2010

An Examination of Type I Errors and Power for Two Differential Item Functioning Indices

Patrick Carl Clark Jr.
Wright State University

AN EXAMINATION OF TYPE I

ERRORS AND POWER FOR TWO

DIFFERENTIAL ITEM FUNCTIONING INDICES

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

By

PATRICK CARL CLARK
B.A., Indiana University of Pennsylvania, 2007

2010
Wright State University

WRIGHT STATE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

June 9, 2010

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Patrick Carl Clark ENTITLED An Examination of Type I Errors and Power for Two Differential Item Functioning Indices BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science

David LaHuis, Ph.D.
Thesis Director

John Flach, Ph.D.
Department Chair

Committee on
Final Examination

David LaHuis, Ph.D.

Debra Steele-Johnson, Ph.D.

Gary Burns, Ph.D.

Andrew T. Hsu, Ph.D.
Dean, School of Graduate Studies

ABSTRACT

Clark, Patrick Carl. M.S., Department of Psychology, Wright State University, 2010. An Examination of Type I Errors and Power for Two Differential Item Functioning Indices

.

This study examined two methods for detecting differential item functioning (DIF): Raju, van der Linden, and Fleer's (1995) differential functioning of items and tests procedure (DFIT) and Thissen, Steinberg, and Wainer's (1988) likelihood ratio test (LRT). The major research questions concerned which test provides the best balance of Type I errors and power and if the tests differ in terms of detecting different types of DIF. Monte Carlo simulations were conducted to address these questions. Equal and unequal sample size conditions were fully crossed with test lengths of 10 and 20 items. In addition, α and β parameters were manipulated in order to simulate DIF. Findings indicate that the DFIT and LRT both had acceptable Type I error rates when sample sizes were equal, but that DFIT produced too many Type I errors when sample sizes were unequal. Overall, the LRT exhibited greater power to detect both α and β parameter DIF than DFIT. However, DFIT was more powerful than LRT when the last two β parameters had DIF as opposed to the extreme β parameters. Therefore, it is recommended under most circumstances to use the LRT for DIF detection, unless there is reason to believe that the DIF is at the high end of the scale.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Introduction

Differential item functioning (DIF) is an important research topic for psychologists because people with the same underlying ability from different groups may have different probabilities of endorsing the same item. Differential item functioning has been used to examine applicant and incumbent differences (e.g., Robie, Zickar, & Schmit, 2001), male and female differences (e.g., Braddy, Meade, and Johnson, 2006; Collins, Raju, & Edwards, 2000), different test versions (e.g., Donvan, Drasgow, & Probst, 2000) and black and white differences (e.g., Braddy, et al., 2006; Collins, et al., 2000; Raju, Laffitte, & Byrne, 2002). As such, there are a number of different techniques used to examine DIF. Some examples of these different methods include the area between two item response functions (Raju, 1988, 1990), Lord's (1980) χ^2 test, the Mantel-Haenszel technique (Holland & Thayer, 1988), the delta method (Angoff & Ford, 1973), the differential functioning of items and tests procedure (DFIT; Raju, van der Linden, & Fleer, 1995), and Thissen, Steinberg, and Wainer's (1988) likelihood ratio test (LRT). Unfortunately, there has been a lack of empirical work on the issue of DIF detection, more specifically determining what DIF detection method to use under certain circumstances.

In the present study I examined the Type I error rates and power of two IRT-based fit indices: the differential functioning of items and tests (DFIT) framework (Raju et al., 1995) and the likelihood ratio test (LRT: Thissen et al., 1988) using the graded response model (GRM). I manipulated a number of conditions in order to identify which

fit index should be preferred overall for DIF detection and whether different indices capture different sources of DIF. In the following sections, I briefly review the GRM that is often used to analyze polytomous items and then describe the two DIF indices.

Differential Item Functioning

Technically, DIF occurs when subgroups have different item response functions (IRF's) for the same item. An IRF plots the relationship between the probability of endorsing a response option and the underlying ability level (usually referred to as theta, θ). Figure 1 presents two IRF's for an item exhibiting DIF based on a two-parameter logistic model (2PLM). Researchers typically refer to one group as the focal group and the other group as the reference group. For example, a group of applicants may be referred to as the focal group while the incumbents are called the reference group. The solid line represents the focal group who has a harder time endorsing the item than the reference group. The item is also more discriminating for the focal group, as indicated by the more exponential curve.

The history of DIF research dates back to the 1960s when researchers were interested in discovering the reason for the difference in test scores of Blacks and Hispanics with Whites on cognitive ability tests (Angoff, 1993). The goal of these studies was to identify items or questions that were biased against minority students. That is, the goal was to remove items that minority students did more poorly on than majority students. In addition to these studies, research was taking place on tests used for selection that appeared to be biased as well (Angoff, 1993). These selection tests did a poor job of predicting performance for minorities. The test scores were lower for minorities than would be indicated by their performance on a criterion measure that the tests were

designed to predict. The solution to this was simple, get rid of the test. The solution for the item bias problem was not so simple because there was no external criterion to measure performance on an item against. In order to resolve the item bias problem, mathematical procedures were applied to the testing of items in order to identify the ones that are biased.

There are several techniques designed to determine whether a test item is functioning differentially. Some of these techniques are based on item response theory (IRT), such as the area between two item response functions (Raju, 1988, 1990), Lord's (1980) χ^2 test, Thissen et al.'s (1988) likelihood ratio test (LRT), and the differential functioning of items and tests procedure (DFIT; Raju et al., 1995). Others are not based on IRT, such as the Mantel-Haenszel technique (Holland & Thayer, 1988) and the delta method (Angoff & Ford, 1973).

Item Response Theory

IRT models specify the relationship between item responses and an individual's underlying latent trait (θ). A number of IRT models are available. In the present study, I will focus on the two-parameter logistic model (2PLM) for dichotomous items and the graded response model (GRM) for polytomous items.

Two-parameter Logistic Model

The 2PLM specifies the probability of endorsing item *i* for person *j* as a function of a person's trait level θ and two item parameters: item discrimination α and item difficulty β. The form of the equation is:

$$P_{ij}\,(Y=1|\theta\,j) = \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1+\exp[\alpha_i(\theta_j - \beta_i)]} \qquad (1)$$

Trait levels represent levels of the underlying construct (e.g., Conscientiousness, Cognitive Ability). It is common to scale trait levels to have a mean of zero and a standard deviation of one. Items with higher discrimination parameters indicate that the probability of endorsing that item is much greater for someone of high ability versus someone with low ability. Whereas an item with lower discrimination indicates that there is less discrepancy between the probability of endorsement between individuals with high and low ability levels. In the 2PL model, item difficulty represents the point on the trait scale at which the probability of endorsement equals 50%. Higher (> 0) item difficulty values indicate harder items whereas lower (< 0) values indicate easier items. In the context of ability testing, item difficulty represents how hard it is to answer an item correctly. In the personality context, item difficulty refers to the amount of the underlying construct necessary to endorse the item. For example, an easy Conscientiousness item may require very little of the trait Conscientiousness for individuals to endorse. An example of this might be, "I am a reliable worker". In contrast, individuals would need higher levels of Conscientiousness to endorse more difficult items such as, "I am exacting in my work".

One can use Equation 1 to plot an IRF. For example, Figure 2 plots IRF's for three items. Items 1 and 2 have equal difficulty parameters but different discrimination parameters. The larger discrimination parameter for item 2 produces a steeper slope for its IRF. Items 1 and 3 have equal discrimination parameters but different difficulty parameters. The item difficulty is greater for item 3; therefore the IRF shifts right.

Graded Response Model

In many situations, researchers use polytomous Likert-scale items with greater

than two response options. Samejima's (1969) GRM has been typically applied to analyze this type of item data. The GRM assumes that an item has *m* ordered categories. Estimates are based on *m*-1 boundary response functions (BRFs). Each BRF represents the cumulative probability of selecting a response option greater than the option of interest. For example, a Likert-type scale with five response options would have four BRF's. Figure 3 shows an example of BRFs for an item with five response options. The first BRF is the probability of choosing the lowest response option versus the cumulative probability of choosing the other four options. The second BRF is the probability of choosing the lowest two response options versus the probability of choosing the other three, and so on. The equation for a BRF is similar to the 2PL equation for dichotomous data: Figure 3

$$P_{ik}^{*}(\mathrm{Y}=1\,|\,\theta)=\frac{\exp[\alpha_{\mathrm{i}}(\theta-\beta_{\mathrm{ik}})]}{1+\exp[\alpha_{\mathrm{i}}(\theta-\beta_{\mathrm{ik}})]} \tag{2}$$

In this equation, α is the item discrimination parameter and β_{ik} is a difficulty parameter for option k. There are *m*-1 threshold parameters. For example, the β_{ik} difficulty parameter for choosing greater than the lowest option represents the point on the theta scale where there is 50% probability that the response is greater than the lowest option. Using equations 2 through 6, the probability of responding to each of the five categories (P_{i1} to P_{i5}) may be calculated.

$$\mathrm{P}_{\mathrm{i1}}(\theta)=1-\mathrm{P}_{\mathrm{i1}}^{*}(\theta) \tag{3}$$

$$\mathrm{P}_{\mathrm{i2}}(\theta)=\mathrm{P}_{\mathrm{i1}}^{*}(\theta)-\mathrm{P}_{\mathrm{i2}}^{*}(\theta) \tag{4}$$

$$\mathrm{P}_{\mathrm{i3}}(\theta)=\mathrm{P}_{\mathrm{i2}}^{*}(\theta)-\mathrm{P}_{\mathrm{i3}}^{*}(\theta) \tag{5}$$

$$P_{i4}(\theta) = P^{*}_{i3}(\theta) - P^{*}_{i4}(\theta) \quad (6)$$

$$P_{i5}(\theta) = P^{*}_{i4}(\theta) - 0 \quad (7)$$

One can plot these probabilities to produce option response functions (ORF's) such as that in Figure 4. This figure allows you to determine an individual's probability of endorsing a response option. For example, a person with a theta value of 0 has essentially a probability of 0 of endorsing options 1 and 2, about a probability of .10 of endorsing option 3, a probability of .30 of endorsing option 4, and nearly a probability of .60 of endorsing option 5. A simpler example would be to look at a person with theta value 4 and see that they have a probability of approximately 1.0 of endorsing option 5 with a probability of 0 of endorsing all of the other response options.

