# Deriving Cut-off Scores for Significance Testing of NCDIF Indices Within the DFIT Framework Through Item Parameter Replication Method.

Andrey V. Koval

Middle Tennessee State University

A thesis presented to the

Graduate Faculty of Middle Tennessee State University

in partial fulfillment of the requirements for the degree of Master of Arts.

August 2008

Deriving Cut-off Scores for Significance Testing of NCDIF Indices Within the DFIT Framework Through Item Parameter Replication Method.

APPROVED:

Graduate Committee

_____

Thesis Advisor, Jwa K. Kim, Ph.D.

_____

Committee Member, Dana K. Fuller, Ph.D.

_____

Chair, Psychology Department, Dennis Papini, Ph.D.

_____

Dean, College of Graduate Studies, Michael D. Allen, Ph.D.

Acknowledgement

I would like to thank Dr. Jwa K. Kim, whose endless support and patience made this work possible and, most importantly, enjoyable. I am most fortunate to have seen in him the example of guidance and encouragement that will remain with me in the years to come.

I would also like to thank Dr. Dana Fuller for keeping her door open for me. I owe her the fun and pride of being a quantitative psychologist.

My special thanks to Jacob Seybert and Joe Backer for playing with me in statistics and keeping the game challenging and fulfilling. Thanks to them, the work did not make me dull.

## Abstract

Differential item functioning (DIF) deals with how test items perform in different demographic groups. Although many operationalizations of DIF are proposed, no single approach proves to be exclusively superior. The current study investigated the properties of DIF indices in the differential functioning of items and tests (DFIT) framework and further tested the power of item parameter replication method (IPR) in identifying biased items within the item response theory (IRT) paradigm. The results indicated the IPR method is a useful tool in DIF detection. The source of item parameters covariance structure was found to affect the results only slightly. The results indicated that the proportion of biased items on the test might not be as influential factor as the actual number of biased items. This finding, along with low false negative rates when $99.9^{th}$ percentile cut-off scores is used, supports the view that sequential elimination of biased items is preferable over the simultaneous detection.

Table of Contents

List of Tables

List of Figures

# CHAPTER I

## Introduction

### *Differential Item Functioning (DIF): Concepts and Definitions*

The issue of test fairness has been inevitably united with the development of measurement science. The pivotal question of test bias is related to the accurateness of measurement in a specific aspect: "Does the test measure the ability *equally accurate* across various groups of test takers, who are united by certain socio-demographic characteristics?" Differential item functioning (DIF) analysis seeks to answer this question.

It is important to point out several idiosyncrasies of the early test bias methods as this would assist to see the objectives of the modern DIF research more clearly. Most of the early test bias methods were united by focusing on only two groups of test takers. Hence, there are *focal group* and *reference group;* terms that still imply minority and majority groups, respectively (Zumbo, 2007). Discrepancy between average tests scores for majority and minority groups served as the ground to question the overall test fairness. The political environment of the time ushered the division of test takers only into two groups. This, however, is not a defining characteristic of modern DIF analysis. Questions might arise as to practical necessity or computational cumbersomeness of DIF methods involving more than two groups of subjects. However, the theoretical framework does not make reservations for the exclusivity of the two-group design.

The term "test bias," therefore, did not initially differentiate between the actual group differences (also known as impact) and the test's predisposition to favor a specific group. The term was descriptive of scores' difference, without elaboration on the causes

of such differences. Classical test theory (CTT) was a prevailing measurement theory of the day when the notions of test bias were formulated. Naturally, the theoretical framework of CTT transcended into the concept of test bias on the fundamental and definitional level. Thus, the measurement of test bias was sample and test dependent (Lord, 1980).

With the advent of the item response theory (IRT), which broke loose from the sample and test dependency, the definition of test bias was revised according to the new framework. The focus shifted from comparing the observed performance of the groups on the test to comparing the behavior of items under different group conditions. As the result, purification of terms occurred, creating *item bias* to represent characteristics of an item that assigns different scores to participants matched on the ability, but belonging to two different groups (Zumbo, 2007).

Related to item bias, the term *item impact* came to signify the difference in an item's functioning that existed due to genuine difference of the ability among groups (Millsap & Everson, 1993). Both *item bias* and *item impact* are focused on how an item functions in each group (hence, differential item functioning). The distinction lied in the attribution to the cause of this difference that each label makes. Thus, by labeling the observed difference in item's functioning an *item impact*, the nature of this difference was explained by the underlying true inequality in ability between groups. The label *item bias*, on the other hand, explained the observed difference by the flaws of the test. As Hambleton and Swaminathan (1985) pointed out, the term DIF distinguished the empirical evidence from the conclusion regarding the nature of discrepancies. Therefore, in order to conclude test bias, the presence of DIF is a necessary, but not a sufficient

condition. Throughout this text, the term *DIF* will be used to describe the existence of a difference in functioning of an item between groups that has been caused by the flaws of the item.

*Theoretical Context: DIF Classifications and IRT Basics*

The most recent and comprehensive review of DIF methods (Millsap & Everson, 1993) suggested the classification, in which the type of measurement bias was distinguished on the basis of type of conditional invariance: either observed conditional invariance (OCI) or unobserved (UCI). The distinction between these two broad families of DIF lied in whether the analysis used the observed variables or focused on the latent traits. In contrast to *DIF,* the authors introduced the term *measurement bias* in order to signify that DIF refers to the functioning of a single item specifically. *Measurement bias*, however, may involve the behavior of a testlet or an entire test. Such is the case with a factor analysis approach to DIF detection (Millsap & Everson, 1993). This approach did not make use of parametric indices of IRT, but still investigated the behavior of latent variables. In fact, it is important to understand that DIF analysis is not an exclusively IRT method and should not be thought of as such. Methods that do not rely on unobserved invariance include Loglinear Models (Kok, Mellengbergh, & Van der Flier, 1985), Mantel-Haenzel (Holland, 1985; Holland & Thayer, 1988), Standardization Method (Dorans & Kulick, 1983, 1986; Dorans & Holland, 1993), Logistic Regression Method (Swaminathan & Rogers, 1990), Logistic Discriminant Function Analysis (Miller, Spray, & Wilson, 1992), and others. The Mantel-Haenzel procedure is currently the most popular among practitioners as a computationally simple method that requires little specialized knowledge. Due to the space limitation, it is impossible to review this

family of DIF methods in the current study. In its stead, this study will focus on methods relying on IRT for their theoretical framework. A review of IRT, thus, is necessary before a look is taken at the mathematics of DIF and specifically the Differential Functioning of Items and Tests (DFIT) indices.

Although possessing somewhat formidable reputation among statisticians and measurement practitioners, IRT holds major premises that are quite straightforward, intuitive, and logically obvious. When one speaks of a measurement instrument that attempts to measure a specific trait or ability, one invariably assumes (and has all the reason to do so) that the test in question has been designed to detect and represent the dependency of a person's performance on the test onto his or her ability. In other words, in intelligence tests higher performance would be linked to higher intelligence. Ideally, the performance on the test *is* the measure of the ability in question. Of course, such is never the case, and one naturally assumes the presence of interfering factors (apart from ability) which affect subject's performance on the test. Thus, an assumption must be made that the performance on the test *is a function* of subject's ability and other extraneous factors. In its most general terms, such logic can be written as

$$\mathrm{Y}_s = f(\Omega) \tag{1}$$

where $\mathrm{Y}_s$ is the performance of the subject *s*, on the test, and Ω is a composite of subject's abilities, test's characteristics, and measurement error.

Although a variety of complex models for $\mathrm{Y}_s$ is possible, the most simple and intuitive of them will be selected for current purposes, namely

$$\mathrm{Y}_s = \sum_{i=1}^{n} \gamma_{si}\,, \tag{2}$$

where $\gamma_{si}$ is the performance of subject *s* on item *i* and *n* is the number of items. Subjects' performance on the item can be measured differently, either by pass/fail or partial credit criterion. For ease of demonstration, pass/fail credit for all the models will be assumed, unless specified otherwise.

In respect to this model of $\mathrm{Y}_s$ , formula (1) can be rewritten as:

$$\mathrm{Y}_s = f(\Omega) = \sum_{i=1}^{n} \gamma_{si} = \sum_{i=1}^{n} f(\omega_{si})\,, \tag{3}$$

where $\omega_{si}$ is a composite of abilities of examinee *s*, characteristics of item *i*, and measurement error. In its turn, $\omega_{si}$ can be broken into its individual components. Thus,

$$\omega_{si} = \theta_s, \pi_{1i}, \pi_{2i} \ldots \pi_{ki}, \varepsilon_i\,,$$

where $\theta_s$ is the ability of subject *s*, $\pi$ is certain coefficients quantifying the characteristics of item *i*, *k* is the number of such coefficients, and $\varepsilon_i$ is the measurement error term for item *i*.

