental design, generalizing beyond the values of the fixed factors is not recommended. If one wants to investigate the generalizability of findings from a fixed factor (computer simulation) experiment, one needs to randomly sample the values of the levels of the manipulated factors; hence, in essence, creating a random factor. The present study does just that by sampling item parameter values and magnitudes of DIF to investigate whether the protected Type I error rate when one has only one item exhibiting DIF generalizes to item parameter values other than those used in study one.

Methods

Therefore, different from study one, in which item parameters for the item exhibiting DIF were real parameters from TOEFL, DIF item parameter values and the b-parameter differences between groups in this study were randomly generated from normal and uniform distributions.

Let us first describe the simulation design in broad strokes with the details to follow. One can think of the simulation running in three large steps.

Step 1: I followed study one and used the same number of items and item parameters (1 DIF item out of 38 total items) and sample sizes per group (25, 50, 125, and 250).

Step 2: For each sample size condition, I generated 50 random item parameter values for the DIF item – recall that the other 37 item parameters were the same as those in study one. This resulted, in essence, in 50 runs for each sample size condition. For each of these runs, as in study one, 5000 replications were conducted using IRT to generate the data to compute the resultant Type I error rate for that run. Note that there are 50 runs for each sample size condition resulting in a total of 200 Type I error rates (one for each run) from the simulation.

Step 3: The resultant Type I error rates and their respective DIF item parameter values for the 200 runs (50 runs for each sample size combination) were then read into a statistical software package SPSS for statistical analysis.

Following similar approaches in the research literature (e.g., Hambleton & Rovinelli, 1986; Hambleton & Swaminathan, 1985; Zwick, Donoghue, & Grima, 1993), item parameter values were selected from probability distributions with specified means and variances – e.g., a-parameters were selected from a uniform distribution. For each run, the DIF item a-parameter values were generated from a uniform distribution (*M IN*= 0, *MAX* = 2), b-parameter values were generated from normal distribution (*M* = 0, *SD* = 0.75), and c-parameter values were generated from uniform distribution (*MIN* = 0, *MAX* = 0.50). Note that, as in study one, the θ values were generated from a normal distribution (*M* = 0, *SD* = 1) for each group; hence, like study one, the resultant proportion of rejected t-tests was the empirical Type I error rate.

Again, as in study one, given that this study focuses on uniform DIF, another factor manipulated in this study is the difference in *b*-parameter values between the focal and reference groups. The differences in b-parameters, Δb, were generated from a normal distribution (*M* = 0, *SD* = 0.50). With *b, c* and Δb, Raju's areas were calculated using equation (4) to quantify the magnitude of the uniform DIF.

$$Area = (1-c)\left|b_2 - b_1\right| \qquad (4)$$

As a descriptive summary of the simulation data, the generated values of the b-parameter ranged from -2.691 to 2.066 (M=-0.106, SD=0.863). Likewise, the a-parameter values ranged from 0.006 to 1.993 (M= 0.987, SD= 0.594); and c-parameters ranged from 0.002 to 0.300 (M=0.161, SD=0.085). Furthermore, Raju's area ranged from 0.004 to 1.156 (M=0.327, SD=0.254), and the difference in b-parameters as an index of DIF (the Δb) ranged from -1.068 to 1.300 (M=-0.031, SD=0.494). Finally, the a-, b-, and c-parameter values were not statistically significantly correlated with each other; ranging from -0.048 to 0.036.

The impact of varying item parameters on Type I error rate was then analyzed by regression modeling using the resultant data from the above simulation. The dependent variable for these analyses was the Type I error rate whereas the explanatory variables were: a-parameter, b-parameter, c-parameter, Δb, sample size, and the magnitude of DIF calculated by Raju's formula in equation (4).

Results and Conclusions

Table 5 listed the minimum and maximum values, means, and standard deviations of the resultant Type I error rates for different sample sizes. The minimum and maximum values, means, and standard deviations of the Type I error rate for the sample sizes of 25 and 50 per group are almost same. As the sample size increases to 125 per group and above, the maximum values of Type I error rate tend to be inflated beyond the Bradley's moderate and liberal criteria. The means and the standard deviations, however, are the same as those of small samples.

Table 5

Type I Error Rates for Different Sample Sizes

n (per group)	Minimum	Maximum	M	SD
25	0.044	0.057	0.050	0.003
50	0.044	0.058	0.051	0.003
125	0.042	0.065	0.051	0.004
250	0.044	0.070	0.052	0.004

Table 6 provides the percentage of Type I error rates, out of 50 runs in that cell, for each sample size that meet Bradley's (1978) various criteria for acceptable Type I error rates. As an example to guide the reader in how to interpret Table 6, for a sample size of 25 per group the Type I error rate met the moderate criterion with all (100%) Type I error rates less than .060, and 47 of 50 (94%) met stringent criterion with values less than .055. In general, from Table 6, it is clear, that with increasing sample size the percentage of the Type I error rate values that met the moderate criterion decreased – this is also true of the stringent criterion.