Linking

In the case of assessing DIF, item parameters from two different groups are independently estimated and will essentially be on a different scale. This causes a problem when assessing DIF using IRT because item parameter estimates from independent calibrations must be in the same scale (linked). Stocking and Lord (1983) describe a method of putting item parameters from two separate calibrations on the same scale using a characteristic curve method to determine the two constants (a multiplicative and an additive constant) involved in the linear transformation. A program called Equate has been developed by Baker (1993) using Stocking and Lord's (1983) test characteristic curve method. It uses the item parameter estimates of both groups to obtain the constants necessary to make the linear transformation. These constants are then applied to the item parameter estimates of one group in order to convert them to the scale of the second group, thus "linking" the parameters.

The DFIT Framework

The DFIT procedure includes three measures: differential test functioning (DTF), compensatory DIF (CDIF), and noncompensatory DIF (NCDIF). The DFIT indices allow for the determination of whether individual items exhibit DIF (CDIF and NCDIF) and/or whether the test as a whole exhibits DIF (DTF). The DTF index is based on the notion that although individual items may function differentially, these differences can cancel each other out across items so the test may not exhibit DTF. If there is no DTF and the researcher is not worried about individual item DIF, then all of the items in the test may remain. The DTF index is defined as the average squared difference between the test characteristic functions (TCFs) for the focal and reference groups. For a dichotomous or polytomous model, the difference is based on comparing the true test scores of examinees using the focal and reference groups item parameters, respecitivley.

$$D(\theta_s) = T_F(\theta_s) - T_R(\theta_s) \quad (7)$$

Where T_F is the true score based on the focal group's parameters and T_R is the true score based on the reference group's parameters. DTF is the expected value of the squared difference between the focal and the reference groups across the θ distribution from the focal group (E_F).

$$DTF = E_F[\mathrm{D}\ (\theta_s)^2] \quad (8)$$

CDIF and NCDIF are the two item-level DIF indices in the DFIT framework. One can use the CDIF index to identify items that contribute to DTF. Under this index, all of the items on the test are considered and items that if removed would lower DTF are noted. When one sums CDIF values across items, the outcome is equal to DTF. CDIF

can be calculated by finding the covariance of D and *d* (probability of success on an item for a person in the focal group probability of success on an item for a person in the reference group) and adding that to the product of the mean of *d* and the mean of *D*.

$$CDIF_i = Cov(d_i, D) + \mu_{di}\mu_D. \quad (9)$$

The NCDIF index can be used to identify individual items exhibiting DIF independent of other items on the test. This means that NCDIF will flag an item even if the item's DIF is cancelled out at the test level. NCDIF is similar to DTF, except that it is based on comparing IRF's across groups. The NCDIF index is the expectation over the focal group (E_F) of squared differences between the probability of endorsing an item using the focal item parameters, $P_{iF}(\theta)$, and the probability of endorsing an item using the reference group parameters, $P_{iR}(\theta)$. If d_i equals the difference between the probabilities of item endorsement under the focal and reference group parameters then

$$NCDIF_i = E_F[P_{iF}(\theta) - P_{iR}(\theta)]^2 = E_F(d_i)^2 = \sigma_{di}^2 + \mu_{di}^2 \quad (10)$$

where σ and μ are the standard deviations and means of d_i, respectively (Raju et al., 1995).

Item Parameter Replication Method

Raju et al. (1995) originally proposed chi-square tests to test for the DTF and NCDIF indices. However, these tests tended to commit too many Type I errors. The focus then shifted to establishing cutoff values for the two indices. Raju et al. (1995) and Flowers, Oshima, and Raju (1999) proposed overall cutoff values of .006 and .016, respectively for the NCDIF index. The cutoff values suggested by Raju et al. (1995) and Flowers et al. (1999) were not generalizable to other items and were also specific to the IRT model being used.

Research has examined empirically derived cutoff values for the NCDIF index. Flowers et. (1999) simulated datasets of 20 and 40 items, with a sample size of 1000 and had conditions of 0%, 5%, 10%, and 20% DIF items. They used a value of .016 for the NCDIF index and found both reasonable power and Type I errors. Bolt (2002) established empirical cutoffs as well and ended up with .032 for (N = 300) and .009 for (N = 1000). Meade, Lautenschlager, and Johnson (2007) simulated data for 12 items and used sample sizes of 500 and 1000. To introduce DIF he manipulated the a parameter by subtracting .25 from the reference groups value and manipulated the b parameters by adding or subtracting either .4 or 1 to the reference group values. Meade et al. found that using large cutoff values, .096, causes the NCDIF and DTF indices to have less power for detecting small to moderate amounts of DIF. Using empirically derived cutoff values (.016, .0115, and .009), however, resulted in the NCDIF index having optimal power and Type I error rates and the DTF index showing improved power and Type I error rates.

Despite the promise in using empirically derived cutoff values, most researchers have neither the expertise nor the time to do these sorts of calculations so a new method was needed. Researchers first developed the item parameter replication (IPR) method for use with dichotomous data (Oshima, Raju, & Nanda, 2006), and recently that research has been extended to polytomous data (Raju, Fortmann-Johnson, Kim, Morris, Nering, & Oshima, 2009). The item parameter replication method utilizes estimates of item parameters for the focal group, as well as variances and covariances for those estimates in order to obtain a distribution of NCDIF values that can be rank ordered. Based on these estimates, a number of replications of item parameters are then generated using the same sampling variance and covariance as the original item parameters. This assumes that any

differences between the original parameters and the newly generated parameters are due to sampling error. After a large number of replications are complete, an empirical sampling distribution of NCDIF under the null hypothesis that both groups (the focal and reference) are equal is produced.

The IPR method consists of a number of steps that will be described for a single polytomous item *i* with five response options and using the *GRM*.

A column vector M_i represents item parameters.

$$M_i = \begin{bmatrix} a_i \\ b_{i1} \\ b_{i2} \\ b_{i3} \\ b_{i4} \end{bmatrix} \tag{13}$$

Each item is also associated with a matrix consisting of the sampling variances and covariances of the item parameters, represented by V_i.

$$V_i = \begin{bmatrix} \sigma^2_{ai} & \sigma_{ab_{i1}} & \sigma_{ab_{i2}} & \sigma_{ab_{i3}} & \sigma_{ab_{i4}} \\ \sigma_{b_{i1}a} & \sigma^2_{b_{i1}} & \sigma_{b_{i1}b_{i12}} & \sigma_{b_{i1}b_{i3}} & \sigma_{b_{i1}b_{i4}} \\ \sigma_{b_{i2}a} & \sigma_{b_{i2}b_{i1}} & \sigma^2_{b_{i2}} & \sigma_{b_{i2}b_{i3}} & \sigma_{b_{i2}b_{i4}} \\ \sigma_{b_{i3}a} & \sigma_{b_{i3}b_{i1}} & \sigma_{b_{i3}b_{i2}} & \sigma^2_{b_{i3}} & \sigma_{b_{i3}b_{i4}} \\ \sigma_{b_{i4}a} & \sigma_{b_{i4}b_{i1}} & \sigma_{b_{i4}b_{i2}} & \sigma_{b_{i4}b_{i3}} & \sigma^2_{b_{i4}} \end{bmatrix} \tag{14}$$

V_i can be used to generate a correlation matrix, R_i

$$R_i = \begin{bmatrix} 1 & \rho_{ab_{i1}} & \rho_{ab_{i2}} & \rho_{ab_{i3}} & \rho_{ab_{i4}} \\ \rho_{b_{i1}a} & 1 & \rho_{b_{i1}b_{i12}} & \rho_{b_{i1}b_{i3}} & \rho_{b_{i1}b_{i4}} \\ \rho_{b_{i2}a} & \rho_{b_{i2}b_{i1}} & 1 & \rho_{b_{i2}b_{i3}} & \rho_{b_{i2}b_{i4}} \\ \rho_{b_{i3}a} & \rho_{b_{i3}b_{i1}} & \rho_{b_{i3}b_{i2}} & 1 & \rho_{b_{i3}b_{i4}} \\ \rho_{b_{i4}a} & \rho_{b_{i4}b_{i1}} & \rho_{b_{i4}b_{i2}} & \rho_{b_{i4}b_{i3}} & 1 \end{bmatrix} \quad (15)$$

R_i may also be expressed as the product of a triangular matrix, T_i, and its transpose, T'_i.

$$R_i = (T_i)(T'_i) \quad (16)$$

Second, two column vectors X_{1i} and X_{2i} are created. These column vectors contain an element for each item parameter. In the current example, each vector will contain five values. These values are drawn randomly from a standard normal distribution.