However, in contrast to CTT, in which the observed score is a composite of true score and the error of measurement ($\mathrm{X}_{IP} = \mathrm{T}_{IP} + \mathrm{E}_{IP}$) for test I and person P, IRT deals with error differently. IRT allocates the error term to the estimation of the function, thus distributing the error among the function's agents. Under the condition of local independence, when controlling for ability, the probabilities of answering items correctly are independent of each other. This implies that the only factor influencing the measure

of subject's ability ($\theta_s$) is the ability itself. In this way, by resorting to the assumptions of IRT it is adopted that

$$\hat{\omega}_{si} = \hat{\theta}_s, \hat{\pi}_{1i}, \hat{\pi}_{2i} \ldots \hat{\pi}_{ki}. \quad (4)$$

Coefficients $\hat{\pi}_{ki}$ are the forerunners of items parameters, which will be discussed in due course. However, now it is time to point out that no definite quantification of these parameters is possible, only estimation. In this way, the only known in the equation (1) is $\mathrm{Y}_s$, which is the overall performance of the subject on the test and ideally is equivalent to subject's ability. Thus, equations (3) and (4) can be rewritten for the item level as

$$\gamma_{si} = f(\hat{\omega}_{si}) = f(\hat{\theta}_s, \hat{\pi}_{1i}, \hat{\pi}_{2i}, \ldots \hat{\pi}_{ki}). \quad (5)$$

It is only natural to assume that items would vary among themselves in respect to various characteristics such as difficulty, discriminatory power, degree to which the item measures purported trait, chances of guessing the correct response, and etc. In fact, if

$$\pi_{11} = \pi_{12} = \ldots = \pi_{1n}$$

$$\pi_{21} = \pi_{22} = \ldots = \pi_{2n}$$

-
-
-

$$\pi_{k1} = \pi_{k2} = \ldots = \pi_{kn},$$

where *n* is the number of items, the coefficients become constants across items and make $f(\theta_s)$ to be linearly dependent on $f(\theta_s, \pi_1, \pi_2, \ldots \pi_k)$, thus reducing equation (5) to $\gamma_{si} = f(\hat{\theta}_s)$. This implies the absolute absence of measurement error and identity of

items' characteristics, both of which are contradictory to reality and common sense. This also implies that each item contributes to the overall score equally, which is another violation of practical sense. Some weights are needed to account for discrepancies of items' characteristics. The aforementioned considerations define two essential questions that IRT attempts to answer:

1. What characteristics of an item should be included in $f(\omega_{si})$?
2. What is the definition of $f(\omega_{si})$? In other words, what is the nature of relationship between the observed response, person's ability, and item's characteristics?

Now, that it has been made clear that the observed response to an item is a *function* of subjects' ability and item characteristics, the existing answers to the two fundamental questions of IRT are ready to be briefly reviewed.

The current state of IRT widely recognizes only three item characteristics (parameters) as influential agents in the function of examinee's responses. Known as *b*, *a*, and *c* parameters, they respectively represent item's difficulty, discriminatory ability, and guessing. One of the underlying postulates of IRT states that the relationship between performance of an examinee on an item and his or her ability can be described by a monotonically increasing function called the item characteristic curve (ICC) (Hambleton & Swaminathan, 1985). ICC is a mathematical function that relates the ability measured by the item and its characteristics to the probability of answering this item correctly. The function that builds an ICC curve depends on the IRT model that is being applied to the data.

While it is possible to describe this function in an infinite number of ways, only a few models are currently being used. The models are one parameter logistic (1PL), two parameter logistic (2PL), and three parameter logistic (3PL) models. These models incorporate different number of parameters as evident from their names. The 1PL model describes the ICC function as

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, \tag{6}$$

where $P_i(\theta)$ is the probability that an examinee chosen at random with ability (θ) will answer item $i$ correctly, and $b_i$ is the difficulty parameter of item $i$. The difficulty parameter is quantified as the point on the ability scale where there is 50% chance of getting the item correct. Therefore, higher values of $b$ parameter signify smaller probability of answering the item correctly. The 1PL model is also known as the Rasch model (Rasch, 1960). Although the form of 1PL model is different from the original model proposed by Rasch, it is mathematically equivalent to the Rasch model, and therefore bares his name.

Lord (1952) proposed the two parameter model based on the cumulative normal distribution (normal ogive). Birnbaum (1968) adapted the two-parameter normal ogive function to a logistic distribution, thus formulating what is now known as 2PL:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \tag{7}$$

where $a_i$ is the item discrimination parameter and $D$ is the scaling constant equal to 1.701. Item discrimination parameter is the slope at the inflection point ($b$ parameter).

The items with higher values of $a_i$ are more useful in differentiating high-ability examinees from low-ability examinees.

In cases where possible answers to the test items are defined in terms of multiple choices, one faces with a possibility of correct answers based solely on chance. Depending on the number of response categories, this probability would be different. To incorporate this distortion in response interpretation, the third parameter *c*, known as the pseudo-chance parameter, has been added to the model. With this incorporation, 3PL has a form of

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}. \quad (8)$$

Depending on what parameters are incorporated in a given IRT model, the ICCs would reflect different properties of the item.

If two examinees that are matched on the ability have different ICCs, the presence of DIF is concluded. DIF would be manifested in the difference between item parameters and, consequently, the area between two ICCs. The two key questions in DIF research regard the issue of quantifying the difference between ICCs and determining whether this difference possesses statistical significance. The DFIT framework, which will be discussed next, provides an approach to measure the difference between two ICCs and to determine whether this difference is statistically significant.

*DFIT framework and its indices*

Differential Test and Item Functioning (DFIT) was originally proposed by Raju, van der Linden, and Fleer (1995). The framework uses the IRT parameters to define and

measure DIF. The use of IRT parameters compares DFIT to other measures of DIF in the IRT family.

The DFIT framework proposes two indices of differential item functioning: non-compensatory (NCDIF) and compensatory (CDIF). The DFIT framework also provides the measure of the total bias in the test: differential test functioning (DTF). The computational constitution of each of these indices is examined next.

Assuming that there exist observed responses to a test of some ability, the responses constitute matrix $\mathbf{A}$ with the dimension of $n \times N$, where $n$ is the number of items on the test, and $N$ is the number of examinees. The sample of subjects is composed of at least two groups of people differing on a specific socio-demographic characteristic (e.g. gender, race, and etc.). Traditional DIF practices assign the minority group into the focal group and the majority group to the reference group. The logic is that the focus of the analysis is on determining whether a test and its items behave differently for members of this group compared to the reference group.

In the first computational step, the IRT software, such as BILOG-MG, MULTILOG, or PARSCALE is used to estimate item parameters (*a, b, c*) for each item and the ability ($\theta_s$) for each subject. One can use a number of algorithms for this estimation (maximum likelihood estimation (MLE), joint maximum likelihood estimation (JMLE), marginal maximum likelihood estimation (MMLE), Bayesian methods, and etc.). However, MMLE seems to be the most frequently used, even when alternative algorithms are available (Lee & Terry, 2004). The MMLE is more accurate than other estimation methods because it relies on the properties of ability distribution for the initial

estimates. However, it is more computationally intense than other methods and requires a large number of examinees to approximate the distribution of ability properly (Hambleton & Swaminathan, 1985).

After the initial estimates for $\theta_s$ are obtained from the total sample, the data file is split based on the group membership. In their original paper, Raju, van der Linden, and Fleer (1995) were vague on this issue. They only indicated that the item responses for each group were *simulated* by using equal $\theta$. No indication was given pertaining the estimates of which group (total, reference, or focal) should be used in the analysis of the real data. However, due to the fact that abilities of both focal and reference group come from the same normal distribution, it seems fair to assume that total group estimation of person parameters would be the most appropriate. The two resulting groups (focal and reference) must be checked for the equivalence of the ability distribution in order to avoid possible impact items. The problem of inequality of distribution between groups is still a threatening issue for theoreticians; this problem is avoided in simulated data.

Next, item parameters are estimated for each groups individually and brought on the same metrics. The $\theta_s$-values from the total group are used as initial estimates for parameter calibration. Thus, each item would have two sets of item parameters: one estimated for the focal group and another for the reference group. Ability estimates, as it was noted, should be invariant across groups and are not dependent on the group membership.

After ability and item parameters are estimated and brought onto the same scale, one can calculate probabilities of $\theta_s$ based on the model selected for the data. Two parameter logistic model (2PL) will be used for this illustration:

$$P_i(\theta_s) = \frac{1}{1 + e^{-Da_i(\theta_s - b_i)}},$$

where $P_i(\theta_s)$ is the probability of success for examinee *s,* with trait level $\theta$ on item *i*, $a_i$ is the discriminating parameter for item *i,* and $b_i$ is the difficulty parameter for item *i. D* is a scaling constant and equals to 1.701 in the 2PL and 3PL models. All the notation and equations (and derivatives of them) to follow are taken from the original paper by Raju et al. (1995), unless specified otherwise.

Naturally, with two sets of item parameters two equivalent probability estimates ( $P_i(\theta_s)$ ) for each examinee per item would be available: one calculated with parameters for the focal group and one for the reference group, namely

$$P_{iF}(\theta_s) = \frac{1}{1 + e^{-Da_{iF}(\theta_s - b_{iF})}} \qquad \text{and} \qquad P_{iR}(\theta_s) = \frac{1}{1 + e^{-Da_{iR}(\theta_s - b_{iR})}}.$$

Notice that $\theta_s$ is the same for both formulae. As it was pointed out earlier, since abilities of both focal and reference groups belong to the same distribution, there is only one $\theta_s$-estimate per examinee. Therefore, the remaining source of possible discrepancy can belong only to the properties of the item. As shown in the discussion of principles of IRT, these properties are quantified by item parameters.

The difference between these two probabilities for each item would constitute the fundamental form of DIF measurement. All indices in the DFIT framework will rely on this difference for computation. Let this difference be represented by

$$d_{is} = P_{iF}(\theta_s) - P_{iR}(\theta_s)\,. \tag{9}$$

The value of $d_{is}$ tells whether item performs differently when the examinee is treated as a member of the focal group rather than the reference group. In other words, it would tell what the difference in probabilities of answering this item correctly was if the subject were treated as a member of different groups. It is possible then to compute such index for every item for a particular examinee.