Table 6

For Different Sample Sizes, the Percentage of Type I Error Rates That Meet Bradley's (1978) Criterion

	Sample size (per group)			
	25	20	125	250
Stringent	94	88	88	88
Moderate	100	100	98	94
Liberal	100	100	100	100

To investigate the effect of sample size on the Type I error rate in this study, a one-way ANOVA was conducted with sample size as the independent variable. The effect of sample size was not significant, F (3, 196) = 2.075, p = 0.105. This result indicates that the sample size effect was trivial in this study situation. To investigate the association among the dependent and explanatory variables in this study, I conducted correlation analyses. Table 7 provides the Pearson correlation and Spearman's *rho* between Type I error rate and *a, b, c*, Δb, and Raju' s area. The results indicated that the Type I error rate only statistically significantly correlates with Raju's area.

It should be noted that the finding in this study, that sample size was not a significant contribution to Type I error rate that is not the same as the finding in study one. The different finding is most likely due to the fact that in study three, I only investigated one DIF item, whereas in study one I investigated from one to 16 DIF items. Furthermore, focusing on the case of one DIF item in study one, see Table 3, it can be seen that the findings are similar to study three –effect of sample size is trivial. Only in the case of large magnitude of DIF and number of DIF items, sample size has a nontrivial contribution to the inflation of Type I error rate.

The above descriptive information indicated that, in general, (a) the magnitude of DIF (Raju's area) was the only factor that significantly correlated with Type I error rate, (b) a-, b-, and c-parameters were not significantly related to Type I error rate. It should be noted that, because the Type I error rate is a proportion; I investigated whether using logit transformation would change the conclusions. The transformation did not change the conclusions, so the analysis was reported using untransformed data.

Study three was conducted to investigate the effects of varying item parameter (*a, b,* and *c*) values, Δb, sample size and magnitude of DIF (quantified by Raju's area) on the Type I error rate in the one DIF item situation. The results indicated that, in this study, the values of item parameters were not related to the inflation of Type I error rate. The only influential factor is the magnitude of DIF (Raju's area). This result confirms (a) what I found in study one: that the magnitude of DIF is a significant explanatory variable for increases in subsequent Type I error rates for the independent samples t-test based on the observed total test score, and (b) the generalizability of results in study one: the Type I error may be protected with one small DIF item retained in the test.

Table 7

Pearson Correlations and Spearman's rho Between Type I Error Rate and a-, b-, c-*parameter,* Δb*, and Raju's Area*

	Type I error rate	
	Pearson correlation	Spearman's rho
a	0.07	0.02
b	-0.01	-0.05
c	-0.09	-0.09
Δb	0.03	-0.04
Raju's area	0.35*	0.29*

*P < 0.01.

Section One Discussion

It was found, as predicted, that DIF did have an effect on the statistical conclusions; both the Type I error rate and effect size index of the observed score differences tended to be inflated. The effect size results are informative for the Type I error findings because they, in essence, show that, when one has DIF the observed score effect sizes are non-zero when they should be zero since the latent means were generated equal between groups in the Type I error situation. Without DIF, the examinees with the same amount of trait (e.g., ability) should have the same observed test scores. That is, the observed score effect sizes are inflated by the DIF. This highlights the earlier statement that the Type I error rates (and effect sizes) reflect the probability of rejecting the null hypothesis of equal observed test score means when the latent means are equal but DIF is present in the test. The Type I error rate (and zero effect size) are defined relative to the latent variable in the IRT model, not the observed variable; hence the inflation in Type I error rate. In short, DIF creates mean differences in the observed scores when none exist on the latent variable.

However, remarkably and not predicted from research, DIF also had little to no effect in some conditions. That is, when one had one DIF item out of 38 items, the Type I error rate of the subsequent hypothesis test was not substantially inflated above the nominal level. Furthermore, the subsequent effect size from the mean comparison was only inflated less than 0.03 on a standardized metric. In fact, this small effect of DIF also held up when there were four (out of 38) DIF items with small magnitude of DIF. Study three shows that the conclusions are not restricted to the specific item parameter values for the DIF item used in study one.

The first section of this paper only addresses the matter of amplification, unidirectional DIF. The next section moves to the question of what happens to the Type I error of subsequent hypothesis tests when DIF items show cancellation patterns.

Section Two: Impact of Differential Item Functioning (DIF) on Statistical Conclusions, II: Cancellation Effects

The results in Section one were based on an amplification view of DIF – i.e., all the DIF items were in the same direction. Section two will build on section one's findings and focus on investigating potential cancellation effects. A cancellation effect (also called multi-directional DIF) occurs when some DIF items favor one group and other DIF items favor the other group and, the DIF effects cancel each other out. Of course, one can have partial cancellation wherein the DIF effects do not cancel out entirely but rather to some degree.

Building on the previous three studies, two computer simulation studies are reported below. Study four reports on the Type I error rates and study five reports on the effect sizes of subsequent statistical tests of mean differences when some degree of cancellation effects are present among non-unidirectional DIF items. Therefore, the general simulation methodology is the same as the one used in studies one and two, respectively, for studies four and five.

The research question addressed in this section was concerned with identifying, under what conditions, the Type I error rate of subsequent hypothesis tests of equality of observed score means was inflated above the nominal level (i.e., 0.05) and, under what conditions, the effect size was biased.