Third, the T'_i matrix is used to transform the two X vectors into two Z vectors.

$$Z_{1i} = (X_{1i})(T'_i) \quad (17)$$

$$Z_{2i} = (X_{2i})(T'_i) \quad (18)$$

The result of this is that each Z vector now contains 5 values drawn from a standardized multivariate distribution with correlational structure as the R_i matrix.

Fourth, each Z vector is transformed into a Y vector in order to put the parameters back into the original metric. Let D_i represent a diagonal matrix of the variances in D_i.

$$Y_{1i} = (\sqrt{D_i})(Z_{1i}) + M_i \quad (19)$$

$$Y_{2i} = (\sqrt{D_i})(Z_{2i}) + M_i \quad (20)$$

Vectors Y_{1i} and Y_{2i} now represent the item parameter estimates of the focal and reference group when no DIF is present and any differences between these two estimates is due to sampling error. An NCDIF value for the item in question can then be obtained using both Y vectors as well as the theta estimates from the focal group.

This process is repeated for as many replications as desired (usually 1000).

The NCDIF values may now be rank ordered such that the 90^{th}, 95^{th}, 99^{th}, 99.5^{th}, and 99.9^{th} percentile rank scores are the cutoff values for alpha levels of .10, .05, .01, .005, and ,.001, respectively. The cutoff value associated with the chosen alpha is used to assess the statistical significance of the NCDIF value obtained for the item. If the NCDIF value exceeds the cutoff, then the item is assumed to have DIF. The process is repeated for each item in the test. Consistent with this, it is to be expected that each item may have different cutoff values.

Substantive Tests of DIF using DFIT

The DFIT framework has been used to assess both DIF and DTF in several published studies (e.g., Collins et al., 2000; Donovan et al., 2000; Henry & Raju, 2006; Maurer, Raju, & Collins, 1998; Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006; Raju, Laffitte, & Byrne, 2002). Maurer et al. (1998) assessed the DIF of peer and subordinate ratings of managers on a teambuilding skill dimension and found the two sets of ratings are comparable. Collins et al. (2000) used DFIT to analyze the DIF of a Likert-style satisfaction scale for Black vs. White and male vs. female. They found the DFIT procedure functioned as expected. Donovan et al. (2000) used DFIT to examine DIF for a

computerized version of a satisfaction scale. Raju et al. (2002) assessed Black-White DIF on a 10-item satisfaction with work scale and found that only one item was DIF according to the DFIT method. Henry and Raju (2006) used DFIT to assess the difference between trait and situational impression management behavior. Morales et al. (2006) identified 9 items with DIF on the Mini-Mental State Examination but found that the test didn't have DTF.

The Likelihood Ratio Test

The likelihood ratio procedure is based on Lord's (1980) χ^2 test in which an item exhibits DIF if the IRF for the reference group differs from that of the focal group. Essentially, because the base of the IRF is on the item parameters, this means that if the item parameters for the two groups are significantly different then the item has DIF according to the LRT. In order to compare the two groups, two types of models are tested, compact and augmented. Under the compact model, all item parameters are constrained to be equal across the two groups. An augmented model allows the item parameters for a focal item to vary across groups while constraining the parameters for the other items. The fit of the two models is compared using a LR test:

$$G^2(df) = -2\log\left[\frac{L[A]}{L[C]}\right] \quad (20)$$

where *L* represents the likelihood of the data given the maximum likelihood estimates of the parameters of the model, A refers to the augmented model, and C refers to the compact model. G^2 is distributed as a $\chi^2$with df equal to the difference between the number of parameters in the augmented model and the number of parameters in the compact model.

Simulation studies have demonstrated the utility of using the LRT for DIF. For example, Kim and Cohen (1998) simulated data for 30 items and had three conditions: a sample size of 300 in both the reference and focal group, 1000 in both groups, and 1000 in the reference group and 300 in the focal group. They also had a matched (reference and focal groups have the same mean ability level) and an unmatched (reference group has a higher mean ability level) condition for each of the above conditions. Their results indicated acceptable Type I error rates for all six combinations of sample sizes and ability matching conditions. They didn't assess the power of the LRT. Ankenmann, Witt, and Dunbar (1999) simulated a 26-item mixed format test (20 items were dichotomous, six were polytomous) with three conditions: a sample size of 2000 in both the reference and focal group, 500 in each group, and 2000 in the reference group and 500 in the focal group. Like Kim and Cohen (1998), Ankenmann et al. had a matched and an unmatched condition for each sample size pairing. In addition, Ankenmann et al. manipulated the α and β parameters for one of the groups to introduce DIF. Ankenmann et al. found the LRT had acceptable Type I error rates across most of the simulated conditions but that the test may lack power when sample sizes are around 500.

Substantive Tests of DIF using LRT

As opposed to the DFIT framework, I am only aware of one study in which researchers use the LRT in a substantive test to detect DIF. Kim (2001) used the LRT to assess DIF in a speaking test for nonnative English speakers. The use of the DFIT framework for substantive tests is likely preferred over the LRT because of its ability to not only detect DIF at the item level but at the test level as well. With DFIT you can also assess the impact of removing an item on the overall DTF using the CDIF index.

Research on DFIT and LRT

The DFIT framework and LRT have been two of the most frequently used ways of testing for DIF, and there has been some research investigating their performance. For example, numerous studies have shown the DFIT framework to be successful at detecting items and tests that may be biased (e.g., Flowers et al., 1999; Meade et al., 2007; Raju et al., 1995). A number of researchers have examined the LRT and shown that the test has enough power to detect DIF at the item level (e.g., Ankenmann et al., 1999; Kim & Cohen, 1998; Thissen et al., 1988).

Several studies have compared DFIT and/or LRT to other DIF detection methods. Kim and Cohen (1995) compared Lord's χ^2 test, Raju's area measures, and the LRT for detecting DIF. They used data from 28 math items and used the procedures to detect differences in item functioning that could be credited to using a calculator. Lord's χ^2 test and Raju's area measures identified the same 5 items as differentially functioning. LRT identified 6 items total as having DIF, the 5 identified by the other two indices plus one other item previously unidentified. Raju et al. (1995) used a two-parameter logistic model to generate simulated datasets of 40 items, with a sample size of either 500 or 1000 and had conditions of 0, 2, 4, or 8 differentially functioning items. The NCDIF procedure identified 82% of DIF items in the N = 500 condition and 89% of DIF items in the N = 1,000 condition. Using the signed area technique allowed the researchers to identify 61% of DIF items in the N = 500 condition and 68% of DIF items in the N = 1,000 condition. The signed area technique allowed the researchers to identify 75% of DIF items in the N = 500 condition and 89% of DIF items in the N = 1,000 condition. The researchers

identified 79% of DIF items in the N = 500 condition and 96% of DIF items in the N = 1,000 condition using Lord's χ^2 test.

Similarly, Finch and French (2007) used simulated datasets with 30 items and sample sizes of 500, 750, 1000, and 1500 and had conditions of 0, 3, or 6 DIF items. The Type I error rates for all four DIF detection indices (simultaneous item bias test or SIBTEST, logistic regression, LRT, and confirmatory factor analysis) compared were between .048 and .056, which is acceptable. Of the four fit indices examined, SIBTEST had the highest power, followed by LRT. Kim, Cohen, Alagoz, and Kim (2007) analyzed 105,731 Kindergarten Assessment measures, a 10-item scale, for male-female DIF. Kim et al. compared the LRT to four other DIF detection methods on a performance assessment and found that four of the five methods identified all ten items as having DIF whereas the other method, the Mantel-Haenszel technique, identified nine of the ten items as having DIF.

Little research has directly compared DFIT with LRT. One exception is Bolt (2002). He generated 30 items with three different IRT models, the graded response model, the generalized partial credit model, and the two-parameter sequential response model for sample sizes of 300 and 1000. Type I error rates for the LRT were in the acceptable 5% range when Bolt used the GRM to generate the data but went up to 20% when he used one of the other models to generate the data. Type I error rates for DFIT were acceptable regardless of the generating model. Both procedures had acceptable power across most conditions, but the LRT exhibited more power overall than the DFIT procedure. Although Bolt (2002) compared the two methods I am interested in for this study, he did not assess DIF when varying the item parameters.

Braddy et al. (2006) also compared LRT and DFIT directly. In this study, the authors used a data set of 5396 employees who took a 21-item Likert measure of leadership. Braddy et al. created samples of 200, 300, 500, 1000, 1500, and 1800 for male/female comparisons and samples of 200, 300, 500, and 800 for the White/Black comparisons. The LRT identified 3, 6, 4, 6, 8, and 14 DIF items respectively for the six male/female comparison and 1, 4, 5, and 7 DIF items respectively for the four White/Black comparisons. The DFIT procedure concluded that the test as a whole did not have DTF in either comparison and only identified 1 item in all of the conditions as DIF. On the basis of the results from the LRT, the conclusion about the test would be to remove a majority of the items in order to eliminate DIF. On the other hand, the DFIT procedure implied that the test had virtually no DIF at the item and test level and therefore should remain unchanged. These results do not indicate which DIF detection method was more accurate at assessing DIF because the true magnitude of population DIF in this study was unknown. The drastically different conclusions implied from Braddy et al.'s study indicates that it is crucial to know which of these fit indices to use in order to avoid producing a test with extensive DIF. It may seem logical to err on the side of caution, use the LRT, and just remove the items that the test flagged as exhibiting DIF. The problem with this approach is that researchers may be removing quality items that may not be DIF after all. Thus, there are still a number of questions concerning which DIF test should be preferred overall and how various conditions affect the DIF tests.