In IRT, the "true" score, $T_s$, of the examinee on the test, also known as the expected proportion correct (EPC), can be expressed as

$$T_s = \sum_{i=1}^{n} P_i(\theta_s)\,, \tag{10}$$

where $n$ is the total number of items on the test. Thus, each examinee would have two true scores,

$$T_{sF} = \sum_{i=1}^{n} P_{iF}(\theta_s)$$

in the focal group, and

$$T_{sR} = \sum_{i=1}^{n} P_{iR}(\theta_s)$$

in the reference group. Although, the subject from the focal group has only one $\theta$ - estimate, two "true" scores are available because two different sets of item parameters

may be used in calculating the probability ( $P_{iF}(\theta_s)$ and $P_{iR}(\theta_s)$ ). The difference between $T_{sF}$ and $T_{sR}$ is defined by Raju et al. as $D_s$. Thus,

$$D_s = T_{sF} - T_{sR}. \tag{11}$$

If $T_{sF} = T_{sR}$, it is concluded that the "true" score (EPC) of an examinee does not depend on the group membership. In other words, parameters estimated individually for each group do not change the function of $P_i(\theta_s)$. The properties of test and items are equivalent in respect to both groups. Therefore, one can see the evidence that the test does not function differently (favors or disfavors) on the examinee in question. In this way, the measure of DTF on the examinee level can be defined as $(T_{sF} - T_{sR})^2$. The measure of DTF across all examinees is defined as

$$DTF = \underset{F}{E}(T_{sF} - T_{sR})^2, \tag{12}$$

where expectation is taken over the focal group. The subject of expectation, however, is arbitrary because in simulation the number of examinees in the focal group will always equal to the number of examinees in the reference group. Using formula (11), this definition can be rewritten as

$$DTF = \underset{F}{E} D_s^2 = \int_{\theta} D_s^2 \, f_F(\theta) d\theta = \sigma_D^2 + (\mu_{TF} - \mu_{TR})^2 = \sigma_D^2 + \mu_D^2, \tag{13}$$

where $f_F(\theta)$ is the density function of $\theta$ in the focal group, and $\mu_{TF}$ and $\mu_{TR}$ are the mean (EPC) of examinees in the focal and reference groups, respectively. $D_s$ can also be defined as $\sum_{i=1}^{n} d_{is}$. Thus, the definition of DTF can be rewritten as

$$DTF = E[(\sum_{i=1}^{n} d_{is})^2] = E[\sum_{i=1}^{n} (P_{iF} - P_{iR})]^2 = \\ = E[(P_{1F} - P_{1R}) + (P_{2F} - P_{2R}) + ... + (P_{nF} - P_{nR})]^2. \quad (14)$$

Using formula (13), DTF can be transformed into

$$DTF = \sum_{i=1}^{n} [Cov(d_i, D_s) + \mu_{d_i}\mu_{D_s}],$$

where $Cov(d_i, D_s)$ is the covariance between the difference in item probabilities for item $i$ ($d_i$) computed for each examinee and the difference between the two EPCs ($D_s$) for each examinee. The $\mu_{d_i}$ and $\mu_{D_s}$ are the means of $d_i$ and $D_s$, respectively. By removing the summation across items, one receives a definition of DTF on the item level, which Raju et at.(1995) called compensatory differential item functioning (CDIF). CDIF is thus defined as

$$CDIF_i = E(D_s d_i) = Cov(d_i, D_s) + \mu_{d_i}\mu_{D_s}. \quad (15)$$

Two important concepts regarding the CDIF index must be immediately brought to attention. First is its compensatory or additive nature. As seen from formula (14), the difference between item probabilities ($d_i$) may take on negative values in cases where item favors the member of the focal group. These values are summed across items for an examinee to produce $D_s$, which is only then squared to produce DTF index at the examinee level ($D_s^2$). Because the squaring occurred after the addition, the negative and positive values of $d_i$ can cancel each other out. Therefore, it is theoretically possible to

have DTF index of 0, indicating the absence of bias in the test and yet to observe bias on the item level. Of course, such biases would have to be symmetric and of matching magnitude.

The second property of CDIF is obvious from equation (15). The computation utilizes both the total bias of an item across examinees ( $d_i$ ) and the total bias associated with each examinee across all items ( $D_s$ ). Therefore, it does not assume that all other items on the test are unbiased. This sets CDIF apart from other IRT-based DIF measures, which all share this assumption (Flowers, Oshima, & Raju, 1999).

Non-compensatory differential item functioning (NCDIF), on the other hand, does hold assumption that all items in the test other than item $i$ are completely unbiased. It does not include information from other items, and concerns solely the difference between probabilities $d_i$ for item $i$ across all examinees. Thus, Raju et al.(1995) define NCDIF as

$$NCDIF_i = \int_{-\infty}^{\infty} d_{is}^2 \, f_F(\theta) d\theta = E(d_i^2) = \sigma_{d_i}^2 + \mu_{d_i}^2 , \qquad (16)$$

where $\sigma_{d_i}^2$ is the variance of difference in probability for item $i$ across all examinees, and $\mu_{d_i}$ is its mean. The computational mechanics of CDIF and NCDIF are presented in Figure 1. Notice that the values of $D_s^2$ do not change from item to item.

Three indices in the DFIT framework have been introduced: DFT, CDIF, and NCDIF. The logic of these indices and their relative computational simplicity make DFIT an attractive choice for practitioners in investigating measurement bias in tests. However,

the issue of significance testing has been at the heart of criticism and research in the DFIT framework. Raju et al. (1995) showed that NCDIF would have a $\chi^2$ distribution. This allows one to determine whether the observed magnitude of NCDIF differs significantly from chance. The next section will examine the logic of significance testing of NCDIF index and problems associated with such testing.

Item: i=1

| Subject | $D_s^2$ | $d_i$ | $D_s d_{is}$ | $d_i^2$ |
|---|---|---|---|---|
| 1 | 0.09970 | 0.024 | 0.0024 | 0.0006 |
| 2 | 0.05373 | -0.014 | -0.0007 | 0.0002 |
| 3 | 0.15744 | 0.008 | 0.0012 | 0.0001 |
| 4 | 0.20579 | 0.025 | 0.0052 | 0.0006 |
| 5 | 0.04646 | -0.016 | -0.0007 | 0.0002 |
| | | $\in$ = | 0.0015 | 0.0003 |
| | | | $CDIF_i$ | $NCDIF_i$ |

Item: i=1

| Subject | $D_s^2$ | $d_i$ | $D_s d_{is}$ | $d_i^2$ |
|---|---|---|---|---|
| 1 | 0.09970 | -0.003 | -0.0003 | 0.0000 |
| 2 | 0.05373 | 0.028 | 0.0015 | 0.0008 |
| 3 | 0.15744 | 0.018 | 0.0028 | 0.0003 |
| 4 | 0.20579 | 0.002 | 0.0004 | 0.0000 |
| 5 | 0.04646 | 0.029 | 0.0013 | 0.0008 |
| | | $\in$ = | 0.0012 | 0.0004 |
| | | | $CDIF_i$ | $NCDIF_i$ |

FIGURE 1. Demonstration of computing CDIF and NCDIF indices.

*Significance testing of NCDIF*

In their original paper introducing the DFIT framework, Raju et al. (1995) suggested using the $\chi^2$-test or the t-test for determining significance of the NCDIF index. Thus, given that $d_i$ is normally distributed and has a finite variance, the $\chi^2$-test for NCDIF is defined as

$$\chi^2_{N_F} = \frac{N_F(NCDIF_i)}{\sigma_{d_i}}$$

with $N_F$ degrees of freedom, where $N_F$ is the number of examinees in the focal group. Under the same assumptions the t-test for NCDIF is defined as

$$t = \frac{(N_F)^{1/2}(\mu_{d_i})}{\sigma_{d_i}}$$

with $N_F$-1 degrees of freedom.

However, as exploratory Monte Carlo studies have shown, the NCDIF index in $\chi^2$-tests is overly sensitive to large sample sizes (Fleer, 1993). Under a zero-bias condition and .01 significance level the rate of false positive significantly exceeded 1%. To avoid this problem, Fleer identified the size (.006) of NCDIF index that would yield approximately 1% false positives at .01 alpha level in $\chi^2$-test. Thus, a two-fold rule was developed: the item was considered to exhibit differential functioning if the absolute value of the NCDIF index exceeded .006 and produced significance in the $\chi^2$-test. The significance level refers to the alpha level that is adopted in the $\chi^2$-test.

Studies that investigated extensions of NCDIF to polytomous IRT models were forced to develop the cutoff scores in a similar fashion through exploratory simulations (Oshima, Raju, & Flowers, 1997). As Oshima, Raju, and Nanda (2006) imply, the appropriate cutoff score of NCDIF would be idiosyncratic to each case. It is obviously impossible for practitioners to conform their tests to the examined IRT paradigm in regards to models, data type, test length, and sample sizes, which all supposedly affect the cutoff value. Chamblee (1998) has demonstrated that in a dichotomous case such varying factors as sample size and the type of IRT model (1PL, 2PL, or 3PL) will influence the cutoff scores. Both the size of sample and the number of parameters in IRT model were found to be directly related to the magnitude of the cutoff scores.