Study Four: Impact of Cancellation DIF on Type I error rates

Method

Given that this is the first study of its kind and part of a larger series of studies, I limited the simulation to some idealized situations in which the magnitude of DIF is the same for each of the items. Future studies will be able to build on this idealized experimental work to more generalized situations. I chose to focus on the case of 16 DIF items (out of a total of 38) because (a) study one showed that this number of DIF items substantially inflated the Type I error rate, and (b) the 16 items allowed me to investigate a large number of degrees of partial cancellation -- as compared to, for example, four DIF items which would only allow me to investigate a quarter, half, or three quarters of the items favoring one group and the remaining items favoring the other group. Therefore, for example, in this simulation design for a small magnitude of DIF (Raju's area of 0.40), I simulated the situation in which eight items favored one group (e.g., boys) and eight items favored a second group (e.g., girls). I denoted this situation as 8:8; which is the balanced DIF situation in which there is complete cancellation. I expect in this situation that the DIF effects will be balanced out. Next, I simulated the same situation except for the DIF items being distributed as 7:9, 6:10, 5:11, 4:12, 3:13, 2:14, 1:15, and 0:16 DIF items per group. For each of these nine simulation conditions, the same simulation procedures were conducted to generate the item response data for this study as in study one. Continuing with the same data analysis strategies, the descriptive information is presented based on Bradley's (1978) criterion followed by regression analysis.

Results and Conclusion

Tables 8 lists the Type I error rates and the robustness information based on Bradley's criterion for small (top portion of Table 8), moderate (middle portion of Table 8) and large (bottom portion) magnitude of DIF, respectively. One can see from this table that, when one has complete cancellation (i.e., 8:8), the Type I error rate is, as expected, not inflated. One can also see that, as in study one, as the sample size and magnitude of DIF increase, so does the Type I error rate. However, depending on the magnitude of DIF, the Type I error rate can be protected for some partial cancellation conditions. For example, one can see from the top portion of Table 8 (small DIF, Raju's area of 0.40) that, for a sample size of 50 per group, the subsequent t-test of equal means has a protected Type I error rate in partial cancellation of five DIF items favoring one group and 11 items favoring the other group (i.e., a six item difference in the number of DIF items). Furthermore, for the middle portion and bottom portion of Tables 8 (i.e., moderate and large DIF), for a sample size of 25 per group the t-test is protected for as much as 6 items favoring one group and 10 items favoring the other (i.e., a 4 item difference in the number of DIF items).

Table 8

Type I Error Rates of t-test under Cancellation Effect

	Number of DIF items against reference and focal groups (reference/focal)								
N	8:8	7:9	6:10	5:11	4:12	3:13	2:14	1:15	0:16
	Small DIF (Raju’s area of 0.40)								
25	0.049	0.052	0.053	0.057	0.074↑	0.065↑	0.076↑↑	0.097↑↑	0.104↑↑
50	0.048	0.054	0.057	0.062↑	0.077↑↑	0.093↑↑	0.108↑↑	0.133↑↑	0.165↑↑
125	0.055	0.060↑	0.069↑	0.082↑↑	0.107↑↑	0.168↑↑	0.215↑↑	0.285↑↑	0.341↑↑
250	0.049	0.052	0.093↑↑	0.127↑↑	0.184↑↑	0.288↑↑	0.393↑↑	0.497↑↑	0.601↑↑
	Moderate DIF (Raju’s area of 0.60)								
25	0.046	0.049	0.055	0.066↑	0.084↑↑	0.102↑↑	0.121↑↑	0.148↑↑	0.172↑↑
50	0.049	0.056	0.075↑↑	0.084↑↑	0.109↑↑	0.163↑↑	0.209↑↑	0.255↑↑	0.308↑↑
125	0.052	0.058	0.100↑↑	0.136↑↑	0.214↑↑	0.334↑↑	0.436↑↑	0.535↑↑	0.628↑↑
250	0.049	0.056	0.134↑↑	0.227↑↑	0.376↑↑	0.562↑↑	0.730↑	0.837↑↑	0.898↑↑
	Large DIF (Raju’s area of 0.80)								
25	0.051	0.051	0.060↑	0.076↑↑	0.098↑↑	0.145↑↑	0.166↑↑	0.214↑↑	0.265↑↑
50	0.048	0.053	0.067↑	0.101↑↑	0.170↑↑	0.259↑↑	0.280↑↑	0.381↑↑	0.488↑↑
125	0.048	0.062↑	0.098↑↑	0.173↑↑	0.356↑↑	0.531↑↑	0.588↑↑	0.747↑↑	0.853↑↑
250	0.055	0.073↑	0.135↑↑	0.318↑↑	0.595↑↑	0.819↑↑	0.866↑↑	0.963↑↑	0.990↑↑