Present Study

In the present study, I examined the Type I error rates and power for the NCDIF index and the LRT across a number of conditions. Specifically, I varied the test length

and source of DIF by manipulating the discrimination and threshold parameters. I was interested primarily in (a) how the manipulations affected the Type I errors across tests, (b) how the manipulations affected the power estimates across tests, (c) which test had more acceptable Type I error rates, and (d) which test had better power estimates.

Method

Design

I generated item data for a simulation study with 116 conditions. The study design was based primarily on Meade et al. (2007). Type I errors were assessed using four control conditions: 10-item equal samples, 10-item unequal samples, 20-item equal samples, and 20-item unequal samples. No parameters were manipulated, and thus no DIF was present in these control conditions. Meade et al. (2007) found that manipulating sample size ($\omega^2 = .001$) and the number of DIF items ($\omega^2 = .002$) had little impact on power for DFIT. Based on this, for half of the conditions I used a sample size of 500 for the reference and 500 for the focal group. In addition, we as well as 4 DIF items in each DIF condition. In most cases in real word settings, one of the groups will have fewer individuals than the other group. To address this, I also used unequal sample size conditions with a sample size of 500 for the reference group and 250 for the focal group. For each manipulation of DIF, I included a no DIF cancellation condition and a DIF cancellation condition. In the no DIF cancellation condition I subtracted values for all four DIF items. In the DIF cancellation condition, I subtracted for two of the DIF items and added for the other two DIF items. Finally, I conducted 100 repetitions of each of the conditions.

Data Generation

I generated the item data using SPSS 12.0. First, population values for the α and β parameter estimates were generated. The α parameters were sampled from a random normal distribution with a mean of 1.25 and a standard deviation of 0.07. I generated four β values for each item to simulate items with five response options. I sampled the β value for the lowest of the four boundary response functions (BRFs) from a random normal distribution with a mean of -1.7 and a standard deviation of 0.45. In order to create the three other β values I added constants of 1.2, 2.4, and 3.6 to the lowest difficulty (Meade et al., 2007). Second, I sampled θ values for each simulee from a random normal distribution with a mean of 0 and a standard deviation of 1. Third, I calculated the probability for endorsing each option for each item by plugging the generated item parameters and θ values into the GRM equations. I used the probabilities for endorsing each option to calculate a cumulative probability of endorsing each response option. Finally, I generated item responses by comparing previously generated random numbers (between 0 and 1) with the cumulative probabilities of endorsing each response option. The lowest response option for which the cumulative probability exceeds the random number is a simulee's item response.

I introduced DIF in one of two ways. First, items' α parameters were simulated to differ across groups by subtracting either 0.25 or 0.50 from the reference group's α parameter in order to create the focal group's α parameters. This gave the effect of the item not distinguishing as well between high and low performers for the focal group. Second, items' β parameters were simulated to differ across groups in one of four ways: 1) adding or subtracting 0.4 or 1.0 from the last two β values, 2) adding or subtracting 0.4

or 1.0 from the two most extreme β values, 3) adding or subtracting 0.4 or 1.0 from the last two β values and having the DIF cancel, and 4) adding or subtracting 0.4 or 1.0 from the two most extreme β values and having the DIF cancel. Adding a constant to the β parameter simulates the effect of the focal group having a lower likelihood of endorsing the option (i.e., the item becomes more difficult). Subtracting a constant from the β parameter simulates the effect of the focal group having a greater likelihood of endorsing the option (i.e., the item becomes easier).

Data Analysis

The NCDIF analysis requires item parameters for the reference and focal group to be estimated and put on the same metric using a linking procedure. Item parameters were estimated using PARSCALE (Muraki & Bock, 2003), and I linked the parameters using the Equate 2.1 program (Baker, 1995). I conducted the NCDIF procedure using the DFIT8 program (Oshima, Kushubar, Scott, & Raju, 2009) and the LRT analysis with the IRTLRDIF (Item response theory likelihood ratio differential item functioning) program (Thissen, 2001).

To evaluate the effectiveness of each of the DIF detection methods, I evaluated Type I errors and power. The Type I error rates were calculated by dividing the number of non-DIF items that are flagged as DIF by the index by the total number of items simulated not to contain DIF for that particular sample. In the conditions where α and β were not manipulated, Type I error rates were based on all of the items. For conditions with DIF, Type I error rates were based on the items that did not have their parameters changed. I assessed power by calculating the number of differentially functioning items that are successfully detected as DIF items divided by the total number of DIF items

generated for the sample. I then averaged these Type I error rates and power across 100 replications for each condition. In addition, to determine the effects of study variables, a mixed-model analysis of variance (ANOVA) was conducted with type of test (LRT vs. NCDIF) as a within group variable and the manipulations as between group variables. Because of the large sample sizes, I focused on interpreting effect sizes rather than statistical significance.

Results

Tables 1 and 2 present effect sizes for the mixed model ANOVA's. I conducted simple effects analyses for two-way interactions with nontrivial effect sizes. As shown in Table 1, the type of test explained 29% of the variance in Type I error rates with the LRT being associated with lower Type I errors. The interaction between the type of test and sample equality ($\eta^2 = .13$) had the next largest effect. The simple effects analyses revealed that sample equality had a larger effect on Type I errors for NCDIF (partial $\eta^2 = .16$) compared with LRT (partial $\eta^2 = .00$). As shown in Figure 5, the Type I error rates are similar across test type for equal sample sizes. However, when the samples are unequal, the Type I error rates for NCDIF (18%) were much higher than the Type I error rates for LRT (3%). None of the other manipulations had sizable interaction effects on Type I error rates.

As shown in Table 2, the type of test explained 8% of the variance in power with the LRT being associated with higher power. The interaction between type of test and magnitude of α DIF was 4%. Simple effects analyses revealed that the α DIF manipulation had a smaller effect on the power of NCDIF (partial $\eta^2 = .12$) than on the power of LRT (partial $\eta^2 = .40$). The effect size for the interaction between type of test

and magnitude of β dif was smaller (partial η^2 = .02). The interaction between the type of test and which β parameters were manipulated (η^2 = .17) had the largest effect. Figure 6 graphs this interaction. The effect was weaker for the NCDIF (partial η^2 = .06) than for the LRT (partial η^2 = .12). Similarly, the three-way interaction between type of test, magnitude of β dif, which β parameters were manipulated had a large effect (partial η^2 = .16). The NCDIF does a better job of detecting the DIF when the last two β parameters were changed by 0.40. In contrast, the LRT does a better job of detecting the DIF when the two most extreme β parameters were manipulated changed by 0.40. Little differences were observed when the β parameters were changed by 1.0.

The interactions between the number of items (partial η^2 = .02) or cancellation of DIF (partial η^2 = .01) with test type did not explain much variance in power. The type of test and sample equality interaction resulted in a partial η^2 of .04. Simple effects analyses revealed a smaller effect on the power for NCDIF (partial η^2 = .00) compared to LRT (partial η^2 = .06). It appears that the LRT is more powerful when sample sizes are equal.

Type I Error Rates and Power Estimates

Tables 3 and 4 present the Type I error rates and power estimates for the NCDIF index for equal and unequal sample sizes, respectively. All Type I error rates were less than or equal to 7% in the equal samples sizes condition for the NCDIF index. Type I error rates for the unequal sample sizes condition were all between 16 and 21%. Overall, for equal sample sizes, the NCDIF index had acceptable Type I error rates ($\leq$ 5) in 49 of the 58 conditions but did not have acceptable error rates in any of the conditions for unequal sample sizes.

Overall, out of the 112 conditions the NCDIF index had acceptable power levels ($\geq$ 80%), in 70 of the conditions. Acceptable levels of power for the NCDIF index were not found when only the α parameter was manipulated in either condition. Similarly, acceptable power levels were not found when the last two β parameters were changed by 0.40 in either condition. Power levels near 100% were found when the β parameters were changed by 1.00, for both conditions where β parameters were manipulated. Combining α and β DIF generally produced acceptable levels of power with the exception of when the α parameter was changed by 0.25 and the extreme β parameters were changed by 0.40.

Tables 5 and 6 present the Type I error rates and power estimates for the LRT index for equal and unequal sample sizes, respectively. All Type I error rates were less than or equal to 8% across both sample size conditions except when the highest two β parameters were manipulated, the Type I error rates went up considerably. Overall, the LRT had acceptable Type I error rates in 54 of the 58 conditions in the equal sample sizes condition, and in 52 of the 58 conditions in the unequal sample sizes condition.

The LRT had acceptable power levels in 87 of the conditions. Acceptable levels of power were not found when only the α parameter was changed by 0.25 in either condition. Unlike the NCDIF index, acceptable power was found when only the α parameter was changed by 0.50. Acceptable power levels were also found in all conditions in which the two extreme β parameters were changed by 0.4 and the α parameter was changed by at least 0.25. In the unequal sample size condition when there was no α parameter DIF and the β parameter was changed by 0.40, acceptable power was not found. However, in the equal sample size condition when there was no α parameter

DIF and the β parameter was changed by 0.40, acceptable power was found in 7 of 8 conditions. Power was also low when the highest two β parameters were changed by 0.4 and there was α parameter DIF of 0.25.