The viable solution would be to identify appropriate cutoff values for each particular case. However, the paradigm proposed by Fleer (1993) for such purposes and adopted afterwards by other researches (Bolt, 2002; Flowers, Oshima, & Raju, 1999) appears to be overly complex and technically challenging for typical practitioners. Such methods involved simulating the pairs of datasets with observed response that exhibited the characteristics of zero DIF and cycling through the entire DFIT computational sequence to achieve a single NCDIF index. Naturally, a large number of simulations would be necessary to approximate the true properties of NCDIF distribution for the given case and derive the appropriate cutoff score. Such methods would typically produce a single cut-off score for all items. Oshima et al. (2006) proposed a new method of deriving the cutoff values from simulations of the given case called item parameter replication (IPR) method. Instead of simulating response sets, it generated item parameters with similar properties of distribution and covariance structure.

*Oshima, et al. (2006) study*

In their original study, Oshima, Raju, and Nanda (2006) introduced item parameter replication method (IPR) for determining NCDIF cutoff values and conducted a simulation study investigating different conditions that would influence these cutoff values. The SAS macro "DIFCUT" (Nanda, Oshima, & Gagne, in press) was used to simulate the cutoff scores and to compute the DFIT indices.

The conditions investigated in the study were as follows:

- Two test length (n = 20, n = 40),
- Three sample size combination (500:500, 1000:1000, 500:1000, focal and reference groups, respectively),
- Three levels of DIF (0%, 10%, and 20%),
- Three IRT models (1PL, 2PL, and 3PL),
- Two ability distributions: No impact (the mean of theta $= \theta_{focal} = \theta_{reference} = 0$) and impact condition ($\theta_{focal} = -.5$, $\theta_{reference} = 0$).

The design would result in a total of 108 test conditions ($2\times 3\times 3\times 3\times 2$). However, the 20% DIF condition was not applied to the 1PL model because nonuniform bias items were already included in this condition. Thus the study considered 96 conditions. The items parameters that were used for generating item responses are the same as in the Raju at al.(1995) and Oshima et al.(2006) and can be found in the corresponding papers.

Although, the overall performance of the new methods seemed to be an improvement over the traditional $\chi^2$-methods used earlier, certain problems were encountered that required further investigation. One of the most evident problems

concerned a high false positive (FP) and false negative (FN) rate for the 3PL model in the 20% DIF condition. False positives are items that were recognized as exhibiting significant NCDIF, but in reality did not. False negatives are the items that did in fact possess NCDIF, but were missed as such by the analysis. The authors indicate poor performance of the 3PL model only for the 40-item condition. However, in relation to the number of biased items and test's length, 3PL performs only slightly better in the 20-item condition than in the 40-item condition. Only half of the biased items (50%) were correctly identified in the N = 500:500 and N = 500:1000 conditions for the 40-items test, and one forth (25%) in the N = 1000:1000 condition.

The authors speculatively attribute such poor performance to several factors including smaller magnitude of DIF in the condition (20% bias, n = 40), small sample size of the focal group (N = 500), and the way the latter influence simulation and estimation of the *c*- parameter. The value of the *c*- parameter might also have exerted a certain influence on the accuracy of the result. Oshima et al. used a fixed *c*-value of .20 for all conditions employing 3PL.

Another important issue concerns the source of item parameters and their covariance structures used for simulating the zero-DIF condition and deriving the cutoff scores. The study used the focal group only with item parameters and covariance structures drawn from the focal group as well. Based on their and the study by Chamblee (1998), the authors do agree that it is unclear which source would provide more appropriate cutoff scores: the focal, reference, total groups, or some combination of them. Such judgment seems impossible from theoretical perspective and must rely on empirical exploration (Oshima et al., 2006). Questions have also been raised regarding the

influence that test's difficulty level and model fit exert onto the item standard errors obtained in estimation.

Lastly, after examination of DIFCUT programming code, a small inaccuracy has been discovered. In computing probability values, the authors used the formula for 3PL across all the IRT models (1PL, 2PL, and 3PL). Under the condition of stable null *c* parameter ($c = 0$), this formula would be reduced to and produce identical results as the 2PL formula. This, however, is not the case with 1PL. The 3PL formula does not have the capacity to be reduced under $c = 0$ and $a = 1$ to 1PL formula due to the different scaling constant (*D)* value of 1 for the 1PL model. In 2PL and 3PL this constant equals to 1.701, but in 1PL such constant is not applied. After applying both formulas to the identical sample data, a discrepancy in probability values amounted to the average of .025 ($M = .025$, $SD = .011$, $MAX = .04$.) in 1,000 paired calculations.

*The Purpose of the Current Study*

The current study aims to replicate the paradigm and proceedings of Oshima et al. (2006) and to introduce some new manipulations of conditions in order to further investigate significance testing of NCDIF index in the DFIT framework. Specifically, the purpose of the current study will include:

1) Two new conditions will be introduced to the existing paradigm with different values of *c* parameter (.1 and .5) in order to investigate the effect of the c-parameter on significance testing of NCDIF.
2) The sources of item parameters and their covariance structures will be manipulated to determine the most appropriate combination for deriving NCDIF cutoff scores. The new conditions will be applied: all (person and

item) estimates and covariance structure come from reference group, instead of the focal group, as in the original study.

3) The original study in respect to the 1PL model will be replicated with the corrected formula.

# CHAPTER II

## Method

The item parameter replication method differs from previous methods of determining significance of NCDIF indices in a number of important respects. First, instead of generating multiple datasets with specified IRT characteristics, only the item parameters are replicated based on the estimates from a single IRT analysis. Second, IPR allows determining the NCDIF cutoff score individually for each item, thus furthering the accuracy of DIF detection. Since the IPR procedures are identical for all items, the steps of IPR for a single item will be described next after the description of data generation.

### *Generating Data for Original Conditions*

Based on the characteristics of DIF to be investigated, the differences in item parameters for the reference and focal groups are quantified. Figure 2 presents a partial recreation of the DIF conditions used in both the original study by Raju et al. (1995) and Oshima et al.(2006). The blank spaces left in the focal group table signify that the parameters of the corresponding items were the same for both groups. The discrepancies in item parameters are arbitrary and depend on the research objectives. Both item 5 and 10 in Figure 2 represent uniform bias. Uniform bias is the condition when the focal and reference groups differ only on the *b* parameter. Higher values of *b* in the focal group indicate that items 5 and 10 were more difficult to pass for members of the focal group relative to the reference group.

| | Reference | | Focal | |
|---|---|---|---|---|
| Item | a | b | a | b |
| 1 | 0.55 | 0.00 | | |
| 2 | 0.55 | 0.00 | | |
| 3 | 0.73 | -1.04 | | |
| 4 | 0.73 | 1.04 | | |
| 5 | 0.73 | 0.00 | 0.73 | 1 |
| 6 | 0.73 | 0.00 | | |
| 7 | 0.73 | 0.00 | | |
| 8 | 0.73 | 0.00 | | |
| 9 | 0.73 | 1.04 | | |
| 10 | 0.73 | 1.04 | 0.73 | 1.54 |

FIGURE 2. Partial recreation of item parameters used in Oshima et al. (2006)

A column vector with dimensions of $N \times 1$, where $N$ is the number of examinees, is created. Each element is drawn at random from a normal distribution with $\mu = 0$, and $\sigma = 1$. Thus, the vector represents a selection of underlying trait levels ($\theta$) for $N$ examinees drawn from the population.

Then, $P_i(\theta_s)$ is calculated for each item and examinee with respect to the focal or reference group parameters. Figure 3 shows the sequence of simulating item responses for item 10 for the first ten examinees.

A random number from a normal distribution is drawn and used for comparison to the calculated probability value ( $P_i(\theta_s)$ ). If a number taken at random from a uniform distribution is smaller than the $P_i(\theta_s)$-value, the value of "1" is assigned as the observed response produced by this examinee on this item. This sequence is executed for all examinees and items, producing an $N \times n$ matrix, where $N$ is the number of examinees and $n$ is the number of items. The matrix contains simulated responses of examinees with specified abilities and parameters.

*Item Parameter Replication Method*

After the dataset for each desired study condition has been generated, an IRT analysis is run through one of the available IRT software such as PARSCALE or BILOG. These procedures are not discussed in this work. The IRT analysis produces output files containing item parameters and their covariance structures for both reference and focal groups. The following 9 steps are adopted from the original study of Oshima et al.(2006) and describe the IPR method.

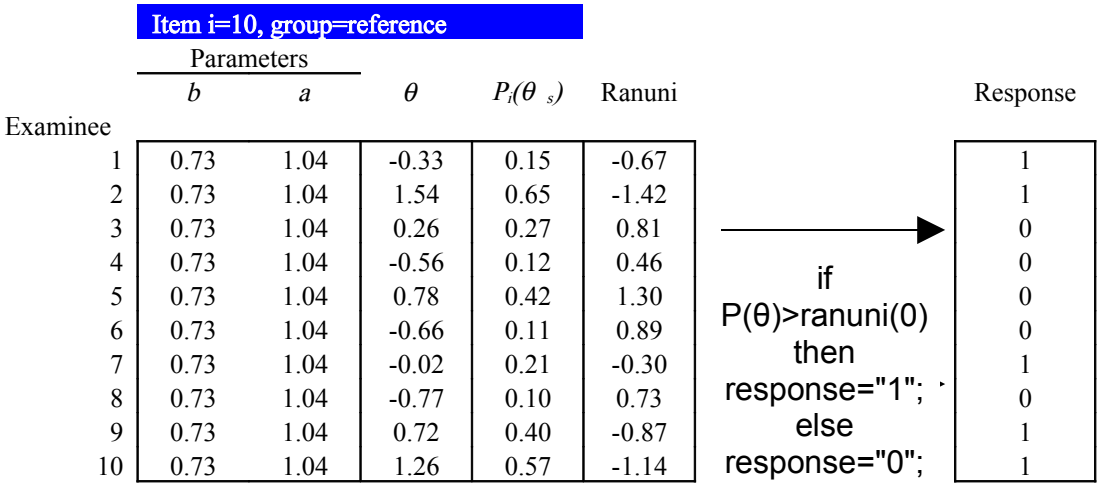


FIGURE 3. Demonstration of generating item responses from set conditions.