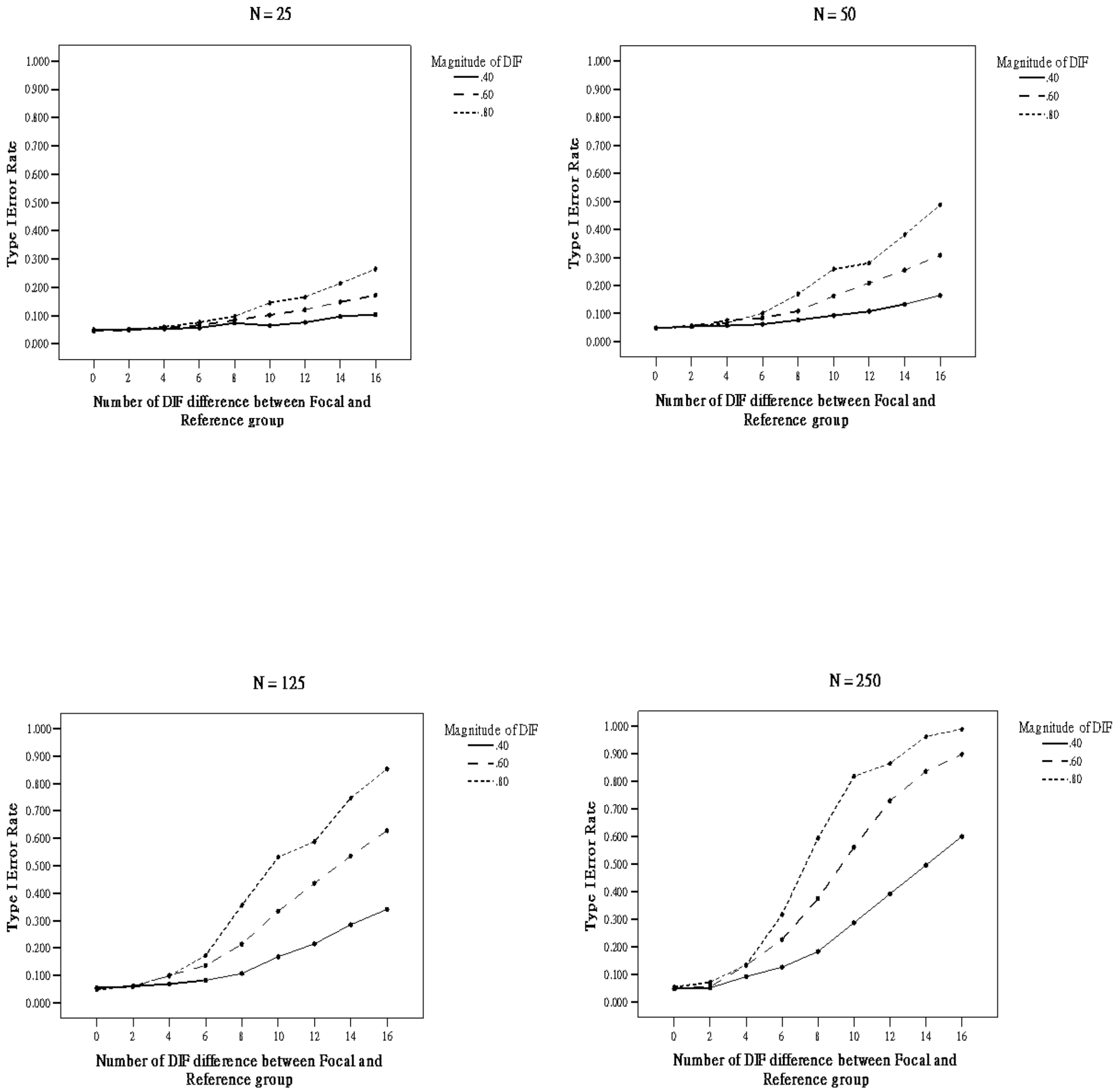
Note:

Type I error rate	Bradley (1978) criterion
α < 0.055	Meet the stringent criterion
0.055 ≤ α < 0.060	Meet the moderate criterion
0.060 ≤ α < 0.075, ↑	Violates the moderate criterion but meet the liberal
α ≥ 0.075 ↑↑	Violates the liberal criterion, Inflated

To investigate which experimental factors influence the Type I error rate, a regression analysis was conducted with magnitude of DIF, sample size, and difference in the number of DIF items between the two groups as independent variables. The resultant model, with two-way and three-way interactions was statistically significant, F (7, 100) = 274.1, p < 0.0001, R-squared = 0.950. All of the main effects and the interactions were statistically significant. A plot of the three-way interaction is provided in Figure 5. One can see from Figure 5 that the relationship between the difference in the number of DIF items (i.e., a proxy for the degree of partial cancellation) is more pronounced (i.e., a higher correlation) for larger magnitudes of DIF, and this relationship increases with increasing sample size.

Clearly then, the Type I error rate depends not only on the degree of partial cancellation but also on magnitude of DIF and sample size, and that, in some cases, the Type I error rate is protected even when one has partial cancellation.

Figure 5. A plot of the effect of number of DIF item differences, sample size and magnitude of DIF on Type I error rate.



Study Five: Impact of Cancellation DIF on Effect Size

Study five was designed to investigate the cancellation effect of DIF on the population effect size of the independent samples t-test. As in study two, Cohen's d was used as the measure of effect size and, the effect size difference, ΔES, was computed as the difference between the observed mean effect size and latent mean effect size so that a positive difference means that the effect size was larger for the observed scores. The research question in this study is: What is the effect size for the observed test score means when the latent means are equal, but a DIF cancellation effect is present in the test?