Discussion

Our first research question concerned how the manipulations affected the Type I error rates across tests. Overall, the LRT resulted in fewer Type I errors compared with NCDIF. In addition, the interaction between test type and sample equality suggested that the LRT had acceptable Type I error rates when sample sizes were equal or unequal. The NCDIF index performed adequately when sample sizes were equal, but resulted in too many Type I errors when sample sizes were unequal. None of the other manipulations explained considerable variance in Type I errors. This suggests that for both NCDIF and LRT, the tests for items without DIF are not affected by the presence of items with DIF in the test. This is similar to Meade et al. (2007) who found acceptable Type I error rates for non-DIF items when there were DIF items on the test.

Our second research question concerned how the manipulations affected the power estimates across tests. Overall, the LRT resulted in greater power to detect α parameter DIF compared with NCDIF. The power estimates suggest that when there was a small change in the α parameter, one test was not better at detecting the DIF than the other test. However, when there was a large change in the α parameter, the LRT had more power than NCDIF. This suggests that NCDIF does not have the power to detect DIF when the item is more discriminating for one group versus another group. LRT on the other hand had acceptable power in all conditions when just the α parameter was changed by a large amount. This indicates that when an item is more discriminating for

one group, the LRT is better at detecting this difference. This is consistent with Meade et al. (2007) who found that the NCDIF index did not consistently detect DIF when there was a small change in α parameter DIF.

Overall, the LRT exhibited greater power to detect β parameter DIF than NCDIF. However, the interaction between test type and which β parameters DIF suggested that the NCDIF index was better able to detect the DIF when the last two β parameters were manipulated. The LRT was better able to detect the DIF caused by manipulating the two most extreme β parameters. This may have implications for researchers who expect that group differences occur only at the higher end of the scale. For example, applicants who fake may be more likely than incumbents to endorse the higher response options. In this situation, the NCDIF index may more applicable because it is more sensitive this type of DIF. None of the other manipulations explained considerable variance in power.

Our results suggest that the NCDIF and LRT both had acceptable Type I error rates when sample sizes were equal, but that NCDIF produced too many Type I errors when sample sizes were unequal. One puzzling finding was that the LRT had inflated Type I error rates when the last two β parameters were manipulated. It appears that not only does the LRT have less power to detect DIF when the change is in the last two β parameters, but it also commits more Type I errors than NCDIF. This reinforces the suggestion made earlier that when a researcher suspects that DIF occurs primarily because of differences at the higher end of the scale, the NCDIF index should be used. It is noteworthy that the Type I error rate for the LRT decreased across this condition as α parameter DIF increased.

Finally, the LRT demonstrated more power across conditions. In particular, LRT had greater power to detect DIF caused by different α parameters. This is important because previous research has suggested that NCDIF is not sensitive to this type of DIF (Meade et al., 2007). The LRT also demonstrated acceptable power to detect β DIF in most conditions. However, the NCDIF index has higher power when the last two β parameters are manipulated, but LRT has higher power when the two most extreme β parameters are manipulated.

It is interesting to compare the results of the DFIT analysis with that of Meade et al. (2007). In general, our results are consistent with the findings of Meade et al. Similar to Meade et al. I found that the Type I error rates for the NCDIF index were acceptable across most conditions for equal sample sizes, and that the NCDIF index is not sensitive to α parameter DIF when there is no β parameter DIF. Meade et al. found that when DIF cancelled the NCDIF index was better able to detect it. I found this to be the case as well. Finally, similar to Meade et al. I found that NCDIF was more powerful when the last two β parameters had DIF as opposed to the extreme β parameters. However, I found the magnitude of the β DIF had less of an impact on power than Meade et al. found. This may be due to the fact that, unlike Meade et al., I did not use the condition of changing all four β parameters.

It is more difficult to compare the results for the LRT to previous research because there have been few studies examining it as I did. In general, the previous research has suggested that the LRT has acceptable Type I errors across most situations and it is powerful when sample sizes are adequate (e.g., Ankenmann et al., 1999; Kim & Cohen, 1998; Meade & Lautenschlager, 2004). Our results are consistent with this. In

addition, I found that it compared favorably to the NCDIF index in terms of Type I errors and power.

Limitations and Future Research

I only investigated how well NCDIF and LRT could detect DIF when using the GRM as the generating model, but researchers should investigate other models. Generating the data using the Generalized Partial Credit Model (Muraki, 1992) or the Generalized Graded Unfolding Model (Roberts, Donoghue, & Laughlin, 2000), for example, could result in different results. In addition, I generated all of the theta values using a standard normal distribution. Future research could examine the effects of other distributions of theta. It may also be interesting to generate theta values with different distributions for the focal and reference group. For example, it may be useful to assess the performance of tests for DIF when one of the groups has a restricted theta distribution. This may reflect a common situation in comparing applicants and incumbents where one of the groups may have a restricted range.

Researchers should examine different sample sizes and test lengths to ensure that the results from this study are constant across conditions. Using smaller sample sizes that are more likely to occur in the real world would be yield useful information. Previous research has found that the LRT does not have optimal power in conditions where sample sizes are less than 500 (e.g., Ankenmann, et al., 1999; Meade & Lautenschlager, 2004). More research should examine how the NCDIF handles smaller sample sizes. Finally, there are a number of other DIF detection indices that should be examined in addition to LRT and DFIT. Poly-SIBTEST (Chang, Mazzeo, & Roussos, 1996) and Mantel-

Haenszel (Mantel & Haenszel, 1959) are two examples that could be more closely examined and directly compared with the two indices in the present study.

Implications and Recommendations

Our results have several implications for using tests of DIF. First, the LRT should be preferred when there are unequal sample sizes. I found that when the sample sizes are unequal the NCDIF index commits more Type I errors than the LRT, and that the LRT still had acceptable Type I errors. This is important because it is likely that DIF studies will have unequal sample sizes. In addition, the LRT has higher power estimates than NCDIF when sample sizes are unequal.

The LRT has better power across most conditions. In particular, LRT had more power to detect DIF caused by different item discrimination parameters. However, the LRT had less power than NCDIF to detect DIF when the two highest difficulty values were changed by a moderate amount. This suggests that the LRT and NCDIF indices may be able to capture different patterns of DIF and the preference of which test to use depends on the type of DIF present. In practice, researchers will not know what type of DIF, but they may have some expectations. For example, it may be that applicants are more likely than incumbents to endorse the highest two response options on Conscientiousness items. Our findings suggest that the NCDIF index would be better able to detect this and therefore might be preferred for that situation.

Although there are some specific situations where the NCDIF index may be preferred, I suggest the LRT should be preferred overall because of its consistent Type I error rates and overall power. In addition, I found the LRT is easier to implement than the NCDIF index. The LRT output indicates the χ^2 statistic associated with the overall item

as well as the χ^2 statistic associated with α parameter and the β parameter. This allows the researchers to identify where the DIF is located for a particular item. In addition, the LRT does not require the linking or calculation of empirical cutoffs needed by the NCDIF index.

References

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-29). Hillsdale, NJ: Erlbaum.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, 10, 95-105. doi: 10.1111/j.1745-3984.1973.tb00787.x

Ankenmann, R.D., Witt, E.A., & Dunbar, S.B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, 36, 279-300. doi: 10.1111/j.1745-3984.1999.tb00558.x

Baker, F. B. (1995). EQUATE 2.1: Computer program for equating two metrics in item response theory [computer program]. Madison, University of Wisconsin, Laboratory of Experimental Design.

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, 15(2), 113-141. doi: 10.1207/S15324818AME1502_01

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

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational*

Measurement, 33, 333-353.

Collins, W.C., Raju, N.S., & Edwards, J.E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology,* 85, 451-461. doi: 10.1037/0021-9010.85.3.451

Donvan, M.A., Drasgow, F., & Probst, T.M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology,* 85, 305-313. doi: 10.1037/0021-9010.85.2.305

Finch, W.H., & French, B.F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement,* 67, 565-582. doi: 10.1177/0013164406296975

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. doi: 10.1177/01466219922031437

Henry, M.S., & Raju, N.S. (2006). The effects of trailed and situational impression management on a personality test: An empirical analysis. *Psychological Science*, 48, 247-267.

Holland, P.W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 139-145). Hillsdale, NJ: Erlbaum.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing,* 18, 89-114. doi: 10.1191/026553201675366418

Kim, S.-H., & Cohen, A.S. (1995). A comparison of Lord's chi-square, Raju's area

measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, 8, 291-312. doi: 10.1207/s15324818ame0804_2

Kim, S.-H., & Cohen, A.S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, 22, 345-355. doi: 10.1177/014662169802200403

Kim, S.-H., Cohen, A.S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Education Management*, 44, 93-116. doi: 10.1111/j.1745-3984.2007.00029.x

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

Maurer, T.J., Raju, N.S., & Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, 83, 693-702. doi: 10.1037/0021-9010.83.5.693

Meade, A.W., Lautenschlager, G.J., & Johnson, E.C. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement*, 31, 430-455. doi: 10.1177/0146621606297316

Morales, L.S., Flowers, C., Gutierrez, P., Kleinman, M., & Teresi, J.A. (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the differential item and test functioning (DFIT) framework. *Medical Care*, 44, 143-151. doi: 10.1097/01.mlr.0000245141.70946.29

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [computer program]. Chicago, IL: Scientific Software.

Oshima, T. C., Kushubar, S., Scott, J. C., & Raju, N. S. (2009). DFIT8 for Window User's Manual: Differential functioning of items and tests. St. Paul MN: Assessment Systems Corporation.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43,* 1-17. doi: 10.1111/j.1745-3984.2006.00001.x

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika,* 53, 495-502. doi: 10.1007/BF02294403

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14, 197-207. doi: 10.1177/014662169001400208

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33,* 133-147.

Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87, 517-529. doi: 10.1037/0021-9010.87.3.517

Raju, N., van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement*, 19, 353-368. doi: 10.1177/014662169501900405

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32. doi: 10.1177/01466216000241001

Robie, C., Zickar, M. J. , & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*(2), 187-207. doi: 10.1207/S15327043HUP1402_04

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrica Monographs*, 34.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210. doi: 10.1177/014662168300700208

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. University of North Carolina at Chapel Hill.

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

Table 1

Mixed Model ANOVA (Type I)

Source	df	F	η^2
Type of Test (T)	1	4715.1	.29
Type of Test (T) * Magnitude of β DIF (M)	1	37.2	.00
Type of Test (T) * Magnitude of α DIF (A)	2	18.0	.00
Type of Test (T) * Which β parameters DIF (W)	1	36.9	.00
Type of Test (T) * Number of items (N)	1	0.3	.00
Type of Test (T) * Cancellation of DIF (C)	1	91.7	.01
Type of Test (T) * Equal Samples (E)	1	1638.7	.13
T * A * N * C	2	38.8	.0

Note. Percentages out of 100 samples and across all items. Only higher order interactions explaining at least 1 percent of the variance were included.

Table 2

Mixed Model ANOVA (power)

Source	df	F	η^2
Type of Test (T)	1	973.6	.08
Type of Test (T) * Magnitude of α DIF (A)	2	257.9	.04
Type of Test (T) * Magnitude of β DIF (M)	1	261.9	.02
Type of Test (T) * Which β parameters DIF (W)	1	2362.6	.17
Type of Test (T) * Number of items (N)	1	173.0	.02
Type of Test (T) * Cancellation of DIF (C)	1	141.1	.01
Type of Test (T) * Equal Samples (E)	1	439.9	.04
T * A * M	2	81.2	.01
T * A * W	2	324.0	.05
T * A * E	2	186.8	.03
T * M * W	1	2092.4	.15
T * M * C	1	109.3	.01
T * M * E	1	218.8	.02
T * A * M * W	2	324.0	.05
T * A * M * E	2	313.5	.01
T * A * W * C	2	48.6	.01
T * A * M * W * C	2	39.2	.01

Note. Percentages out of 100 samples and across all items. Only higher order interactions explaining at least 1 percent of the variance were included.

Table 3

Type I Errors and Power for equal sample sizes (NCDIF)

β DIF	Items	Cancellation of DIF	No α DIF TIE (P)	0.25 α DIF TIE (P)	0.50 α DIF TIE (P)
None	10	No	5	5 (23)	3 (72)
	10	Yes	-	5 (22)	5 (51)
	20	No	5	5 (20)	4 (75)
	20	Yes	-	5 (15)	4 (51)
0.4 Highest 2	10	No	5 (83)	5 (77)	6 (92)
	10	Yes	5 (86)	3 (86)	4 (89)
	20	No	4 (81)	5 (80)	4 (91)
	20	Yes	4 (85)	5 (86)	4 (92)
0.4 Extremes	10	No	3 (56)	4 (61)	5 (84)
	10	Yes	3 (60)	6 (60)	5 (76)
	20	No	4 (53)	4 (59)	4 (85)
	20	Yes	4 (58)	5 (56)	4 (71)
1.0 Highest 2	10	No	4 (100)	5 (99)	5 (99)
	10	Yes	4 (100)	4 (100)	4 (99)
	20	No	5 (100)	5 (100)	6 (99)
	20	Yes	5 (100)	5 (100)	6 (100)
1.0 Extremes	10	No	6 (100)	4 (100)	7 (99)
	10	Yes	4 (100)	6 (99)	6 (98)
	20	No	5 (100)	4 (100)	4 (98)
	20	Yes	4 (100)	5 (100)	7 (99)

Note. Percentages out of 100 samples and across all items. DIF = differential item functioning. NCDIF = noncompensatory DIF. P = power; TIE = Type I error.

Table 4

Type I Errors and Power for unequal sample sizes (NCDIF)

β DIF	Items	Cancellation of DIF	No α DIF TIE (P)	0.25 α DIF TIE (P)	0.50 α DIF TIE (P)
None	10	No	17	17 (40)	17 (76)
	10	Yes	-	16 (38)	16 (60)
	20	No	18	18 (41)	17 (72)
	20	Yes	-	18 (27)	17 (54)
0.4 Highest 2	10	No	16 (74)	17 (76)	20 (89)
	10	Yes	19 (80)	16 (83)	18 (87)
	20	No	18 (72)	18 (79)	17 (87)
	20	Yes	18 (77)	20 (80)	16 (89)
0.4 Extremes	10	No	18 (62)	17 (67)	21 (81)
	10	Yes	15 (66)	17 (66)	20 (71)
	20	No	18 (58)	18 (66)	16 (82)
	20	Yes	17 (58)	18 (63)	18 (73)
1.0 Highest 2	10	No	19 (99)	18 (99)	21 (98)
	10	Yes	18 (100)	19 (100)	17 (99)
	20	No	20 (99)	18 (99)	19 (99)
	20	Yes	19 (100)	18 (99)	19 (99)
1.0 Extremes	10	No	18 (99)	19 (99)	19 (97)
	10	Yes	19 (99)	20 (99)	20 (96)
	20	No	17 (100)	17 (99)	19 (96)
	20	Yes	16 (100)	19 (99)	21 (97)

Note. Percentages out of 100 samples and across all items. DIF = differential item functioning. NCDIF = noncompensatory DIF. P = power; TIE = Type I error.

Table 5

Type I Errors and Power for equal sample sizes (LRT)

β DIF	Items	Cancellation of DIF	No α DIF TIE (P)	0.25 α DIF TIE (P)	0.50 α DIF TIE (P)
None	10	No	6	1 (67)	1 (100)
	10	Yes	-	1 (45)	1 (100)
	20	No	1	1 (71)	1 (100)
	20	Yes	-	1 (53)	1 (100)
0.4 Highest 2	10	No	3 (57)	2 (42)	1 (100)
	10	Yes	1 (89)	1 (75)	2 (97)
	20	No	1 (71)	1 (56)	1 (100)
	20	Yes	1 (89)	2 (80)	1 (100)
0.4 Extremes	10	No	1 (86)	1 (99)	1 (100)
	10	Yes	1 (91)	1 (100)	1 (100)
	20	No	1 (87)	2 (100)	1 (100)
	20	Yes	1 (93)	1 (100)	1 (100)
1.0 Highest 2	10	No	24 (100)	13 (99)	4 (100)
	10	Yes	1 (100)	1 (100)	4 (100)
	20	No	3 (100)	3 (100)	2 (100)
	20	Yes	1 (100)	1 (100)	1 (100)
1.0 Extremes	10	No	6 (100)	5 (100)	2 (100)
	10	Yes	1 (100)	1 (100)	2 (100)
	20	No	1 (100)	1 (100)	1 (100)
	20	Yes	1 (100)	1 (100)	1 (100)

Note. Percentages out of 100 samples and across all items. DIF = differential item functioning. LRT = likelihood ratio test. P = power; TIE = Type I error.

Table 6

Type I Errors and Power for unequal sample sizes (LRT)

β DIF	Items	Cancellation of DIF	No α DIF TIE (P)	0.25 α DIF TIE (P)	0.50 α DIF TIE (P)
None	10	No	1	3 (38)	2 (100)
	10	Yes	-	2 (35)	3 (93)
	20	No	3	3 (44)	3 (100)
	20	Yes	-	3 (33)	2 (95)
0.4 Highest 2	10	No	4 (35)	4 (27)	3 (97)
	10	Yes	2 (68)	2 (53)	3 (92)
	20	No	3 (49)	2 (37)	2 (99)
	20	Yes	2 (64)	3 (57)	2 (94)
0.4 Extremes	10	No	3 (64)	2 (92)	2 (100)
	10	Yes	3 (73)	2 (93)	2 (100)
	20	No	2 (63)	3 (92)	3 (100)
	20	Yes	2 (70)	3 (92)	2 (100)
1.0 Highest 2	10	No	18 (100)	13 (92)	10 (99)
	10	Yes	3 (100)	3 (100)	4 (100)
	20	No	5 (100)	4 (95)	4 (100)
	20	Yes	3 (100)	2 (100)	3 (100)
1.0 Extremes	10	No	8 (100)	6 (100)	6 (100)
	10	Yes	3 (100)	2 (100)	3 (100)
	20	No	3 (100)	3 (100)	3 (100)
	20	Yes	3 (100)	3 (100)	2 (100)

Note. Percentages out of 100 samples and across all items. DIF = differential item functioning.

LRT = likelihood ratio test. P = power; TIE = Type I error.