*Step 1.* Let column vector $M_i$ with dimensions $k\times1$ contains item parameters of item $i$ for the focal group, where $k$ is number of parameters according to the selected IRT model. Thus, for a 3PL model, vector $M_i$ has the form of

$$M_i = \begin{bmatrix} b_i \\ a_i \\ c_i \end{bmatrix}.$$

The variances and covariances of item parameters for each item, obtained through iterative estimation, can be grouped in a variance/covariance matrix $\mathbf{V}_i$:

$$V_i = \begin{bmatrix} \sigma^2_{b_i} & \sigma_{b_i a_i} & \sigma_{b_i c_i} \\ \sigma_{b_i a_i} & \sigma^2_{a_i} & \sigma_{a_i c_i} \\ \sigma_{b_i c_i} & \sigma_{a_i c_i} & \sigma^2_{c_i} \end{bmatrix}.$$

This covariance matrix can be easily transformed into correlation matrix

$$R_i = \begin{bmatrix} 1 & \rho_{b_i a_i} & \rho_{b_i c_i} \\ \rho_{b_i a_i} & 1 & \rho_{a_i c_i} \\ \rho_{b_i c_i} & \rho_{a_i c_i} & 1 \end{bmatrix}.$$

By assuming that $\mathbf{R}_i$ is positive definite, one can decompose the $\mathbf{R}_i$ matrix into the product of triangular matrix $\mathbf{T}$ and its inverse. Thus, in case of the 3PL model, the upper triangular matrix $\mathbf{T}_i$ will have the form of

$$T_i = \begin{bmatrix} 1 & \rho_{b_i a_i} & \rho_{b_i c_i} \\ 0 & \sqrt{1-\rho^2_{b_i a_i}} & \dfrac{\rho_{a_i c_i} - \rho_{b_i c_i}\rho_{b_i c_i}}{\sqrt{1-\rho^2_{b_i a_i}}} \\ 0 & 0 & \sqrt{1-\left[\rho^2_{b_i c_i} + \dfrac{(\rho_{a_i c_i} - \rho_{b_i c_i}\rho_{b_i c_i})^2}{(1-\rho^2_{b_i a_i})}\right]} \end{bmatrix}.$$

*Step 2.* Let $\mathbf{x}_{1i}$ represent a column vector with dimensions $k \times 1$, where $k$ is the number parameters of the corresponding IRT model. For the 3PL, the vector $\mathbf{x}_{1i}$ will contain three elements. Each element is drawn at random from an independent, standardized ($M = 0$, $SD = 1$) and normally distributed population. Let $\mathbf{x}_{2i}$ represent a second vector, of which dimensions and elements were defined in a similar fashion.

*Step 3.* Each vector $\mathbf{x}$ must contain random elements from a $k$-dimensional standardized multivariate normal distribution with a correlation structure for the k dimensions that are equivalent to the structure of $\mathbf{R}_i$ matrix. In order to achieve this transformation each vector is pre-multiplied by the transpose of triangular matrix $\mathbf{T}$, which contains this correlation structure for item $i$. The result is two column $\mathbf{z}$ vectors possessing the qualities of the desired multivariate distribution. Thus,

$$z_{1i} = T_i' x_{1i},$$

$$z_{2i} = T_i' x_{2i}.$$

*Step 4.* It is necessary to obtain a vector of $k$ values that come at random from sets of populations, in which interrelating properties are defined by original parameters' covariance structure. This has been achieved by step 3. Now $\mathbf{z}$ vectors must be adapted to reflect the means and variances of the original item parameters. This is done by a simple linear transformation:

$$y_{1i} = D_i^{1/2} z_{1i} + M_i,$$

$$y_{2i} = D_i^{1/2} z_{2i} + M_i.$$

In this way, vectors **y** represent item parameters that came from standardized multivariate distributions with $\mu_{1i} = b_i$, $\mu_{2i} = a_i$, and $\mu_{3i} = c_i$, and covariance structure defined by $\mathbf{R}_i$ matrix. Vector $\mathbf{y}_{1i}$ would represent simulated item parameters for the reference group for item i, and $\mathbf{y}_{2i}$, respectively, item parameters for the focal group. Please note, that the distribution from which both vectors come are identical. Therefore, any discrepancy between them must be attributed to sampling error exclusively. This would represent a condition when the true DIF is zero.

*Step 5.* Using the computations described earlier, it is now possible to compute the NCDIF indices for each item. With theta estimates from the focal group and two **y** vectors probability functions are computed. The differences in these pairs of probability functions will constitute the basis for computation of the NCDIF.

*Step 6.* Step 5 can be replicated as many times as necessary according to the needs of the study. However, Oshima et al. (2006) investigated the number of replications necessary to achieve stable cutoff scores. The indices seem to stabilize after 600 replications and do not differ substantially from results from 10,000 replications. The 1,000 replications were chosen as the appropriate number for their study. This study will follow their guidelines.

*Step 7.* All values of NCDIF obtained from the step 6 are sorted in ascending order and ranked. The $90^{th}$, $95^{th}$, $99^{th}$, and $99.9^{th}$ percentile of the scores are recorded to represent the cutoff values at significance levels of .10, .05, .01, and .001, respectively.

*Step 8.* Now the NCDIF index for item *i* obtained from the actual data is compared to the cutoff value at the selected alpha level to determine significance.

*Step 9.* Steps 1 through 8 are repeated for every item of the test. Because the algorithm is applied to each item individually, it allows establishing possibly different cutoff scores for each item according to the properties of its parameters.

*Generation of response data*

Based on the item parameters used by Oshima et al.(2006) the response data were generated with the SAS environment via the procedure spelled out in Raju et al. (1995). The procedure was identical to the RANGEN (Fleer et al., 1991) computer program that was described earlier. Each unique combination of item parameters generated a separate response dataset, which was used in the subsequent IRT analysis.

*IRT Analysis*

The generated datasets with simulated responses were read into the BILOG software. The results of BILOG analysis included ability estimates, item parameters estimates, and covariance structure of the latter for the total, reference, and focal group.

*NCDIF computation*

The DIFCUT SAS program, obtained from the Nanda et al. (in press), was used to derive the NCDIF cutoff scores, to compute the observed NCDIF, and to make a decision in regards to its significance. The rates of false positives and false negatives under the new conditions were compared to the results of the original study.

# CHAPTER III

## Results

Overall, the results of the current study were in accordance with the results of Oshima et al. (2006). Although certain differences were peculiar and would be discussed separately, the results of this replication of IPR method did support the usefulness of IPR in determining the significance of NCDIF indices in the DFIT framework.

Table 1 showed the number of false positives (FPs: identifying non-DIF items as having significant DIF) and false negatives (FNs; identifying DIF items as having no DIF). In the N = 500:500 condition, only 3PL with c = .20 and 20% DIF produced 1 false positive and 1 false negative cases. The N = 1000:1000 condition provided somewhat aberrant results, giving 2 false positives in the 1PL 10% DIF condition. The N = 500:1000 condition gave results similar to the results of the previous study in corresponding conditions; 1PL with 10% DIF produced 1 false positive and 3PL with 20% DIF produced 1 false positive and 1 false positive.

For the ease of convenience comparing these results to the results of the original study, the formatting was replicated. The IPR method performed well at identifying DIF items in all sample size conditions with 0% and 10% DIF: 5 FP cases and 4 FN cases in total. However, it did not perform well in the 20% DIF condition in both 20 and 40-item tests with 5 FP cases and 25 FN cases. The majority of FN cases came from the 40-item test condition. Oshima et al. (2006) admitted poor performance of DIFCUT macro only with the 3PL model for the 40-item 20% DIF test. The results of this study, however, indicated that poor detection accuracy of DIFCUT might not be localized to the 3PL case, but be inherent to the entire 20% DIF condition.