Methods

The simulation factors manipulated in this study, as well as the simulation methodology, were the same as those in study two except for the experimental factor, number of DIF items. That is, within each magnitude (small, moderate and large) of DIF, I manipulated the number of DIF items (out 16 DIF items) against focal group and reference groups. As in study four, the simulated number (out 16 DIF items) of DIF items against reference and focal groups were as follows: 8:8, 7:9, 6:10, 5:11, 4:12, 3:13, 2:14, 1:15, and 0:16. It should be noted that the 0:16 condition does, of course, not reflect cancellation but was included for comparison purposes. Three levels of magnitude of DIF -- small, moderate, and large as defined by Raju’s 1988 area statistic of .4, .6, and .8 were investigated. This resulted in a 9x3 completely crossed factorial design resulting in a total of 27 cells in the simulation design (including completely balanced, 8:8, and completely unbalanced cases, 0:16). As in study two, I simulated 10,000 examinees in each cell of the simulation design for the pseudo-populations. For each cell, I computed the effect size for the observed total test score mean difference and for the latent mean difference (and their corresponding standard deviations). Because both the observed score and latent variable effect size values are on the same metric (both being standardized), I computed the difference between them (ΔES) as an index of how much the DIF biases the effect size.

Results and Conclusions

Table 9 lists differences between the effect sizes of the observed and the latent variable scores for the three magnitudes of DIF and the nine DIF situations. One can see that, when the number of DIF items present in each group are totally balanced (8:8), the effect sizes of the latent mean and observed mean are, as expected, almost equal – i.e., -0.008. However, as the number of DIF items against each group was not balanced, the ΔES increased; the more unbalanced, and the larger the magnitude of DIF, the greater the ΔES – i.e., the observed mean differences being spuriously inflated by the presence of DIF.

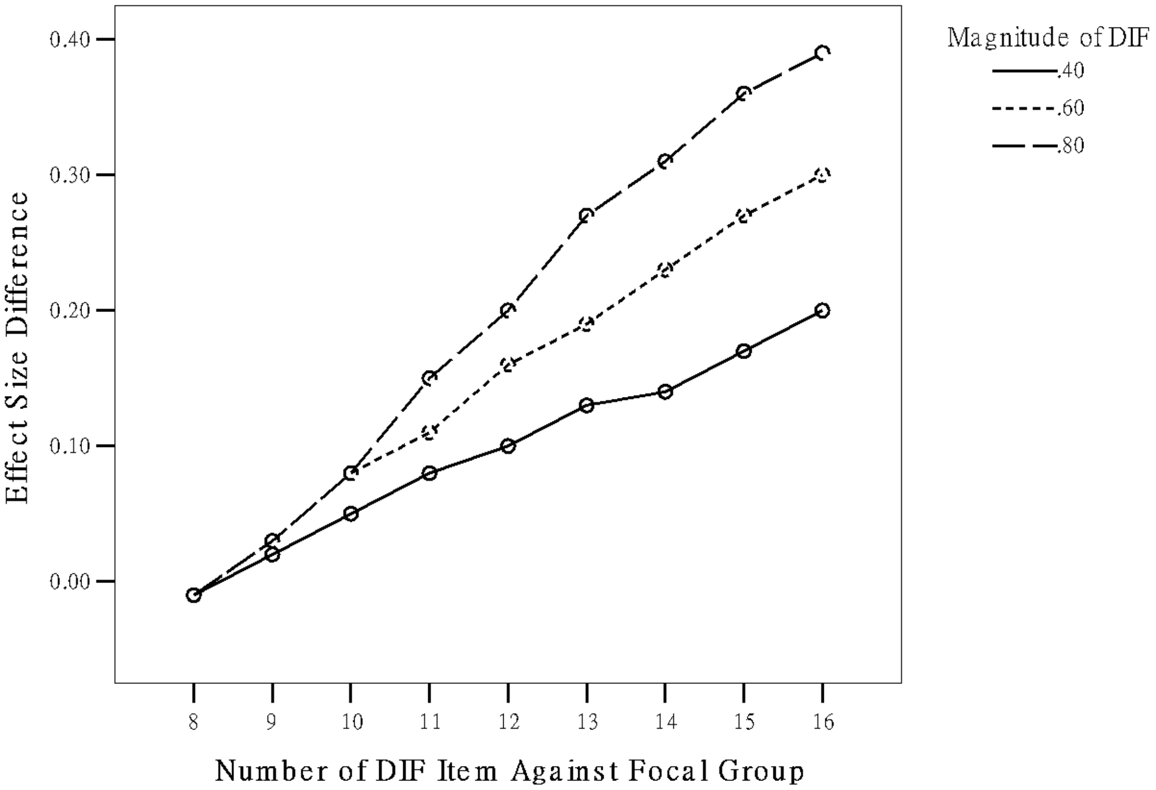
Table 9

Differences Between Observed and Latent Mean Effect Sizes for Varying Number of DIF Items in Reference and Focal Group for Different Magnitudes of DIF

Number of DIF items in each group	ES difference (ΔES)		
	Magnitude of DIF		
Reference vs. focal	Small	Moderate	Large
8 : 8	-0.008	-0.010	-0.009
7 : 9	0.023	0.032	0.035
6 : 10	0.052	0.077	0.082
5 : 11	0.083	0.109	0.147
4 : 12	0.100	0.158	0.198
3 : 13	0.124	0.193	0.266
2 : 14	0.149	0.225	0.315
1 : 15	0.174	0.270	0.354
0 : 16	0.196	0.297	0.396

The simulation results were analyzed using regression analysis with ∆ES as the dependent variable, and magnitude of DIF, number of DIF item difference between groups, and the two-way interaction between magnitude of DIF and number of DIF item difference as independent variables. The model is statistically significant (F (3, 23) = 1621.08, p < .0001) with an R-squared of 0.995. All of the predictors were statistically significant, including the main effects and the interaction term. The interactions among the independent variables can be seen in Figure 6. Clearly, ∆ES increased as the imbalance in DIF and magnitude of DIF increased.