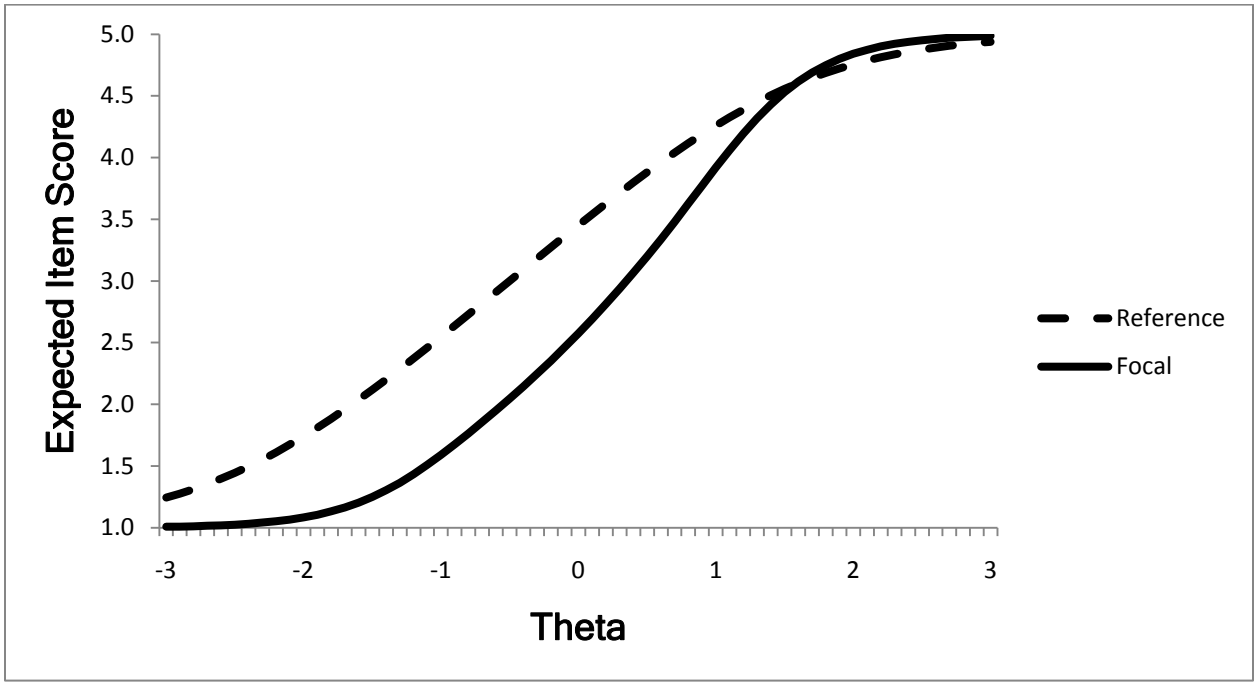

Figure 1. Example of an item with DIF.

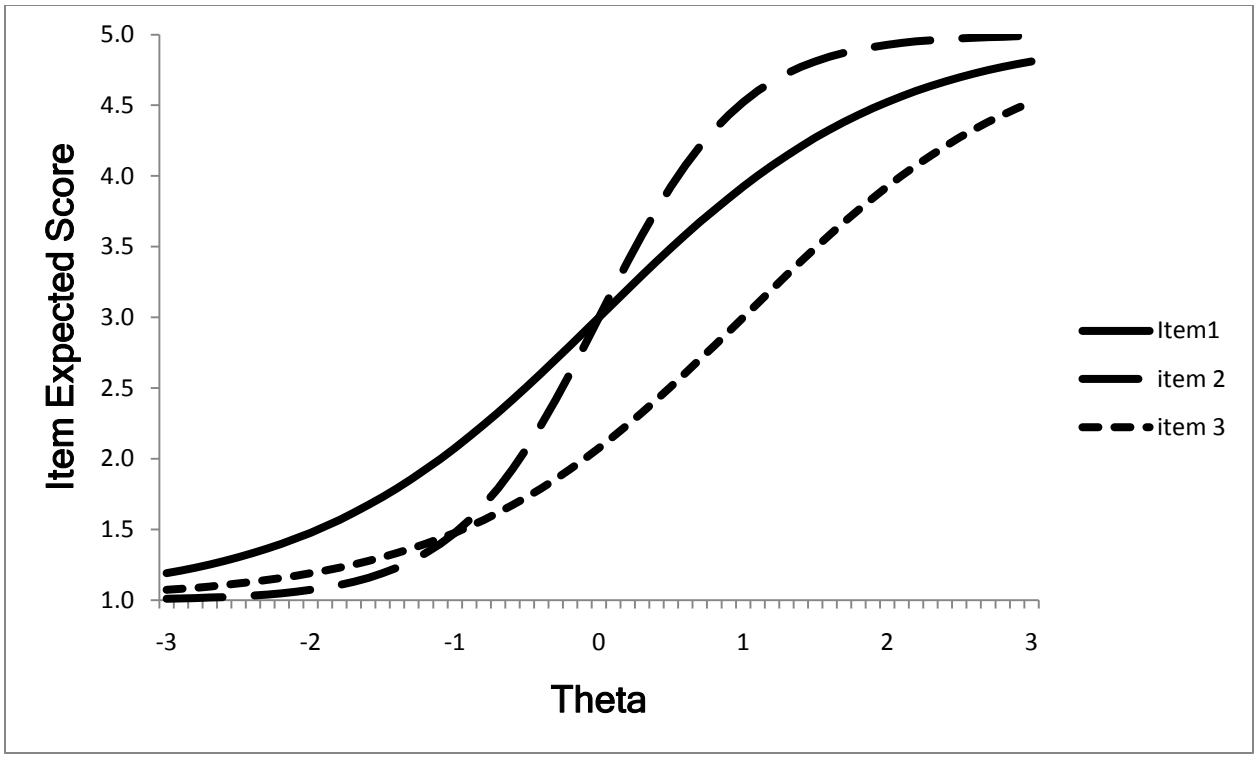

Figure 2. Example of Item Response Curves

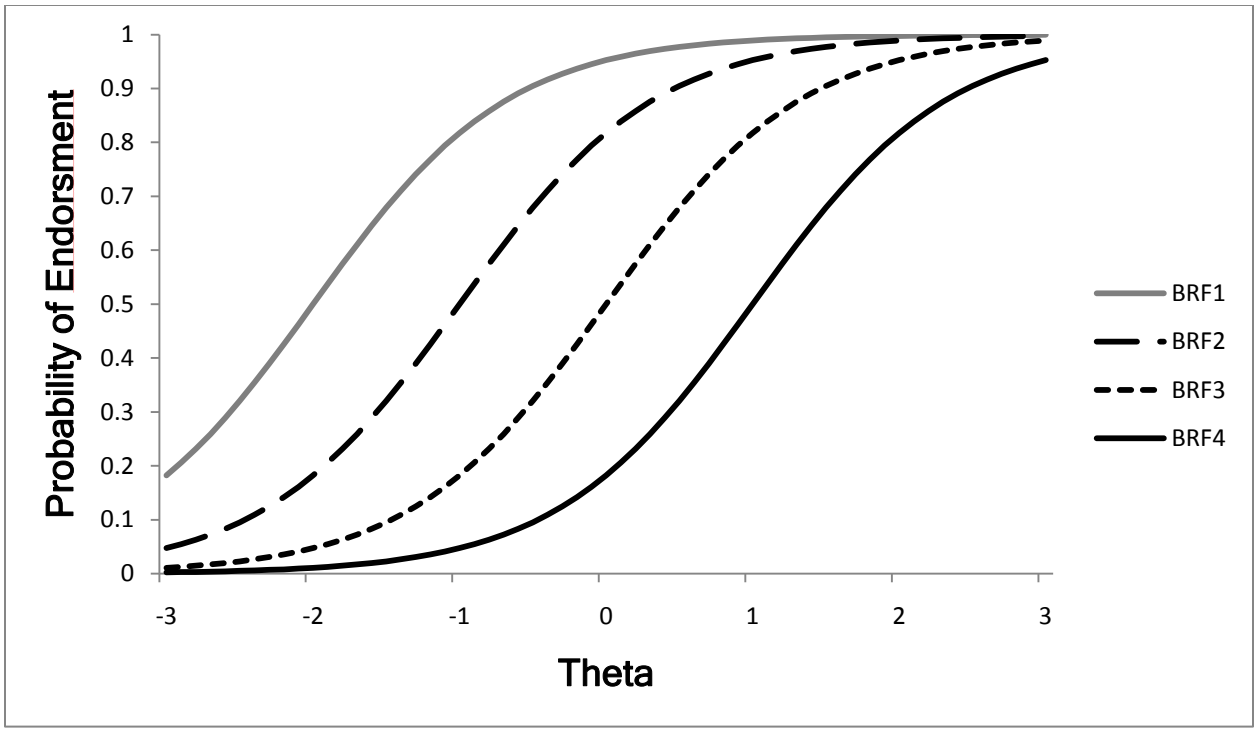

Figure 3. Boundary Response Functions for a 5-reponse option item.