Table 1

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1PL | 2PL | 3PL(.2) | 1PL | 2PL | 3PL (.2) | 1PL | 2PL | 3PL (.2) |
| | | | | (a) Test length = 20 items** | | | | | |
| 0% | 0.0117 | 0.0127 | 0.0272 | 0.0060 | 0.0063 | 0.0191 | 0.0117 | 0.0127 | 0.0272 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0121 | 0.0130 | 0.0278 | 0.0060 | 0.0062 | 0.0240 | 0.0121 | 0.0130 | 0.0278 |
| FP | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0.0119 | 0.0134 | 0.0457 | 0.0059 | 0.0065 | 0.0229 | 0.0119 | 0.0134 | 0.0457 |
| FP | NA | 0 | 1 | NA | 1 | 0 | NA | 0 | 1 |
| FN | NA | 0 | 1 | NA | 0 | 0 | NA | 0 | 1 |
| | | | | (b) Test length = 40 items** | | | | | |
| 0% | 0.0114 | 0.0122 | 0.0244 | 0.0059 | 0.0059 | 0.0177 | 0.0114 | 0.0122 | 0.0244 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0117 | 0.0123 | 0.0250 | 0.0059 | 0.0059 | 0.0167 | 0.0117 | 0.0123 | 0.0250 |
| FP | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 20% | 0.0117 | 0.0121 | 0.0257 | 0.0058 | 0.0060 | 0.0154 | 0.0117 | 0.0121 | 0.0257 |
| FP | NA | 0 | 0 | NA | 1 | 0 | NA | 1 | 0 |
| FN | NA | 3 | 5 | NA | 2 | 4 | NA | 4 | 5 |

* NCDIF cutoff at 99$^{th}$ percentile

** Not linked

Table 2 showed FPs and PNs for two new experimental conditions: 3PL with c = 0.1 and 3PL with c = 0.5. The results indicated that with decreased pseudo-guessing parameter (c = 0.1) the increase in accuracy of detection did not follow. In fact, when c = 0.1 and c = 0.2 conditions were compared, the latter outperformed in terms of both FPs and FNs in many cases. The discussion of this finding is provided later. The extreme condition of c = 0.5, however, does suggest that extreme values of pseudo-guessing parameter undermine accuracy of the model. This is consistent with the general research in the field; 3PL models have been consistently found to be the most problematic.

Although the numbers of FPs and FNs in this study are comparable to their counterparts in the Oshima et al.'s study, the overall NCDIF cutoff points were considerably higher. Oshima et al.(2006) reported the NCDIF cutoff values of .0054, . 0074, and .0100 for 1PL, 2PL, and 3PL, respectively for the 0% DIF condition with N = 500:500. The values for corresponding conditions in the current study were .0117, .0127, and .0272. This observation is important in a number of respects. First, the magnitude of cutoff scores did not influence the accuracy of DIF detection. In previous studies, higher values of cutoff scores were presented as possible cause of increased FNs in general, particularly in the 3PL models with 40-item tests. Second, the effects of sample size and the type of IRT models on cutoff scores were similar to those of Oshima et al., in spite of the increased cutoff values. The smaller sample size resulted in higher cutoff values. It was also noted that the more parameters in the IRT model, the higher cutoff values.

Table 2

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0272 | 0.0284 | 0.0437 | 0.0191 | 0.0210 | 0.0623 | 0.0272 | 0.0284 | 0.0437 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0278 | 0.0256 | 0.0355 | 0.0240 | 0.0202 | 0.0538 | 0.0278 | 0.0256 | 0.0355 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 |
| 20% | 0.0457 | 0.0274 | 0.0449 | 0.0229 | 0.0164 | 0.0624 | 0.0457 | 0.0274 | 0.0449 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 2 | 3 | 4 | 1 | 1 | 3 | 2 | 3 | 4 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0244 | 0.0264 | 0.0407 | 0.0177 | 0.0152 | 0.0372 | 0.0244 | 0.0264 | 0.0407 |
| FP | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0250 | 0.0259 | 0.0399 | 0.0167 | 0.0186 | 0.0600 | 0.0250 | 0.0259 | 0.0399 |
| FP | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 |
| 20% | 0.0257 | 0.0308 | 0.0529 | 0.0154 | 0.0198 | 0.0558 | 0.0257 | 0.0308 | 0.0529 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 5 | 5 | 8 | 4 | 3 | 7 | 5 | 6 | 7 |

* NCDIF cutoff at 99$^{th}$ percentile

** Not linked

Tables 3 and 4 presented FPs and FNs for the same conditions as in Tables 1 and 2, respectively, with the 99.9$^{th}$ percentile of NCDIF as the cutoff score. Although the 99.9$^{th}$ percentile as cutoff score reduced the number of biased items identified, the FP cases virtually disappeared. A more conservative cutoff score may be useful in determining the most reliable DIF detection. Using a higher percentile rank for NCDIF cutoff score, however, considerably increased the FNs rate. On the other hand, items marked as biased at 99.9$^{th}$ NCDIF percentile rank would be most probably identified correctly. In other words, the probability of identifying an item as biased when it is in fact unbiased would be minimal.

Another important aspect must be pointed out: linking of item parameters. The algorithm of the DIFCUT macro and the logic of the NCDIF index assumed that after the cutoff scores of NCDIF have been determined through item replication, the actual NCDIF indices are computed using the originally estimated item parameters. Due to the fact that the data from focal and reference groups were analyzed in BILOG separately, the item parameters from these two estimations would not be on the same scale. For this purpose, the IPLink software was utilized to bring the item parameters of the reference group on the same scale with the item parameters of the focal group. After parameters were brought on the same scale, the actual NCDIF indices were computed by applying two sets of item parameters (focal and reference) to each member of the focal group. As the means to explore possible effects of different scales of item parameters on the NCDIF detection, a separate condition was run in which the item parameters in the computation of the actual NCDIF remained on the original scales. Unlinked parameters, surprisingly, yielded equally good results.

Table 3

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1PL | 2PL | 3PL(.2) | 1PL | 2PL | 3PL (.2) | 1PL | 2PL | 3PL (.2) |
| | | | | (a) Test length = 20 items** | | | | | |
| 0% | 0.0117 | 0.0127 | 0.0272 | 0.0060 | 0.0063 | 0.0191 | 0.0117 | 0.0127 | 0.0272 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0121 | 0.0130 | 0.0278 | 0.0060 | 0.0062 | 0.0240 | 0.0121 | 0.0130 | 0.0278 |
| FP | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0.0119 | 0.0134 | 0.0457 | 0.0059 | 0.0065 | 0.0229 | 0.0119 | 0.0134 | 0.0457 |
| FP | NA | 0 | 0 | NA | 0 | 0 | NA | 0 | 0 |
| FN | NA | 0 | 2 | NA | 0 | 1 | NA | 0 | 2 |
| | | | | (b) Test length = 40 items** | | | | | |
| 0% | 0.0114 | 0.0122 | 0.0244 | 0.0059 | 0.0059 | 0.0177 | 0.0114 | 0.0122 | 0.0244 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0117 | 0.0123 | 0.0250 | 0.0059 | 0.0059 | 0.0167 | 0.0117 | 0.0123 | 0.0250 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 2 |
| 20% | 0.0117 | 0.0121 | 0.0257 | 0.0058 | 0.0060 | 0.0154 | 0.0117 | 0.0121 | 0.0257 |
| FP | NA | 0 | 0 | NA | 0 | 0 | NA | 1 | 0 |
| FN | NA | 3 | 7 | NA | 2 | 4 | NA | 4 | 7 |

* NCDIF cutoff at 99.9$^{th}$ percentile

** Not linked

Table 4

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0272 | 0.0284 | 0.0437 | 0.0191 | 0.0210 | 0.0623 | 0.0272 | 0.0284 | 0.0437 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0278 | 0.0256 | 0.0355 | 0.0240 | 0.0202 | 0.0538 | 0.0278 | 0.0256 | 0.0355 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 |
| 20% | 0.0457 | 0.0274 | 0.0449 | 0.0229 | 0.0164 | 0.0624 | 0.0457 | 0.0274 | 0.0449 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 2 | 3 | 4 | 1 | 1 | 3 | 2 | 3 | 4 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0244 | 0.0264 | 0.0407 | 0.0177 | 0.0152 | 0.0372 | 0.0244 | 0.0264 | 0.0407 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0250 | 0.0259 | 0.0399 | 0.0167 | 0.0186 | 0.0600 | 0.0250 | 0.0259 | 0.0399 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 1 | 1 | 3 | 1 | 1 | 2 | 2 | 1 | 3 |
| 20% | 0.0257 | 0.0308 | 0.0529 | 0.0154 | 0.0198 | 0.0558 | 0.0257 | 0.0308 | 0.0529 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 7 | 6 | 8 | 4 | 6 | 8 | 7 | 6 | 8 |

* NCDIF cutoff at $99.9^{th}$ percentile

** Not linked

Tables 1 through 4 contained results from unlinked data; tables 5 through 8 recreated the same conditions for linked data. Linking of the item parameters did decrease the number of FNs, however it also increased the number of FPs by a larger proportion. The increase in FPs, it must be noted, was primarily localized in the unequal sample size conditions. The overall conclusions, however, regarding the performance of IPR were not affected by linking.

Instead of generating a separate response dataset for each cycle of the simulation, the IPR method created a large number of sets of parameters, which were used in computing the NCDIF cutoff values. Each of such parameter sets was created using the original item parameters and the covariance structure. One of the manipulations of the current study was the use of parameters and their covariance structures from the reference group. Such condition had already been tested by Oshima et al., but only under specific circumstances. The current study applied this manipulation to the all design conditions.

Table 9 and 10 showed FPs and FNs for conditions in which the reference group parameters were used for parameter replication and deriving the NCDIF cutoff scores at the $99^{th}$ percentile. Numbers of FPs and FNs for the $99.9^{th}$ percentile cutoff were presented in tables 11 and 12. Results indicated that using the reference group information to simulate parameters for the NCDIF cutoff score was as effective as using the focal group. It yielded slightly lower number of FNs and slightly higher number of FPs. These findings supported similar conclusions drawn in Oshima et al.

One of the objectives of the current study was to replicate Oshima et al. with the correction to the formula used by DIFCUT for the 1PL model. In the original macro, the general formula of 3PL was used for each model, with the anticipation of its reduction

under conditions of constant parameters. Formula for 1PL was rewritten as described earlier. No significant differences were observed after the implementation of the correction. The resulting numbers were virtually identical and therefore not reported. The only benefit of the corrected formula was the reduced computer time: by reducing the number of mathematical operations performed it decreased the amount of computer time necessary for the entire simulation.