Figure 6. A plot of the effect of number of DIF item differences and magnitude of DIF on effect size differences.



Section Two Discussion

The results confirm the hypothesis that when there is a balanced number of DIF items between groups, the Type I error rate is protected and ΔES was not biased no matter how large the magnitude of DIF and number of DIF items present. On the other hand, as the number of DIF items become more unbalanced between groups, both the Type I error rate and the ΔES were inflated. Furthermore, the effect of imbalance was even more inflated by the magnitude of DIF.

General Discussion

As noted in the introduction, it is common for researchers to either not test for DIF before comparing groups, or they decide to leave DIF items in the test. Of course, DIF is a statistical characteristic of a sample so it is possible that DIF items are simply not detected (e.g., Type II error) during item analysis. In short, this leaves us with the question of the impact of DIF items on the eventual statistical tests conducted on the observed test (or scale) scores. To answer this question, five related simulation studies were conducted. To my knowledge, this is the first of a line of research that directly answers the often heard question: What is the impact of having DIF on the eventual statistical conclusions from test scores. The results of this dissertation offer advice to practicing researchers about when and how much the presence of DIF will affect their statistical conclusions based on the total observed test scores. Although, simulated in idealized situations deliberately (e.g., I did not look at the cases of mixed magnitude of DIF, nonuniform DIF which are more common in real test data), the five related simulation studies provide researchers and practitioners with general guidelines.

In the case of Section I, wherein the DIF items were all in one direction (e.g., the test favors girls consistently, amplification DIF), as expected, DIF results in inflation in Type I error rates of the eventual t-test. Likewise, of course, reflecting the inflation in Type I error rates, the observed score effect size is also inflated, sometimes substantially. The inflation of the effect size was important because it is now widely recommended that effect sizes be reported with statistical hypothesis test results. What was not expected, however, was that the Type I error rate and effect sizes were not biased by the presence of DIF when the number of DIF items is small (i.e., one DIF item out of 38 items, and even four DIF items out of a total of 38 items when the magnitude of DIF is small to moderate). This is important, and comforting, to researchers who do not typically screen for DIF or ones who do not remove DIF items from the test. However, what is not yet known is the impact of DIF, in these situations when the Type I error rate is protected, on the eventual statistical power. I did not investigate the statistical power (i.e., the results under the case when the population means are not equal) due to the fact that Type I error rates need to be established before one can interpret the results of statistical power. The issue of impact of DIF on statistical power will be investigated in forthcoming studies. Likewise, the present studies should not be interpreted to suggest that one need not screen items for DIF. In fact, the conclusions are quite to the contrary. DIF analyses are needed because, under many situations, the Type I error rate and effect sizes are severely biased by the presence of DIF.

In Section II, wherein one has an imbalance of DIF items, for example, some items favoring girls and others favoring boys, the effect of DIF depends on the degree of imbalance. As expected, when the DIF is balanced (e.g., eight items favoring boys and eight items favoring girls), the DIF effect cancels out and the Type I error rate and effect sizes were not biased by DIF. However, the degree of imbalance and the magnitude of DIF interact to inflate both the Type I error rate and effect size. Again, the t-test was surprisingly robust in terms of Type I error rate and effect size with a small amount of imbalance (e.g., the t-test was not greatly affected when six items favored one group and 10 items favored the other group).

Overall, these findings highlight why it is important to use DIF screening procedures before conducting group comparisons because one may find oneself in the situation wherein the Type I error rate of the hypothesis test, and the corresponding effect size reported, are highly inflated, declaring group differences where none exist. Likewise, retaining DIF items in the test may also have significant effects on other psychometric procedures, such as equating results when used in concert with DIF detection or more broadly in the use of linking and equating. Although several studies have investigated the effects of linking or equating methods on DIF detection (e.g., Candell & Drasgow, 1988; Cohen & Kim, 1992; Hidalgo-Montesinos & Lopez-Pina, 2002; Miller & Oshima, 1992), there is a need for more research on the effect of DIF on equating or linking (e.g., Chu & Kamata, 2005) in its more general use in large-scale testing, much like I do for significance testing herein.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 34,* 144-152.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.)*, Statistical theories of mental test scores* (pp. 397 - 424). Reading, MA: Addison-Wesley.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed). Mahwah, NJ: Lawrence Erlbaum.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer, Nijhoff.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

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

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

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.

Shealy, R., & Stout, W. (1991). *An item response theory model for test bias* (Office of Naval Research Tech. Rep. No. 4421-548). Urbana: University of Illinois, Department of Statistics.

Shealy, R., & Stout, W. (1993). An item response theory model for test and differential item functioning. In H. Wainer & P. Holland (Eds.), *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum.

Zumbo, B. D., & Harwell, M. R. (1999). *The methodology of methodological research: Analyzing the results of simulation experiments*. (Paper No. ESQBS-99-2). Prince George, B. C.: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*, 136-147.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45-79). Elsevier Science B.V.: The Netherlands.