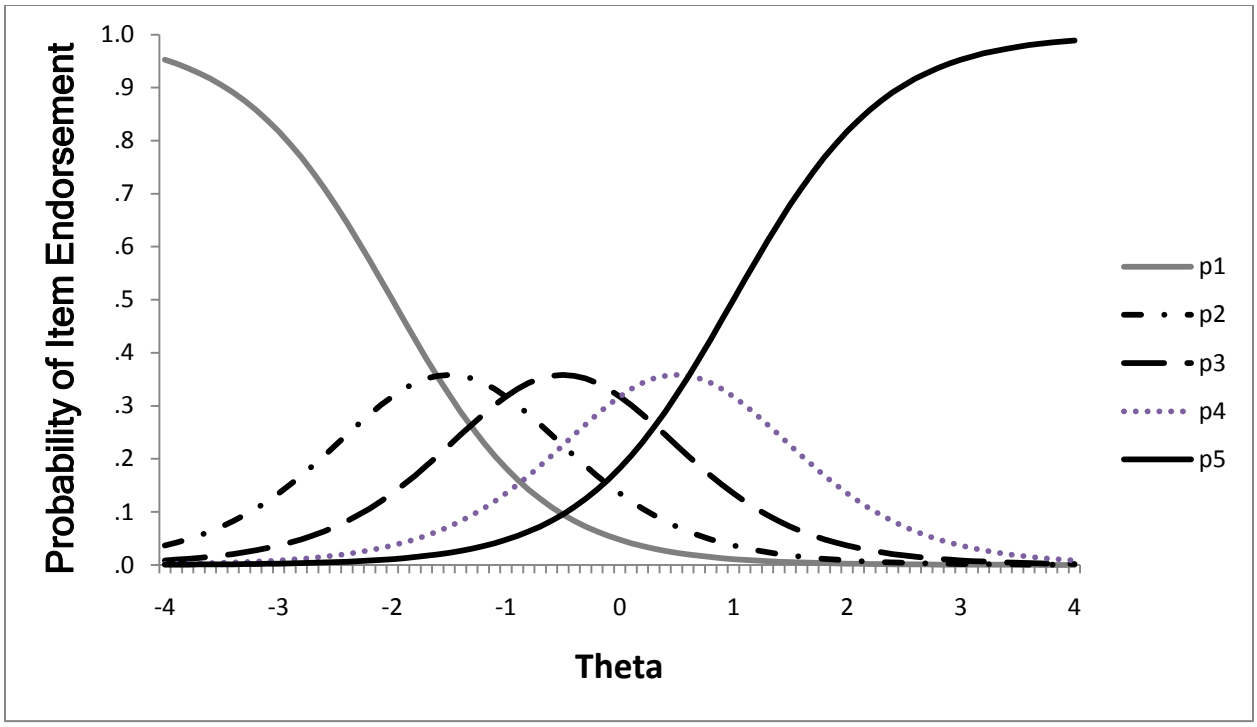

Figure 4. Option Response Function for a 5-response option item.

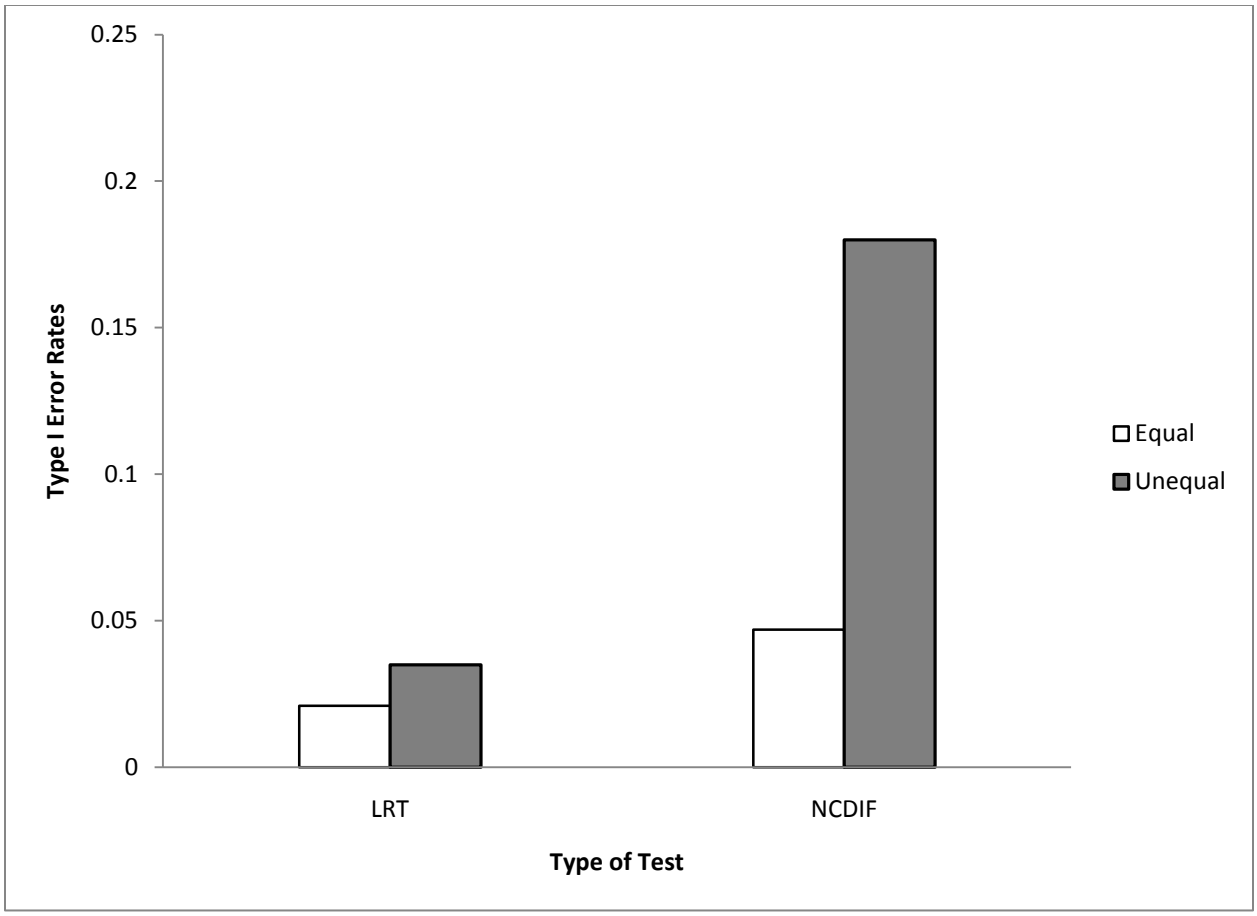

Figure 5. The interaction between the type of test and sample equality.

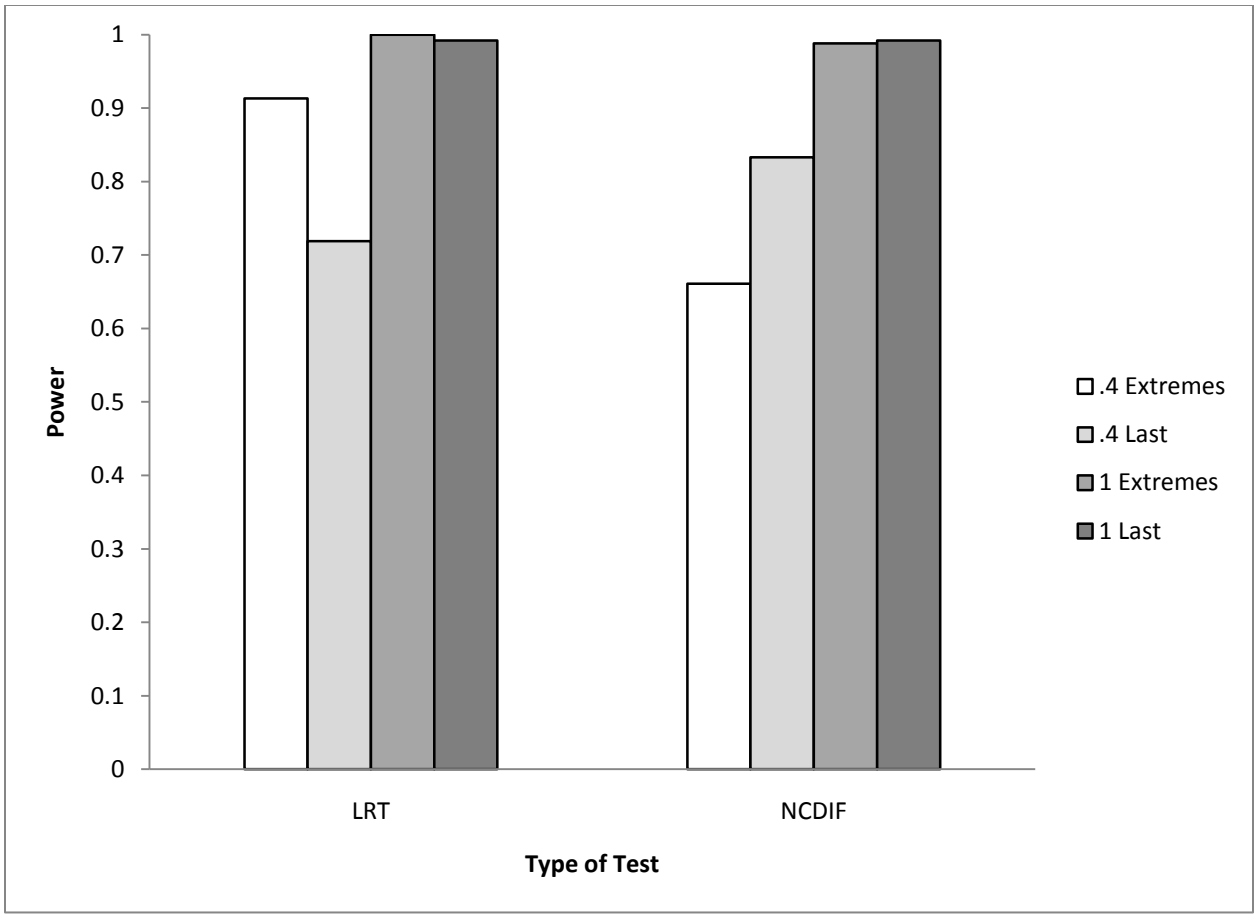

Figure 6. The interaction between the type of test and which β parameters were manipulated.