Table 5

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1PL | 2PL | 3PL(.2) | 1PL | 2PL | 3PL (.2) | 1PL | 2PL | 3PL (.2) |
| | | | | (a) Test length = 20 items** | | | | | |
| 0% | 0.0117 | 0.0127 | 0.0272 | 0.0060 | 0.0063 | 0.0191 | 0.0117 | 0.0127 | 0.0272 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 2 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0121 | 0.0130 | 0.0278 | 0.0060 | 0.0062 | 0.0240 | 0.0121 | 0.0130 | 0.0278 |
| FP | 0 | 1 | 0 | 1 | 2 | 0 | 3 | 1 | 1 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0.0119 | 0.0134 | 0.0457 | 0.0059 | 0.0065 | 0.0229 | 0.0119 | 0.0134 | 0.0457 |
| FP | NA | 6 | 0 | NA | 5 | 0 | NA | 3 | 1 |
| FN | NA | 0 | 1 | NA | 0 | 0 | NA | 0 | 1 |
| | | | | (b) Test length = 40 items** | | | | | |
| 0% | 0.0114 | 0.0122 | 0.0244 | 0.0059 | 0.0059 | 0.0177 | 0.0114 | 0.0122 | 0.0244 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 3 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0117 | 0.0123 | 0.0250 | 0.0059 | 0.0059 | 0.0167 | 0.0117 | 0.0123 | 0.0250 |
| FP | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 1 | 1 |
| FN | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 20% | 0.0117 | 0.0121 | 0.0257 | 0.0058 | 0.0060 | 0.0154 | 0.0117 | 0.0121 | 0.0257 |
| FP | NA | 2 | 2 | NA | 6 | 0 | NA | 1 | 0 |
| FN | NA | 2 | 4 | NA | 0 | 3 | NA | 3 | 4 |

* NCDIF cutoff at 99$^{th}$ percentile

** Linked

Table 6

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0272 | 0.0284 | 0.0437 | 0.0191 | 0.0210 | 0.0623 | 0.0272 | 0.0284 | 0.0437 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 1 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0278 | 0.0256 | 0.0355 | 0.0240 | 0.0202 | 0.0538 | 0.0278 | 0.0256 | 0.0355 |
| FP | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0.0457 | 0.0274 | 0.0449 | 0.0229 | 0.0164 | 0.0624 | 0.0457 | 0.0274 | 0.0449 |
| FP | NA | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| FN | NA | 2 | 2 | 0 | 1 | 2 | 1 | 2 | 3 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0244 | 0.0264 | 0.0407 | 0.0177 | 0.0152 | 0.0372 | 0.0244 | 0.0264 | 0.0407 |
| FP | 0 | 0 | 0 | 0 | 2 | 3 | 3 | 1 | 1 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0250 | 0.0259 | 0.0399 | 0.0167 | 0.0186 | 0.0600 | 0.0250 | 0.0259 | 0.0399 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| FN | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 |
| 20% | 0.0257 | 0.0308 | 0.0529 | 0.0154 | 0.0198 | 0.0558 | 0.0257 | 0.0308 | 0.0529 |
| FP | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| FN | 4 | 5 | 7 | 3 | 3 | 6 | 4 | 5 | 7 |

* NCDIF cutoff at 99$^{th}$ percentile

**Linked

Table 7

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1PL | 2PL | 3PL(.2) | 1PL | 2PL | 3PL (.2) | 1PL | 2PL | 3PL (.2) |
| | | | | (a) Test length = 20 items** | | | | | |
| 0% | 0.0117 | 0.0127 | 0.0272 | 0.0060 | 0.0063 | 0.0191 | 0.0117 | 0.0127 | 0.0272 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0121 | 0.0130 | 0.0278 | 0.0060 | 0.0062 | 0.0240 | 0.0121 | 0.0130 | 0.0278 |
| FP | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0.0119 | 0.0134 | 0.0457 | 0.0059 | 0.0065 | 0.0229 | 0.0119 | 0.0134 | 0.0457 |
| FP | NA | 1 | 0 | NA | 2 | 0 | NA | 2 | 1 |
| FN | NA | 0 | 1 | NA | 0 | 1 | NA | 0 | 1 |
| | | | | (b) Test length = 40 items** | | | | | |
| 0% | 0.0114 | 0.0122 | 0.0244 | 0.0059 | 0.0059 | 0.0177 | 0.0114 | 0.0122 | 0.0244 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0117 | 0.0123 | 0.0250 | 0.0059 | 0.0059 | 0.0167 | 0.0117 | 0.0123 | 0.0250 |
| FP | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 2 |
| 20% | 0.0117 | 0.0121 | 0.0257 | 0.0058 | 0.0060 | 0.0154 | 0.0117 | 0.0121 | 0.0257 |
| FP | NA | 0 | 0 | NA | 1 | 0 | NA | 1 | 0 |
| FN | NA | 3 | 6 | NA | 2 | 4 | NA | 4 | 6 |

* NCDIF cutoff at 99.9$^{th}$ percentile

** Linked

Table 8

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0272 | 0.0284 | 0.0437 | 0.0191 | 0.0210 | 0.0623 | 0.0272 | 0.0284 | 0.0437 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0278 | 0.0256 | 0.0355 | 0.0240 | 0.0202 | 0.0538 | 0.0278 | 0.0256 | 0.0355 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 |
| 20% | 0.0457 | 0.0274 | 0.0449 | 0.0229 | 0.0164 | 0.0624 | 0.0457 | 0.0274 | 0.0449 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| FN | 1 | 3 | 3 | 1 | 1 | 3 | 1 | 2 | 4 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0244 | 0.0264 | 0.0407 | 0.0177 | 0.0152 | 0.0372 | 0.0244 | 0.0264 | 0.0407 |
| FP | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0250 | 0.0259 | 0.0399 | 0.0167 | 0.0186 | 0.0600 | 0.0250 | 0.0259 | 0.0399 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 4 |
| 20% | 0.0257 | 0.0308 | 0.0529 | 0.0154 | 0.0198 | 0.0558 | 0.0257 | 0.0308 | 0.0529 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 6 | 5 | 8 | 4 | 5 | 8 | 6 | 7 | 7 |

* NCDIF cutoff at 99.9$^{th}$ percentile

**Linked

Table 9

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0118 | 0.0126 | 0.0293 | 0.0060 | 0.0062 | 0.0200 | 0.0060 | 0.0063 | 0.0201 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0118 | 0.0126 | 0.0293 | 0.0060 | 0.0062 | 0.0200 | 0.0060 | 0.0063 | 0.0201 |
| FP | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 2 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0.0118 | 0.0126 | 0.0293 | 0.0060 | 0.0062 | 0.0200 | 0.0060 | 0.0063 | 0.0201 |
| FP | NA | 0 | 0 | NA | 1 | 0 | NA | 2 | 0 |
| FN | NA | 0 | 0 | NA | 0 | 0 | NA | 0 | 0 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0114 | 0.0121 | 0.0223 | 0.0098 | 0.0060 | 0.0162 | 0.0058 | 0.0060 | 0.0160 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0114 | 0.0121 | 0.0223 | 0.0059 | 0.0060 | 0.0162 | 0.0058 | 0.0060 | 0.0160 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5 | 2 |
| FN | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 20% | 0.0114 | 0.0121 | 0.0223 | 0.0059 | 0.0060 | 0.0162 | 0.0058 | 0.0060 | 0.0160 |
| FP | NA | 1 | 0 | NA | 1 | 0 | NA | 4 | 0 |
| FN | NA | 3 | 3 | NA | 2 | 4 | NA | 3 | 3 |

* NCDIF cutoff at 99$^{th}$ percentile

** Not linked

Table 10

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0293 | 0.0271 | 0.0298 | 0.0200 | 0.0179 | 0.0592 | 0.0201 | 0.0181 | 0.0592 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0293 | 0.0271 | 0.0298 | 0.0200 | 0.0179 | 0.0592 | 0.0201 | 0.0181 | 0.0592 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 20% | 0.0293 | 0.0271 | 0.0298 | 0.0200 | 0.0179 | 0.0592 | 0.0201 | 0.0181 | 0.0592 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 1 | 1 | 0 | 1 | 2 | 0 | 1 | 3 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0223 | 0.0262 | 0.0467 | 0.0162 | 0.0161 | 0.0481 | 0.0160 | 0.0167 | 0.0470 |
| FP | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0223 | 0.0262 | 0.0467 | 0.0162 | 0.0161 | 0.0481 | 0.0160 | 0.0167 | 0.0470 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 |
| FN | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| 20% | 0.0223 | 0.0262 | 0.0467 | 0.0162 | 0.0161 | 0.0481 | 0.0160 | 0.0167 | 0.0470 |
| FP | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 1 |
| FN | 3 | 3 | 7 | 4 | 4 | 6 | 3 | 4 | 7 |