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

CHAPTER THREE: CONCLUSIONS AND FUTURE DIRECTIONS

Revisiting the Research Questions

Based on the empirical research, it can be argued that DIF is found in many tests and testing programs. However, given that DIF is a statistical conclusion from a hypothesis test, the true status of DIF in the test is usually unknown to the researchers or practitioners. Imagine, as suggested in the opening chapter to this dissertation, that DIF analyses were conducted prior to investigating group differences on an outcome variable (e.g., gender differences on a variable such as language proficiency). If DIF is present, the internal validity of the statistical conclusion (i.e., in my example, a conclusion of gender differences) is suspect given that the group difference might just be an artificial effect caused by DIF items. What was unknown prior to the research reported herein, is the effect of DIF on the eventual statistical conclusions based on the observed test scores. This study investigated the impact of DIF on the Type I error rates of the independent samples t-test and the effect size of the observed score mean difference. Five simulation studies were conducted to investigate the effect of DIF in two general conditions: (1) DIF items were generated favoring one group of examinees consistently, and (2) DIF items were generated favoring different groups of examinees (e.g., some items favoring boys while other items favoring girls). The first condition was defined as the DIF amplification effect whereas the second condition was defined as the DIF cancellation effect. For the DIF amplification effect, how and to what extent DIF items acting in the same direction distort the Type I error rate and the effect size index were questions of interest. In the DIF cancellation effect studies, how and to what extent DIF items acting in different directions affect Type I error rate and effect size index were of particular interest. The DIF effect on Type I error rate and effect size were investigated under various magnitudes of DIF, numbers of DIF items, and sample sizes. The items used to generate DIF were sampled from 38 items in a fixed form of a test. Therefore, in addition to the two studies looking at the amplification effect of DIF on Type I error rate and effect size, for the cells where Type I error was protected, an additional study was conducted investigating the impact of varying item parameter values on the Type I error rate as an empirical check to see whether the findings were generalizable to item parameter values other than just the ones under investigation. For all other conditions, wherein Type I error was inflated, the impact of varying item parameter values was not investigated because the Type I error was already inflated and therefore there was not much practical meaning in doing so. The results of condition (1) suggested that DIF, even in a small magnitude, may cumulate at test level, and then inflate the Type I error rate and effect size. On the other hand, as indicated in the results of condition (2) when one has bi-directional DIF items, as expected, when the DIF is balanced the total effects of DIF at the test level cancel out and the Type I error rate and effect sizes were not biased by DIF. However, as the degree of imbalance increases, both the Type I error rate and effect size were biased; that was, the more imbalance and the larger the magnitude of DIF, the more bias present. In addition, the amplification and cancellation effect DIF items also depend on the magnitude of DIF and sample size. Overall, these findings highlight why it is important to use DIF screening procedures before conducting group comparisons because one may find themselves in the situation wherein the Type I error rate of their hypothesis test, and the corresponding effect size reported, are highly inflated. Without DIF analysis one may declare group differences where none exist.

Contribution to the Literature

The novel contribution of this dissertation to the literature is that this study directly answered the question of the impact of DIF on the eventual statistical conclusions based on the total observed test scores. It also offers advice to practicing researchers about when, how, and how much the presence of DIF will effect their statistical conclusions based on the total observed test scores. To my knowledge, this is the first research that directly answers this often heard question. Although deliberately simulated in idealized situations, the five related simulation studies provide researchers and practitioners with general guidelines about how DIF affects statistical conditions.

Limitations and Future Studies

As indicated in the beginning, the study conditions simulated in these five inter-related studies were idealized to help the interpretation, and more importantly, to set up a baseline for proposed series of follow-up studies in this program of research. In other words, the studies conducted put the priority on the internal validity of the simulation studies focusing on identifying the causal effect and a limited number of influential factors that I believed would affect Type I error rates and effect sizes of the independent sample t-test. To obtain the net effect of these factors and ease the interpretation, the study factors were limited to what were studied in this dissertation. In essence, the studies conducted so far are controlled experiments in which only a few factors were studied, such as sample size, number of DIF items, magnitude of DIF, and the interactions among these factors. A number of other factors that certainly may have an effect on the statistical conclusions were not investigated, such as, non-uniform DIF, and mixed magnitude of DIF.

As Zumbo (2007, p. 24) states “Item impact described the situation in which DIF exists, because there were true differences between the groups in the underlying ability of interest being measured by the item. Item bias described the situations in which there is DIF because of some characteristic of the test item that is not relevant to the underlying ability of interest (and hence the test purpose)”. From these definitions offered by Zumbo, it appears that this dissertation was focused on test bias given that the latent variable means were equal. It seems that a study of the statistical power of the t-test, wherein the latent variable means are different, may better characterize item impact.

With these limitations in mind, in forthcoming research the priority will be put on studying the effect of real DIF conditions on the statistical conclusions. That is, I will simulate the data close to the real data to get the full story of effect of DIF on day-to-day practice. In other words, in the future study, the external validity of the simulation studies will be the focus and the priority will be put on the generalizability of the results. Particularly, I will investigate the following factors as suggested by the literature and as mentioned in the first chapter of this dissertation that might have an effect on test level statistics.

DIF with mixed magnitude

In current study, the simulation factors were simulated in idealized conditions. In each study cell, the DIF items were simulated having the same magnitude. This rarely happens in practice; in reality, DIF never shows up with an even magnitude across items and across groups. The DIF effect may come from a variety of magnitudes varying from negligible to significantly large. In the future studies, closer to practice conditions will be studied, particularly, the mixed magnitude of DIF.

Non-uniform DIF

Non-uniform DIF is discussed in the DIF screening literature (e.g., Rogers & Swaminathan, 1991, 1993). Non-uniform DIF is a more complicated situation wherein the amount and/or direction of DIF vary by level of ability within a group of examinees. When non-uniform DIF is present, an item may favor a different group of examinees at different levels of ability or favor the same group of examinees differently at different levels of ability – this may lead to cancellation within an item and thus may reduce the overall effect of DIF on the test level statistics such as the observed mean scores.

Causality of DIF

In practice, DIF comes from different sources. In other words, using the language of factor analysis, these individual DIF items may be caused by (load on) different factors. Several situations may happen: (a) several DIF items load on one prominent factor; in this case all DIF items all caused by one factor, (b) a number of DIF items load on a number of different factors with approximately equal proportions of variance of the total variance; this is the case wherein DIF items come from different sources, (c) some DIF items load on one (or a few) dominant factors and other DIF items load on a number of different secondary factors, and (d) cross loading may also happen (e.g., one item may load on more than one factors). Depending on the influence of (or the magnitude/proportion of variance accounted for by) the secondary factor, it may have a different effect on the observed score differences. Drasgow (1987), Roznowski (1987), and Roznowski and Reith (1999) argued that non-trait variance may overshadow trait variance in an item but no single source of bias predominates in the total score; thus, as long as secondary factors do not overpower the primary dominant factor, the overall DIF effect will not be influential. However, these ideas need to be systematically examined in empirical research and simulation studies of the nature I have reported herein.

Non-normal distributions

The distribution characteristics of the reference and focal groups in this study were simulated identical and normal, which are not realistic in practice. When the distribution is neither identical nor normal (i.e., skewed or kurtotic), the proportion of examinees influenced by DIF items can be varied along the ability continuum and thus can further affect the observed mean score.

Unequal sample sizes

The studying sample sizes were simulated equal in reference and focal group. Type I error rate may be influenced by unequal sample size between groups. The effect of imbalance in samples (e.g., unequal sample size, missing cells) is a main issue in both statistically significant tests and psychometric research (Zwick, Thayer, & Wingersky, 1994) as it is well known that violation of the assumption of equal variances when sample sizes are unequal inflates the Type I error rate substantially (Zimmerman, 2004). In future research, the DIF effect combined with unequal sample sizes will be explored.

Test length

Baker (2001) stated "… in general, longer tests will measure an examinee's ability with greater precision than will shorter tests" (p. 107). Thus, with less estimation precision, plus any proportion of DIF, a short test might be more likely to distort test level statistics than a longer test. A longer test might be more robust to the effect of DIF than a short test. In the future research, tests will be generated with various lengths combined with different proportions of DIF to examine how DIF affects statistical conclusions based on total test scores of tests of different length.

Impact of DIF on other statistical and psychometrical procedures

Likewise, retaining DIF items in the test may also have a significant effect on other statistical and psychometric procedures, such as statistical power, regression, and test score equating.

The effect of DIF on statistical power will also be investigated as a continuation of this dissertation to find out how DIF inflates or deflates the statistical power of hypothesis test. This is the situation wherein latent means between groups are not equal. DIF as an artificial effect of test might artificially exaggerate mean differences between groups by favoring the group with higher ability or minimize the group difference by favoring the group with lower ability. Same as the studies reported in this dissertation, the impact of DIF will be considered in two directions: (1) DIF amplification effect wherein DIF items will be simulated favoring one group of examinees consistently and investigate under what conditions and to what extent the statistical power will be affected, and (2) DIF cancellation effect, wherein DIF items will be simulated favoring both groups of examinees. The impact of DIF in both of these two conditions will be investigated combined with other simulation factors mentioned above.

In addition to the influence of DIF on hypothesis tests, the impact of DIF on psychometric analysis is also worth investigating. That is, previous studies have investigated the effects of linking or equating methods on DIF detection (e.g., Candell & Drasgow, 1988; Kim & Cohen, 1992; Hidalgo-Montesinos & Lopez-Pina, 2002; Miller & Oshima, 1992); however, there is a need for more research on the effect of DIF on equating and linking (e.g., Chu & Kamata, 2005) in its more general use in large-scale testing much like I do for significance testing herein. For example, item parameter values may change over time due to a number of different reasons. Including these items in the anchor test will have a significant effect on the equating results. The procedures used to detect item parameter drift are statistical procedures which may make Type I or Type II errors. No matter the parameter drift unbeknownst to the researchers or not, more research needs to be conducted to see how much the change in item parameters affects equating results.

In general, the impact of DIF is an important topic in psychometric and applied statistical research. It affects the final conclusion of applied practitioners and researchers who use measures to collection data in their practical clinical (counseling) or diagnostic work and their research. The presence of DIF does not have negligible effect on the statistical conclusions!

References

Baker, F. B. (2001). *The basics of item response theory (2nd ed)*. ERIC Clearinghouse on Assessment and Evaluation. http://echo.edres.org:8080/irt/baker/final.pdf.

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

Chu, K.L., & Kamata, A. (2005). Test equating with the presence of DIF. *Journal of Applied Measurement, 6*, 342-354.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.

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

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

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

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72*, 480-483.

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*, 248-269.

Zimmerman, D. W. (2004). Inflation of Type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation test. *Psicol ógica, 25*, 103-113.

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