* NCDIF cutoff at 99$^{th}$ percentile

**Not linked

Table 11

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0118 | 0.0126 | 0.0293 | 0.0060 | 0.0062 | 0.0200 | 0.0060 | 0.0063 | 0.0201 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0118 | 0.0126 | 0.0293 | 0.0060 | 0.0062 | 0.0200 | 0.0060 | 0.0063 | 0.0201 |
| FP | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20% | 0.0118 | 0.0126 | 0.0293 | 0.0060 | 0.0062 | 0.0200 | 0.0060 | 0.0063 | 0.0201 |
| FP | NA | 0 | 0 | NA | 0 | 0 | NA | 1 | 0 |
| FN | NA | 0 | 0 | NA | 0 | 1 | NA | 0 | 1 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0114 | 0.0121 | 0.0223 | 0.0098 | 0.0060 | 0.0162 | 0.0058 | 0.0060 | 0.0160 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0114 | 0.0121 | 0.0223 | 0.0059 | 0.0060 | 0.0162 | 0.0058 | 0.0060 | 0.0160 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 |
| FN | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 20% | 0.0114 | 0.0121 | 0.0223 | 0.0059 | 0.0060 | 0.0162 | 0.0058 | 0.0060 | 0.0160 |
| FP | NA | 0 | 0 | NA | 0 | 0 | NA | 2 | 0 |
| FN | NA | 3 | 5 | NA | 2 | 5 | NA | 3 | 5 |

* NCDIF cutoff at 99.9$^{th}$ percentile

** Not linked

Table 12

False Positive (FP) and False Negative (FN) for the No-Impact Condition*

| | N = 500:500 | | | N = 1000:1000 | | | N = 500:1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) | 3PL (.2) | 3PL (.1) | 3PL (.5) |
| | (a) Test length = 20 items** | | | | | | | | |
| 0% | 0.0293 | 0.0271 | 0.0298 | 0.0200 | 0.0179 | 0.0592 | 0.0201 | 0.0181 | 0.0592 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0293 | 0.0271 | 0.0298 | 0.0200 | 0.0179 | 0.0592 | 0.0201 | 0.0181 | 0.0592 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 1 |
| 20% | 0.0293 | 0.0271 | 0.0298 | 0.0200 | 0.0179 | 0.0592 | 0.0201 | 0.0181 | 0.0592 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 2 | 3 | 1 | 1 | 3 | 1 | 1 | 3 |
| | (b) Test length = 40 items** | | | | | | | | |
| 0% | 0.0223 | 0.0262 | 0.0467 | 0.0162 | 0.0161 | 0.0481 | 0.0160 | 0.0167 | 0.0470 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10% | 0.0223 | 0.0262 | 0.0467 | 0.0162 | 0.0161 | 0.0481 | 0.0160 | 0.0167 | 0.0470 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 1 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 3 |
| 20% | 0.0223 | 0.0262 | 0.0467 | 0.0162 | 0.0161 | 0.0481 | 0.0160 | 0.0167 | 0.0470 |
| FP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FN | 5 | 5 | 7 | 5 | 6 | 6 | 5 | 6 | 7 |

* NCDIF cutoff at 99.9$^{th}$ percentile

**Not linked

# CHAPTER IV

## Discussion

The current study supports the usefulness of the IPR method in identifying biased items. Although certain issues still remain unresolved, the DFIT framework proves to be an effective and flexible mechanism for detecting bias in tests, equipped well to rival with the conventional DIF detection methods and even outperform them.

The study found that the correction in the computational formula from Nanda et al. (in press) did not affect the results noticeably. The only advantage was the decrease in computation time, which may be of essence for future simulation studies. The source of variance for the item parameters (focal or reference) did not seem to influence the detection results significantly. Some differences were found in FNs and FPs of two studies. However, the nature of these differences is yet to be explored. The current study also supports the view that when the values of *c*-parameter approach its extreme of 0.5, the DIF detection models become increasingly unstable. The study also found that compared to the value of 0.1, the *c*-parameter of 0.2 performed better in DIFCUT.

The authors of the original paper introducing the DIFCUT macro pointed out its poor performance in the 3PL models under the 20% DIF conditions with 40-item tests (Oshima, et al., 2006). Their results also indicated an overall poor performance in both 2PL and 3PL in the 20% DIF condition with 40-item tests in all sample sizes. This trend can be found even more prominent in the results of the current study than the original study. The NCDIF index, as the majority of DIF indices, relies on the assumption that no other items in a given test are biased. When the conditions of DIF in the current study, as in Oshima et al.(2006), were considered, the main focus of the experimental conditions

lied in the ratio of biased items to unbiased ones (10%DIF, 20% DIF), not the actual number of biased items on the test (2, 4, or 8 items). To illustrate, the 10% DIF condition had two biased items on the 20-item test while the same 10% DIF condition on 40-item test included four biased items. Naturally, with the higher the number of biased items the greater the chance to miss to identify them as biased. In light of this fact, it might not be fair to compare the conditions with the same percent of DIF items (e.g. 10% DIF) but different number of biased items.

This might suggest that the high FNs rates in the larger DIF conditions are linked to the percentage of biased items only vicariously, primarily depending on the *actual number* of biased items. In order to test this hypothesis, new studies may be conducted to utilize other combinations of percentage and actual number of biased items. Procedural complexity of such tests does not allow this hypothesis to be included into the scope of the current study.

The study also showed that when using the 99$^{th}$ percentile for the NCDIF cutoff value, not a single combination of sample size and variance source gave the detection that was robust across other conditions (IRT model, amount of DIF, and number of items on the test). This implies that the practitioners might not have the necessary precision in identifying biased items and would not know the nature of influence of aforementioned conditions on the accuracy of detection.

However, the study found that by adopting a high percentile value (99.9$^{th}$) for the NCDIF cutoff scores, one can obtain almost perfect FP rates and still have a significant number of correctly identified biased items. From the practitioner's perspective, this trade-off is justifiable: FNs or incorrectly identified as unbiased items would be kept on

the test and analyzed again while a number of items correctly identified as biased would be discarded or rewritten. However, the perfect FP rate would ensure that good and unbiased items would be not taken out from the test.

Although the study of DIF in the IRT models has been going on for a couple of decades, the DFIT framework is a fairly recent development. Published theoretical works dedicated to the DFIT topic are scarce and insufficient for the breadth and complexity of underlying statistical mechanisms. Some of the literature used for theory development is inaccessible, and some leading scientists, who had been working on DFIT are unavailable for consults. Also, due to the number of transitional steps necessary to employ DIFCUT, the level of detail in the description of procedure was sometimes compromised. The lack of details in the DIFCUT module resulted in unforeseen ambiguity in procedural sequences when the procedure was replicated in this study. These limitations somewhat undermine the integrity of results and conclusions of this study. However, without the accretion of exploratory works on DFIT, constructive and collaborative polemics would be stunted by the insufficiency of empirical research.

Among the suggestions for further studies, one can point out to several directions. First is the further exploration of the effects of the pseudo-guessing parameter on the DIF detection. The current study suggested that the magnitude of the *c*-parameter and the accuracy of DIFCUT are linked by a complex, non-linear, and not a unidirectional relationship. Values closer to the maximum and minimum extremes did not produce diametrically opposed accuracy rates. This relationship should be explored and understood in greater depths.

Very low rates of FN at the 99.9$^{th}$ percentile for the NCDIF cutoff values may suggest another direction for further studies. Cyclic DIFCUT analysis in the DFIT framework does not detect the biased items with an acceptable degree of accuracy after a single application of the DIFCUT macro. However, as with other DIF detection methods, multiple runs of simulation macro, each following a test's revision, might provide a systemic approach to the DIF detection. Currently, the revisions to the tests in iterative DIF analyses are governed by subjective judgments and rules of thumb. The decision which items to leave and which to discard or rewrite is done in the field of a certain ambiguity. The DFIT framework has the potential of systematizing and automating the process due to the empiric nature of its major indices.

Another possible improvement lies in the incorporation of all the steps in the simulation sequence into single software package. Although the majority of work in the current study was done in SAS, the flow of the process was interrupted a number of times by the necessity to resort to BILOG and IPLink. With the current developments which bring IRT computing into SAS ( Lee & Terry, 2004), such transition into a single software seems only logical. This would allow minimizing human error and bringing about the convenience in the variable manipulation, which by its complexity hindered realization of studies in the past.

# References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, *Statistical Theories of Mental Test Scores* (chapters 17-20). Reading, MA: Addison-Wedley.

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

Chamblee, M. C. (1998). *A Monte Carlo investigation of conditions that impact Type I error rates of DFIT*. Unpublished doctoral dissertation, Georgia State University.

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

Fleer, P. F., Kiley, K. A., & Raju, N. S. (1991).RANGEN [Computer program]. Unpublished computer program, Illinois Institute of Technology, Chicago.

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

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

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the 27th Annual Conference of the Military Testing Association* (Vol. 1; pp. 282-287). San Diego CA: Navy Personnel Research and Development Center.

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

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

Lee, S. & Terry, R. (2004). IRT-FIT: SAS® Macros for Fitting Item Response Theory (IRT) Models, presented at SUGI 30th conference in Philadelphia (The best contributed paper).

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

Lord, F.M. (1952). *A theory of test scores.*(Psychometric Monograph NO.7). Iowa City, IA: Psychometric Society.

Miller, T. R., Spray, J. A., & Wilson, A. (1992, July). *A comparison of three methods for identifying nonuniform DIF in polytomously scored test items.* Paper presented at the Psychometric Society Meeting, Columbus OH.

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

Nanda, A. O., Oshima, T. C., & Gagn´e, P. (in press). DIFCUT: A SAS-IML program for calculating cutoff scores for DFIT. *Applied Psychological Measurement.*

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement, 34,* 253–272.

Oshima, T.C., Raju, N.S., & Nanda, A. O. (2006). A new method for assessing the statistics significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement 43*, pp.1-17.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

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

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.